Open science and Trusted Research Environments

Dundee Data Meetup 26 November 2024

Simon Li

@manics

@penguinoops.bsky.social

Health Informatics Centre

University of Dundee

Who am I?

Research infrastructure engineer at the **Health Informatics Centre**, University of Dundee

Also open-source contributor to JupyterHub and other random projects. Previously worked for the Open Microscopy Environment.

The Health Informatics Centre

The vision:

> To advance data science and its community, simplifying access to sensitive data whilst maintaining security as a global leader in open, reproducible and scalable research platforms

https://www.dundee.ac.uk/hic

Health Informatics Centre

Overview

1. Trusted Research Environments: what are they, why do we need them?
2. Open science: The relevance of open science to TREs
3. What do the public think?

Trusted Research Environments

What are they? Secure workspaces where you can analyse sensitive data with minimal risk of data being leaked.

Health Informatics Centre

Sensitive data?

E.g. Personal data, names/addresses, healthcare records, financial statements, commercial secrets, raw census data.

I'll mostly talk about healthcare data (HIC)

Sensitive or "Special category" data

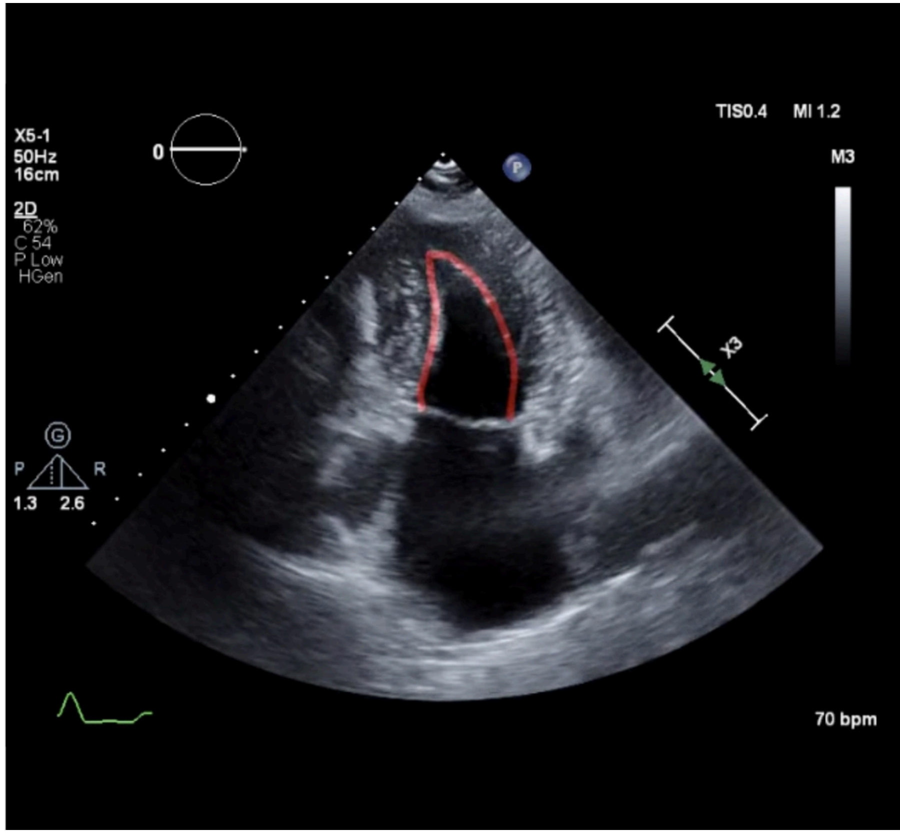
> - personal data revealing racial or ethnic origin
> - personal data revealing political opinions
> - personal data revealing religious or philosophical beliefs
> - personal data revealing trade union membership
> - genetic data
> - biometric data (where used for identification purposes)
> - data concerning health
> - data concerning a person’s sex life
> - data concerning a person’s sexual orientation

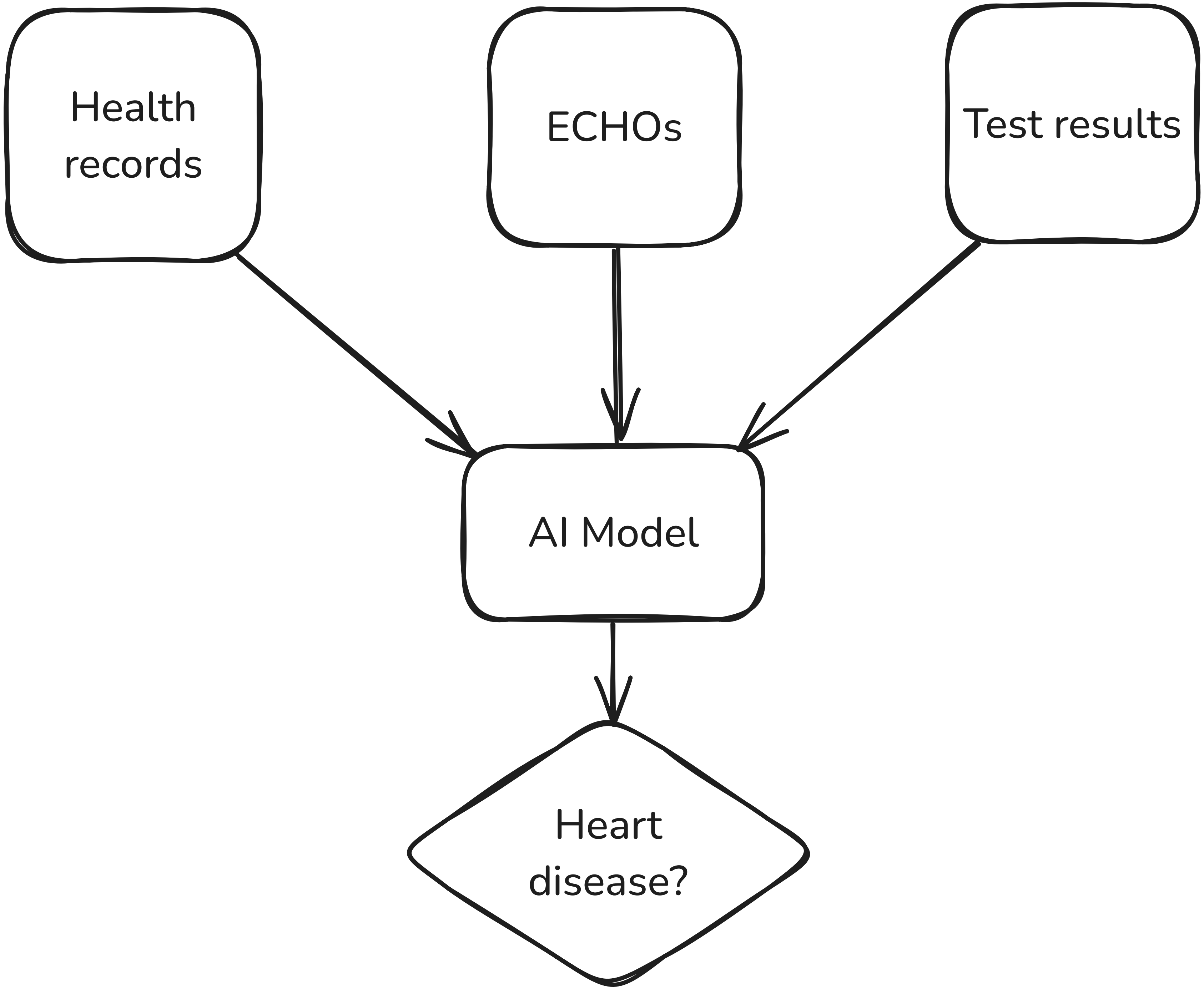
https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/special-category-data/

Example: Can undiagnosed heart failure be detected from routine medical records?

Artificial Intelligence has the capability to detect people living with the risk of experiencing heart failure, new University of Dundee research has discovered.

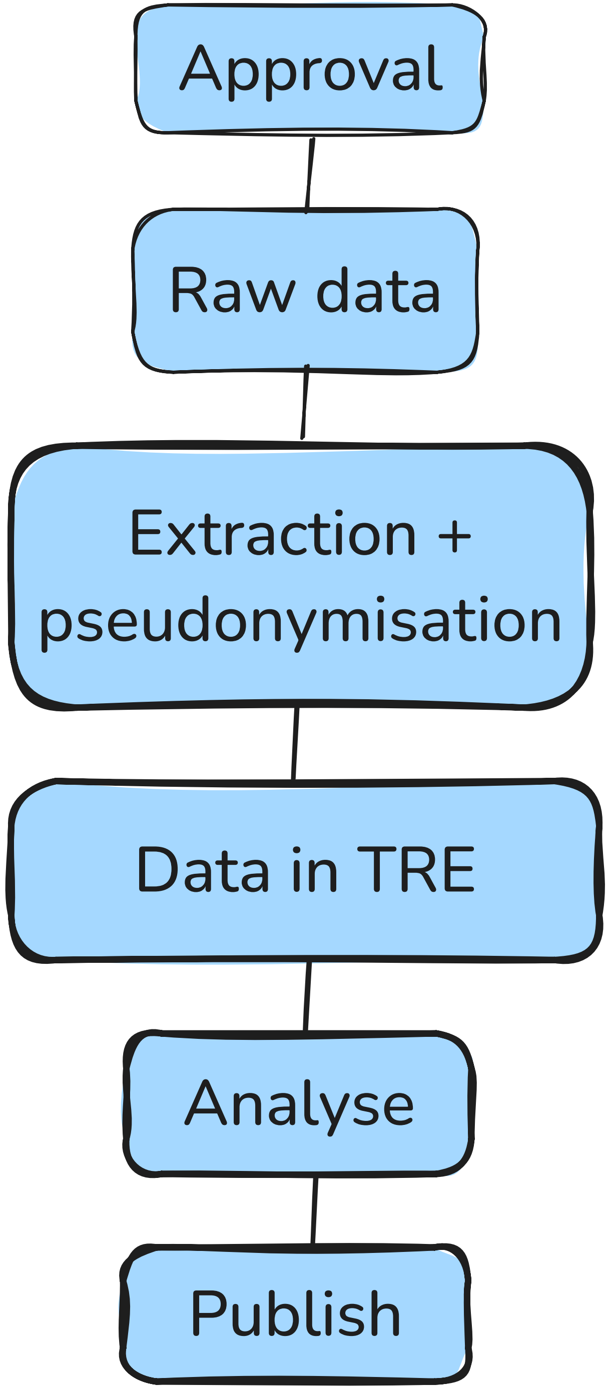
https://www.dundee.ac.uk/stories/heart-failure-risk-detected-ai-pioneering-study





