



# Shaping the future of data research: Meet the DARE UK Next-Gen Catalysts

Wednesday 4<sup>th</sup> March 2026



## Housekeeping

- Please remain **muted** at all times, unless invited to ask a question during the Q&A session
- To ask a question during the Q&A session: please use the **'raise hand' function** if you wish to ask a question out loud; or you can ask your question via the **Chat box** before and during the Q&A session
- During the discussion, please remain respectful, do not interrupt others and ensure any questions or comments remain on-topic
- Please keep your **video** off during presentations – you are encouraged to turn your video on during the Q&A/discussion session
- The webinar is being **recorded**

# Agenda for today

- Welcome and overview of the Next-Gen Catalysts
- **10 minute** presentations from the projects, followed by **5-minute Q&A**
  - Running order:
    - The Data Matryoshka: Progressive Synthetic OMOP Data Layers for Secure Health Research Collaboration Across UK TRES
    - STAR-TRE: Safe and Trustworthy Assessment of Risk in TRES for Sensitive Free-Text Access
    - TRESS: Trusted Research Environments Software Stewardship
    - FIRE DANSE: Federated Infrastructure for Digital Pathology Reporting and Expert Data Annotation in a Secure Environment
    - FORTRESS-TeHR: Federated, Open and Reliable TRES for Synthetic Textual Healthcare Records
    - TRUSTEE: Trusted Research on Untrusted Systems using Trusted Execution Environments
    - TRExt: TRE Text Analytics
    - GROVE: Governance for Household Level Environment and Health Data

**Sensitive data:** affecting people, communities, and populations...



**Data linkage:** safely connecting sensitive data to benefit people, communities and populations...



## Vision

For all research and innovation to benefit from seamless, secure use of diverse sensitive data, at a pace, efficiency and scale approaching that of the open data ecosystem, that revolutionises research productivity and accelerates research to deliver public good

## Mission

To put the UK at the forefront of sensitive data research and innovation by assembling the tools, technologies and standards needed to streamline secure data linkage and use

## The purpose of the DARE UK Next-Gen Catalysts

- developing early-stage prototypes that test ideas **beyond DARE UK's current programme of work**
- the projects will generate evidence, working prototypes, and practical guidance to inform future national infrastructure and policy decisions

# Stay in touch! – QR codes

Visit our website



 SCAN ME

Subscribe to our newsletter



 SCAN ME

Follow us on  
LinkedIn



 SCAN ME

Follow us on X  
(Formerly Twitter)



 SCAN ME



# Federated Infrastructure for digital pathology Reporting and Expert Data Annotation in a Secure Environment

Simon Doran and the FIRE DANSE project team



# Context of the FIRE DANSE PROJECT



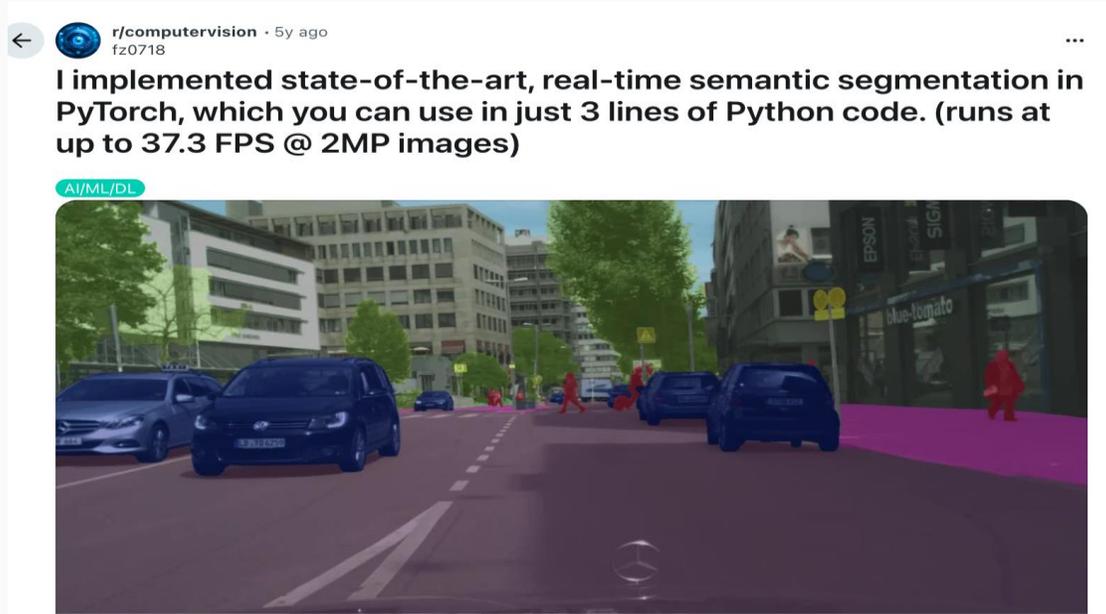
**0 → Cat**



**1 → Dog**

- AI for image analysis has come a long way in the 15 years since we were trying to decide what was a dog and what was a cat ...

# Context of the FIRE DANSE project



Video credit: Reddit fz0718

- Automatic identification of objects and semantic segmentation have become ubiquitous in everyday life.
- But how did we get here? How does a computer “learn” to do this?
- Crucial to the advances has been the generation of large quantities of **ground truth data** by human annotators.

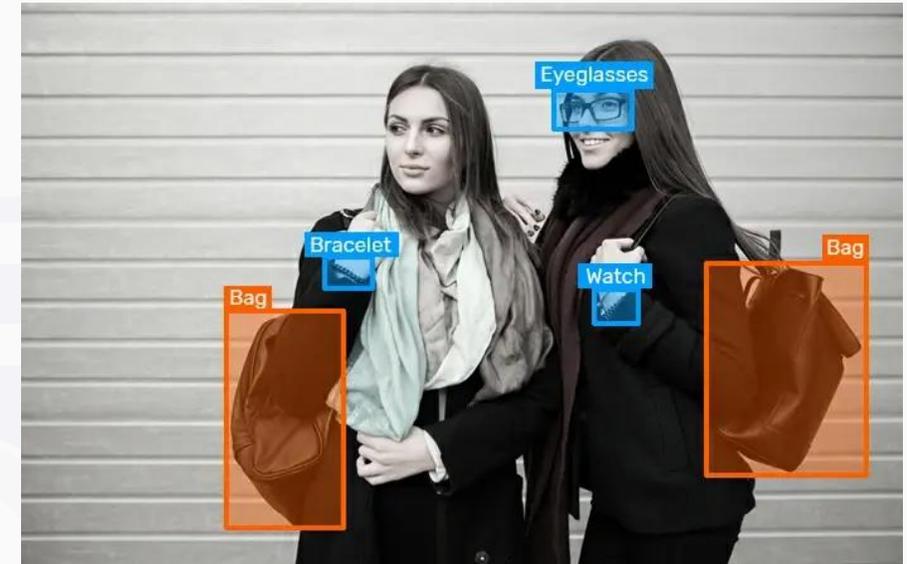


Image credit: Analytics

## Bounding Box Annotation Services

Enclosing objects in boxes for object localization in images, videos, and even 3D sensor data meets the training data requirements of machine learning models.

[Get In Touch →](#)

**15+**  
Years' Experience

**1500+**  
Annotators working

**24X7**  
Availability

# Context of the FIRE DANSE project



Image credit: Katharina von Loga / Owkin / Bioptimus

- Huge progress has also been made in medical imaging, both radiology and digital pathology.
- But we still need "ground truth" data to train models.
- At it's heart, the issue is still the same as with the dogs and cats.

# Context of the FIRE DANSE project

From: [Deep learning for multi-class semantic segmentation enables colorectal cancer detection and classification in digital pathology images](#)

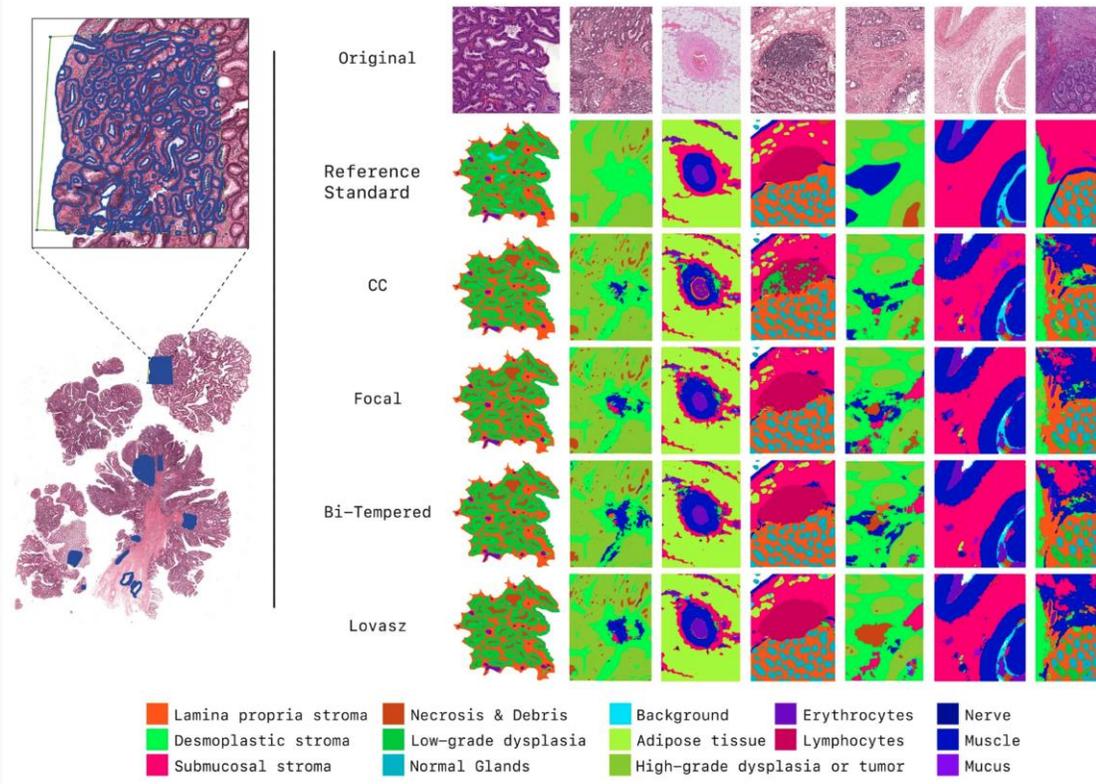


Image credit: Bockhorst *et al. Scientific Reports* 13, 8398 (2023)

- Huge progress has also been made in medical imaging, both radiology and digital pathology.
- But we still need "ground truth" data to train models.
- At it's heart, the issue is still the same as with the dogs and cats.
- But now, the questions we ask are more complicated.
- The types of annotation we need experts to perform are more complex and time-consuming.
- The humans who can do the annotation are in short supply, time poor and (much!) more expensive...
- ... and the data we want them to annotate are both more sensitive and located in separate TREs requiring federation.

# User stories: What issues does FIRE DANSE address?

Alice is a member of the public and potential cancer patient.



I WANT TO GET THE RESULT OF MY TEST BACK QUICKLY.  
I CAN SEE HOW AI MIGHT IMPROVE THIS PROCESS AND I'M HAPPY FOR MY DATA TO BE USED TO TRAIN IT.  
I NEED TO BE SURE THAT AI IS TELLING THE TRUTH. A HUMAN EXPERT MUST CHECK THAT IT'S CORRECT.

Chidi is the IG Lead at a large teaching hospital.



IT'S MY JOB TO KEEP THE DATA SAFE, AND I HAVE TO BE CONVINCED THAT OUR RESEARCHERS ARE TRAINING THEIR AI IN A WAY THAT'S ACCORDING TO HOSPITAL GOVERNANCE.  
I NEED TO UNDERSTAND WHERE OUR DATA ARE GOING IN FEDERATED PROJECTS.

Bob is that expert. He's a skilled but overworked pathology consultant in a busy District General Hospital.



