Creating a Highly Restricted Data Service

Gary Leeming, University of Manchester

MANCHESTER
1824

What is Restricted Data?

- Special category data under GDPR (Managing personally identifable data should be BAU?)

- Data defined as sensitive/secret/restricted by the data owner

- Definition usually risk-based

- UoM has 3 categories: Unrestricted, Restricted and Highly Restricted

Move from Compliance-based to Commitment-based Culture

“We also believe in building a culture of security. Employees are your first line of defense; **none of them leave their houses in the morning without locking the door**, and none of them should leave their worksites at night without locking their computer and sensitive documents away. If you really want your employees to be your first line of defense, you need to teach them how, and you must be readily available, helpful, and responsive when they call.”

- Stackpole, Bill, and Eric Oksendahl. Security strategy: from requirements to reality. Auerbach Publications, 2010.

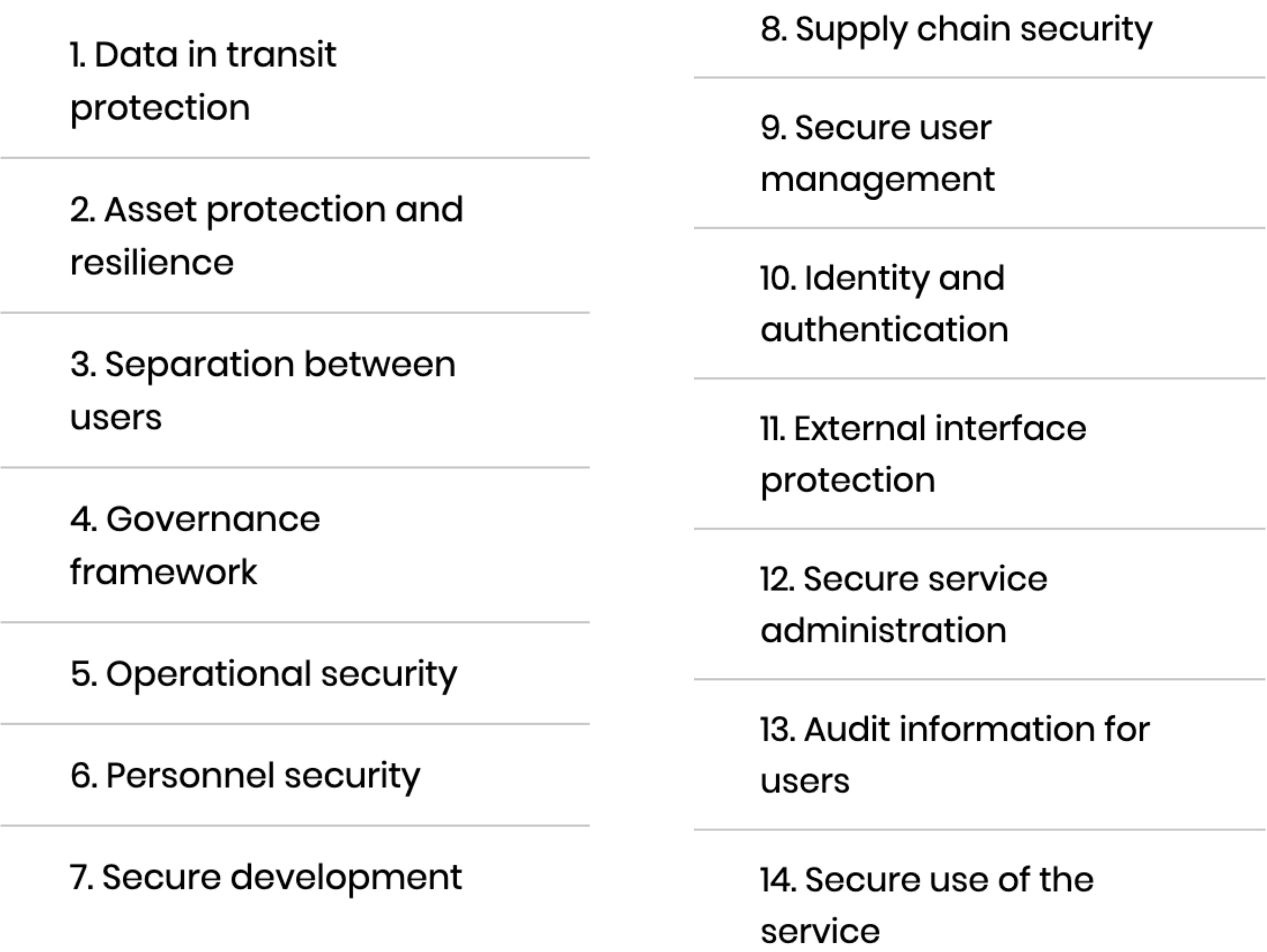
Information and Security Standards and Guidance

