Sensitive Data Handling at the Turing - Overview for Data Providers

Introduction

Secure Environments for analysis of sensitive datasets are essential for research.

Such "data safe havens" are a vital part of the research infrastructure.

It is essential that sensitive or confidential datasets are kept secure, both to enable analysis of personal data in a manner that is capable of being compliant with data protection law, and to avoid jeopardising the consent of society for research activities with personal data (called 'social license').

To create and operate these Environments safely and efficiently whilst ensuring usability, requires, as with many sociotechnical systems, a complex stack of interacting business process and design choices. This document describes the approaches taken by the Alan Turing Institute when building and managing Environments for productive, secure, collaborative research projects.

We propose choices for the security controls that should be applied in the areas of: * data classification * data ingress (data entering a secure Environment from an external source) * data egress (data leaving a secure Environment to an external recipient) * software ingress (software entering a secure Environment from an external source) * user access * user device management * analysis Environments

We do this for each of a small set of security "Tiers" - noting that the choice of security controls depends on the sensitivity of the data.

Why classify?

One of the major drivers for usability or security problems is over- or under-classification, that is, treating data as more or less sensitive than it deserves.

Regulatory and commercial compliance requirements place constraints on the use of datasets; implementation of that compliance must be set in the context of the threat and risk profile and balanced with researcher productivity.

Almost all security measures can be circumvented, security can almost always be improved by adding additional barriers, and improvements to security almost always carry a cost in usability and performance.

Misclassification is seriously costly for research organisations and their partners: overclassification results not just in lost researcher productivity, but also a loss of scientific engagement, as researchers choose not to take part in a project with cumbersome security requirements. Systematic overclassification increases data risk by encouraging workaround breach.

The risks of under-classification include not only legal and financial sanction, but the loss of the social licence to operate of the whole community of data science researchers.

Document structure

This document describes our approach to handling research data. It does not cover the Turing's core enterprise information security practices, which are described elsewhere. Nor do we cover the data-centre level or organisational management security practices which are fundamental to any secure computing facility - we do not operate our own data centres, but rely on upstream data centre

provision, such as Microsoft Azure and the Edinburgh Parallel Computing Centre, compliant with ISO 27001 (Information Security Management System Requirements).

The document is structured as follows: we begin by defining terms which are used throughout the document. We then discuss some aspects of the design, before describing our 'model' for secure research Environments. Next, we discuss the possible choices for each security control around each of the areas bullet-pointed above, while leaving open the question of which controls are appropriate at which tiers. Finally, we make specific choices assigning controls to security tiers.

Definitions - a model for secure data research projects

Work Packages

Assessing the sensitivity of a dataset requires an understanding of both the base sensitivity of the information contained in the dataset and of the impact on that base sensitivity of the operations that it will undergo in the research project. The classification exercise therefore relates to each stage of a project and not simply to the datasets as they are introduced into it.

Classification to a tier is therefore **not** a property of a dataset, because a dataset's sensitivity depends on the data it can be combined with, and the use to which it is put.

In our model, projects are divided into **work packages**, which we use here to refer to the activities carried out within a distinct phase of work carried out as part of a project, with a specific outcome in mind. A work package can make use of one or more datasets, and includes an idea of the analysis which the research team intends to carry out, the potential outputs they are expecting, and the tools they intend to use – all important factors affecting the data sensitivity.

Classification is carried out on work packages rather than individual datasets.

Environments and Platforms

Once a work package has been classified, an appropriate secure analysis Environment is instantiated depending on the tier assigned.

For the initial work package in a project, a new Environment must always be deployed. For additional work packages, the project may deploy a new environment per work package or, where appropriate, add the new work package to an existing Environment deployed for the project.

When considering adding a work package to an existing environment, the **combination** of the new work package plus all existing work packages the Environment has already been used for must be considered as the effective work package when making classification decisions. The classification tier of a combination of work package(s) can never be lower than the highest classification tier of any of the individual work packages, but may be higher due to additional risks introduced by combining datasets and activities across work packages. If the combined classification is higher than the tier associated with the existing Environment, a new Environment must be deployed. The classification tier of an Environment cannot be upgraded or downgraded "in place".

Depending on the classification assigned, an Environment may be instantiated on one of several supported Platforms. The Turing currently supports secure deployments to Microsoft's Azure cloud platform.