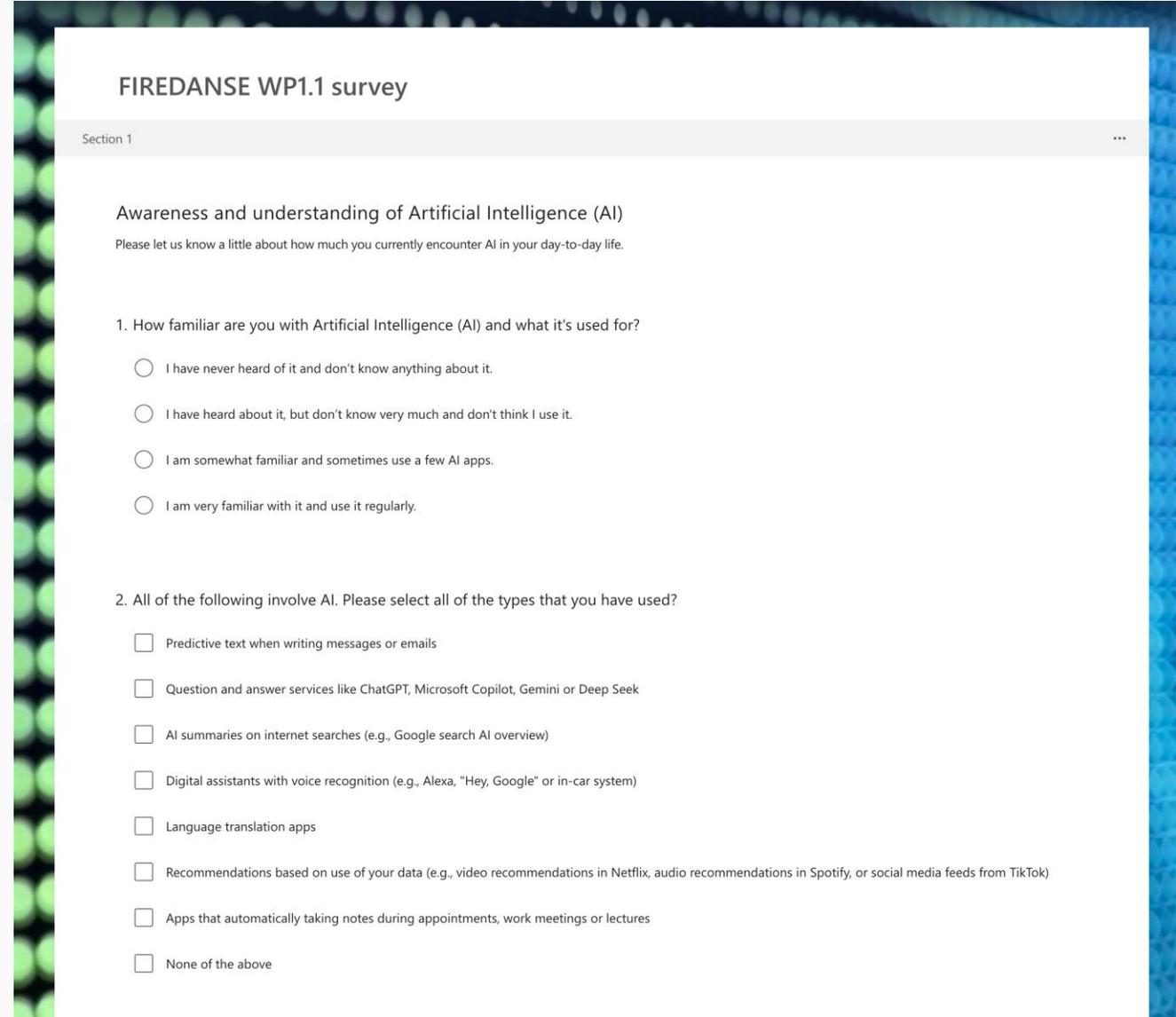
HOW CAN I USE MY SPECIALIST KNOWLEDGE OF RARE DISEASE PATHOLOGY TO HELP TRAIN AN AI?  
THE CASES COME FROM ALL OVER THE COUNTRY AND ALL SIT IN DIFFERENT TREES.  
IT'S JUST TOO TIME-CONSUMING AND TOO COMPLICATED TO TAKE PART IN RESEARCH FROM BASIC SCIENCE TO CLINICAL TRIALS.

# The programme of work at a glance

- WP1: Public involvement and engagement
  - Are the assumptions we made about whether Alice represents “what the public think about AI” really true?
  - Survey (project start) – builds on substantial existing PIE resources to achieve diverse representation
  - Education session and focus group (project midpoint)
  - Final showcase and feedback
  - Development of resources
- WP2 – 4: Innovative federation technologies
  - WP2 evaluation of DARE UK Phase 1 TELEPORT paradigm
  - WP3 federated worklist and task manager
  - WP4 pathology review tool with focus on productivity and end-user experience
- WP5: Platform testing using samples from ICR-Imperial study
- WP6: Governance review of final system

# Progress to date: Months 1 – 3

- WP1: Public involvement and engagement
  - We have worked with Project Co-Leads who are also patient members of the Royal Marsden's PPIE Team (Patient and Public Involvement and Engagement).
  - Initial consultations on information governance for the survey have taken place.
  - Survey questions are now decided for WP1.1.
  - Our survey is now in the review / piloting stage.
  - We hope to start distributing the survey to potential participants by mid-March.
- WP2: Implementing previous DARE UK methodology
  - Currently in consultation and planning stages



**FIRE DANSE WP1.1 survey**

Section 1

Awareness and understanding of Artificial Intelligence (AI)  
Please let us know a little about how much you currently encounter AI in your day-to-day life.

1. How familiar are you with Artificial Intelligence (AI) and what it's used for?

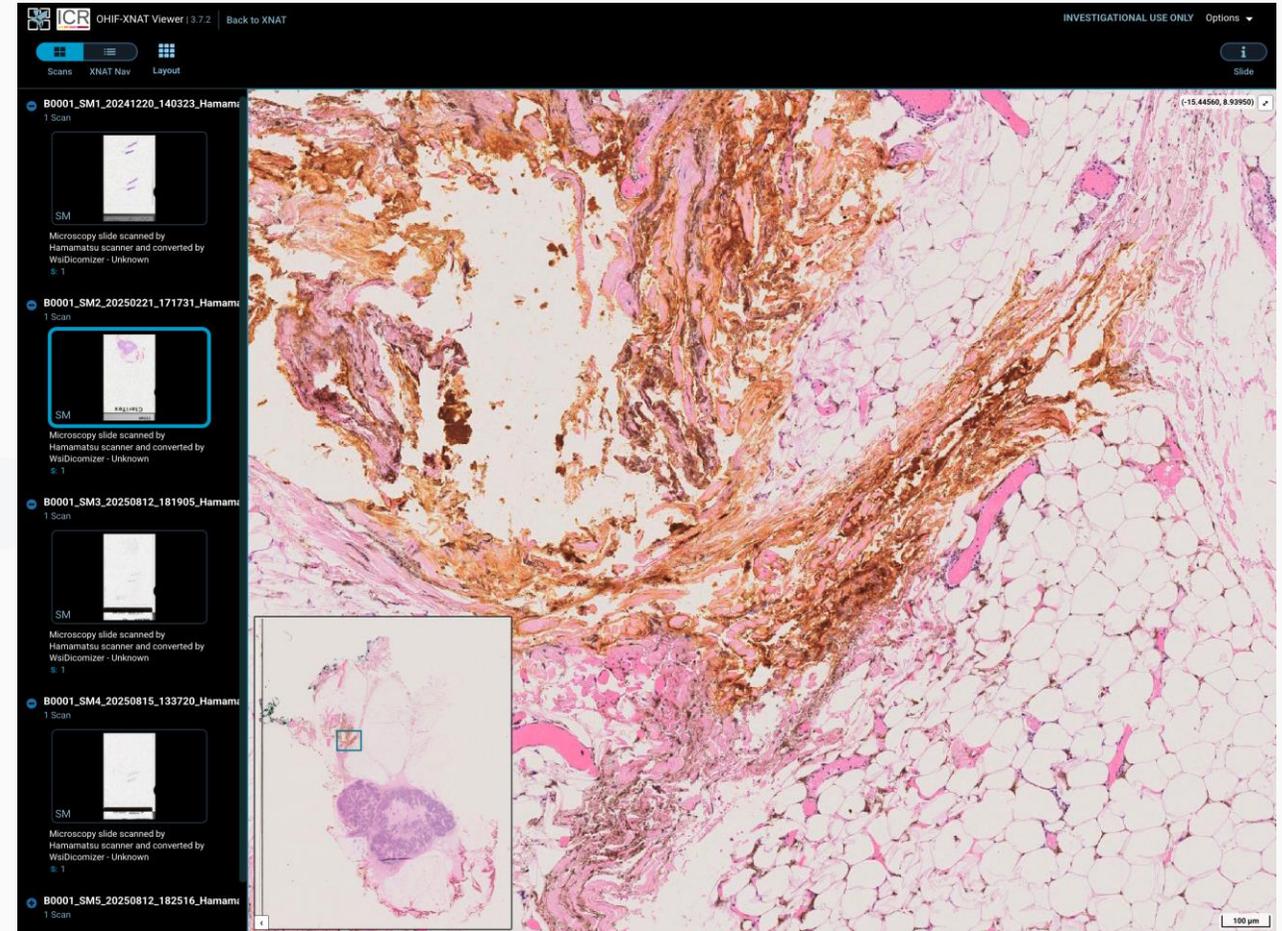
- I have never heard of it and don't know anything about it.
- I have heard about it, but don't know very much and don't think I use it.
- I am somewhat familiar and sometimes use a few AI apps.
- I am very familiar with it and use it regularly.

2. All of the following involve AI. Please select all of the types that you have used?

- Predictive text when writing messages or emails
- Question and answer services like ChatGPT, Microsoft Copilot, Gemini or Deep Seek
- AI summaries on internet searches (e.g., Google search AI overview)
- Digital assistants with voice recognition (e.g., Alexa, "Hey, Google" or in-car system)
- Language translation apps
- Recommendations based on use of your data (e.g., video recommendations in Netflix, audio recommendations in Spotify, or social media feeds from TikTok)
- Apps that automatically taking notes during appointments, work meetings or lectures
- None of the above

# Progress to date: Months 1 – 3

- WP3: Federated worklist and task manager
  - Initial research on appropriate methodologies / coding language.
  - Version 1.0 of task manager software has been created and tested.
  - First deployment of the task manager on a “production” system is underway.
- WP4: Digital pathology viewer
  - Good progress has already been made on integrating the microscopy viewer with our customised forms, but this is still a work in progress.
- WP5: Platform testing
  - Work is underway to define a dataset and set of annotations required to test the finished platform.



# The Balinese fire dance: inspiration for our project's name



- Exchanging data between Trusted Research Environments is a complex process.
- Preventing data from being accessed by the wrong people or the accidental disclosure of personal information is a high priority for the public.
- The Balinese fire dance symbolises the careful choreography required to move sensitive data safely – a balancing act of jeopardy, precision, coordination and technical mastery.
- Our goal is to provide smooth, seamless and secure access to sensitive image data stored in multiple different repositories around the country.
- The practical outcome will be to make it easier to train computers to analyse pathology images...
- ... and for human experts to verify that the new AI algorithms are working properly, so that patients can trust diagnoses that they make.