- ISO27001
- NHS Digital Data Security and Protection Toolkit
- CyberEssentials (plus)
- NCSC 14 Principles
- Five Safes (UK Data Service)
- STRIDE
- CVSS
- Mitre Att&ck

Cyber Essentials

- Firewall
- Secure Configuration
- Access Control
- Malware Protection
- Patch Management

NCSC 14 principles



1. Data in transit protection
2. Asset protection and resilience
3. Separation between users
4. Governance framework
5. Operational security
6. Personnel security
7. Secure development
8. Supply chain security
9. Secure user management
10. Identity and authentication
11. External interface protection
12. Secure service administration
13. Audit information for users
14. Secure use of the service

Turing proposal

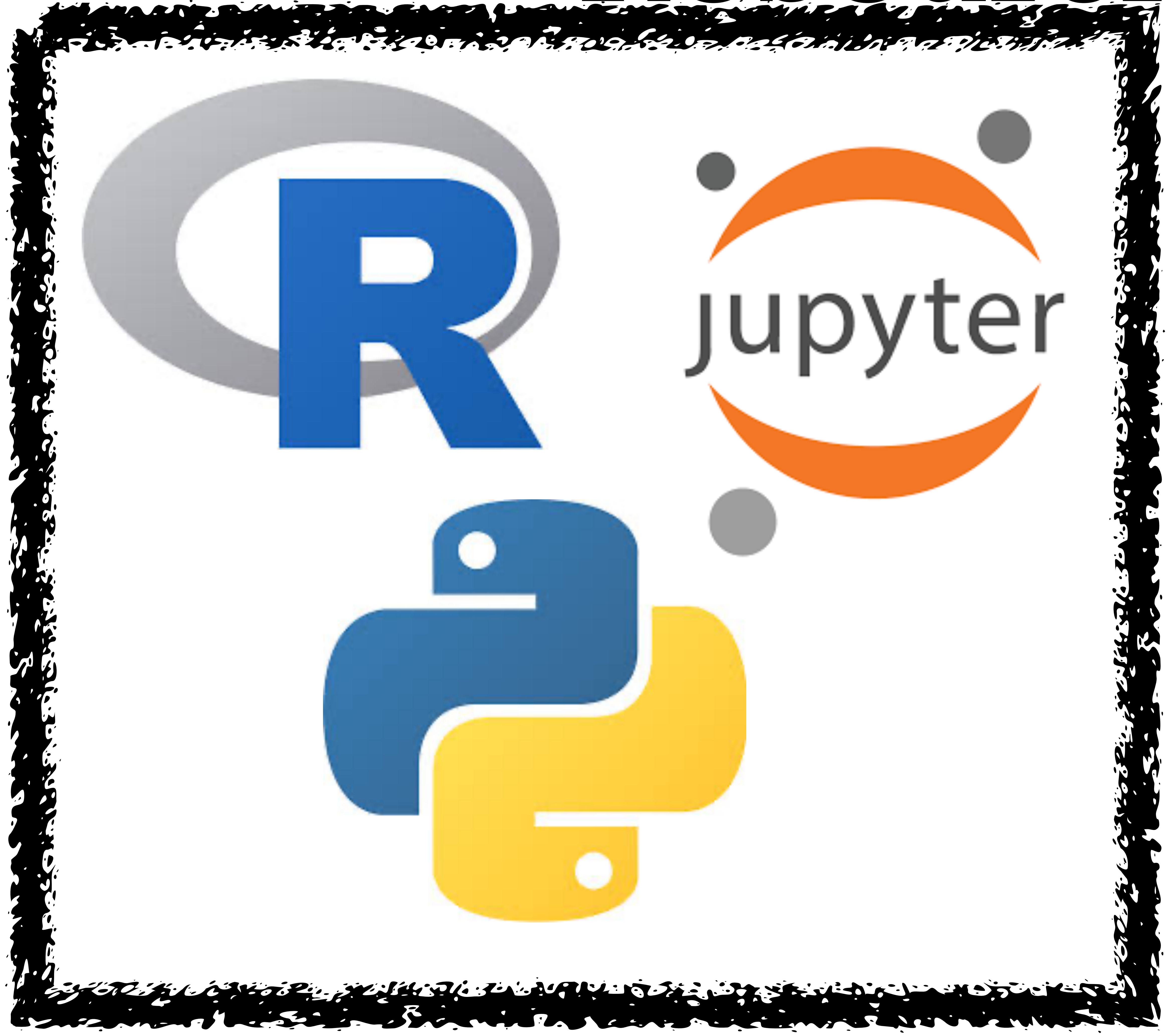
Turing Classification	University Classification	Risk (Reputation, legal, commercial, political)	Examples
Tier 0	Unrestricted	No risk if accessed by non-authorised actor	Public dataset, published paper
Tier 1	Unrestricted	Low risk if accessed by non-authorised actor	Research output intended for publication, non-personal research data
Tier 2	Restricted	Medium risk if accessed by non-authorised actor.	CPRD data extract, low-risk commercial in confidence data, low-risk IP
Tier 3	Highly Restricted	High risk if accessed by non-authorised actor, low-medium risk of attack	Detailed but anonymised hospital data, politically sensitive data, personal data where low risk of harm to the data subject
Tier 4	Highly Restricted	High risk if accessed by non-authorised actor, high risk of attack	Highly sensitive data, e.g. nuclear or pharmaceutical industry, personal data where high risk of harm to the data subject, e.g. refugee data.