Researcher

A project member, who analyses data to produce results. We reserve the capitalised term "Researcher" for this role in our user model. We use the lower case term when considering the population of researchers more widely.

Investigator

The research project lead, this individual is responsible for ensuring that project staff comply with the Environment's security policies. A single lead Investigator must be responsible for a project. Multiple collaborating institutions may have their own lead academic staff, and academic staff might delegate to a researcher the leadership as far as interaction with the Environment is concerned. In both cases, the term Investigator here is independent of this - regardless of academic status or institutional collaboration, this individual accepts responsibility for the conduct of the project and its members.

Referee

A Referee volunteers to review code or derived data (data which is computed from the original dataset), providing evidence to the Investigator and Dataset Provider Representative that the researchers are complying with data handling practices. Referees also play a role in classifying work packages at Tiers 2 or above, when they should be consulted by either the research team or the Dataset Provider Representative (see below).

Dataset Provider and Representative

The **Dataset Provider** is the organisation who provided the dataset under analysis. The Dataset Provider will designate a single representative contact to liaise with the Turing. This individual is the **Dataset Provider Representative**. They are authorised to act on behalf of the Dataset Provider with respect to the dataset and must be in a position to certify that the Dataset Provider is authorised to share the dataset with the Turing.

There may be additional people at the Dataset Provider who will have input in discussions around data sharing and data classification. It is the duty of the Dataset Provider Representative to manage this set of stakeholders at the Dataset Provider.

Programme Manager

A designated staff member in the research institution who is responsible for creating and monitoring projects and Environments and overseeing a portfolio of projects. This should be a member of professional staff with oversight for data handling in one or more research domains.

The Programme Manager can add new users to the system and assign users to specific projects. They assign Project Managers and can, if they wish, take on this role themselves.

Project Manager

A staff member with responsibility for running a particular project. This role could be filled by the Programme Manager, or a different nominated member of staff within the research institution.

While the Programme Manager should maintain responsibility for adding users to the user list, and can add users to projects, the Project Manager should also have the authority to assign existing users to their project. To do this they will need to be able to review and search existing users.

System Manager

Members of Turing staff responsible for configuration and maintenance of the Environment.

Software-defined infrastructure

Our approach - separately instantiating an isolated Environment for each project - is made feasible by the advent of "software-defined infrastructure".

It is now possible to specify a whole arrangement of IT infrastructure, servers, storage, access policies and so on, completely as **code**. This code is executed against web services provided by infrastructure providers (the APIs of cloud providers such as Microsoft, Amazon or Google, or an in-house "private cloud" using a technology such as OpenStack), and the infrastructure instantiated.

Our model therefore assumes the availability of a software-defined infrastructure provision offering, in an ISO 27001 compliant data-centre and organisation, the scripted instantiation of virtual machines, storage, and secure virtual networks.

We also assume that "Identification, Authorisation and Authentication" (IAA) is available as a service from this provider, so that they provide user account creation, the creation of security groups, the assignment of users to security groups, the restriction of access to resources by such users, login challenge by password and a second factor, password reset, and other such security considerations.

A software-defined infrastructure platform on which to build, means that the definition of the Environment can be meaningfully audited - as no aspect of it is not described formally in code, it can be fully scrutinised.

Secure data science

We highlight two assumptions about the research user community critical to our design:

Firstly, we must consider not only accidental breach and deliberate attack, but also the possibility of "workaround breach", where well-intentioned researchers, in an apparent attempt to make their scholarly processes easier, circumvent security measures, for example, by copying out datasets to their personal device. Our user community are relatively technically able; the casual use of technical circumvention measures, not by adversaries but by colleagues, must be considered. This can be mitigated by increasing awareness and placing inconvenience barriers in the way of undesired behaviours, even if those barriers are in principle not too hard to circumvent.

Secondly, research institutions need to be open about the research we carry out, and hence, the datasets we hold. This is because of both the need to publish our research as part of our impact cases to funders, and because of the need to maintain the trust of society, which provides our social licence. This means we cannot rely on "security through obscurity": we must make our security decisions assuming that adversaries know what we have, what we are doing with it, and how we secure it.