# Federated Synthetic Text Generation

Goran Nenadic, University of Manchester  
gnenadic@manchester.ac.uk



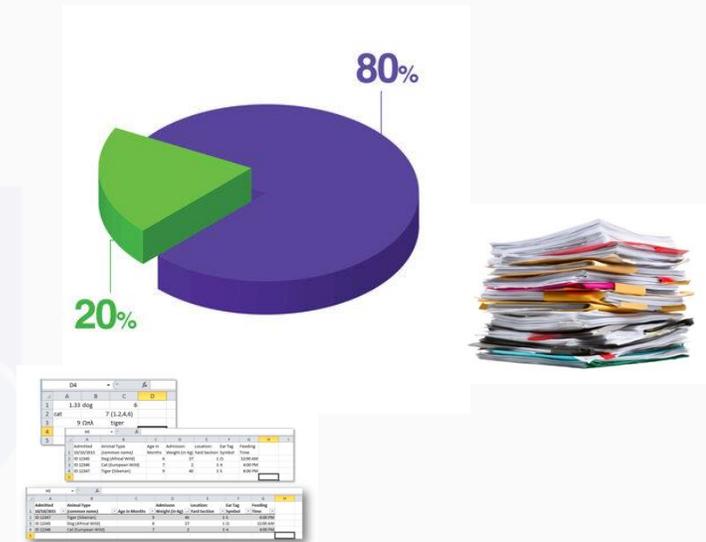


# Federated Synthetic Text Generation



# Focusing on free-text data

- Often most personal, sensitive and risky
  - Healthcare, finance, education, security, . . .
  - Sometimes unexpected things would appear in these.
- How do we provide safe access to sensitive free-text data for AI development?
- Options to reduce the risk of re-identification
  - De-identification and pseudo-anonymisation (STAR-TRE): remove sensitive information and assess risks
  - Synthetic (artificial) free-text generation (FORTRESS): generate realistic data that is not about real people (or events)



## Aims

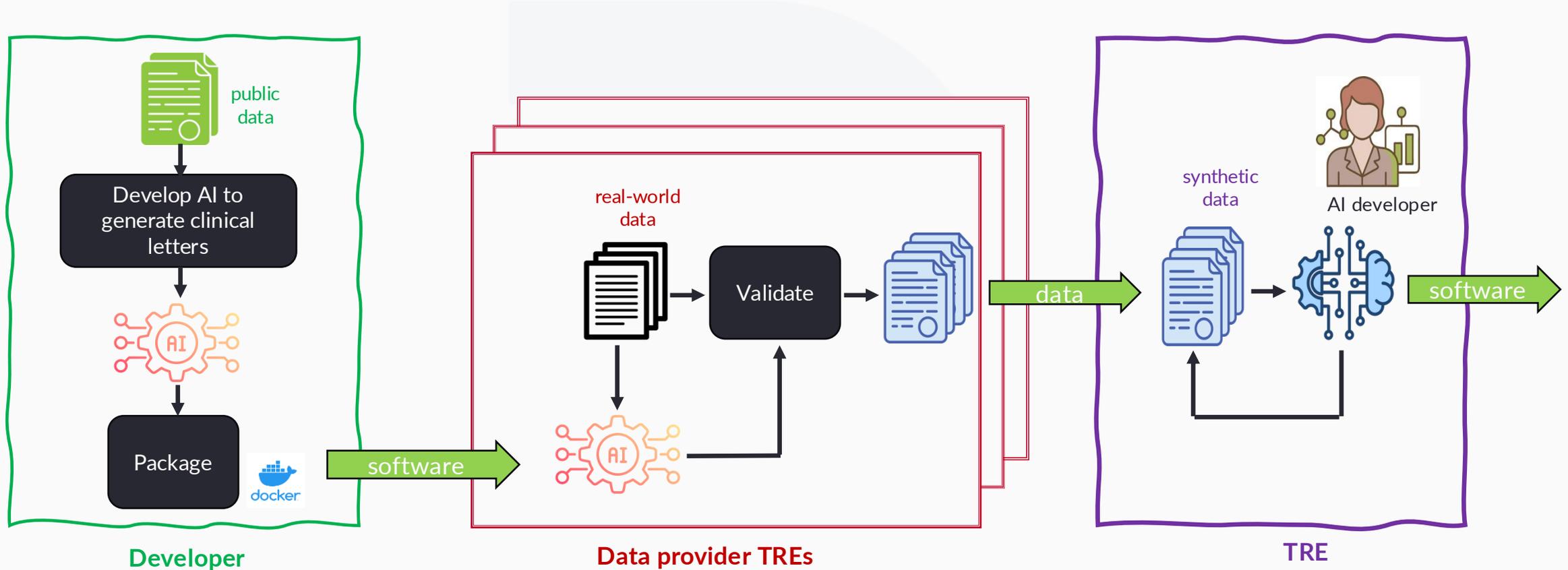
- **What should generation, validation and access to synthetic text within TRE look like?**
  - pain points (technical, governance, quality)
  - safeguards that we need
  - current limitations.
- How to assess whether synthetic free text preserves sufficient analytical value while reducing disclosure risk?
- Making synthetic FAIR: how to document synthetic texts so that it can be found, accessed, interoperable, re-used.

## Work streams

- **Technical development – text generation methods**
  - How to guide large-scale, federated text generation?
- **Public involvement and engagement**
  - What are public views on synthetic text generation?
- **Synthetic text validation**
  - What needs to be validated and how?
- **Meta-data and descriptions**
  - How to ensure transparency and provenance?
- **Federated TRE deployment**
  - How to ensure that training and generation can be run in TREs?



# Work streams: Federated TRE deployment



# Work streams: Technical development



Can you please generate a part of an outpatient letter for a cardiology patient that describes their history of present illness? Use UK style and make it concise. Generate two examples.

Sure — here are **two concise UK-style examples** suitable for an outpatient cardiology letter.

## History of Present Illness (Example 1):

The patient was referred for evaluation of exertional chest pain. They report a 6-week history of central chest tightness occurring on exertion and resolving with rest within a few minutes. There is associated mild breathlessness. They deny radiation of pain, palpitations, syncope, orthopnoea or ankle swelling. Symptoms have remained stable.

## History of Present Illness (Example 2):

The patient presents with intermittent palpitations over the past 3 months. Episodes occur at rest and last up to 30 minutes, occasionally associated with lightheadedness but no syncope or chest pain. There is no clear precipitant. They deny exertional dyspnoea, orthopnoea or peripheral oedema. Frequency of symptoms has increased slightly in recent weeks.



## \*\*History of Present Illness\*\*

The patient reports intermittent exertional chest discomfort over the past several months, described as a dull central pressure occurring primarily when walking uphill. Symptoms resolve with rest and have not been associated with syncope, palpitations, or dyspnoea at rest.



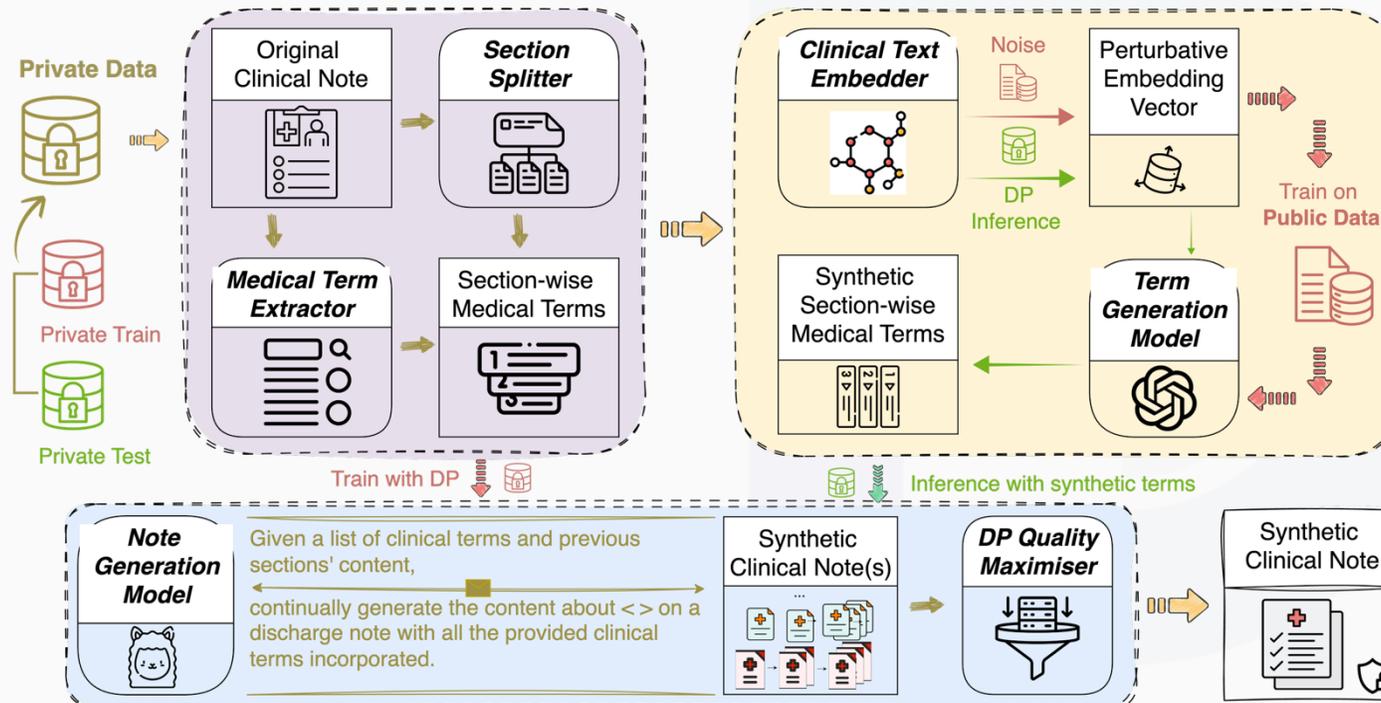
During your stay, you received various treatments aimed at improving your cardiovascular health, addressing concerns about infections and increasing blood flow to organs damaged by strokes or other vascular events.

Your condition stabilized significantly after you stopped smoking cigarettes four weeks ago and started drinking alcohol regularly three months ago. Your weight returned slowly after stopping steroids for asthma treatment.

We recommend continuing to drink beer occasionally (about half pint per night) as recommended by your doctor unless advised against doing so by yourself or others nearby.



# Work streams: Technical development



- Built-in differential privacy
- Separate training for
  - document style
  - content: (perturbed) clinical terms used as a blueprint

<https://arxiv.org/abs/2509.10882>

# Work streams: Public involvement and engagement

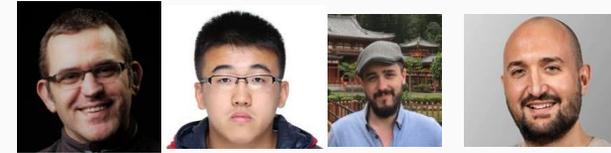
- **Public Advisory Panel (PAP)**

- A diverse group (gender, age, ethnicity, carer status) of 12 people
- To share their views and help shape the research and research outputs
- Two public contributors included in the Project Advisory Board

- **PAP meetings** (January, February, March)

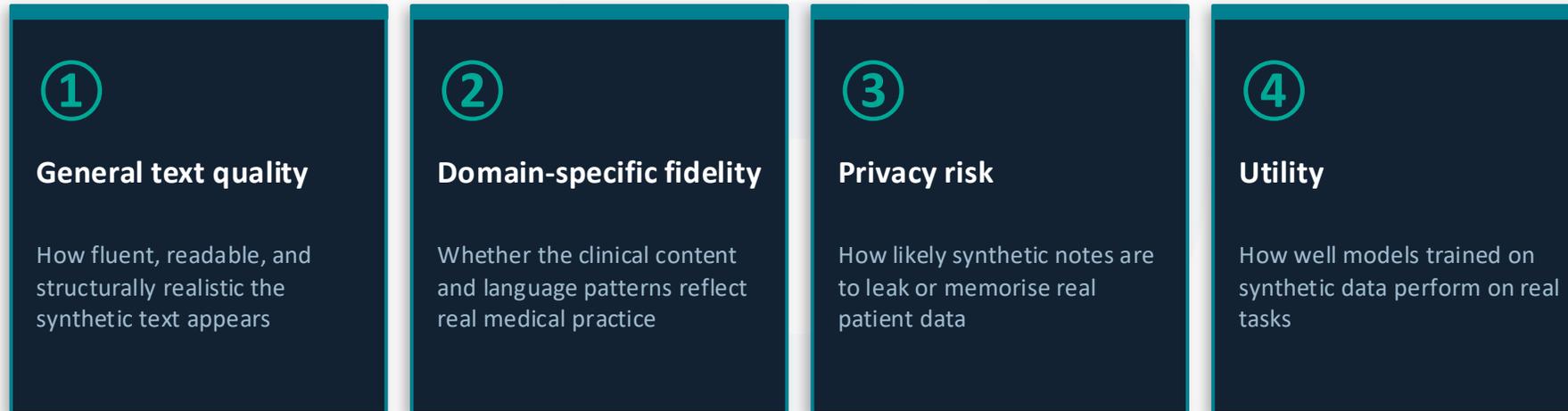
1. What you think of the idea to generate synthetic free-text (health) data to train AI?  
What would be your main concerns?
2. What is important to you and what worries you about how synthetic text may be made?  
What do we share in TREs: the model that generates the synthetic data or the data?
3. What ways of providing access to the model/data are acceptable?  
What needs to be included within meta-data descriptions of synthetic free text health datasets?