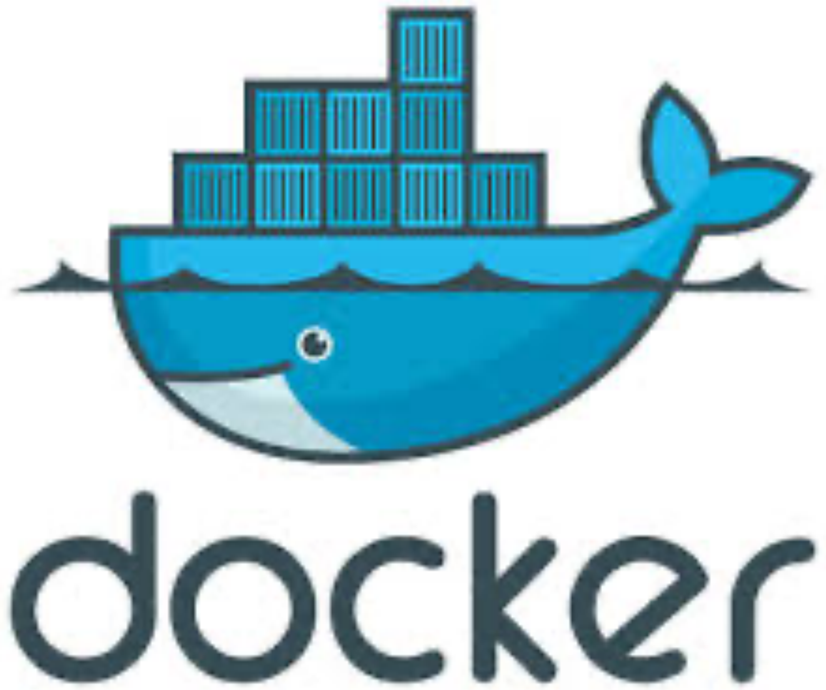
Arenas, D., Atkins, J., Austin, C., Beavan, D., Egea, A. C., Carlysle-Davies, S., ... & Forrest, O. (2019). Design choices for productive, secure, data-intensive research at scale in the cloud. *arXiv preprint arXiv:1908.08737*.

MANCHESTER
1824

What is Research?