Health Informatics Centre

A very rough workflow



Approval

Raw data

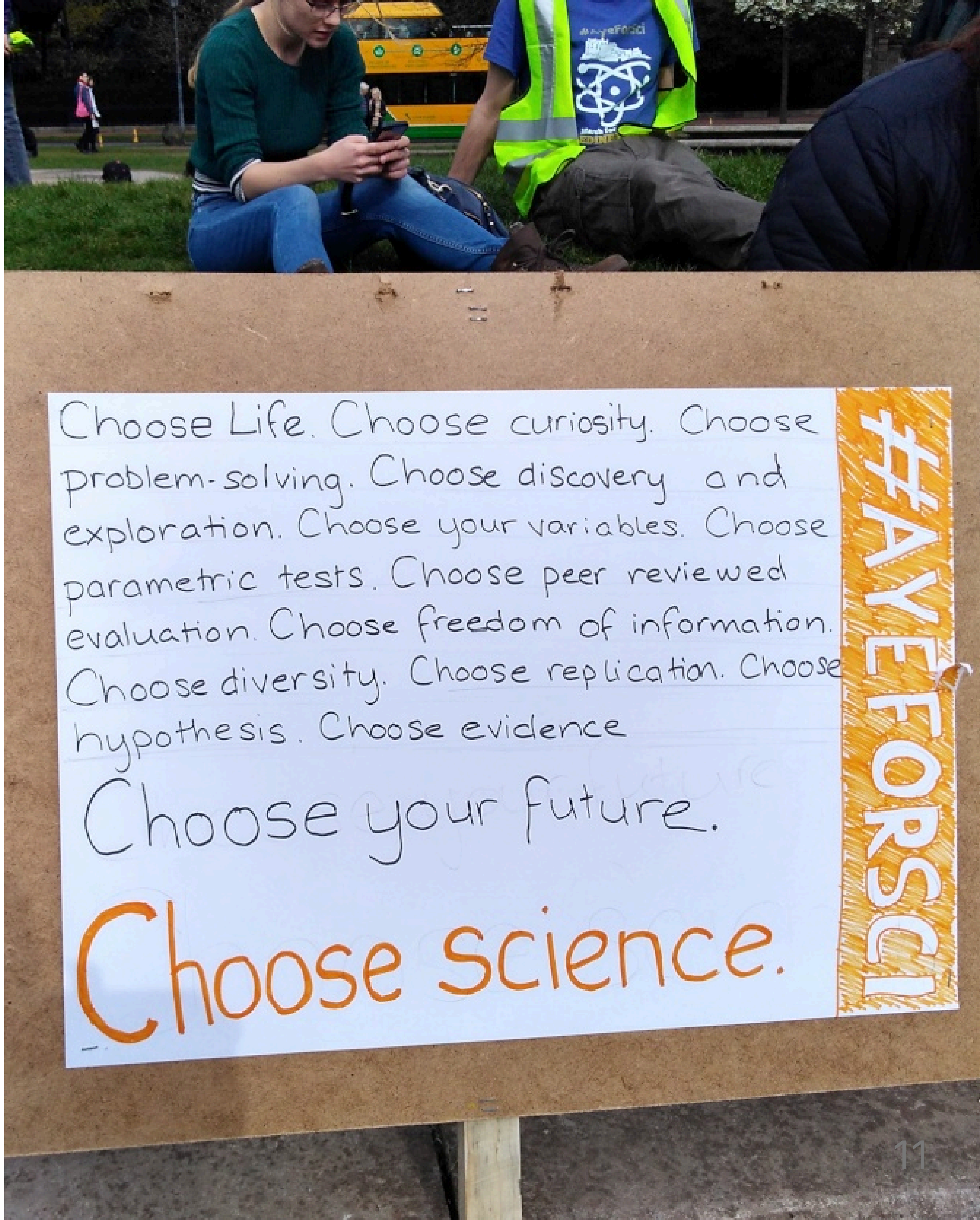
Extraction + pseudonymisation

Data in TRE

Analyse

Publish

1: Obtain approval and funding for the research project

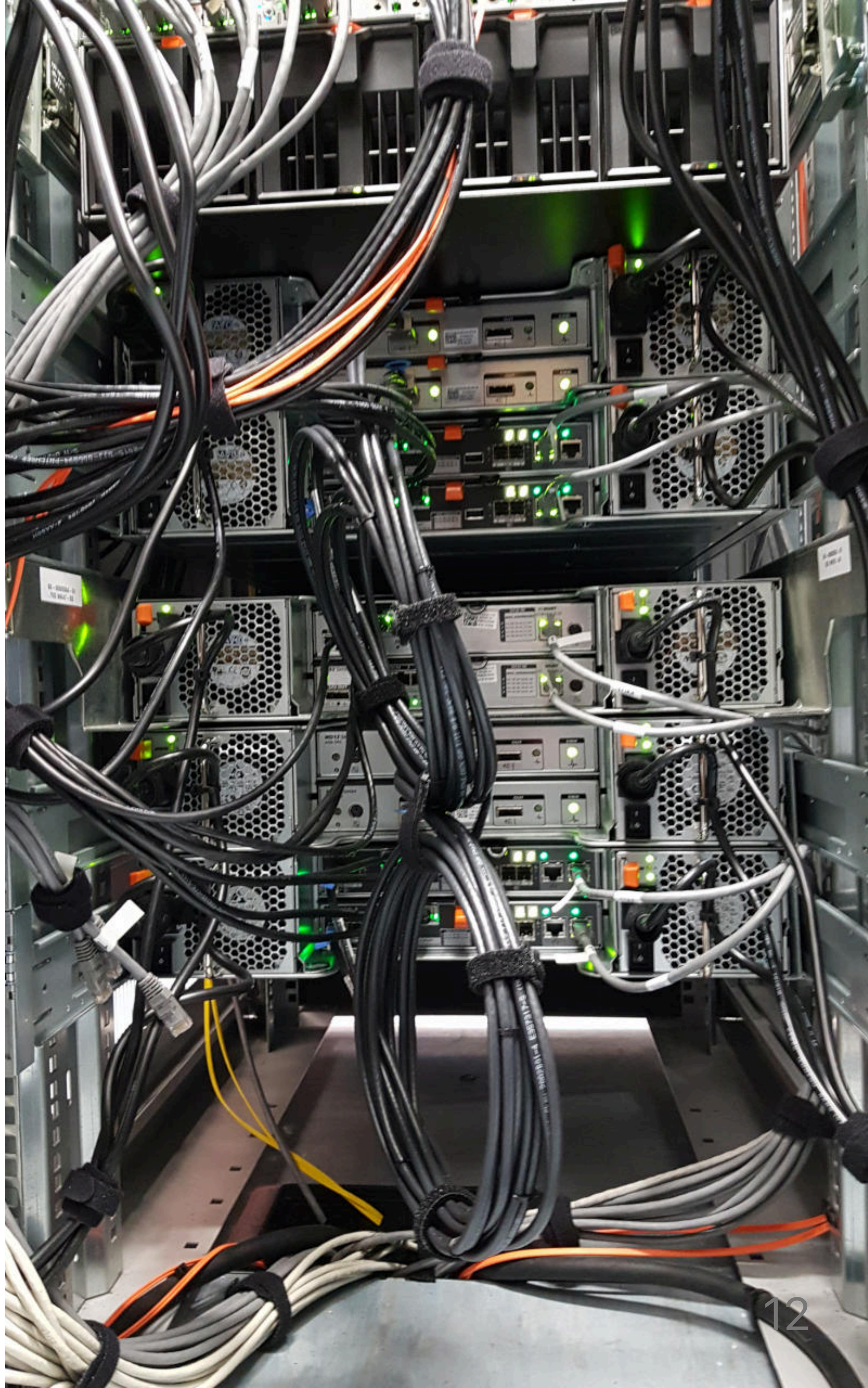

Choose Life. Choose curiosity. Choose problem-solving. Choose discovery and exploration. Choose your variables. Choose parametric tests. Choose peer reviewed evaluation. Choose freedom of information. Choose diversity. Choose replication. Choose hypothesis. Choose evidence
Choose your future.
Choose science.
#AYEFORSCI

2. What raw data do we have?

Sensitive patient data, very few people have access to it.

- Echocardiograms (heart imaging scans), and other medical imaging
- Electronic health records across possibly multiple databases- medical diagnoses, prescription records, going back several years

Researchers ***never*** have access to this data.



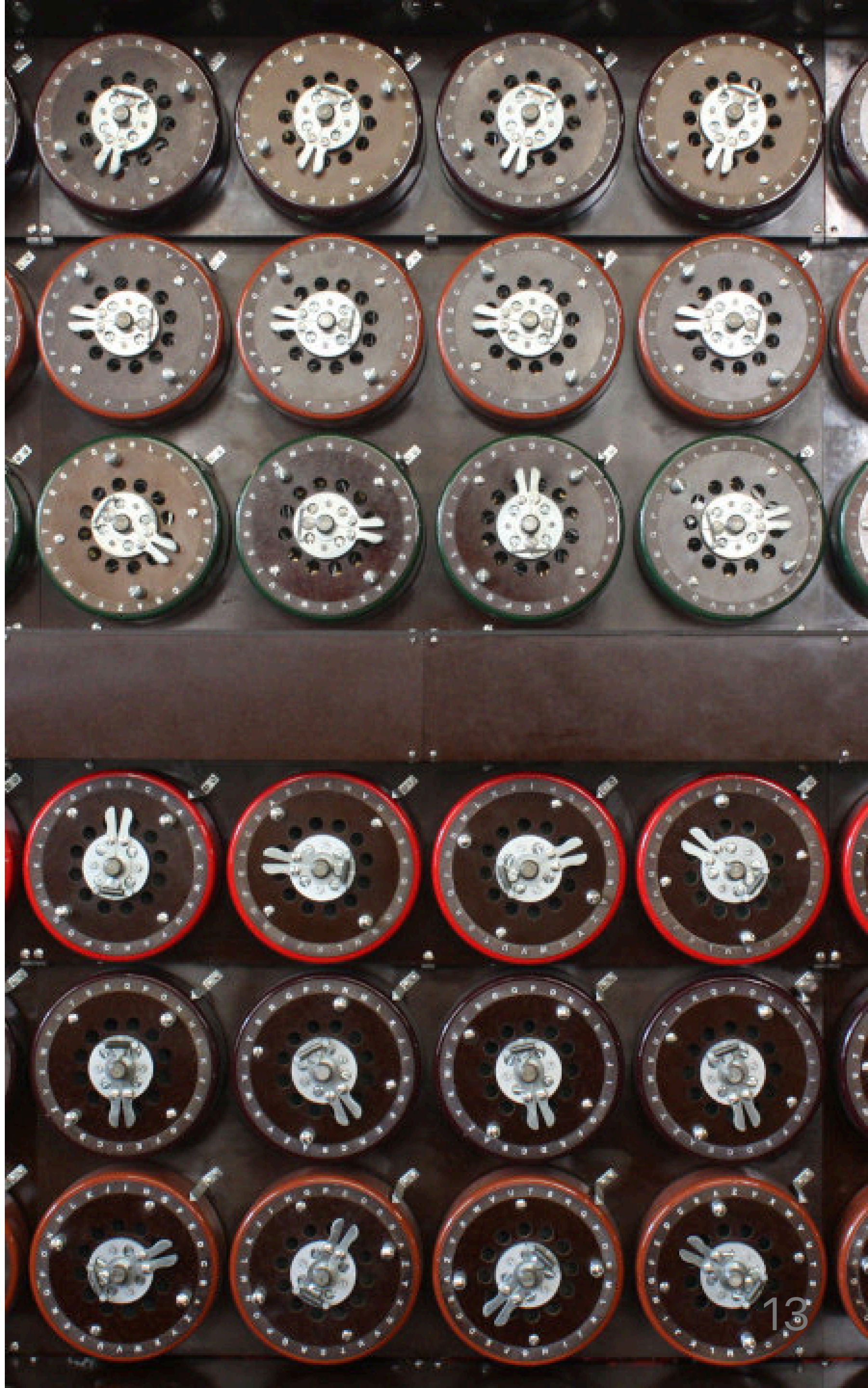
3. Extraction and Pseudonymisation of data

Data minimisation: only extract a subset of records and fields

Identifying information is removed from records

- True anonymisation is almost impossible: ***pseudonymised data***
- Original information might be recoverable by cross-referencing with another dataset.

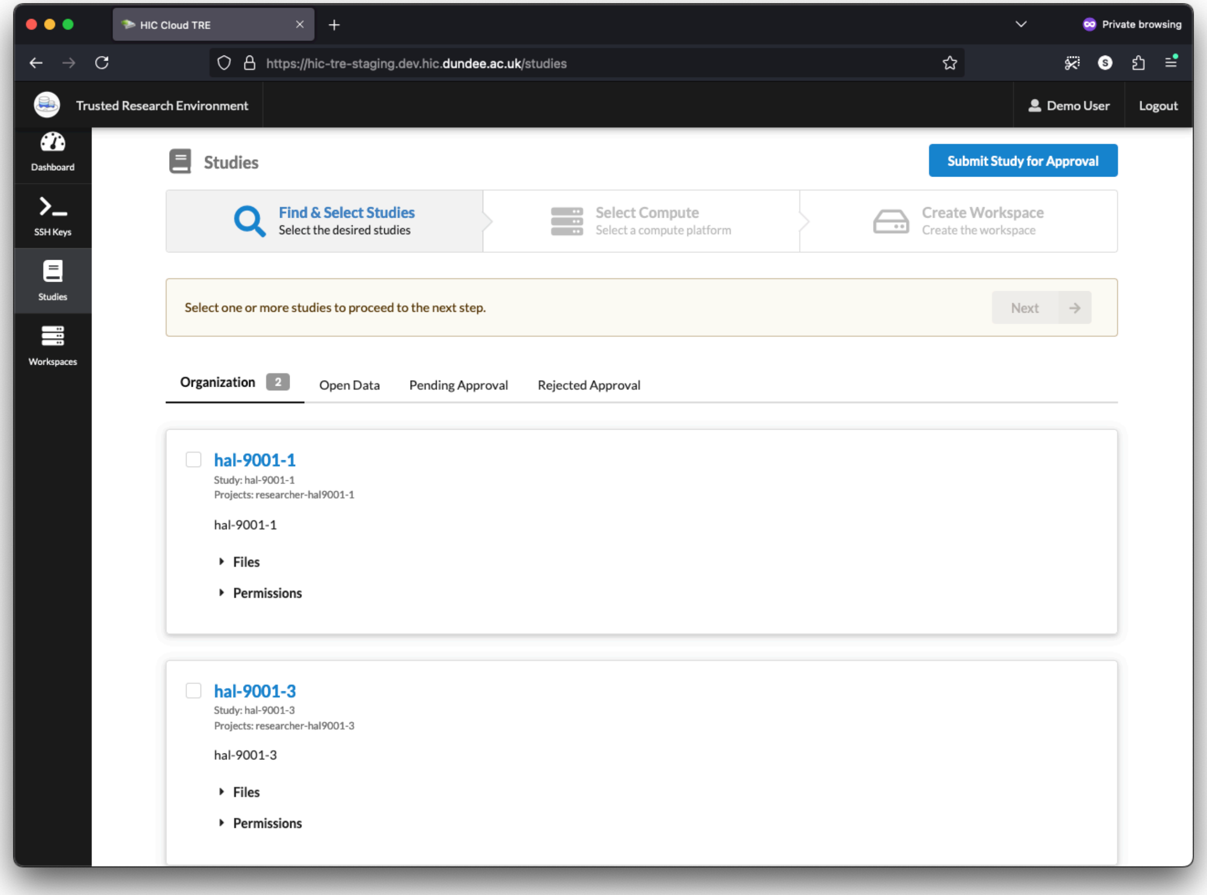
Pseudonymised data can still be classed as personal data.