# Work streams: Synthetic text validation

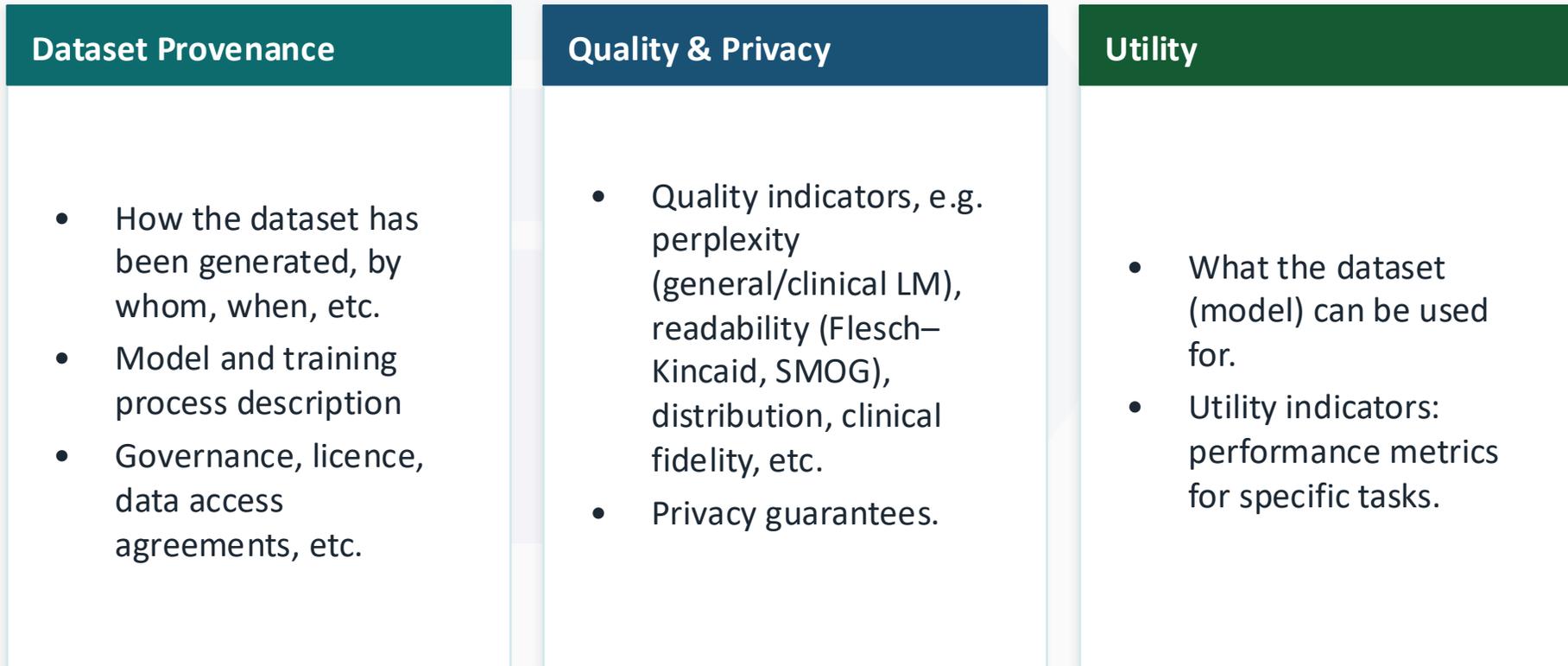
- What needs to be validated and how?
- Comparison to the real-world dataset



Both automated (e.g. distributions, utility) and manual (e.g. clinical fidelity) validation

# Work streams: Meta-data and descriptions

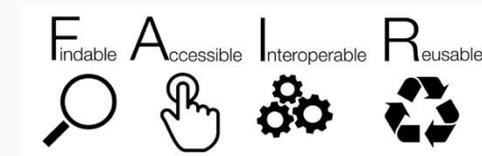
- Synthetic text “data card”



*Results reported for both original and synthetic dataset*

## Project outcomes

- **A synthetic text generation prototype packed as a Docker container**
  - Can it run within a TRE?
  - Can we federate text generation?
- **Report on public views on synthetic text generation**
  - Key concerns and expectations
- **Synthetic text validation protocol**
  - What needs to be validated?
- **Synthetic free-text data card template**
  - What to report to ensure FAIRness



# Get in touch!

- Email [gnenadic@manchester.ac.uk](mailto:gnenadic@manchester.ac.uk)
- DARE UK working group:

*Safetext - Community-led Protocols for the Safe and Responsible Use of De-identified and Synthetic Healthcare Text for AI Development*





# Introduction to GROVE

Dr Olly Butters, University of Liverpool

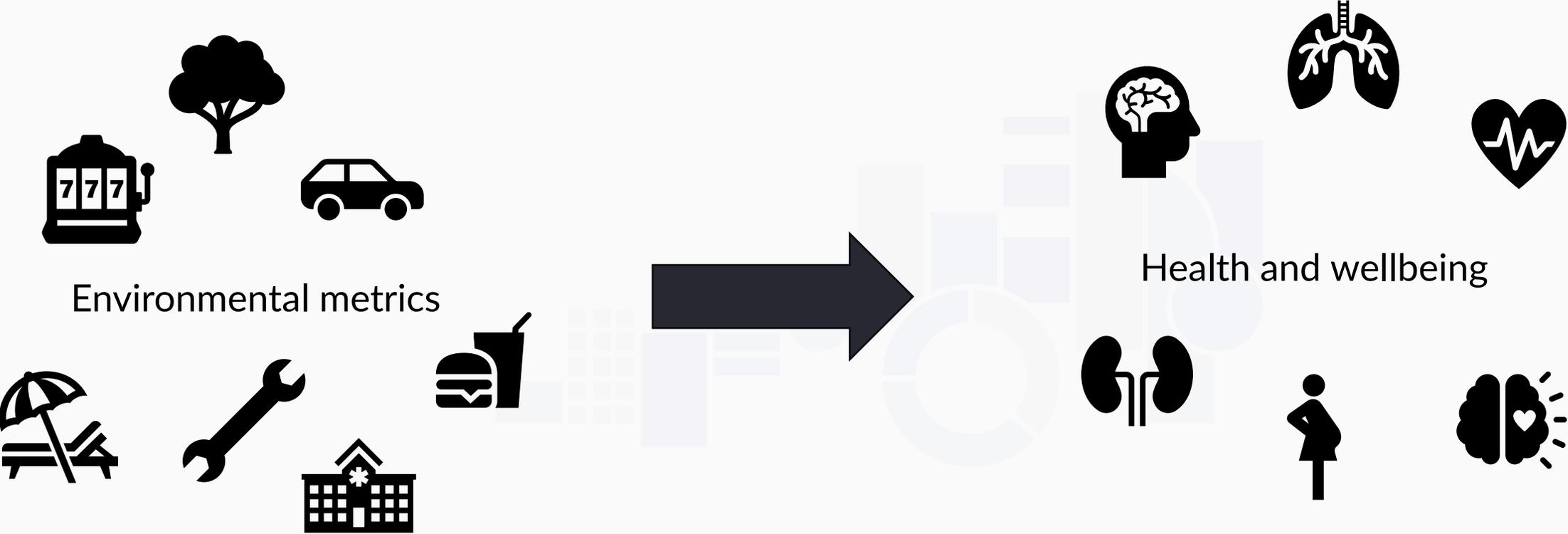


# Where we live profoundly affects our health & wellbeing



We often use neighbourhoods (~1500 people) as the smallest area to describe environmental effects e.g. air pollution

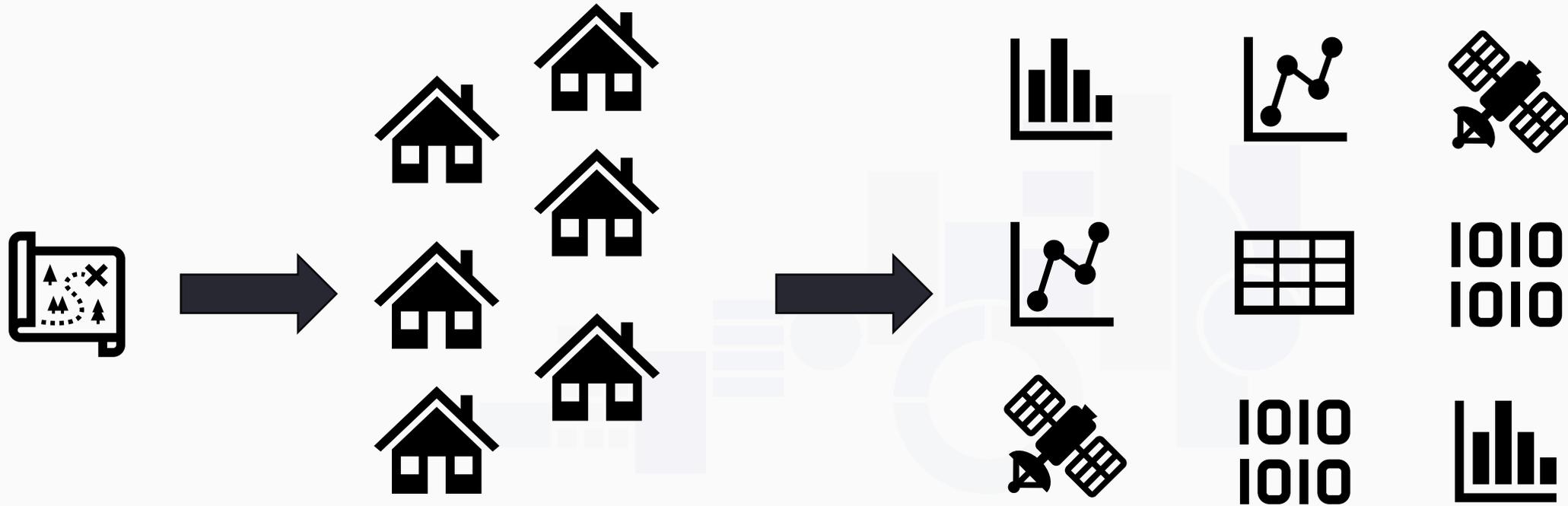
# Not just air pollution that impacts health and wellbeing



# Goal

We need to be able to link environmental metrics at a household (not just neighbourhood) level, to health data, in a safe and secure way.

# This is challenging



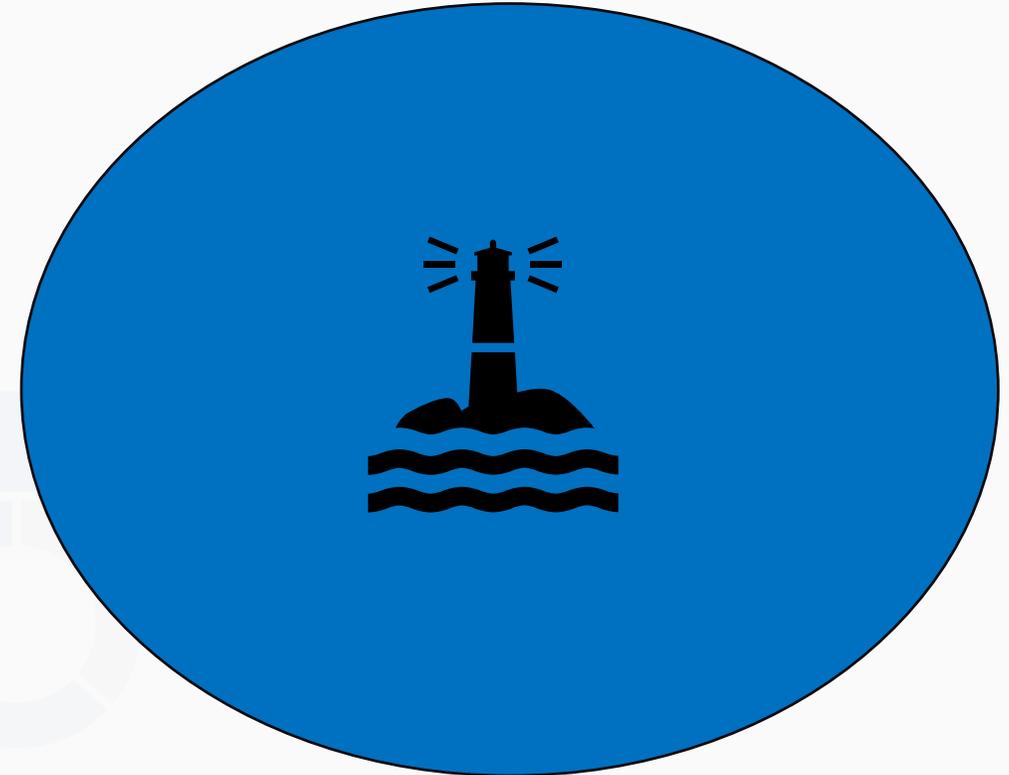
There are a lot of homes in England (~25M)

Lots of data generated for each property means there is a lot of data!

# With big data comes big risks

What if we investigate how the amount of nearby blue space (rivers, canals, the sea etc) affects health?

What about a lone lighthouse keeper?



ID number	Had heart attack	Blue space %
12345	Yes	11
13377	No	34
14873	Yes	100
19832	No	25

← May be the only person with 100% blue space coverage nearby.