Environment Tiers

Our recommendation for secure information processing tiers is based on work which has gone before. We have begun with the UK government classifications, and reconciled these to the definitions of personal data, whether or not something is 'special category' under the GDPR or relates to criminal convictions, and related them to common activities in the research community.

Where input datasets contain personal data, consideration should always be given at the outset to minimising the personal data, including by pseudonymisation or anonymisation.

Pseudonymised data is still personal data, as it can be re-identified by those who hold the key to turn pseudonyms back into individual identifiers. This may include synthetic data derived from personal data, or models trained on personal data, depending on the methods used to synthesise the data or generate the models.

Anonymised data, including pseudonymised data where that key is destroyed, is not personal data when it is impossible to re-identify any living individuals from it. However, if the quality of anonymisation is ambiguous or if individuals can be identified when the anonymised data is combined with another dataset, such data would by definition not be anonymised, and would therefore be personal data. The question as to whether re-identification is possible or not is a very subtle one, and the assessment of this risk is critical to the assignment of security tiers.

We emphasise that this classification is based on considering the sensitivity of all information handled in the project, including information that may be generated by combining or processing input datasets. In every case, the categorisation does not depend only on the input datasets, but on combining information with other information or generated results in a work package.

Derived information may be of higher security tier than the information in the input datasets. (For example, information on the identities of those who are suspected to possess an undiagnosed neurological condition on the basis of analysis of public social media data.) This should form part of the information constituting a work package; when a project team believes this will be the case, the work package should be classified at the higher tier of secure Environment.

If it becomes apparent during the project that intended analysis will produce this effect then the inputs should be treated as a new work package with this extra information, and classified afresh, following the full classification process below. In the below, "personal data" follows the GDPR definition: information from which a living individual is identified or identifiable. It excludes information about individuals who are dead.

Tier 0

Tier 0 Environments are used to handle open information, which is legally available to the general public with no restrictions, where all generated and combined information is also suitable for open handling.

Tier 0 applies where none of the information processed, combined or generated includes personal data, commercially sensitive data, or data which will have legal, political or reputational consequences in the event of unauthorised disclosure.

Tier 0 environments may be used for anonymised or synthetic information generated from personal data, where one has **absolute** confidence in the quality of anonymisation or the privacy preserving nature of the data synthesis. This makes the information no longer personal data. This does **not** include pseudonymised data which can be re-identified in combination with a key or other dataset. This is still considered personal data.

Note that in practice it is extremely difficult (if not impossible) to guarantee that data is truly anonymous, especially when considering the risk of the anonymised data being linked with other datasets that currently exist or may exist in the future, and the potential development of more sophisticated re-identification attacks.

If there is not **absolute** confidence in the anonymous or synthetic data no longer being personal data, then the minimum tier environment this data can be processed in is Tier 2.

Tier 0 data should be considered ready for publication. Although this data is open, there are still advantages to handling it through a managed data analysis infrastructure.

Management of Tier 0 data in a visible, well ordered infrastructure provides confidence to stakeholders as to the handling of more sensitive datasets.

Although analysis may take place on personal devices or in non-managed cloud-based analysis Environments, the data should still therefore be listed through the inventory and curatorial systems of a managed research data Environment.

Finally, audit trails as to the handling of Tier 0 information reduce risks associated with misclassification - if data is mistakenly classified as a lower tier than it should be, we still retain information as to how it was processed during the period of misclassification.

Tier 1

Tier 1 Environments are used to handle, process and generate data that is intended for eventual publication or that could be published without reputational damage.

Information is kept private in order to give the research team a competitive advantage, not due to legal data protection requirements.

Both the datasets and the proposed processing must otherwise meet the criteria for Tier 0.

It may be used for pseudonymised or synthetic information generated from personal data, where one has **absolute** confidence that the personal data cannot be re-identified.

It may also be used for commercial data where commercial consequences of disclosure would be no impact or very low impact, with the agreement of all parties.

Relationships to other classification schemes

Pseudonymised data is considered Personal Data under the GDPR. Anonymised data is not considered Personal Data under the GDPR, but in practice it is extremely difficult (if not impossible) to guarantee that data is truly anonymous. Therefore, unless we are **absolutely** confident in the anonymisation process, we consider all data related to living individuals as Personal Data under the GDPR and therefore at least Tier 2.