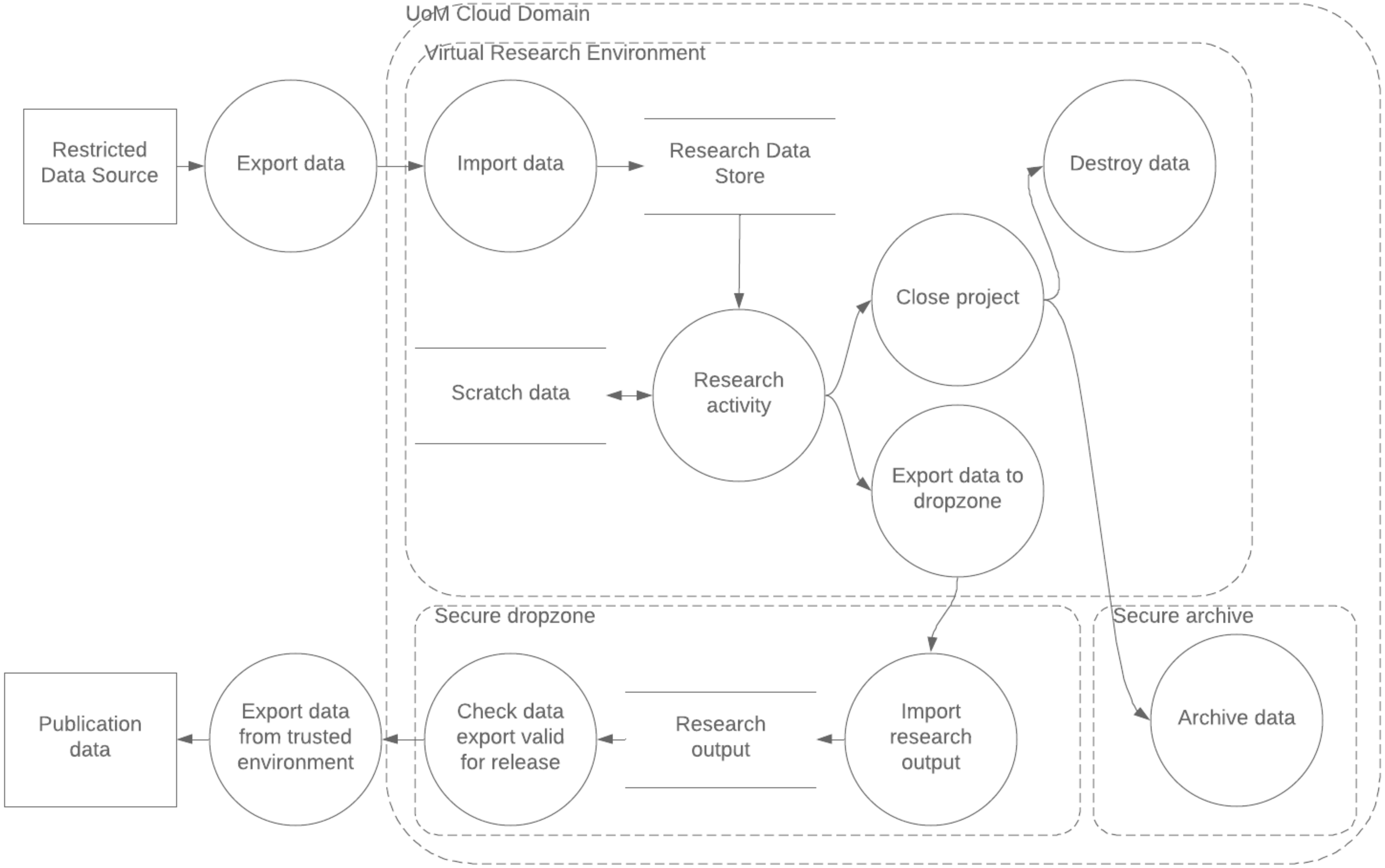
What is Restricted Data Research?