# Governance for household-level environment and health data

- **Aim: establish a system that safely links regional health records and environmental data across England, by:**
  - 1) Engaging with communities – particularly young people, healthcare providers and governments, to understand their viewpoints on linking health and environmental data.
  - 2) Creating a pipeline for linking the regional health data records with environmental data metrics, including from satellite images and mapping of natural infrastructure like green spaces.
  - 3) Developing new sophisticated technology to ensure that our newly linked data cannot be used to identify individual people.

# Public and project partner engagement

Workshops and consultations with the public and project partners

## Public - particularly young people

- Viewpoints on benefits and challenges of linking environmental and health data.
- Opinions on potential types and mechanisms of linking and using this linked data.
- Develop easy to digest summaries (including animation).
- Co-designed with seldom heard communities.

## Project partners

- Understand the different governance frameworks the partners work in.
- Opinions on potential types and mechanisms of linking and using this linked data.
- Develop best practice guidelines – including input from the public – so we can scale to a national level.
- We will establish a Community of Practice to share learning.

# Technological Development

- **Richer linkage, higher disclosure risk**

- **Problem:** More spatial variables can create a unique “fingerprint” and enable location inference after pseudonymisation.
- **Output:** Risk model + safe enrichment rules (aggregation, suppression, spatial k-anonymity / minimum cell sizes).



- **Spatial insight without coordinates**

- **Problem:** Removing coordinates protects privacy but prevents spatial methods (and control for spatial dependence).
- **Output:** Coordinate-free spatial objects (adjacency/proximity graphs; spatial weights matrices) resistant to reverse engineering.



- **Maps that don't disclose**

- **Problem:** Maps can leak identity/location via small areas and low counts.
- **Output:** Disclosure-controlled mapping outputs (generalisation, density/heatmaps, automated output checks) for DataSHIELD-style analysis.



# Meet the team

Sarah Rodgers  
Project Lead

Olly Butters  
Project Co-Lead

Roberto Villegas-Diaz  
DataSHIELD geospatial  
lead Research Software  
Engineer

Hayley Jones  
DataSHIELD geospatial  
Research Software  
Engineer

Terry Bryant  
Public Adviser

Rebecca Geary  
Project Co-Lead

Kimon Krenz  
Project Co-Lead

Konstantinos Daras  
Environmental data  
expert

Elly King  
Comms and  
Impact Lead

Nancy Schumann  
Administrator

Rebecca Wilson  
Project Co-Lead

Emily Rempel  
Community  
Engagement  
Researcher

Anna Head  
Research Fellow,  
Healthy Urban Places

Demin Hu  
Research Fellow,  
Secure Spatial  
Analytics

Louise Jaeger-Wright  
Project Manager

Get in contact to:

- contribute to our stakeholder or public workshops later in 2026
- Join our Community of Practice

[hipr@liverpool.ac.uk](mailto:hipr@liverpool.ac.uk)



**STAR-TRE**  
A DARE UK PROJECT

# Safe and Trustworthy Assessment of Risk in Trusted Research Environments for Sensitive Free-Text Access

Dr Arlene Casey

Senior Vivensa Foundation Fellow, DataLoch NLP Programme Lead

University of Edinburgh



## Unstructured data (free-text)



Conversation captured  
(e.g. during phone  
encounter)

*Mr. A. Campbell from the Isle of Coll  
phoned the helpline. His grandson  
Calum has difficulties in school and  
hasn't visited. Angus feels  
increasingly isolated. Visit arranged  
w Sally thurs 2pm -  
Angus is at the 'the old mill'.*

## Structured data

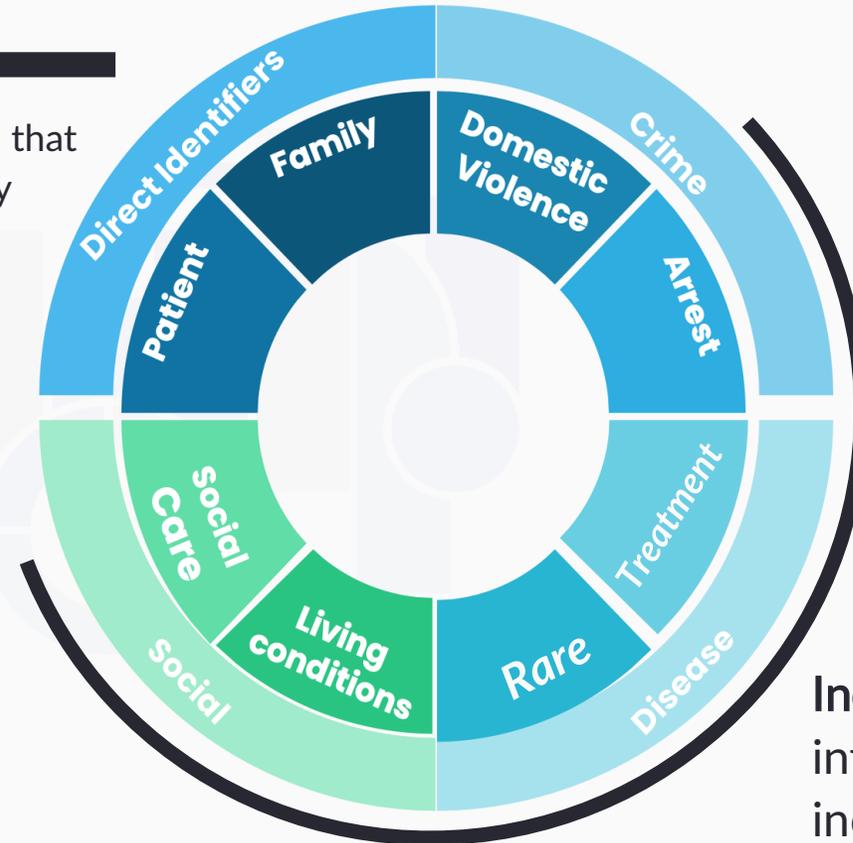
ID	First Name	Address
1	Angus	Old Mill
2	...	...

# Privacy Risks in Sensitive Free-text Notes

-  Names
-  Dates
-  Addresses
-  Phone Numbers
-  Medical Record No

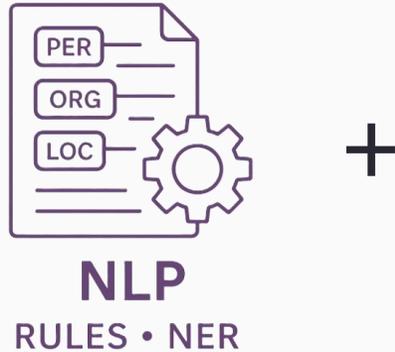


Examples of information that identifies patients, family members, medical professionals directly



Indirect Identifiers: information that increases the risk of re-identification

# Addressing and Managing Privacy Risks with Natural Language Processing (NLP)

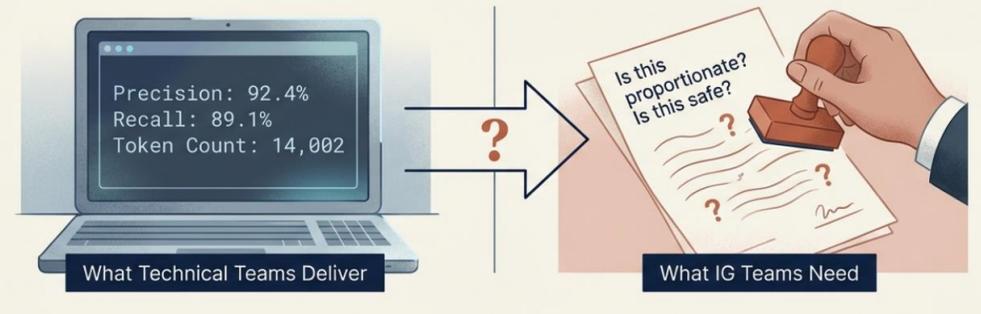


Current privacy risk identification focuses on direct identifiers



Newer technologies may be better at finding privacy risks

Algorithms provide metrics, but IG teams need support for judgment calls

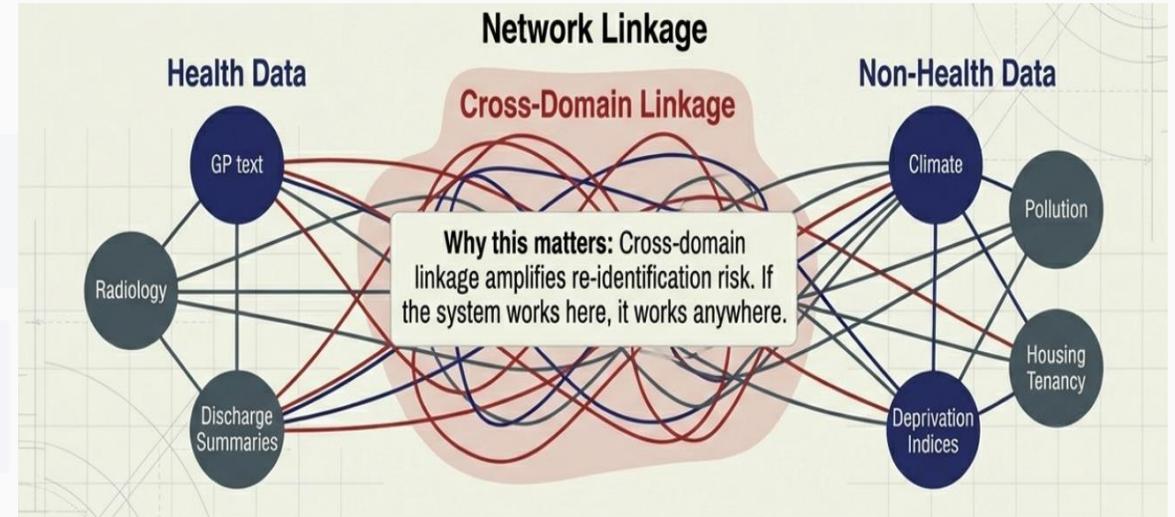


Embedding privacy-risk algorithms into a toolkit for TREs to manage and assess privacy risk transparently, with public support.

# Addressing and Managing Privacy Risks

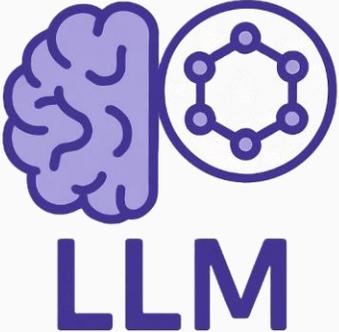


Transparent, auditable decisions with human oversight

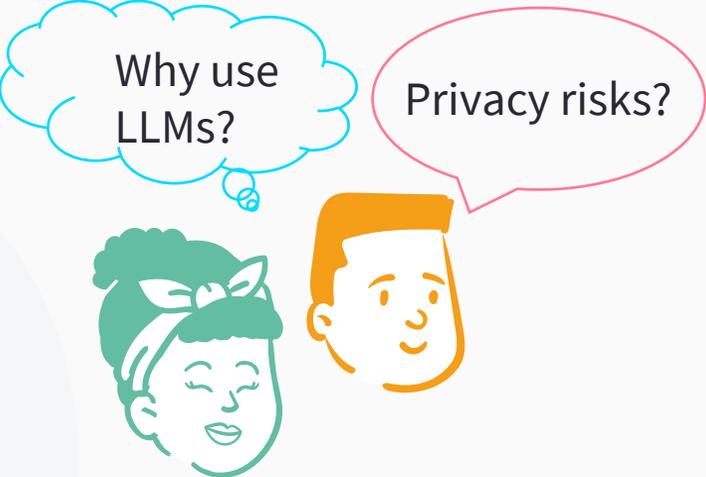


- Complex, linked projects (data) to work through and test
- Community-based approach to building Privacy Risk Management tools, algorithms, and processes through Collaboration Cafes, Public workshops and TRE workshops

# Our Public Involvement and Engagement



Large Language Models (LLMs), still a relatively new technology and we don't want to make assumptions about how they should be used



# Public Consultation

**Names: e.g. Mr Robertson**  
Replace with pseudonyms / remove

**Sensitive conditions: e.g. suspect drug overdose**  
Generalise or remove

**Alcohol / drug liaison**  
Generalise: appropriate medical care

**Specific locations: e.g. council housing in Torry**  
Remove or generalise

**Health Data - Hospital Record**




*Mr Robertson, 40yo M was admitted to MRI, alcohol intoxication suspected drug overdose(?). Seen by alcohol and drug liaison. He was discharged after two days with a referral to Lorso community mental health services.*

*History: Jamie lives in council housing in Torry with his partner and young daughter, (8). History of anxiety and previous attendance at A&E in 2021 for alcohol-intoxication.*

*Mr Robertson reported feeling increasingly isolated after losing his job as a delivery driver in January 2023. His partner, Laura Brown, works part-time at a local supermarket and has expressed concerns about his alcohol use and mood changes. Family are known to mental health and Drug liaison services, recently referred to the Melrose community wellbeing programme.*

*Follow-up with the drug liaison scheduled for 25 March 2023.*

**Referral to Lorso services**  
Generalise: referral to mental health services

**Ages: e.g. 8yo**  
Replace with age range

**Mental health: e.g. history of anxiety**  
Anonymise / remove

**Family known to mental health/drug liaison services**  
Generalise: access mental health support

**Dates: e.g. 25 March 2023**  
Approximate, e.g. late March 2023

**Uncommon job roles: e.g. delivery driver**  
Use broader term

# Stay in touch! – QR codes

Visit our website



 SCAN ME

Subscribe to our newsletter



 SCAN ME

Follow us on  
LinkedIn



 SCAN ME

Follow us on X  
(Formerly Twitter)



 SCAN ME

Arlene Casey  
Arlene.Casey@ed.ac.uk



# Trusted Research Environments Software Stewardship

Click to edit Master subtitle style



A Research Project, funded by Data and Analytics Research Environments (DARE) UK, running until January 2027, and a collaboration between University College London, Thames Valley & Surrey Secure Data Environment and the University of Oxford.