4. Data made available to researchers in TRE

Highly restricted compute environment:

- Prevents re-identification through cross-linking with other datasets
- Ensures sensitive data isn't copied out
- No worries about laptops being left on a train



Select one or more studies to proceed to the next step.

□ hal-9001-1
Study: hal-9001-1
Projects: researcher-hal9001-1

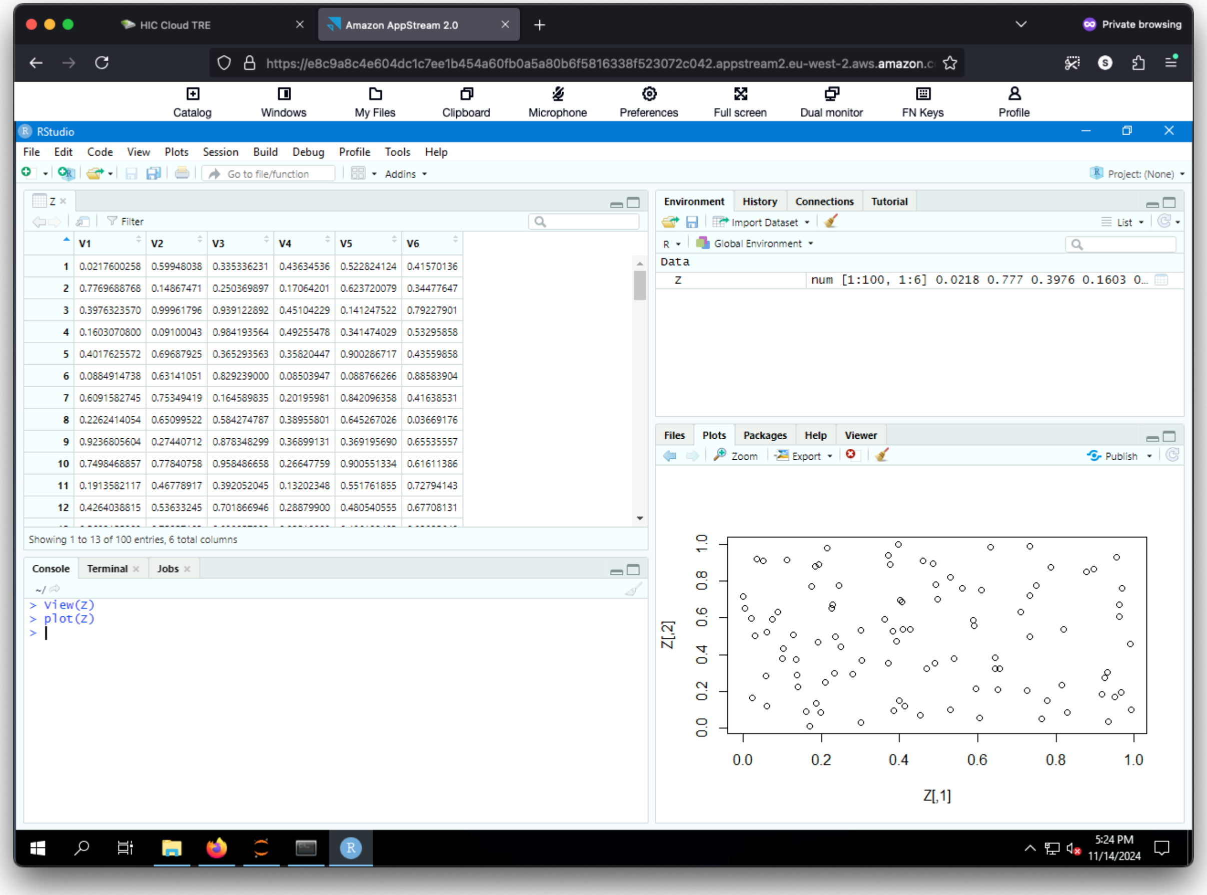
hal-9001-1

□ hal-9001-3
Study: hal-9001-3
Projects: researcher-hal9001-3

hal-9001-3

5. Researchers analyse data

Do past medical records contain information that can predict future heart failure?



	V1	V2	V3	V4	V5	V6
1	0.0217600258	0.59948038	0.335336231	0.43634536	0.522824124	0.41570136
2	0.7769688768	0.14867471	0.250369897	0.17064201	0.623720079	0.34477647
3	0.3976323570	0.99961796	0.939122892	0.45104229	0.141247522	0.79227901
4	0.1603070800	0.09100043	0.984193564	0.49255478	0.341474029	0.53295858
5	0.4017625572	0.69687925	0.365293563	0.35820447	0.900286717	0.43559858
6	0.0884914738	0.63141051	0.829239000	0.08503947	0.088766266	0.88583904
7	0.6091582745	0.75349419	0.164589835	0.20195981	0.842096358	0.41638531
8	0.2262414054	0.65099522	0.584274787	0.38955801	0.645267026	0.03669176
9	0.9236805604	0.27440712	0.878348299	0.36899131	0.369195690	0.65535557
10	0.7498468857	0.77840758	0.958486658	0.26647759	0.900551334	0.61611386
11	0.1913582117	0.46778917	0.392052045	0.13202348	0.551761855	0.72794143
12	0.4264038815	0.53633245	0.701866946	0.28879900	0.480540555	0.67708131

Showing 1 to 13 of 100 entries, 6 total columns

```
Data
  z        num [1:100, 1:6] 0.0218 0.777 0.3976 0.1603 0...
```

```
~/
> View(z)
> plot(z)
> 
```

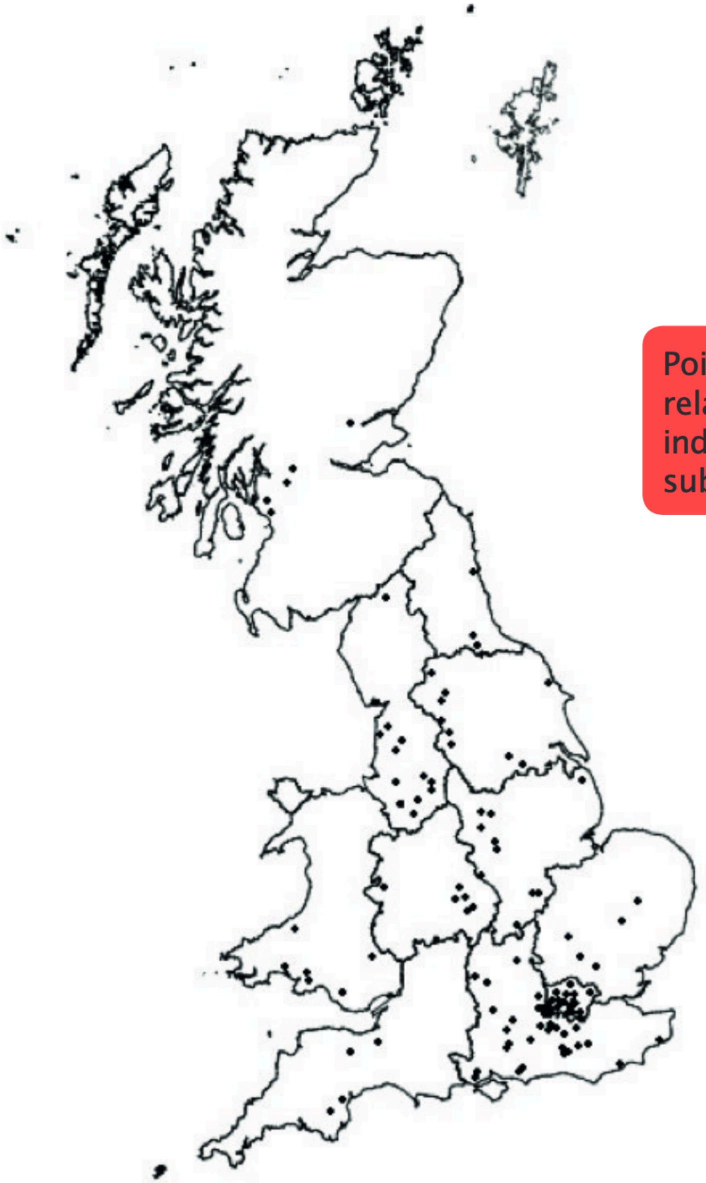
6. Researchers publish results

Analysis results are reviewed for potentially identifying information before they can are released from the TRE

→ Statistical disclosure control

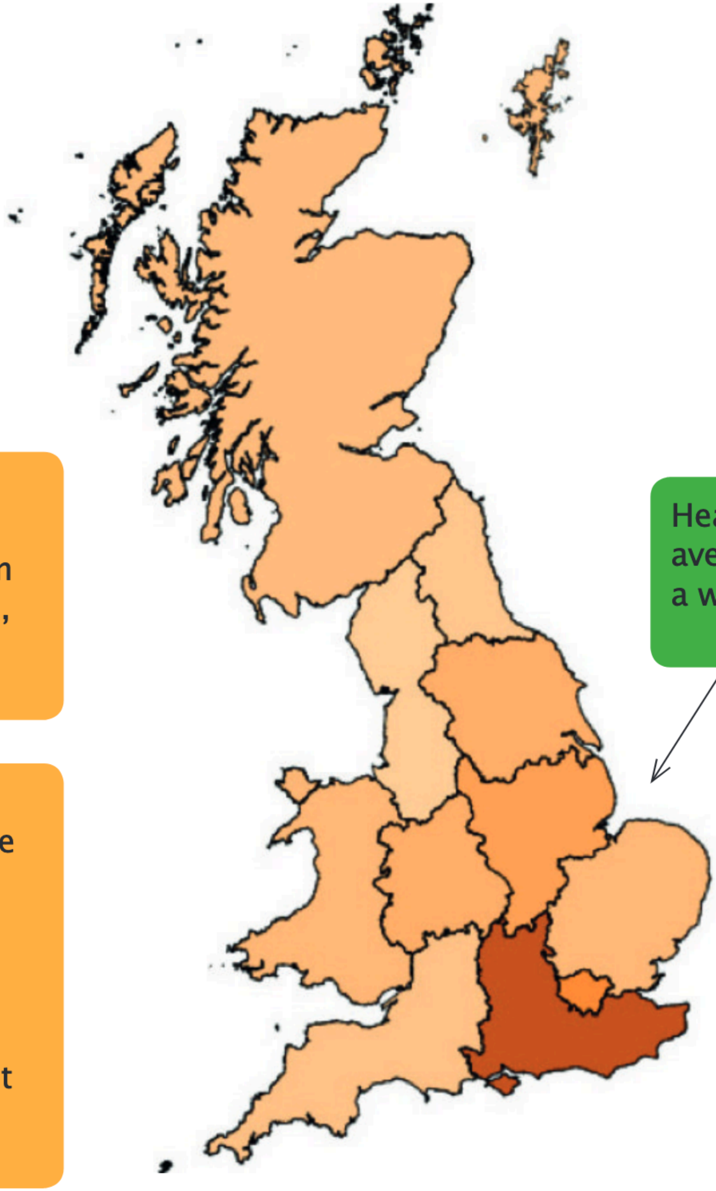
> confidentiality of individuals and individual statistical units must be protected

FIGURE 9:
Point map



Points could relate to individual data subjects

FIGURE 10:
Heat map



Could information be combined from public sources, e.g. Google Maps?