MANCHESTER
1824

Restricted Data VRE



Threat Actors

Error by user	Malicious accounts person	External unskilled hacker
Malicious ex-user	Error by accounts person	Student hacker
Prof or senior staff member who "just wants to get it done"	Malicious hosting company worker	Knowledgeable external attacker
User who has been told by prof to "just get it done"	Error by hosting company worker	Malicious software developer
User who wants to try new plug in/ software to "just get it done"	Organised crime	Error by software developer
Malicious network team member	Nation state (UK, US, China, Russia, Iran)	Malicious package manager
Error by network team member	Competitor (for research or for students)	Error by package manager
Malicious first-line support	Commercial espionage	Error by retention policy definition
Error by first-line support	Malicious ex-network team member	Malicious time service maintainer
Social engineer	Malicious IT manager	Malicious DNS maintainer

MANCHESTER
1824

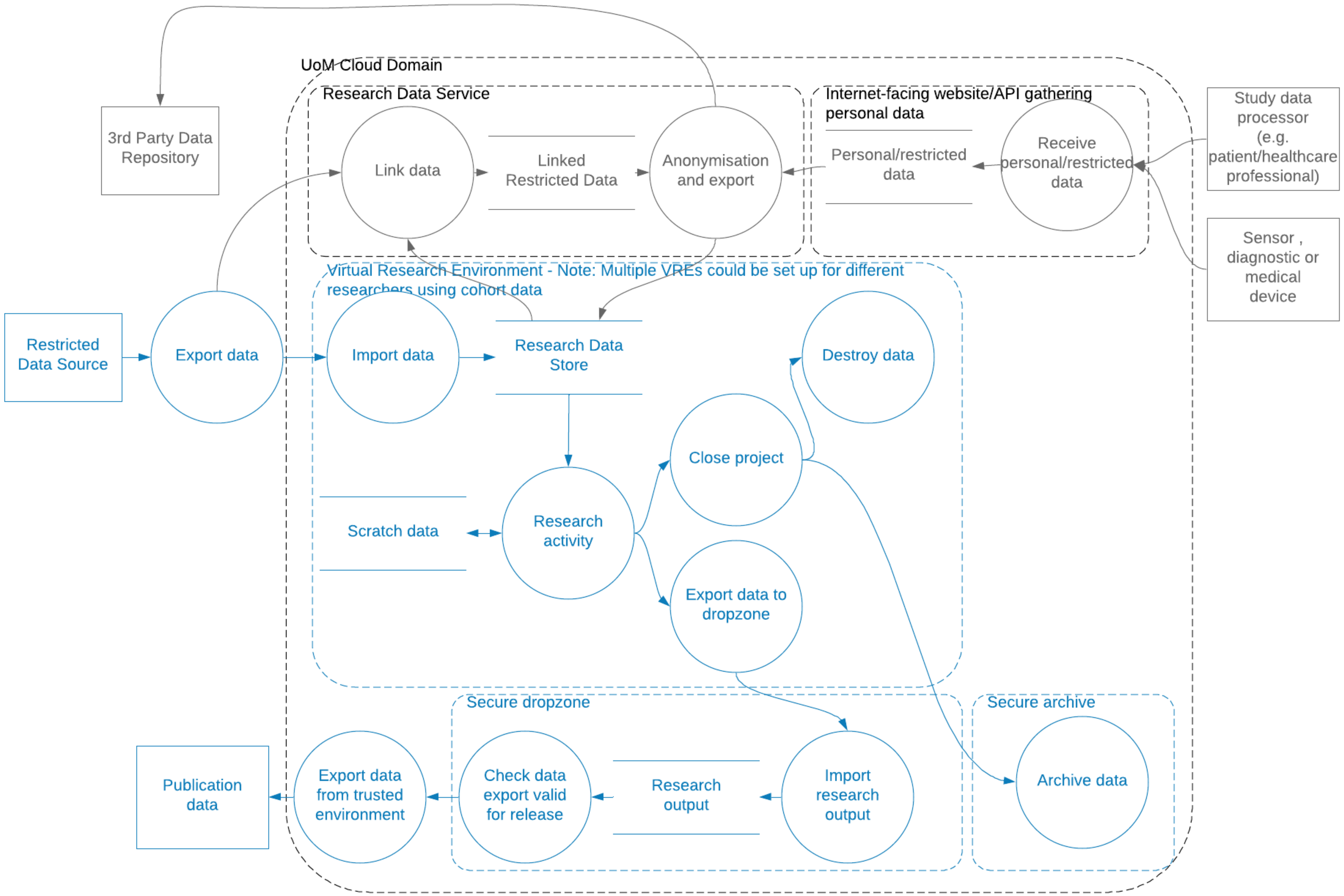
STRIDE Analysis

Component	STRIDE	Example Risk Description	Mitigation
Internet-face website/API			
Receive Personal/ Restricted data	Spoofing	Access could be attempted by Threat actor trying to log on.	Ensure combination of appropriate controls are in place, e.g. whitelisting IP addresses, certificates and 2FA.
	Tampering	Data or input could be falsified, devices could be tampered with to change the results of the research	Data can be checked on receipt to ensure that it is appropriate. Project will need to think about controls and audit of devices
	Repudiation	User or service could deny that an acitivity had taken place.	All activity should be logged including location, time and user account details.
	Information Disclosure	Data store holds restricted data. If access is gained then there could be information disclosure.	The application should have appropriate controls to prevent disclosure. Identifiable data should not be stored within the application
	Denial of Service	As the API is publicly available there could be an opportunity to deny access to the service	Project will need to ensure that data is not lost as it is submitted in this circumstance. Data should not be read unless necessary.
	Elevation	In the case of a website application there could be a risk that a user is able to view data that should not be able to	The management and design of the website must restirct the ability to modify priviliges without validation from the data owners

More Research Requirements

- Sensitive data from aircraft black boxes and airports received regularly for analysis
- Drug registry data collected every six months
- Geodata linked to medical records to be accessed by external researchers
- Good Clinical Practice

Restricted Data Virtual Service Environment (VSE)