George Svarovsky, University College London

- Product Owner of the UCL Advanced Research Computing TRE
- Collaborator on DARE-UK funded projects
- Experienced in software engineer practices, management, & secure infrastructure



Professor David De Roure, University of Oxford

- Academic Director of Digital Scholarship & Laboratory for AI Security Research
- SSI cofounder
- National roles on digital research infrastructures
- Knowledge of data infrastructures across disciplines, reproducibility, data security, software sustainability, AI



Damon Strange, Ex Nihilo/Uni. of Oxford

- Experienced digital project manager
- Expertise in planning & delivery of complex stakeholder projects, running workshops & public events
- Knowledge of research data management and digital infrastructures



Professor Brian Marsden, University of Oxford

- University of Oxford Academic Lead for Research Computing and TREs
- Chair of national NIHR BRC Informatics Forum
- Expert in research informatics and data management in biomedical spaces



Kinga Várnai, OUH NHS FT, Thames Valley and Surrey Secure Data Environment



- Special Projects Lead for TVS SDE
- Experienced programme manager with a background in software engineering
- Expertise in delivering large-scale clinical informatics change programmes in public institutions



“**TRE Software Stewardship** is the responsible management of TRE software (including both **Research** code and **Environment** code) throughout its development and its **operational** lifecycle in **multiple** deployments, including its relationship to the organisational **context** ensuring the maintenance of security and compliance requirements.”

## Why is Software Stewardship Hard?

- TREs inherently require secure coding practices
  - Software changes in shared code must be managed to prevent context-specific vulnerabilities and maintain the 5 Safes
- The context is different between orgs and over time
  - Local policies, services, certifications
  - Information Security Management Systems – what can be shared?
- Scalability is served by layered sharing of responsibility
  - E.g. compute platform, TRE core, research software
  - Responsibilities must be well-defined and well-understood
- Good practice in research software is still emerging
  - Researchers are not (often) software engineers

## Our Approach

*building on the shoulders of giants*

Focus on the Software/Security Development Life Cycle

- *including* the Information Security Management System

For real, in an exemplar collaboration:

- A certified, operational TRE under active DevSecOps
- An in-development derivative of the TRE
- Receiving data from an active SDE
- 5 Pilot Projects with unique software requirements, including AI

## Why this Collaboration?

- Oxford and UCL are aligned in their requirements for supporting research on sensitive data
  - Large *variety* of research domains,
  - Large *volume* of research projects,
  - with support for High Performance and AI Compute,
  - in which the TRE is *not* the Data Provider.
- UCL ARC's TRE and ISMS use a Shared Responsibility Model, ideal for this kind of TRE
  - in operation and certified to ISO 27001 and DSPT at UCL
- Oxford would like to develop and implement best practice in TRE Software Stewardship
- UCL want Oxford's help to continuously improve

## What Does Success Look Like?

- An exemplar of federated TRE software engineering
  - Working with **shared** secure, scalable TRE code, ISMS documentation and Research software
  - Establishing a pattern for scaling to other institutions, using open-source
  - DevOps **in production** and ISO 27001 certified, and **active** research
  - In two large, complex institutions with broad research portfolios and high watermark security requirements
  - Supporting new approaches in TRE federation
- Leading to a unified, open approach to Software Stewardship
  - which we will continue to operate
  - and others can adapt and build on
  - underpinned by P(P)IE activities

## Pilot Project Spotlight

# The ORIGIN Project

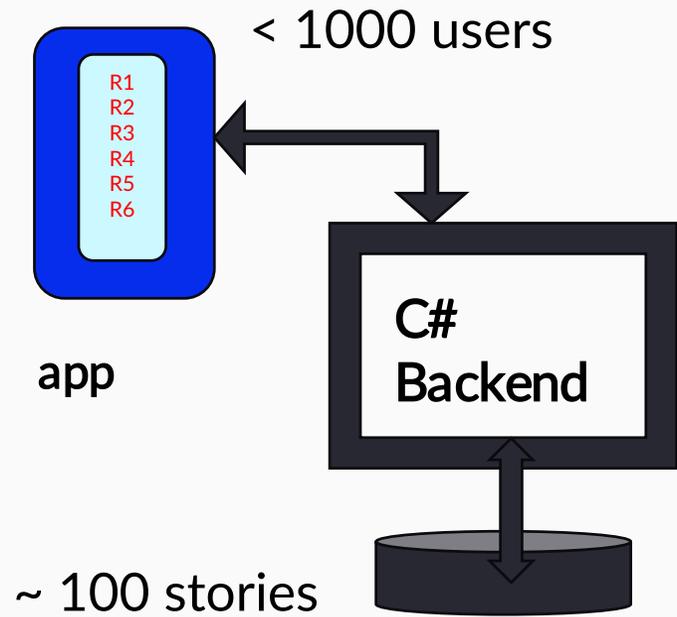
Online arts and culture for young people's mental health

- Online arts and culture for young people's mental health – NIHR funded collaboration between NHS Trusts, multiple universities, museum and charity partners.
- Closely involves Oxford Gardens, Libraries and Museums (GLAM).
- Community engagement and collaboration at the heart - young people who don't feel heard, feel excluded or feel disadvantaged in some way.
  - Recruited Research Advisory Group of 16–24 year-olds.
  - Young people involved in co-producing the interventions.
  - Representatives at every ORIGIN meeting.

<https://fundingawards.nihr.ac.uk/award/NIHR204148>



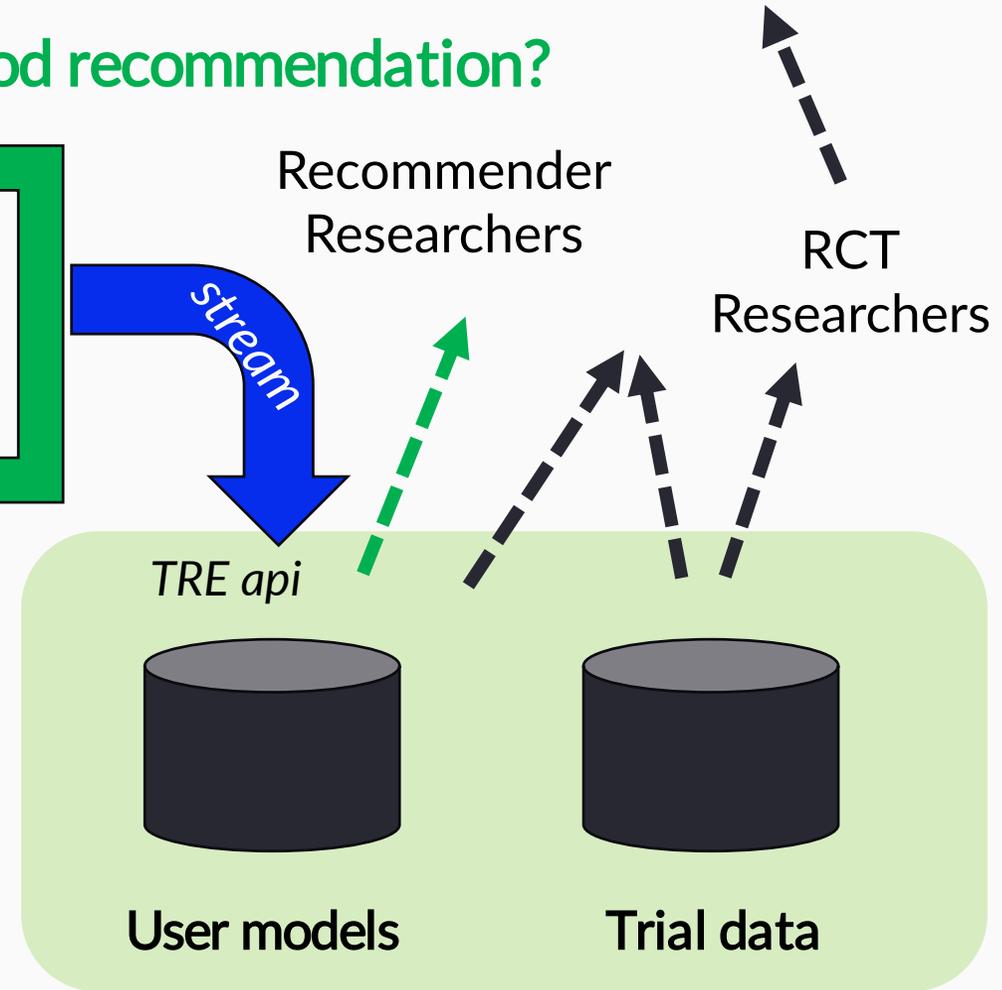
# ORIGIN Architecture



**Codesign Group**

Q. What makes a good recommendation?

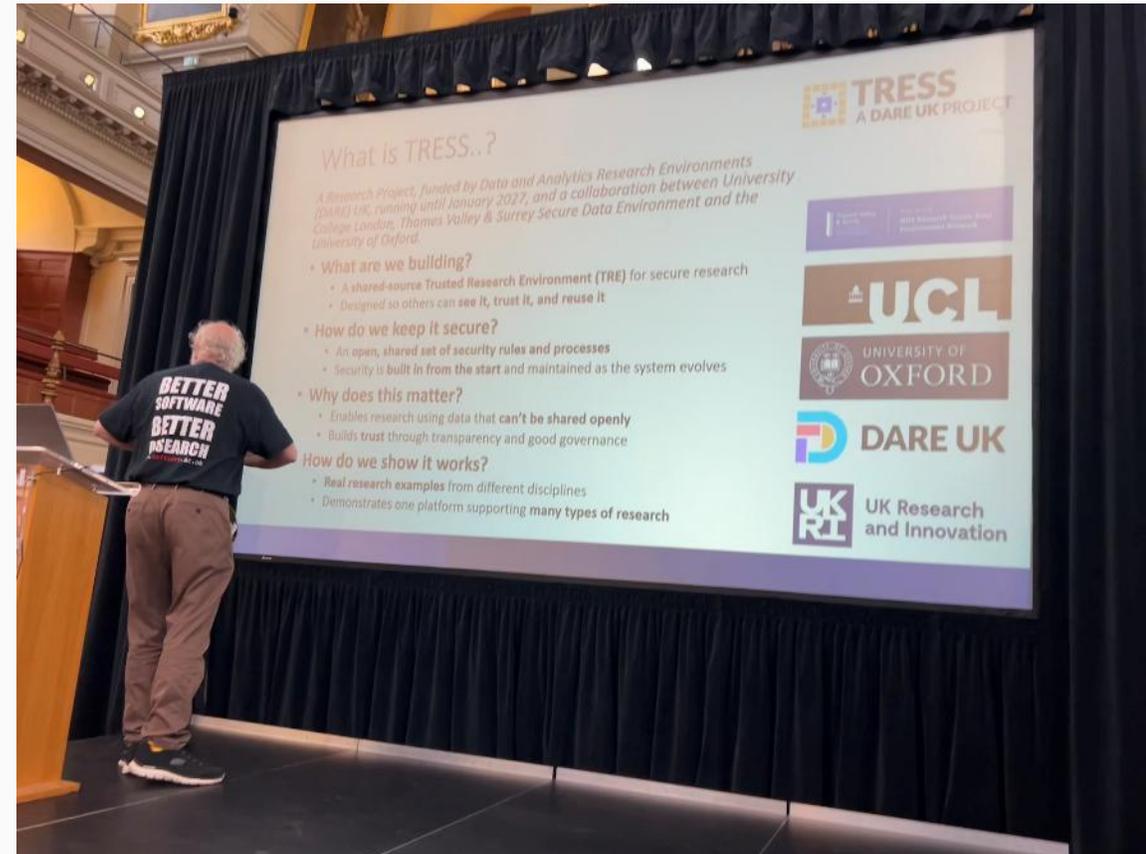
The ORIGIN Project  
Online arts and culture for young people's mental health