Depending on the scale of the map, and the nature of the data subject, it could be possible to release without disclosure occurring.

Heat map averages across a wide area

https://securedatagroup.org/wp-content/uploads/2019/10/sdc-handbook-v1.0.pdf

How do you design a TRE?

Basic principles of a TRE: The 5 safes

Safe data: data is treated to protect any confidentiality concerns.

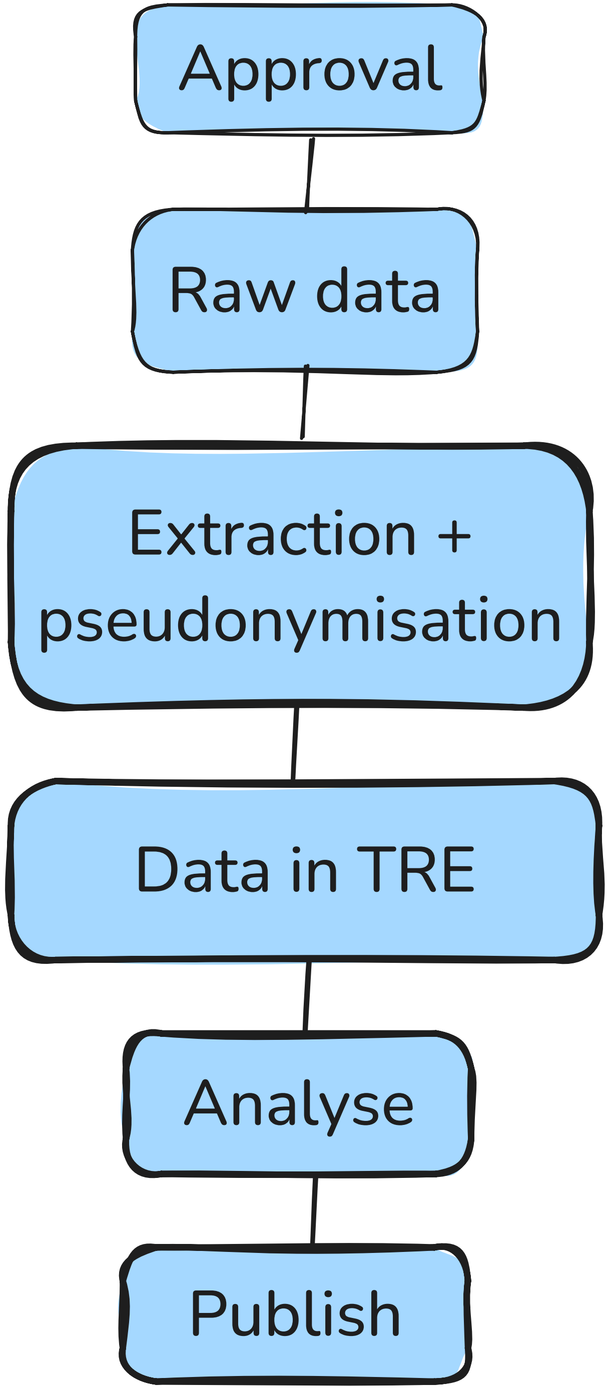
Safe projects: research projects are approved by data owners for the public good.

Safe people: researchers are trained and authorised to use data safely.

Safe settings: a secure environment prevents unauthorised use.

Safe outputs: screened and approved outputs that are non-disclosive.

https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/



Approval

Raw data

Extraction + pseudonymisation

Data in TRE

Analyse

Publish

Safe projects, Safe people

Safe data

Safe settings

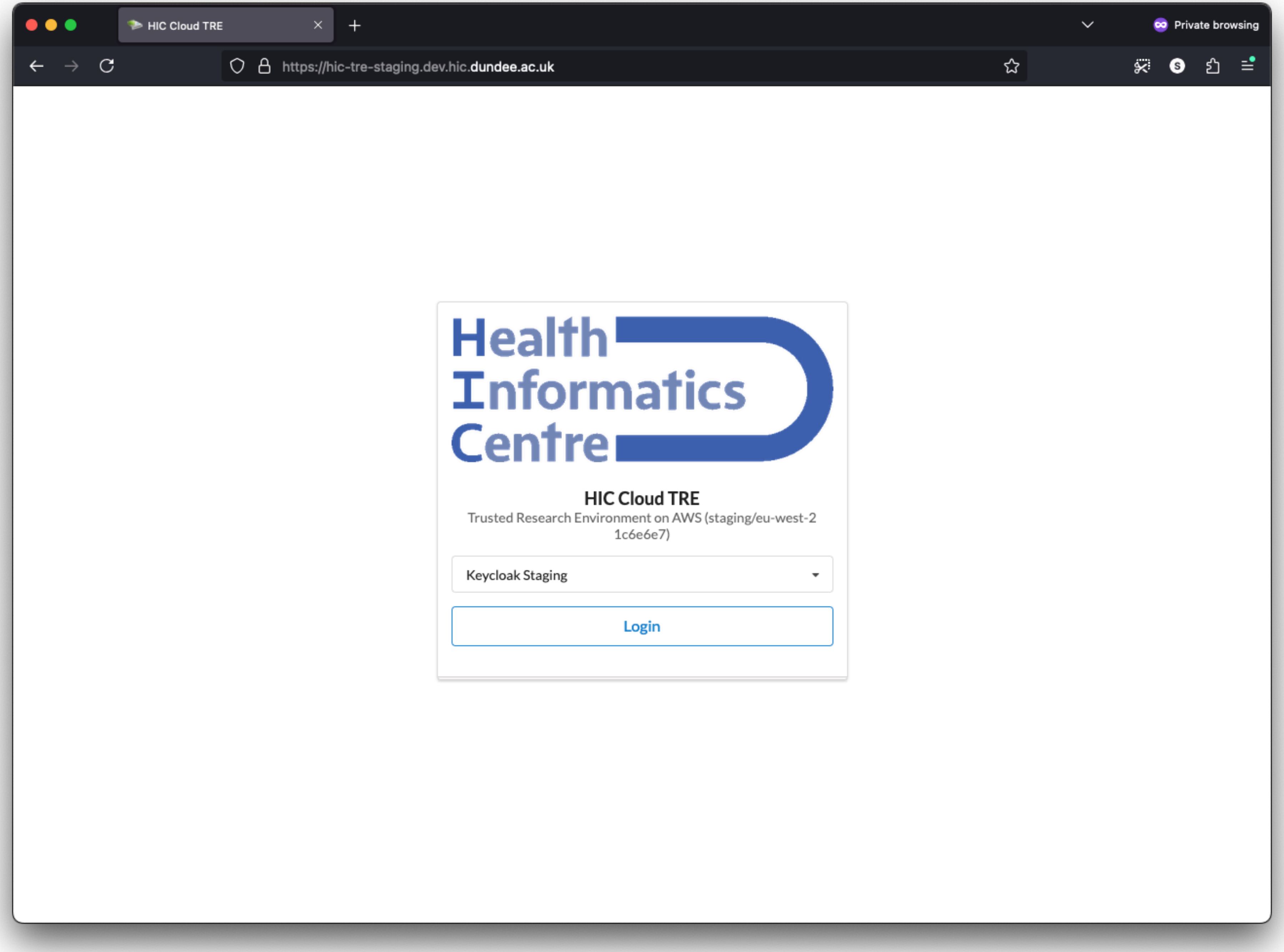
Safe outputs

Balancing the 5 safes

No such thing as 100% secure, need a balance between the 5 safes. Different TREs can make different tradeoffs, e.g.

- Access via a browser from anywhere
- vs must be physically present in a "SafePod"

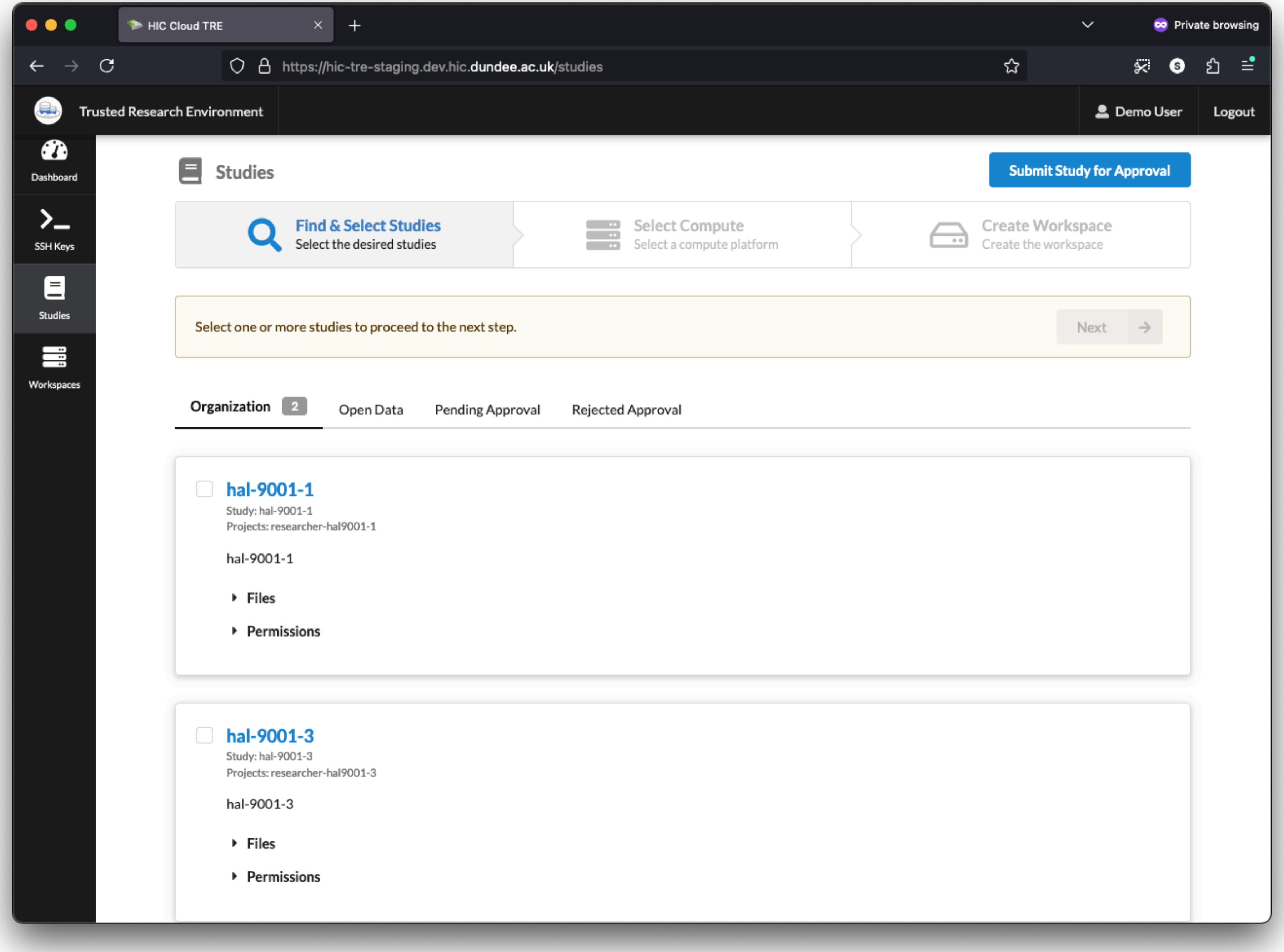
What does a TRE look like?


HIC Cloud TRE
Private browsing
https://hic-tre-staging.dev.hic.dundee.ac.uk
Health Informatics Centre

HIC Cloud TRE

Trusted Research Environment on AWS (staging/eu-west-2 1c6e6e7)

Keycloak Staging


HIC Cloud TRE
Private browsing
https://hic-tre-staging.dev.hic.dundee.ac.uk/studies
Trusted Research Environment
Demo User
Logout
Dashboard
SSH Keys
Studies
Workspaces

Studies

Submit Study for Approval

Find & Select Studies Select the desired studies

Select Compute Select a compute platform

Create Workspace Create the workspace

Select one or more studies to proceed to the next step.