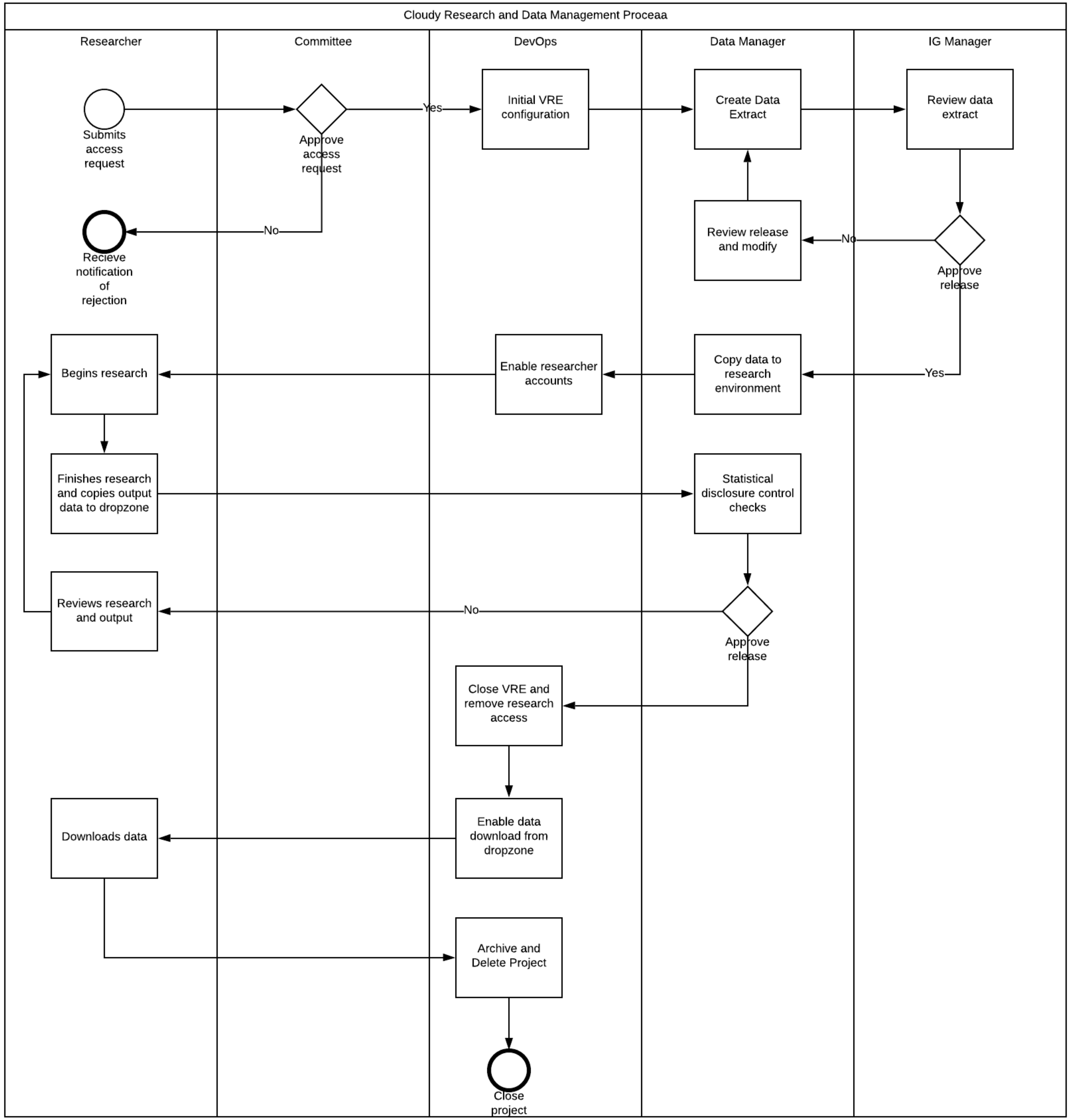

UoM Cloud Domain
Research Data Service
Link data
Linked Restricted Data
Anonymisation and export
Internet-facing website/API gathering personal data
Personal/restricted data
Receive personal/restricted data
Study data processor (e.g. patient/healthcare professional)
Sensor , diagnostic or medical device
3rd Party Data Repository
Restricted Data Source
Export data
Virtual Research Environment - Note: Multiple VREs could be set up for different researchers using cohort data
Import data
Research Data Store
Destroy data
Close project
Scratch data
Research activity
Export data to dropzone
Secure dropzone
Secure archive
Publication data
Export data from trusted environment
Check data export valid for release
Research output
Import research output
Archive data

More requirements

- Can I work from home?
- Tell me what to put in my data management plan to get the grant
- [3rd Party Organisations] over-classifies their data

Managing Data Flows

- Security is about process not technology
- Different parts of the process can be at different levels of risk
- Need to ensure that solution is appropriate to the risk at that point and can move between process requirements, e.g. Anonymisation