In particular, Tier 1 is not suitable for any data derived from personal data that is not otherwise suitable for processing in Tier 0. If this is not the case, then the minimum tier environment such data can be processed in is Tier 2.

Tier 2

Tier 2 Environments are used to handle, combine or generate information which is not linked to identifiable personal data.

It may be used for pseudonymised, synthetic or anonymised information generated from personal data, where we have strong, but not absolute, confidence that the personal data cannot be re-identified. This assessment should consider the risk of processing the data in a manner that permits personal data to be re-identified, including by combining it with other data available within the environment.

The pseudonymisation, synthesis or anonymisation process itself, if carried out in the Turing, should take place in a Tier 3 Environment. A typical model for a project will be to instantiate both Tier 2 and Tier 3 Environments, with pseudonymised, synthetic or anonymised data generated in the Tier 3 Environment and then transferred to the Tier 2 Environment.

Tier 2 Environments are also used to handle, combine or generate information which is confidential but not, in commercial or national security terms, sensitive. This includes commercial-in-confidence datasets or intellectual property where the legal, commercial, political and reputational consequences from disclosure are low. Where such consequences are not low, Tier 3 should be used.

At Tier 2, the most significant risks are "workaround breach" and the risk of mistakenly believing data is robustly pseudonymised or anonymised, when in fact re-identification might be possible.

Relationships to other classification schemes

Almost all data at the baseline UK government OFFICIAL classification is likely to be Tier 2, as well as a large proportion of data at the OFFICIAL-SENSITIVE [COMMERCIAL] classification.

All pseudonymised Personal Data under the GDPR that is not Special Category Personal Data is Tier 2. Note that pseudonymised data is Personal Data under the GDPR. While anonymised data is not considered Personal Data under the GDPR, in practice it is extremely difficult (if not impossible) to guarantee that data is truly anonymous. Therefore, unless we are **absolutel** confident in the anonymisation process, we consider all data related to living individuals as Personal Data under the GDPR.

Tier 3

Tier 3 Environments are used to handle, combine or generate personal data, excluding personal data where there is a risk that disclosure might pose a substantial threat to the personal safety, health or security of the data subjects (which would be Tier 4).

This also includes pseudonymised, synthetic or anonymised information generated from personal data, where we have only weak confidence that the personal data cannot be re-identified.

Tier 3 Environments are also used to handle, combine or generate information, including intellectual property, which is sensitive in commercial, legal, political, or national security terms. This tier anticipates the need to defend against compromise by attackers with bounded capabilities and resources. This may include hacktivists, single-issue pressure groups, investigative journalists, competent individual hackers and the majority of criminal individuals and groups. The threat profile excludes sophisticated, well-resourced and determined threat actors, such as highly capable serious organised crime groups and state actors.

The difference between Tier 2 and Tier 3 Environments is the most significant in this model, both for researcher productivity and organisational risk.

At Tier 3, the risk of hostile actors attempting to break into the Environment becomes significant.

Relationships to other classification schemes

All data at the UK government OFFICIAL-SENSITIVE [PERSONAL] classification will be Tier 3, as well as some data at the OFFICIAL-SENSITIVE [COMMERCIAL] classification, where the consequence of disclosure are particularly high.

All pseudonymised Special Category Personal Data under the GDPR is Tier 3.

All non-pseudonymised Personal Data under the GDPR, whether or not it is Special Category Personal Data is Tier 3.

Tier 4

Tier 4 Environments are used to handle, combine or generate personal data where disclosure poses a substantial threat to the personal safety, health or security of the data subjects.

This also includes handling, combining or generating datasets which are sensitive in commercial or national security terms, and are likely to be subject to attack by sophisticated, well-resourced and determined actors, such as serious organised crime groups and state actors.

It is at Tier 4 that the risk of hostile actors penetrating the project team becomes significant.

Relationships to other classification schemes

All data at the UK government SECRET classification will be Tier 4.

Connections to and from the Environment

At lower tiers direct inbound connections to resources within the Environment may be permitted. At higher tiers inbound connections are only permitted via a secure access node (e.g. Microsoft Remote Desktop Services).

A remote desktop connection allowing access to graphical interface applications should be provided to allow researchers to connect to the remote secure analysis Environment. At all but the lowest tiers, this requires two-factor authentication, and, at some tiers, the copy paste function is disabled.