Next →

Organization 2 | Open Data | Pending Approval | Rejected Approval

□ **hal-9001-1**
Study: hal-9001-1
Projects: researcher-hal9001-1

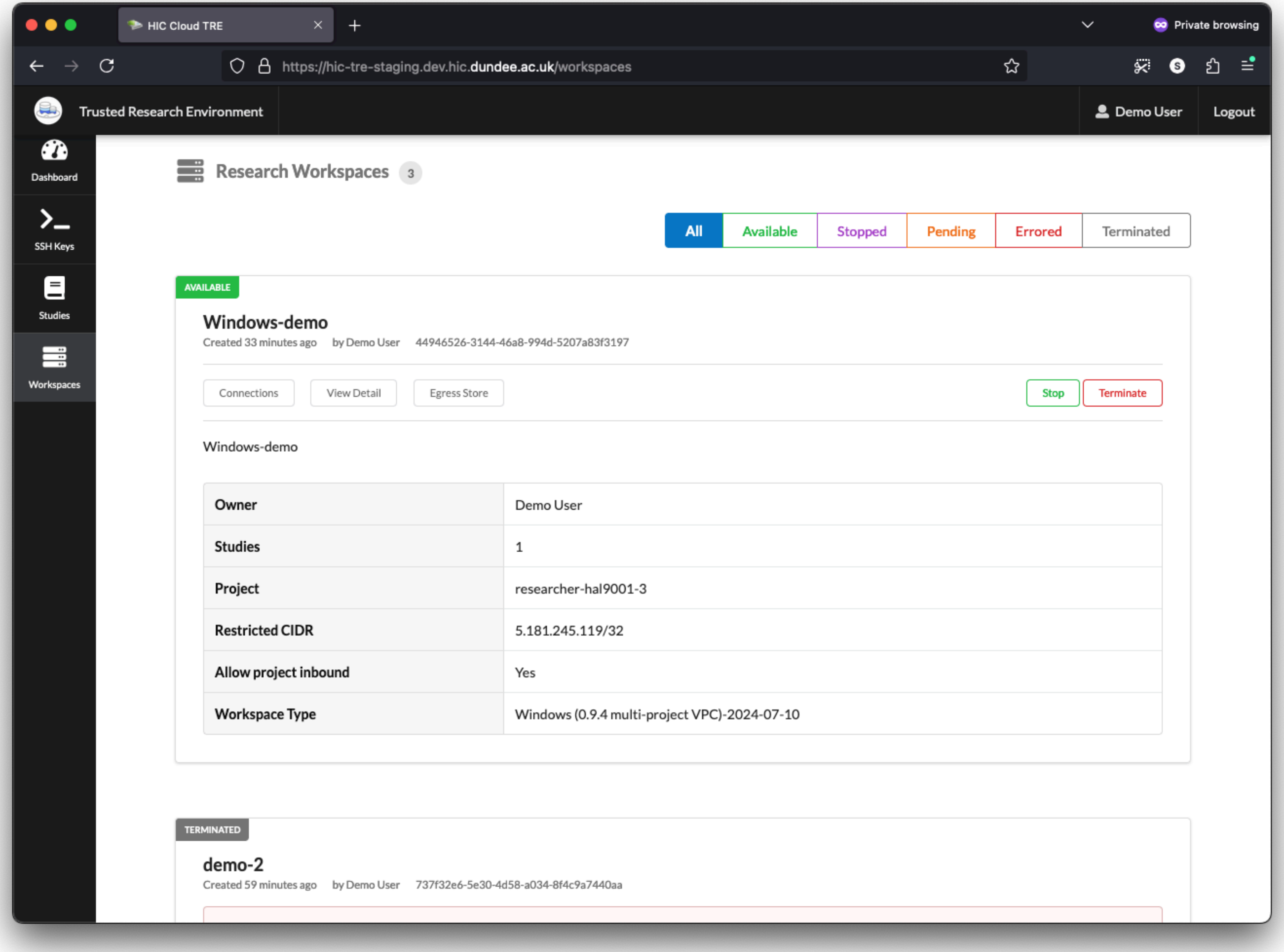
hal-9001-1

- Files
- Permissions

□ **hal-9001-3**
Study: hal-9001-3
Projects: researcher-hal9001-3

hal-9001-3

- Files
- Permissions



Research Workspaces 3

AVAILABLE

Windows-demo

Created 33 minutes ago by Demo User 44946526-3144-46a8-994d-5207a83f3197

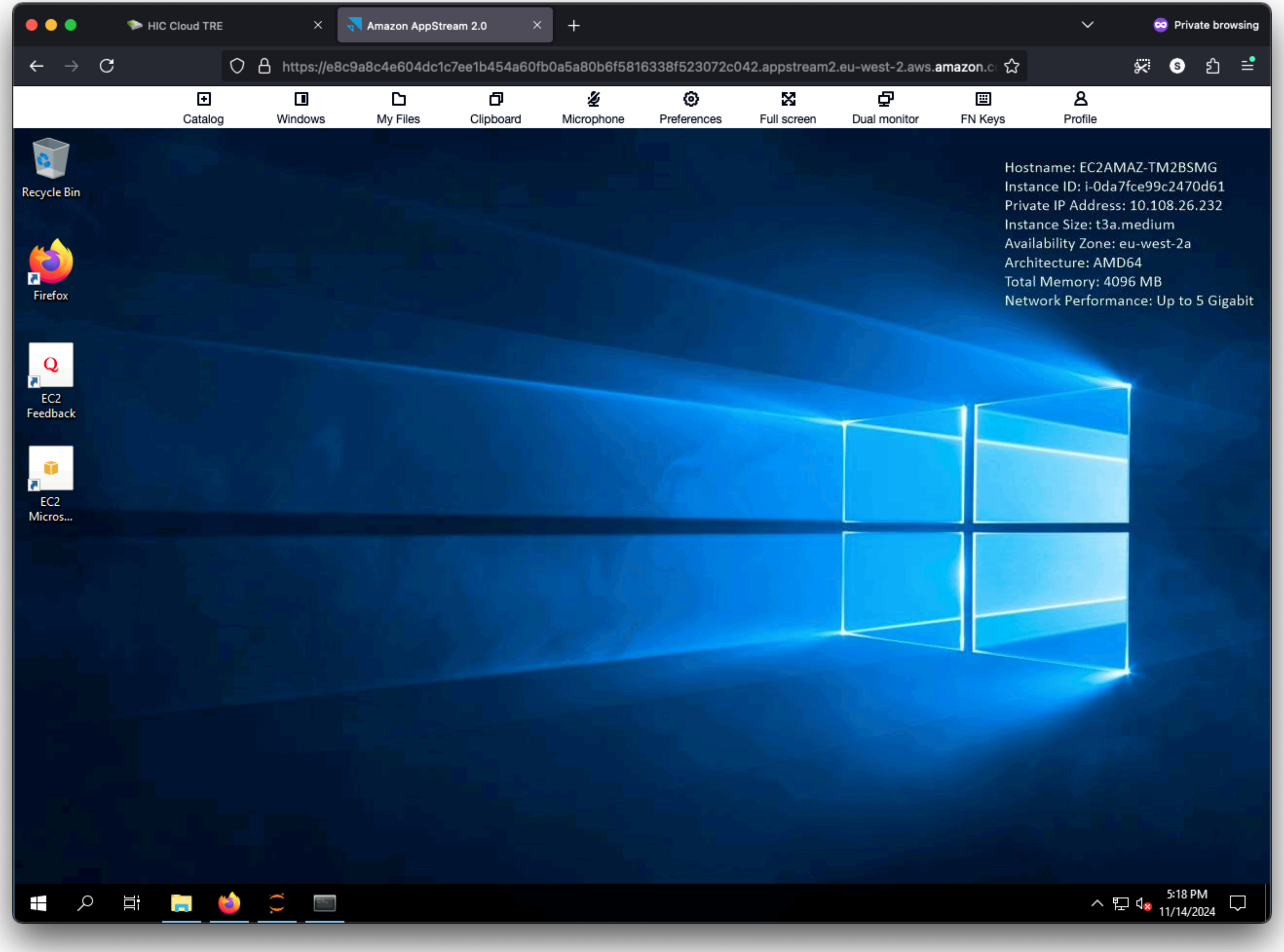
Windows-demo

Owner	Demo User
Studies	1
Project	researcher-hal9001-3
Restricted CIDR	5.181.245.119/32
Allow project inbound	Yes
Workspace Type	Windows (0.9.4 multi-project VPC)-2024-07-10

TERMINATED

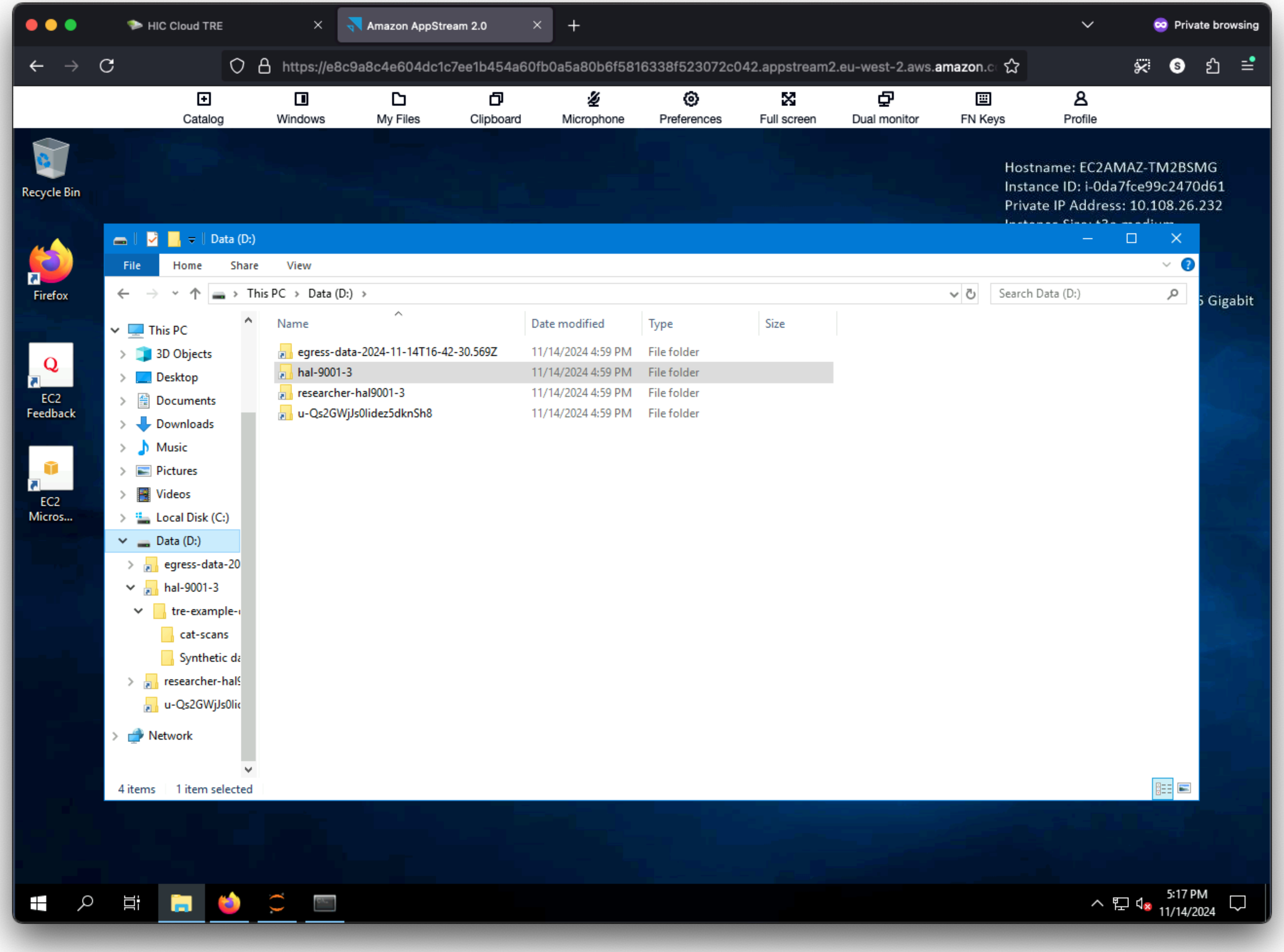
demo-2

Created 59 minutes ago by Demo User 737f32e6-5e30-4d58-a034-8f4c9a7440aa



https://e8c9a8c4e604dc1c7ee1b454a60fb0a5a80b6f5816338f523072c042.appstream2.eu-west-2.aws.amazon.c