# Public Impact and Engagement and the TRESS Project



- 1<sup>st</sup> Session held celebrating International Love Data Week in Oxford – Providing a TRESS introduction
- A series of blogs, interest pieces and webinars are included within the plan
- Co-creation session with the Thames Valley & Surrey Secure Data Environment’s Community of Practice – focused on:
  - The role of academic institution and research, and its relationship with the public
- ‘Summer Showcase’ proposed to take place to provide the opportunity to demo what researchers have achieved using the TRE



# Stay in touch!



Subscribe to our newsletter  
[Click here to subscribe](#)



Visit our website  
[www.dareuk.org.uk](http://www.dareuk.org.uk)

Follow us



in

DARE UK



X

@DARE\_UK1

# Stay in touch! – QR codes

Visit our website



 SCAN ME

Subscribe to our newsletter



 SCAN ME

Follow us on  
LinkedIn



 SCAN ME

Follow us on X  
(Formerly Twitter)



 SCAN ME



# TRExt: TRE Text Analytics

Tim Beck, University of Nottingham  
tim.beck@nottingham.ac.uk



# Project summary

- Demonstrate how researchers can use TREs to analyse unstructured text
- Provide a pipeline for TRE providers to convert unstructured text into structured standard data models
- Show how the state-of-the-art in NLP and data mapping can be used by TRE providers, using technologies with underlying methods that can be applied to medical and other domains
- Capture public trust and concerns that will influence our design decisions

# Project team



Tim Beck



Graziela Figueredo



Phil Quinlan



Angus Roberts



Robert Stewart



Claire Newman



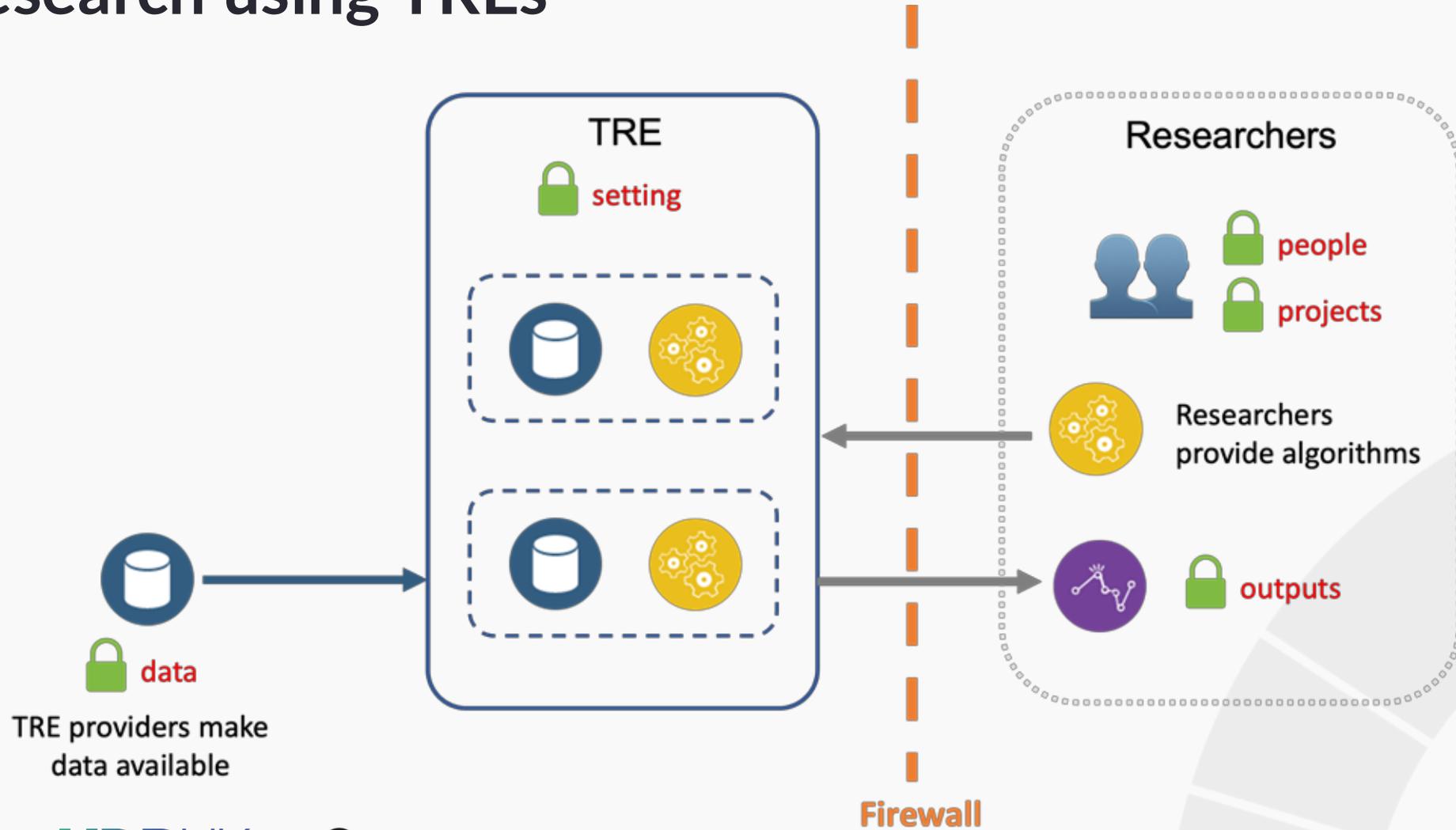
Simon Thompson



Legend for roles:

- Light blue dot: Text analytics
- Yellow dot: TRevolution
- Grey dot: Machine learning
- Green dot: Data mapping
- Blue dot: TRE provider
- Red dot: PIE specialist
- Purple dot: Clinician & product owner

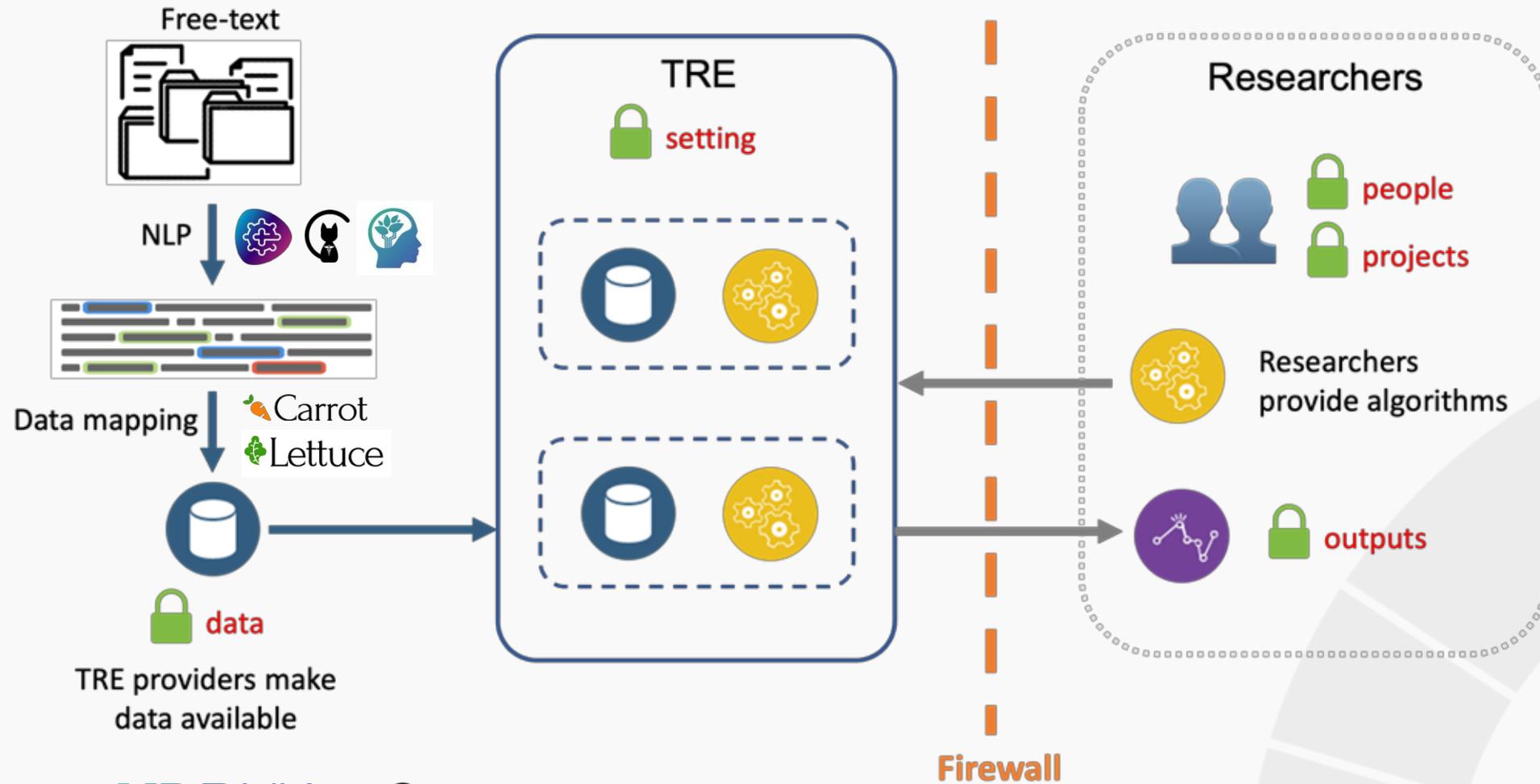
# Safe research using TREs



# What is the problem and what is TRExt doing about it?

- TRE databases can use the same structure to make it easier for computers to understand
- It is difficult for TRE providers to make information contained in free text available for research using these common databases
- TRExt is developing a pipeline that uses NLP and data mapping to convert text into formats that can be stored in databases
- The pipeline is open-source and will be free for TRE providers to use

# TRExt pipeline



# TRExt example

## Sources of clinical narrative text:

Free-text Electronic Health  
Record (EHR) notes

Outpatient letters

Discharge summaries

Radiology reports

Pathology reports

Referral letters

Mental health notes

# TRExt example

## Sources of clinical narrative text:

Free-text Electronic Health Record (EHR) notes

Outpatient letters

**Discharge summaries**

Radiology reports

Pathology reports

Referral letters

Mental health notes

## Reason for Admission

Elective admission for left total hip replacement due to end-stage osteoarthritis of the hip, causing pain and reduced mobility despite conservative management.



# TRExt example

## Sources of clinical narrative text:

Free-text Electronic Health Record (EHR) notes

Outpatient letters

**Discharge summaries**

Radiology reports

Pathology reports

Referral letters

Mental health notes

NLP

Find **diagnoses** and **interventions**

### Reason for Admission

Elective admission for **left total hip replacement** due to end-stage **osteoarthritis of the hip**, causing pain and reduced mobility despite conservative management.



# TRExt example

## Sources of clinical narrative text:

Free-text Electronic Health Record (EHR) notes

Outpatient letters

**Discharge summaries**

Radiology reports

Pathology reports

Referral letters

Mental health notes



**NLP**

Find **diagnoses** and **interventions**

### Reason for Admission

Elective admission for **left total hip replacement** due to end-stage **osteoarthritis of the hip**, causing pain and reduced mobility despite conservative management.