At every tier, long and strong passphrases (for example, at least four randomly chosen dictionary words) should be enforced, and users are trained in the use of keychain managers on their access devices, locked with two-factor authentication, so that the inconvenience of repeatedly typing a long passphrase is mitigated, reducing the risk of users choosing insecure passwords.

At some tiers, we may provide **secure shell** connections using the command line, in addition to the remote desktop.

The text-based access this grants is sufficient for some professional data scientists. The primary driver for this preference is that processes can easily be reproduced based on the commands typed. If not needed, providing a remote desktop interface adds complexity and therefore risk.

At some tiers, specific commands commonly used for copying out data can therefore be blocked for users.

In neither case is the user absolutely prevented from copying out to the device used to access the Environment (with remote desktop software, malicious users can script automated screen-grabs). However, this can be made difficult in order to deter casual workaround risk, and, at the highest tiers, prevented by only permitting access to the Environment from user devices permanently located within a secure physical Environment.

We therefore believe it will be possible to make secure shell access just as secure as remote desktop access, but this remains a work in progress.

At lower tiers outbound connections from the Environment to the internet and other external resources are permitted. At higher tiers connections to resources outside the Environment's private network are not permitted.

The classification process

The Dataset Provider Representative and Investigator must agree on a classification for a work package. If the classification is likely to be Tier 2 or higher, they should also involve an independent Referee. Prior to datasets being transfered to the Turing, only the Dataset Provider Representative will have access to the actual dataset(s). The Investigator (and Refereee if necessary) will need to make their classification judgements based on discussions with the Dataset Provider Representative, alongside a clear description of the dataset and associated metadata such as data dictionaries.

The Dataset Provider Representative, Investigator and Referee (if applicable) should independently classify the work package using the classification web application or classification flowchart. If the flowchart is used, the full path of decisions made should be recorded, not just the final outcome. If the web application is used, this is done automatically.

The project should only proceed if the Investigator, the Dataset Provider Representative, and the Referee (if applicable), can come to a consensus on a work package classification. If consensus cannot be reached, the work package should be reconsidered.

The Turing does not currently have access to a Tier 4 Environment. Therefore, if the work package classification is Tier 4, it should be reconsidered.

If the classification is Tier 3 or below, the dataset(s) should be ingressed into an Environment at that Tier to which the Investigator and Referee (if applicable) have access, so that they can verify the classification based on complete information. If at this point either the Investigator or Referee disagree with the original classification, the consensus seeking process between the Data Provider Representative, Investigator and Referee (if applicable) should be repeated. If consensus cannot be achieved the dataset(s) must be deleted from the Environment.

If, at any point during the project, the research team decides to analyse the data differently or for a different purpose than previously agreed, this constitutes a new work package, and should be newly classified by repeating this process. This is also the case if the team wishes to ingress another dataset in combination, which will require Representatives from all Dataset Providers to arrive at the same consensus as the Investigator and Referee (if applicable).

Data egress and new classification

It is a central premise of our model that any output data is classified as a new work package, separate from the work package that it is derived from.

This is the case whether the output data is for publication (in which case the output data should be Tier 1 or Tier 0) or will be analysed in a new Environment.

As a convenience, if derived data resulting from analysis is in a form which has been agreed with Dataset Provider Representatives at the initial classification stage – for example, a summary statistic which was the intended output for analysis of a sensitive dataset – then this re-classification may be pre-approved.

We recommend that, whenever data egress is conducted with the intent of establishing a new environment for further research, a Referee is consulted to ensure balance.

In all cases, classification of a work package at the point of egress should be done with all parties fully aware of the analytical processes which created the derived data from the initial work package. These processes should be well documented and ideally fully reproducible (e.g. as code that can be run to regenerate the exact output data from the input data).

The initial classification of a work package may be for the purpose of ingress into an initial high-tier environment to carry out anonymisation, pseudonymisation or synthetic data generation work, with the intention of making the data appropriate for treatment in a lower-tier Environment. In this case, the egress review should include validation that the anonymised, pseudonymised or synthetic data undergoes its own classification process for analysis that will be performed in the "downstream" work package. This should include a review of the anonymisation, pseudonymisation or synthetic data generation process, including all associated code.