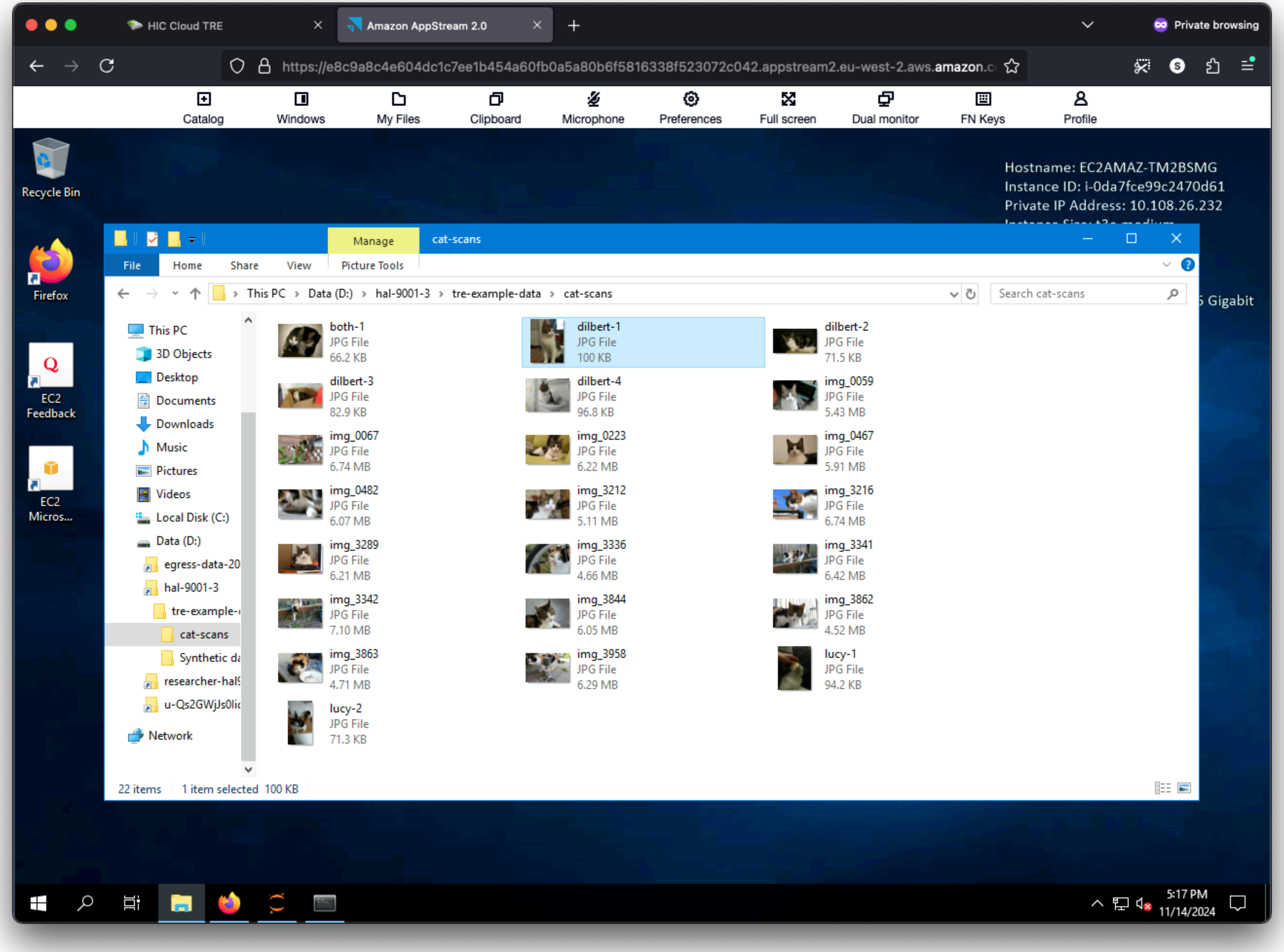
Hostname: EC2AMAZ-TM2BSMG
Instance ID: i-0da7fce99c2470d61
Private IP Address: 10.108.26.232
Instance Size: t3a.medium
Availability Zone: eu-west-2a
Architecture: AMD64
Total Memory: 4096 MB
Network Performance: Up to 5 Gigabit



Hostname: EC2AMAZ-TM2BSMG
Instance ID: i-0da7fce99c2470d61
Private IP Address: 10.108.26.232

Name	Date modified	Type	Size
egress-data-2024-11-14T16-42-30.569Z	11/14/2024 4:59 PM	File folder	
hal-9001-3	11/14/2024 4:59 PM	File folder	
researcher-hal9001-3	11/14/2024 4:59 PM	File folder	
u-Qs2GWjJs0lidez5dknSh8	11/14/2024 4:59 PM	File folder	

4 items 1 item selected



https://e8c9a8c4e604dc1c7ee1b454a60fb0a5a80b6f5816338f523072c042.appstream2.eu-west-2.aws.amazon.c

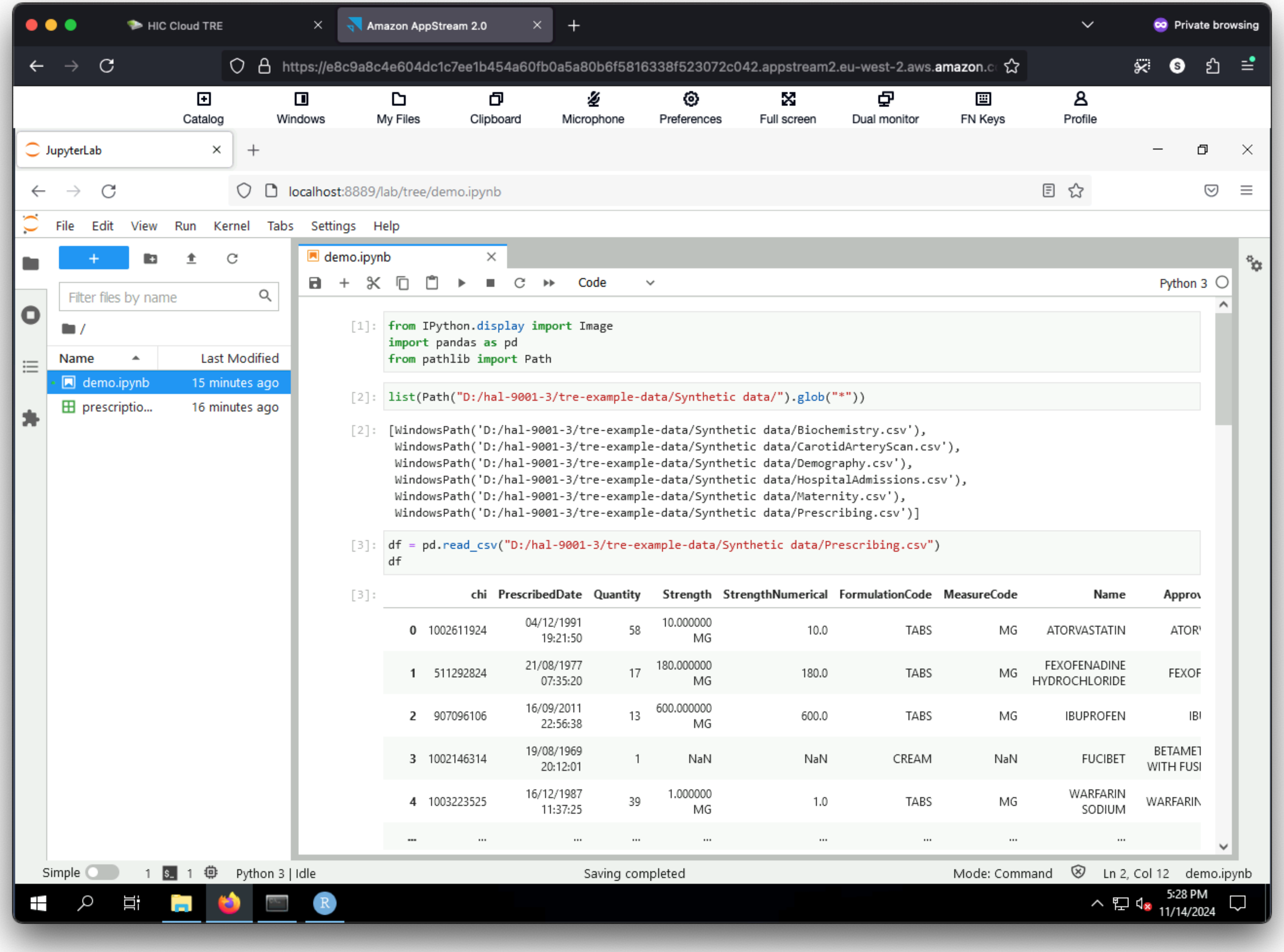
Hostname: EC2AMAZ-TM2BSMG
Instance ID: i-0da7fce99c2470d61
Private IP Address: 10.108.26.232

cat-scans

This PC › Data (D:) › hal-9001-3 › tre-example-data › cat-scans

both-1	JPG File	66.2 KB
dilbert-1	JPG File	100 KB
dilbert-2	JPG File	71.5 KB
dilbert-3	JPG File	82.9 KB
dilbert-4	JPG File	96.8 KB
img_0059	JPG File	5.43 MB
img_0067	JPG File	6.74 MB
img_0223	JPG File	6.22 MB
img_0467	JPG File	5.91 MB
img_0482	JPG File	6.07 MB
img_3212	JPG File	5.11 MB
img_3216	JPG File	6.74 MB
img_3289	JPG File	6.21 MB
img_3336	JPG File	4.66 MB
img_3341	JPG File	6.42 MB
img_3342	JPG File	7.10 MB
img_3844	JPG File	6.05 MB
img_3862	JPG File	4.52 MB
img_3863	JPG File	4.71 MB
img_3958	JPG File	6.29 MB
lucy-1	JPG File	94.2 KB
lucy-2	JPG File	71.3 KB

22 items 1 item selected 100 KB



```
[1]: from IPython.display import Image
     import pandas as pd
     from pathlib import Path
```

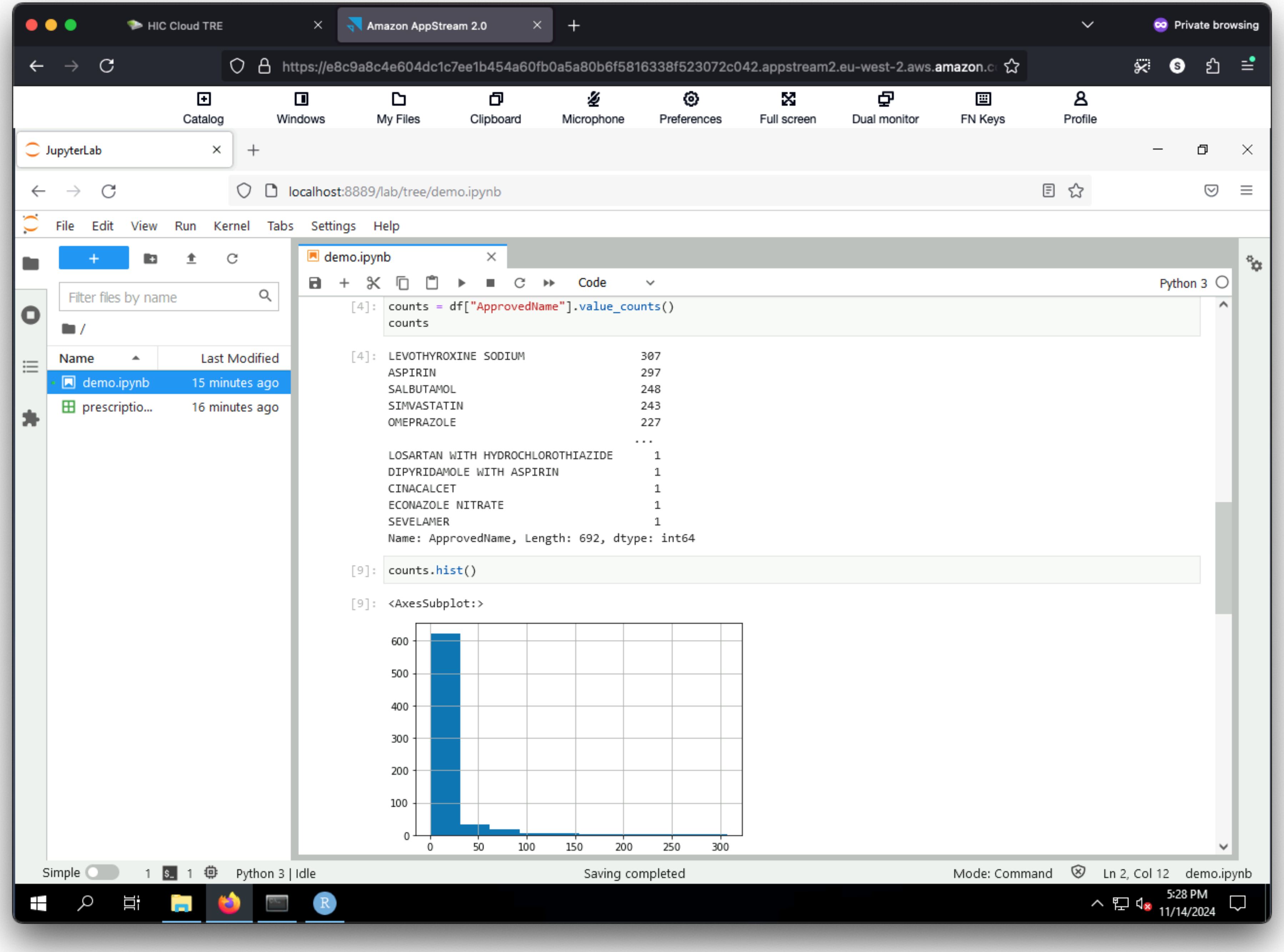
```
[2]: list(Path("D:/hal-9001-3/tre-example-data/Synthetic data/").glob("*"))
```

```
[2]: [WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/Biochemistry.csv'),
      WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/CarotidArteryScan.csv'),
      WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/Demography.csv'),
      WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/HospitalAdmissions.csv'),
      WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/Maternity.csv'),
      WindowsPath('D:/hal-9001-3/tre-example-data/Synthetic data/Prescribing.csv')]
```

```
[3]: df = pd.read_csv("D:/hal-9001-3/tre-example-data/Synthetic data/Prescribing.csv")
     df
```