Cloudy Research and Data Management Proceaa

Researcher

Committee

DevOps

Data Manager

IG Manager

Submits access request

Approve access request

Yes

No

Recieve notification of rejection

Initial VRE configuration

Create Data Extract

Review data extract

Approve release

No

Review release and modify

Yes

Copy data to research environment

Enable researcher accounts

Begins research

Finishes research and copies output data to dropzone

Statistical disclosure control checks

Approve release

No

Reviews research and output

Close VRE and remove research access

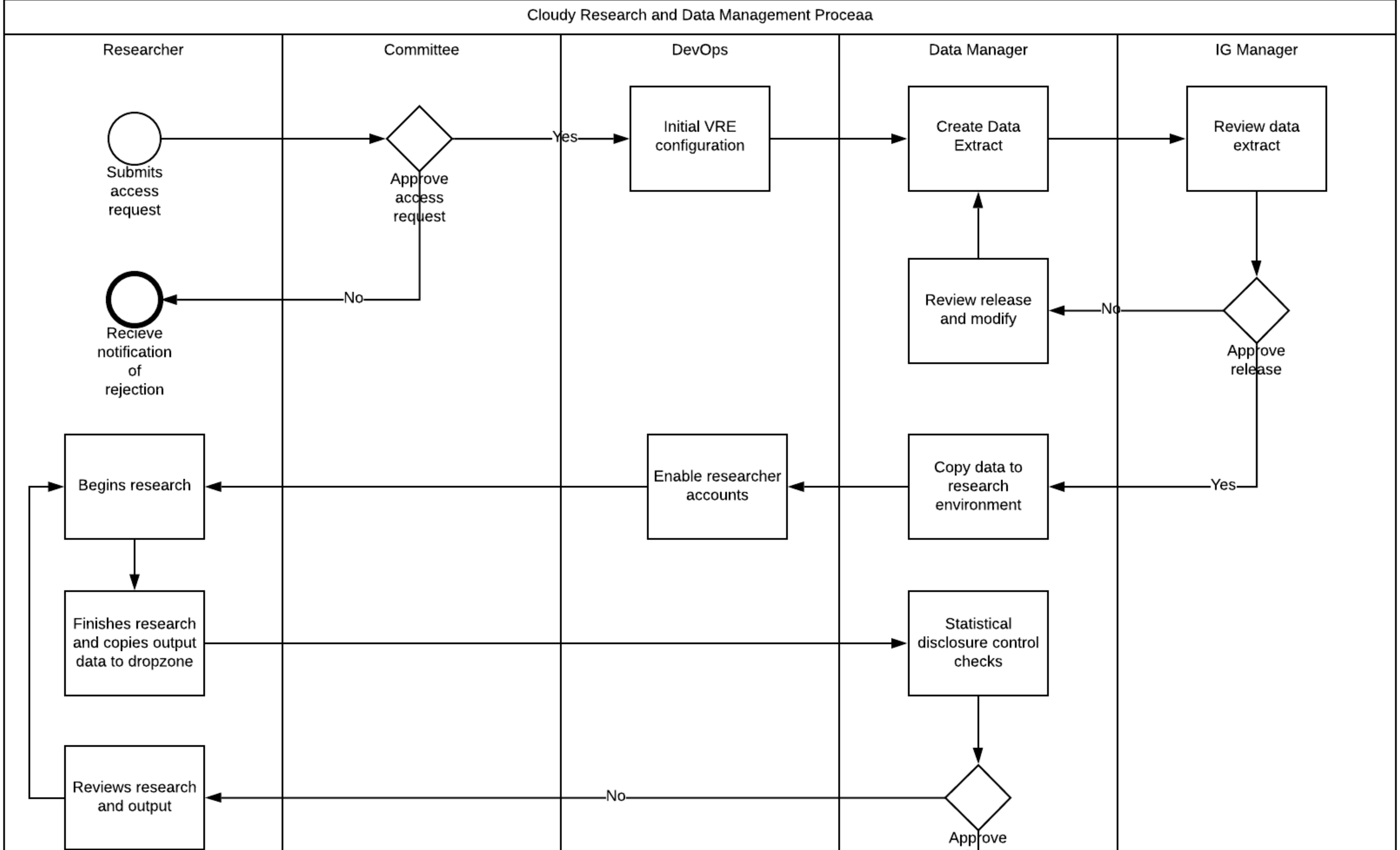
Enable data download from dropzone

Downloads data

Archive and Delete Project

Close project

MANCHESTER
1824


Cloudy Research and Data Management Proceaa
Researcher
Committee
DevOps
Data Manager
IG Manager
Submits access request
Approve access request
Yes
Initial VRE configuration
Create Data Extract
Review data extract
No
Recieve notification of rejection
Review release and modify
No
Approve release
Begins research
Enable researcher accounts
Copy data to research environment
Yes
Finishes research and copies output data to dropzone
Statistical disclosure control checks
Reviews research and output
No
Approve

Key Features

- Management platform
- Secure VRE templates
- Secure Virtual Service Environment (VSE) templates

Services

- Configuration of environments
- Deployment and testing
- Key management - Encryption keys, API keys, data identifiers
- IDAM - User and role management
- Software/VM Repo for approved images
- System health - patching, load
- Security Monitoring & Vulnerability Management
- Network configuration - Firewalls and subnets, no public internet, encryption in transit
- Disaster Recovery

Challenges

- Usual cloud challenges - Supplier management, 3rd party resellers, cost/contract management, etc.
- Identity and roles
- DevOps model of services
- Ingress of software & scripts
- Research governance and finance process integration
- Fixed regions
- Serverless
- Making it easy to use….

Benefits

- Transparency of costs
- More consistent controls
- Better compliance and visibility of risk
- Updates and management of software
- Access to variety of compute and storage
- Collaboration opportunities
- Reproducibility and Audit

Proof of Concepts

- Research access to cohort data
- IoT data flows
- Anonymisation of data