Data sharing agreement

This should be a formal data sharing agreement as required under data protection law, drafted with the benefit of legal advice, and should be signed after the initial classification of a work package but before a dataset is received by the Turing. Where the Dataset Provider is not the owner of all the dataset(s) covered by the data sharing agreement, the agreement must specify the legal basis under which the Dataset Provider is permitted to share this data with the Turing. This agreement should include any specific commitments required from Researchers working with the dataset. The Turing has a template agreement that can be used to minimise the turnaround time and legal effort required.

The classification tier may potentially be raised from that agreed prior to data ingress, once the Investigator and Referee have had a chance to view the actual data. The classification tier for later work packages in a project may also be higher than that for the original work package, depending on the planned analysis and any additional data required. We therefore recommend that the data sharing agreement is worded to permit this.

User lifecycle

Users who wish to have access to the Environment first complete an online form certifying they understand the confidentiality requirements. An account is then created for them within the Turing Environment management system, and the user activates this.

Projects are created in the management system by a Programme Manager, and an Investigator and Project Manager assigned.

Programme Managers and Project Managers may add users to groups corresponding to specific projects or work packages through the management framework.

The Project Manager has the authority to assign Referees and Data Provider Representatives to a project or work package.

At some tiers, new Referees or members of the research team must also be approved by the Dataset Provider Representative.

Before joining a project or work package, Researchers, Investigators and Referees must agree to any additional commitments specific to that project or work package.

Users are removed from a project or work package promptly once their involvement with it ends.

Data ingress (data entering a secure Environment from an external source)

The policies defined here minimise the number of people who have access to restricted information before it is in the Environment.

Datasets must only be transferred from the Dataset Provider to the Turing after an initial classification has been completed and the data sharing agreement executed.

For lower tiers, where the Environment is accessible from the internet, standard secure data transfer mechanisms such as secure copy (SCP) and secure file transfer protocol (SFTP) may be used.

For higher tiers, all data transfer to the Turing should be via our secure data transfer process, which provides the Dataset Provider time-limited, write-only access to a dedicated data ingress volume from a specific location. Prior to access to the ingress volume being provided, the Dataset Provider Representative must provide the IP address(es) from which data will be uploaded and an email address to which a secure upload token can be sent. Once these details have been received, the Turing will open the data ingress volume for upload of data.

To minimise the risk of unauthorised access to the dataset while the ingress volume is open for uploads, the following security measures are in place:

- Access to the ingress volume is restricted to a limited range of IP addresses associated with the Dataset Provider and the Turing.
- The Dataset Provider receives a **write-only** upload token. This allows them to upload, verify and modify the uploaded data, but does not viewing or download of the data. This provides protection against an unauthorised party accessing the data, even they gain access to the upload token.
- The upload token expires after a time-limited upload window.
- The upload token is transferred to the Dataset Provider via a secure email system.

To further minimise the risk of unauthorised access to the dataset during the upload window, the Dataset Provider should take the following precautions.

- Data should always be uploaded directly into the secure volume to avoid the risk of individuals unintentionally retaining the dataset for longer than intended.
- After their dataset has been transferred, the Dataset Provider should immediately indicate that the transfer is complete. In doing so, they lose access to the data volume.

If consensus on data classification cannot be made from metadata, an initial conservative classification may be made to permit the data to be ingressed into a higher tier environment. If the final classification of a work package is lower than the initial classification, the data may be egressed from this higher tier environment Environment to a new Environment matching the final classification tier. The web management workflows should ensure that all parties have reached consensus on the classification tier at this stage before allowing analysis to begin.

Software library distributions

Maintainers of shared research computing environments face a difficult challenge in keeping research algorithm libraries and platforms up to date - and in many cases these conflict. While sophisticated tools to help with this exist, the use of independent virtual environments opens another possibility: downloading the software as needed from package repositories (such as PyPI for Python or CRAN for R), which automate the process of installing and configuring programs.

For lower tiers, with access to the internet, required software packages can be installed directly from their canonical sources on the internet. For higher tiers, without access to the external internet, this requires maintenance of full or partial mirrors (exact copies) of package repositories inside the Environment.

Use of package mirrors inside the Environment means that the set of default installed packages can be kept to a minimum, reducing the likelihood of encountering package-conflict problems (where a package can be prevented from being installed due to the presence of an existing package with the same name) and saving on System Manager time.