[3]:

	chi	PrescribedDate	Quantity	Strength	StrengthNumerical	FormulationCode	MeasureCode	Name	Approv
0	1002611924	04/12/1991 19:21:50	58	10.000000 MG	10.0	TABS	MG	ATORVASTATIN	ATOR'
1	511292824	21/08/1977 07:35:20	17	180.000000 MG	180.0	TABS	MG	FEXOFENADINE HYDROCHLORIDE	FEXOF
2	907096106	16/09/2011 22:56:38	13	600.000000 MG	600.0	TABS	MG	IBUPROFEN	IBI
3	1002146314	19/08/1969 20:12:01	1	NaN	NaN	CREAM	NaN	FUCIBET	BETAMET WITH FUSI
4	1003223525	16/12/1987 11:37:25	39	1.000000 MG	1.0	TABS	MG	WARFARIN SODIUM	WARFARIN
...	...	...	...	...	...	...	...	...	

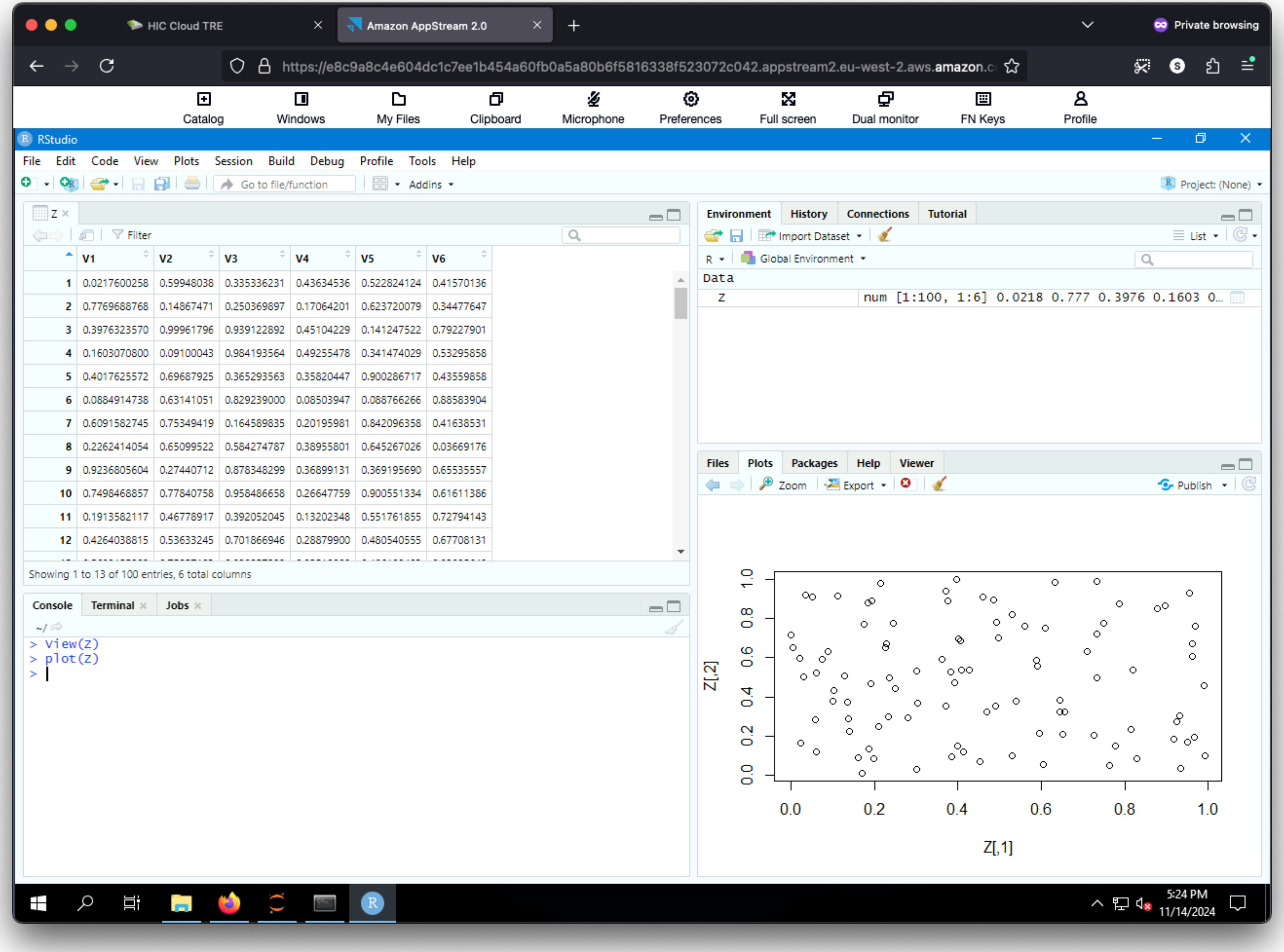


```
[4]: counts = df["ApprovedName"].value_counts()
     counts

[4]: LEVOTHYROXINE SODIUM                307
     ASPIRIN                             297
     SALBUTAMOL                          248
     SIMVASTATIN                         243
     OMEPRAZOLE                          227
                                        ...
     LOSARTAN WITH HYDROCHLOROTHIAZIDE     1
     DIPYRIDAMOLE WITH ASPIRIN             1
     CINACALCET                            1
     ECONAZOLE NITRATE                     1
     SEVELAMER                             1
     Name: ApprovedName, Length: 692, dtype: int64

[9]: counts.hist()

[9]: <AxesSubplot:>
```

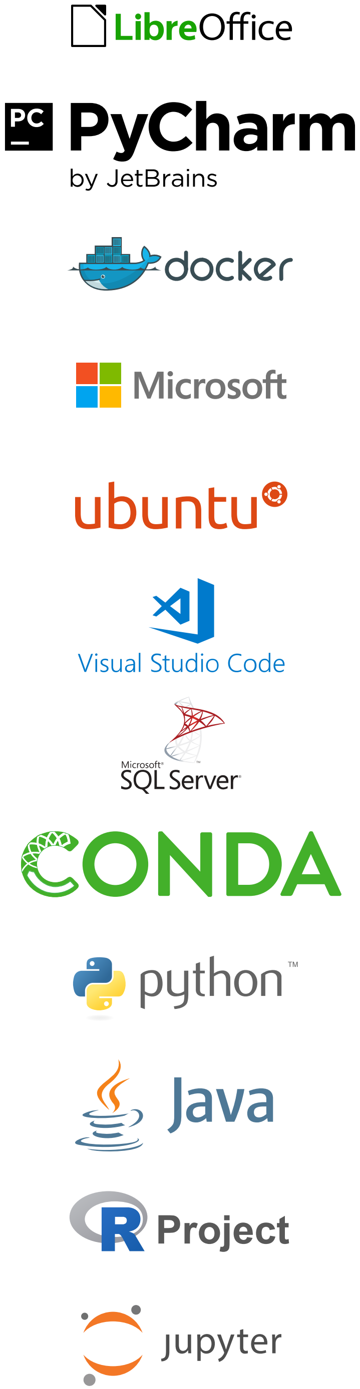


	V1	V2	V3	V4	V5	V6
1	0.0217600258	0.59948038	0.335336231	0.43634536	0.522824124	0.41570136
2	0.7769688768	0.14867471	0.250369897	0.17064201	0.623720079	0.34477647
3	0.3976323570	0.99961796	0.939122892	0.45104229	0.141247522	0.79227901
4	0.1603070800	0.09100043	0.984193564	0.49255478	0.341474029	0.53295858
5	0.4017625572	0.69687925	0.365293563	0.35820447	0.900286717	0.43559858
6	0.0884914738	0.63141051	0.829239000	0.08503947	0.088766266	0.88583904
7	0.6091582745	0.75349419	0.164589835	0.20195981	0.842096358	0.41638531
8	0.2262414054	0.65099522	0.584274787	0.38955801	0.645267026	0.03669176
9	0.9236805604	0.27440712	0.878348299	0.36899131	0.369195690	0.65535557
10	0.7498468857	0.77840758	0.958486658	0.26647759	0.900551334	0.61611386
11	0.1913582117	0.46778917	0.392052045	0.13202348	0.551761855	0.72794143
12	0.4264038815	0.53633245	0.701866946	0.28879900	0.480540555	0.67708131

Showing 1 to 13 of 100 entries, 6 total columns

```
Data
  Z                      num [1:100, 1:6] 0.0218 0.777 0.3976 0.1603 0...
```

```
> View(Z)
> plot(Z)
> 
```