At the the highest tiers a subset of whitelisted packages (packages which are explicitly marked as safe) is maintained. This whitelist can be specific to the work package if required. At other tiers the full package list is mirrored, but with a short delato provide an opportunity for the wider community to catch any malicious code uploaded to the canonical package mirrors.

Storage

Which storage volumes exist in the analysis Environment?

A Secure Data volume is a read-only volume that contains the secure data for use in analyses. It is mounted as read-only in the analysis Environments that must access it. One or more such volumes will be mounted depending on how many managed secure datasets the Environment has access to.

A Secure Document volume contains electronically signed copies of agreements between the Data Provider and the Turing.

A Secure Scratch volume is a read-write volume used for data analysis. Its contents are automatically and regularly deleted. Users can clean and transform the sensitive data with their analysis scripts, and store the transformed data here.

An Output volume is a read-write area intended for the extraction of results, such as figures for publication.

The Software volume is a read-only area which contains software used for analysis.

A Home volume is a smaller read-write volume used for local programming and configuration files. It should not be used for data analysis outputs, though this is enforced only in policy, not technically. Configuration files for software in the software volume point to the Home volume.

User device networks

Our network security model distinguishes three dedicated research networks for user devices.

- The open internet (any network outside a partner institution)
- An Institutional network
- A Restricted network

An Institutional network corresponds to a network managed by a partner institution. Guest access may be permitted on such networks (e.g. eduroam), but these guests should be known users. Access to Environments can be restricted such that access is only allowed by devices which are connected to a particular set of Institutional networks. However, it is assumed that a wide segment of the research community can access these networks. This access may also be remote for authorised users (for example, via VPN).

A Restricted network corresponds to a network managed by a partner institution that can support additional controls such as restricting access to a narrower set of users, devices or locations. Access to Environments can be restricted such that access is only allowed by devices which are connected to a particular set of Restricted networks. Access to a particular Environment may be permitted from multiple Restricted networks at multiple partner organisations. This can permit users from multiple organisations to access an Environment, as well permitting users to access the Environment while away from their home institution at another partner institution. However, remote access to a Restricted network (for example via VPN) is not permitted.

At higher tiers Environment firewall rules permit access only from network IP ranges corresponding to specific Institutional and Restricted networks approved for the Environment. Note that these restrictions on networks that can access Environments relate to inbound connectivity only. Separate controls determine whether outbound connections can be made from an Environment and whether inbound connections are permitted directly to resources within the Environment or must be made via a secure access node. In addition to these netwrok level restrictions, users must additionally authenticate to the Environment in order to access it.

Physical security

Some data requires a physical security layer around not just the data centre, but the physical space users are in when they connect to it.

We distinguish three levels of physical security for research spaces:

- Open research spaces
- Medium security research spaces
- High security research spaces

Open research spaces include university libraries, cafes and common rooms.

Medium security research spaces control the possibility of unauthorised viewing. Card access or other means of restricting entry to only known researchers (such as the signing in of guests on a known list) is required. Screen adaptations or desk partitions can be adopted in open-plan spaces if there is a high risk of "visual eavesdropping".

Secure research spaces control the possibility of the researcher deliberately removing data. Devices will be locked to appropriate desks, and neither enter nor leave the space. Mobile devices should be removed before entering, to block the 'photographic hole', where mobile phones are used to capture secure data from a screen. Only researchers associated with a secure project have access to such a space.

Firewall rules for the Environments can permit access only from Restricted network IP ranges corresponding to these research spaces.

User Devices

What devices should researchers use to connect to the Environment?

We define two types of devices:

- Managed devices
- Open devices

Managed devices

Managed devices are devices provided by a partner institution on which the user does not have administrator/root access, with the device instead administered by the institution's IT team.

Managed devices could be provided by the Turing, or one of the partner organisations for a work package.

They have an extensive suite of research software installed.

This includes the ability to install packages for standard programming environments without the need for administrator access.

Researchers can compile and run executables they code in User Space (the portion of system memory in which user processes run).

Open Devices

These include personal devices such as researcher-owned laptops, but also include devices provided by a partner institution where the user has administrator/root access.

These devices permit the easy use of a wider range of software than managed devices, as well as easier access to peripheral hardware.

However, such devices should not be used to access the highest tier Environments.

Firewall rules for the higher tier Environments can permit access only from Restricted network IP ranges that only permit managed devices to connect.

Software ingress (software entering a secure Environment from an external source)

The base data science virtual machine provided in the secure analysis Environments comes with a wide range of common data science software pre-installed. Package mirrors also allow access to a wide range of libraries for the programming languages for which package mirrors are provided (currently Python and R).

For other languages for which no package mirror is provided, or for software which is not available from a package repository, an alternative method of software ingress must be provided. This includes custom researcher-written code not available via the package mirrors (e.g. code available on a researcher's personal or institutional Github repositories).

For lower tier environments, the data science virtual machine has outbound access to the internet and software can be installed in the usual manner by either a normal user or an administrator as required.

For higher tier environments, the following software ingress options are available.

Installation during virtual machine deployment

Where requirements for additional software are known in advance of the data science virtual machine being deployed to a secure analysis Environment, the additional software sucan be installed during deployment. In this case, software installation is performed while the virtual machine is outside of the Environment with outbound internet access available, but no access to any project data. Once

the additional software has been installed, the virtual machine is ingressed to the Environment via a one-way airlock.

Installation after virtual machine deployment

Once a virtual machine has been deployed into a secure analysis Environment, it cannot be moved outside of the Environment, as is has had access to the data in the Environment and therefore represents an unauthorised data egress risk. As higher tier Environments do not have access to the internet, any additional software required must be brought into the Environment in order to be installed.

Software is ingressed in a similar manner as data, using a software ingress volume:

- In **external mode** the researcher is provided temporary **write-only** access to a software ingress volume.
- Once the Researcher transfers the software source or installation package to this volume, their access is revoked and the software is subject to a level of review appropriate to the Environment tier.
- Once any required review has been passed, the software ingress volume is switched to **internal mode**, where it is made available to Researchers within the analysis Environment with **read-only** access.
- For software that does not require administrative rights to install, the Researcher can then install the software or transfer the source to a version control repository within the Environment as appropriate.
- For software that requires administrative rights to install, a System Manager must run the installation process.

The choices

Having described the full model, processes, and lifecycles, we can now enumerate the list of choices that can be made for each Environment. These are all separately configurable on an environment-by-environment basis. However, we recommend the following at each tier.

Software installation

At Tier 3 and above, package mirrors (copies of external repositories inside the secure Environment) should include only white-listed software packages.

At Tier 2, package mirrors should include all software packages.

At Tier 2 and above, all software not available from a package mirror must be installed either at the time the analysis machine is first deployed or by ingressing the software installer as data, with an associated ingress review.

At Tier 1 and 0, all software installation should be from the internet.

Inbound connections

At Tier 2 and above, analysis machines and other Environment resources are not accessible directly from client devices. Instead, secure "access nodes" provide secure web-based remote desktop facilities used to indirectly access the analysis Environments (e.g. Microsoft Remote Desktop Services).

At Tier 1 and 0, analysis machines and other Environment resources are directly accessible from client devices.

At Tier 3 and above Environment access nodes are only available from approved Restricted networks.

At Tier 2 Environment access nodes are only be accessible from approved Institutional networks.

At Tier 1 and 0 Environment resources are accessible from the open internet

Outbound connections

At Tier 2 and above no connections are permitted from the Environment private network to the internet or other external resources.

At Tier 1 and 0 the internet is accessible from inside the Environment.

Data ingress

At Tier 2 and above, the high-security data transfer process is required (i.e. write only access from particular locations for a limited time).

At Tier 1 and 0 the use of standard secure data transfer processes (e.g. SCP/SFTP) may be permitted.

Data egress

At Tier 3 and above, the Data Provider Representative, Investigator and Referee are all required to sign off all egress of data or code from the Environment.

At Tier 2, only the Investigator and Referee are required to review and approve all egress of data or code from the Environment.

At Tier 1 and 0 users are permitted to copy out data when they believe their local device is secure, with the permission of the Investigator.

Refereeing of classification

Independent Referee scrutiny of data classification is required when the initial classification by the Investigator and Data Provider Representative is Tier 2 or higher.

Two factor authentication

At Tier 2 and higher, two factor authentication is required to access the Environment.