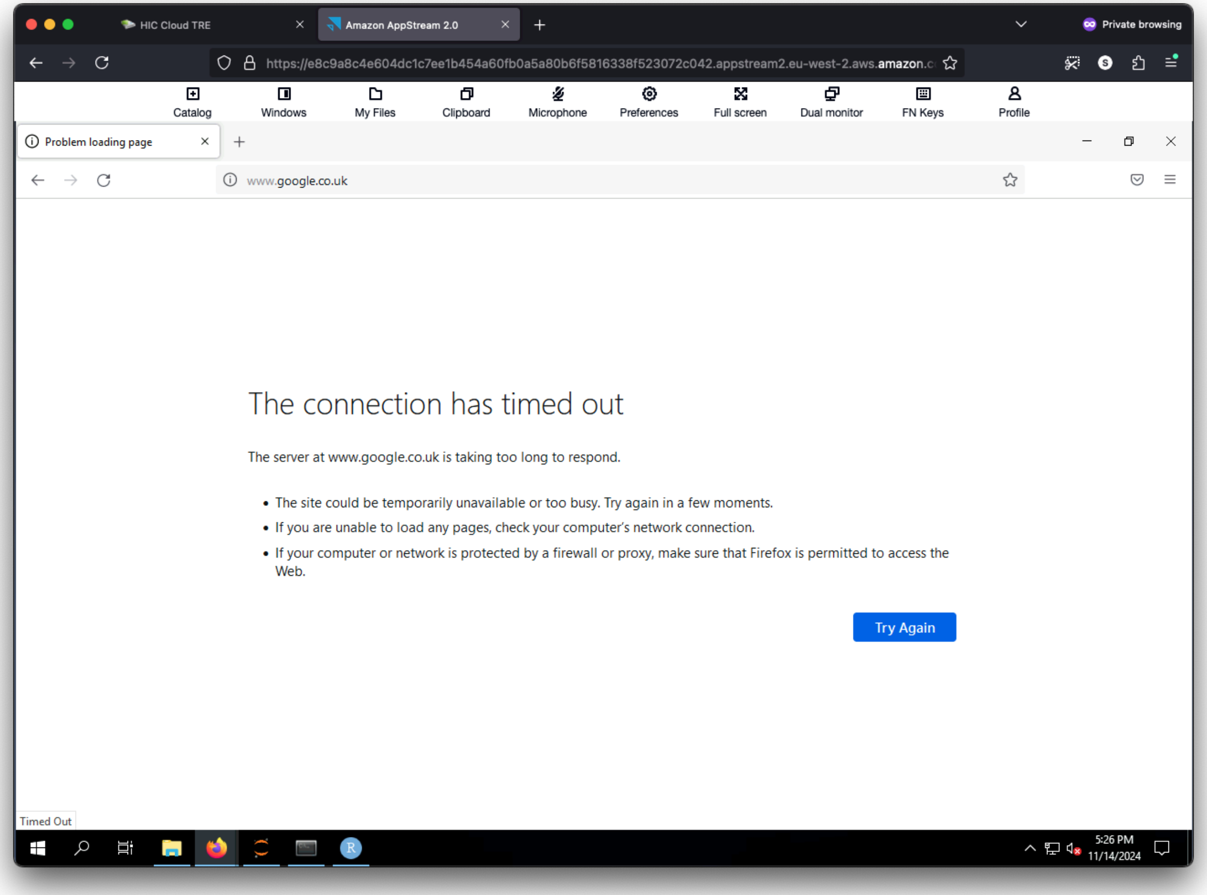


Full desktop environment, with many of the usual analysis applications

- Focus on the analysis, not on how to manage the data

Downsides: Very restrictive environment compare to what you're used to

- No or very limited outbound network access
- Can't install random software packages
- Clipboard is restricted (can't copy out)



The connection has timed out

The server at www.google.co.uk is taking too long to respond.

- The site could be temporarily unavailable or too busy. Try again in a few moments.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

Open Science: what is it?

Scientific research should be freely accessible to everyone

UNESCO Recommendation on Open Science

> " various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. "

Open-science in TREs

❌ Open data

✅ Open access: share your results
✅ Open source: share your code and analyses

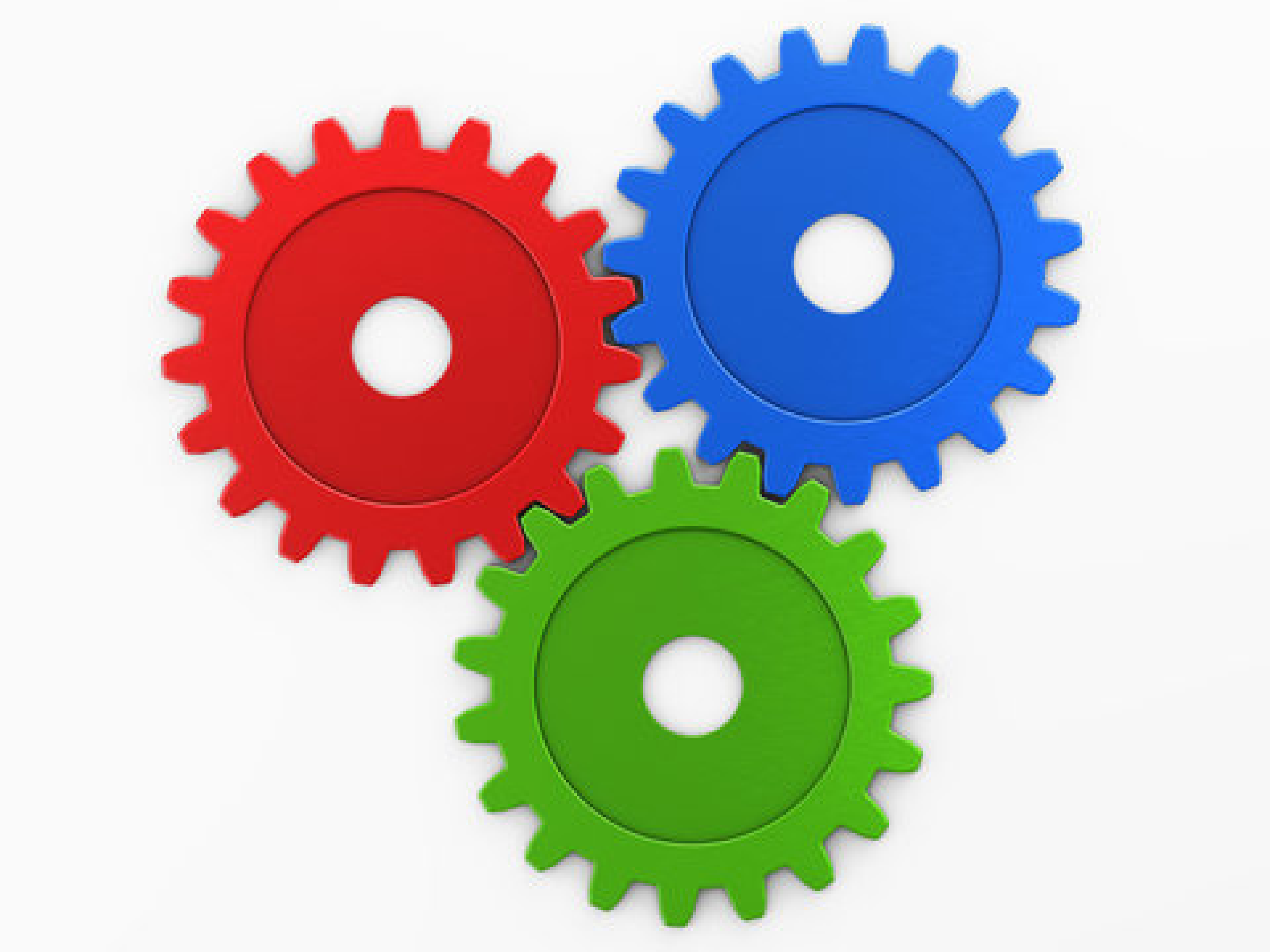
✅ Open infrastructure: show others how to run infrastructure to support science
✅ Open collaboration: share your knowledge and help others

The problem

100s(?) of TREs in the UK

- Every TRE is different
- No best practice on how to design, operate and use a TRE

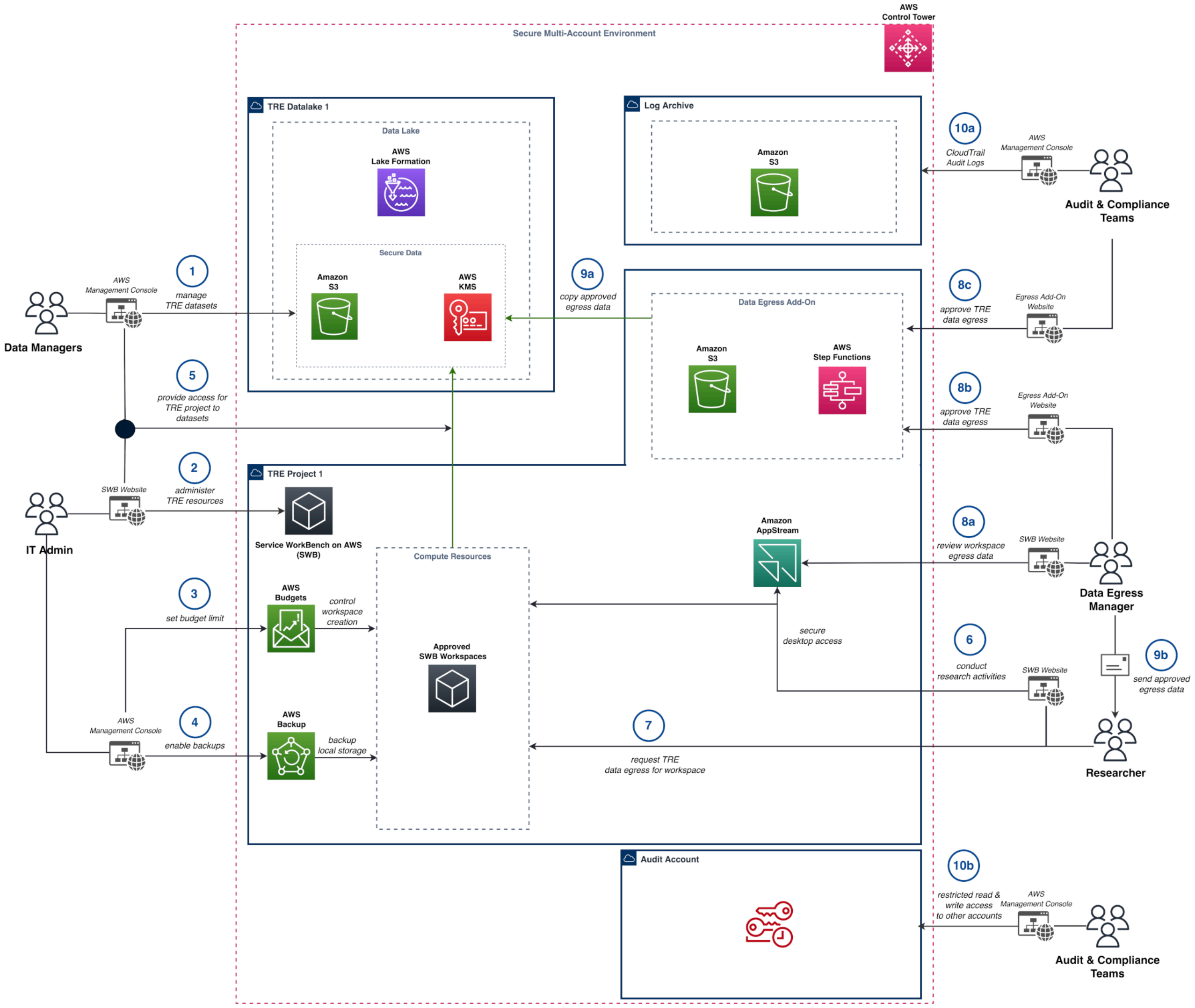
Part of the solution....



Open infrastructure: open-source isn't enough

Running a single application is easy, but how do you deploy and run an entire TRE?

- Share your deployment code and documentation



Federated analysis

Data is currently siloed in different TREs

- We need standards for interoperability!

A review commissioned by the Secretary of State for Health and Social Care

Better, Broader, Safer: Using Health Data for Research and Analysis

April 2022

Uniting the UK's Health Data: A Huge Opportunity for Society

A review of the UK health data landscape commissioned by the Chief Medical Officer for England, the UK National Statistician and NHS England's National Director for Transformation

November 2024

Open collaboration

UK TRE

Health Informatics Centre

Standard Architecture for Trusted Research Environments

Can we make sure all new TREs work in a similar way?

https://satre-specification.readthedocs.io/

What do the public think?

- A cautionary note

2014: Care.data

https://www.bmj.com/content/354/bmj.i3907

> By now, if all had gone to plan, England would be reaping the early benefits of a national database of patients' medical records spanning primary and secondary care. Patients would have the comfort of knowing that their records could soon be accessed wherever they were treated....

> But all did not go to plan. NHS England’s care.data programme failed to win the public’s trust and lost the battle for doctors’ support. Two reports have now condemned the scheme, and last week the government decided to scrap it.

Public trust is really important!

The NHS has learnt from that

"People in the UK overwhelmingly support the use of their health data, with appropriate safeguards, to benefit themselves and others."

TREs and open-science have a big part to play

Uniting the UK's Health Data: A Huge Opportunity for Society

A review of the UK health data landscape commissioned by the Chief Medical Officer for England, the UK National Statistician and NHS England's National Director for Transformation

November 2024

Bedtime reading

Trusted Research Environments

- https://www.dundee.ac.uk/hic/safe-haven
- https://satre-specification.readthedocs.io
- https://dareuk.org.uk

Public views on sensitive data

- https://www.pedri.org.uk
- https://understandingpatientdata.org.uk

Important reports

- https://www.hdruk.ac.uk/helping-with-health-data/the-sudlow-review/
- https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis

Scottish Health Research Register and Biobank (SHARE)

- https://www.registerforshare.org

Health Informatics Centre

University of Dundee

DARE UK

HDRUK Health Data Research UK

ADR UK

UK Research and Innovation

The Alan Turing Institute

UCL

Ulster University

Research Data Scotland