**Data mapping**

Map terms to the SNOMED code list

**left total hip replacement** maps to code SNOMED:770606008

**osteoarthritis of the hip** maps to code SNOMED:239872002

SNOMED:770606008

SNOMED:239872002

Codes added to database



# TRExt pipeline example

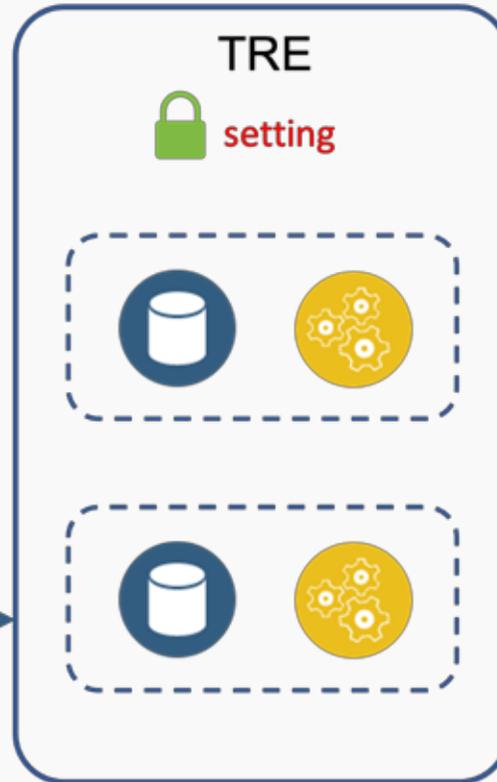
Reason for Admission  
 Elective admission for **left total hip replacement** due to end-stage **osteoarthritis of the hip**, causing pain and reduced mobility despite conservative management.

**left total hip replacement** → SNOMED:770606008  
**osteoarthritis of the hip** → SNOMED:239872002

Codes added to database



TRE providers make data available



Firewall

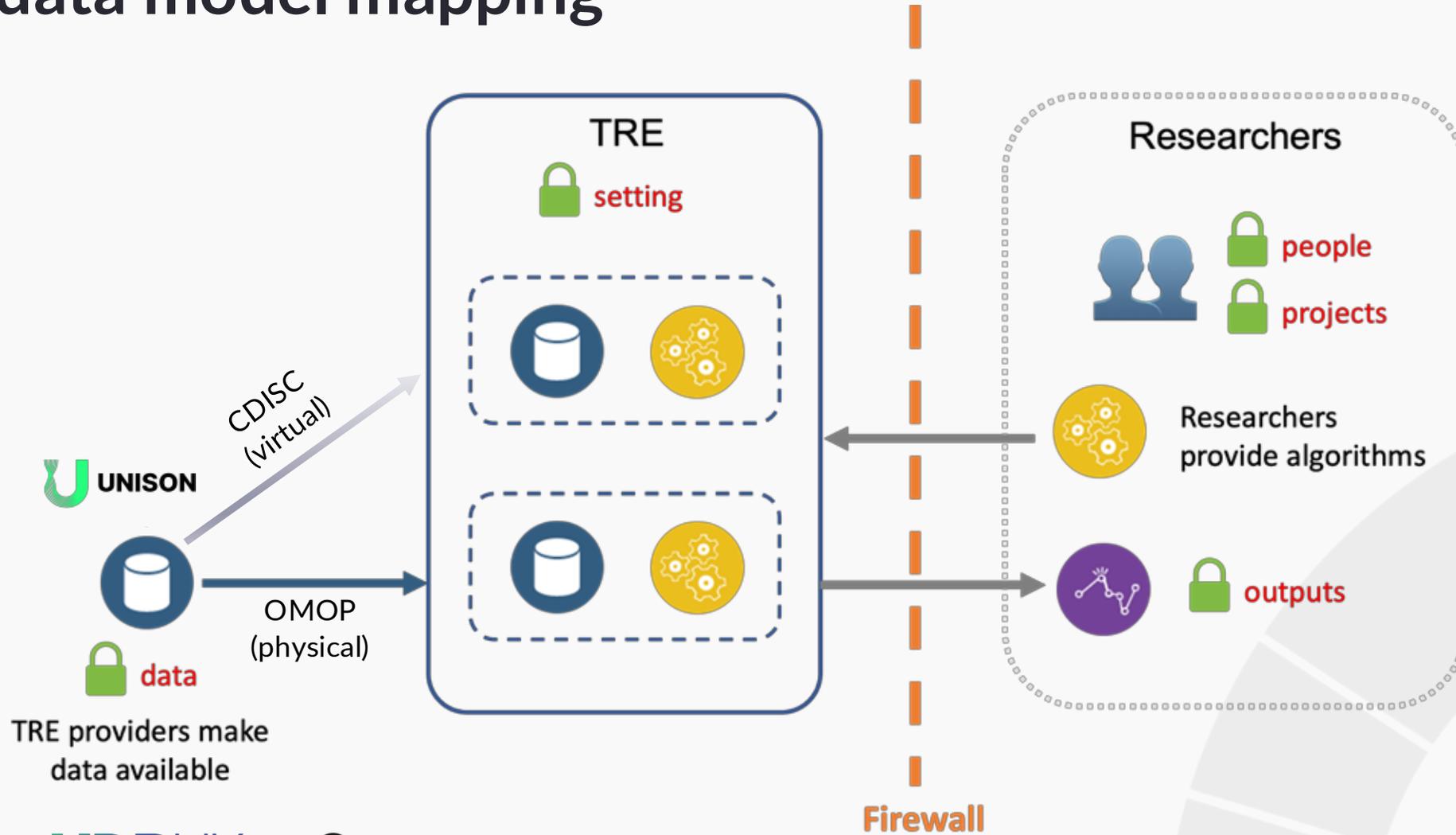
Researchers



Researchers provide algorithms



# TRExt data model mapping



# Public Involvement and Engagement

- Discussions with SAIL Databank's Consumer Panel have shaped our PIE strategy
- Management meetings are chaired by the project's lead public representative
- Focus groups will collect the views and concerns that members of the public have with the use of software and AI to process medical record text



- Outputs and findings from the focus groups will direct TRExt co-design decisions

# Acknowledgements



- Graziela Figueredo
- James Mitchell-White
- Thomas Rowlands
- Jonathan Couldridge
- Phil Quinlan

- Angus Roberts
- Robert Stewart
- Yamiko Msosa

- Claire Newman
- Simon Thompson

- Rory Popert
- Daniel Sozonov
- Artem Naumenko

The  
Alan Turing  
Institute



**TRUSTEE**  
A DARE UK PROJECT

---

## TRUSTEE: Safely using other people's computers as part of a Trusted Research Environment

Martin O'Reilly (Research lead) and Antony Chuter (PIE lead)

04 March 2026 | DARE UK Next-Gen Catalysts webinar



---

The key idea

---

# The key idea

**Goal:** To **safely use** computer systems **managed by others** for research with **sensitive data**

---

# The key idea

**Goal:** To safely use computer systems managed by others for research with sensitive data

**How:** By temporarily **taking control** of those systems and **securing them from their administrators**

---

# The key idea

**Goal:** To safely use computer systems managed by others for research with sensitive data

**How:** By temporarily taking control of those systems and securing them from their administrators

**Why:** To enable **research** with sensitive data, especially with **large-scale data** and **for AI**

---

# Trusted Research Environments (TREs)

TREs are highly secure computer systems ...

TREs are highly secure computer systems ...  
... where approved researchers ...

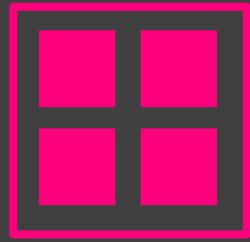
TREs are highly secure computer systems ...

... where approved researchers ...

... can carry out approved research ...

TREs are highly secure computer systems ...  
... where approved researchers ...  
... can carry out approved research ...  
... using sensitive data

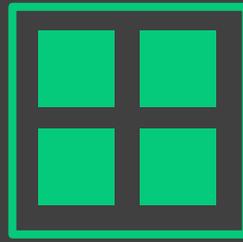
# TREs combine strong **technical** controls...



Project A



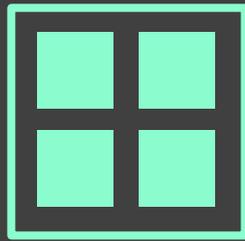
Project B



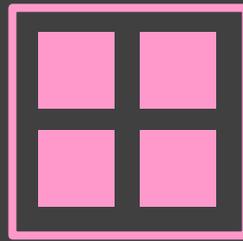
Project C



Project D



Project E

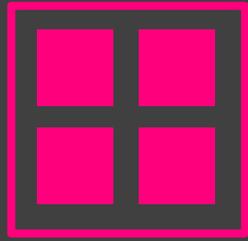


Project F

## Technical controls

- Isolated projects
- De-identified data
- No internet
- No copy + paste
- Secure connections

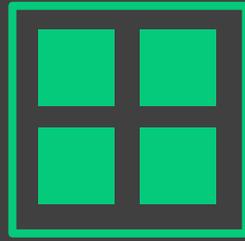
... with strong **rules** for people and processes



Project A



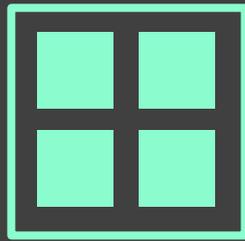
Project B



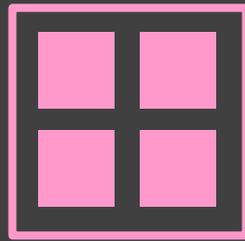
Project C



Project D



Project E



Project F

### Governance controls

- Project approval
- Researcher training
- Approved software
- Review of outputs

---

Why this work is needed

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI
- Large scale / AI computers are **not TRE-secure**

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI
- Large scale / AI computers are **not TRE-secure**
  - National supercomputers

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI
- Large scale / AI computers are **not TRE-secure**
  - National supercomputers
  - Institutional clusters

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI
- Large scale / AI computers are **not TRE-secure**
  - National supercomputers
  - Institutional clusters
  - Lab-hosted clusters

---

# Why this work is needed

- Limited processing power within existing trusted research environments (TREs), especially for AI
- Large scale / AI computers are **not TRE-secure**
  - National supercomputers
  - Institutional clusters
  - Lab-hosted clusters
  - Individual researcher workstations

---

Where we are

---

# Cloud TREs: Shared responsibility model

---

# Cloud TREs: Shared responsibility model

- Cloud service provides **secure building blocks**

---

# Cloud TREs: Shared responsibility model

- Cloud service provides secure building blocks
- TRE operator securely “wires together” these building blocks into a **secure system**

---

# Cloud TREs: Shared responsibility model

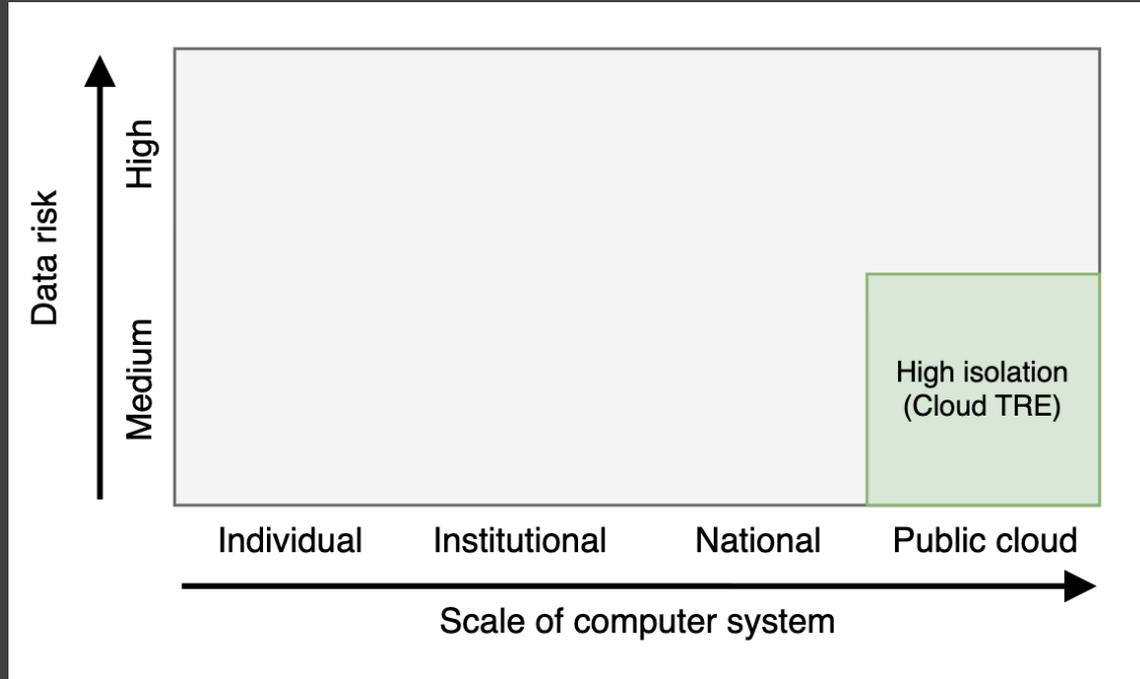
- Cloud service provides secure building blocks
- TRE operator securely “wires together” these building blocks into a secure system
- Cloud provider **promises not to look at data** stored and processed on their systems

---

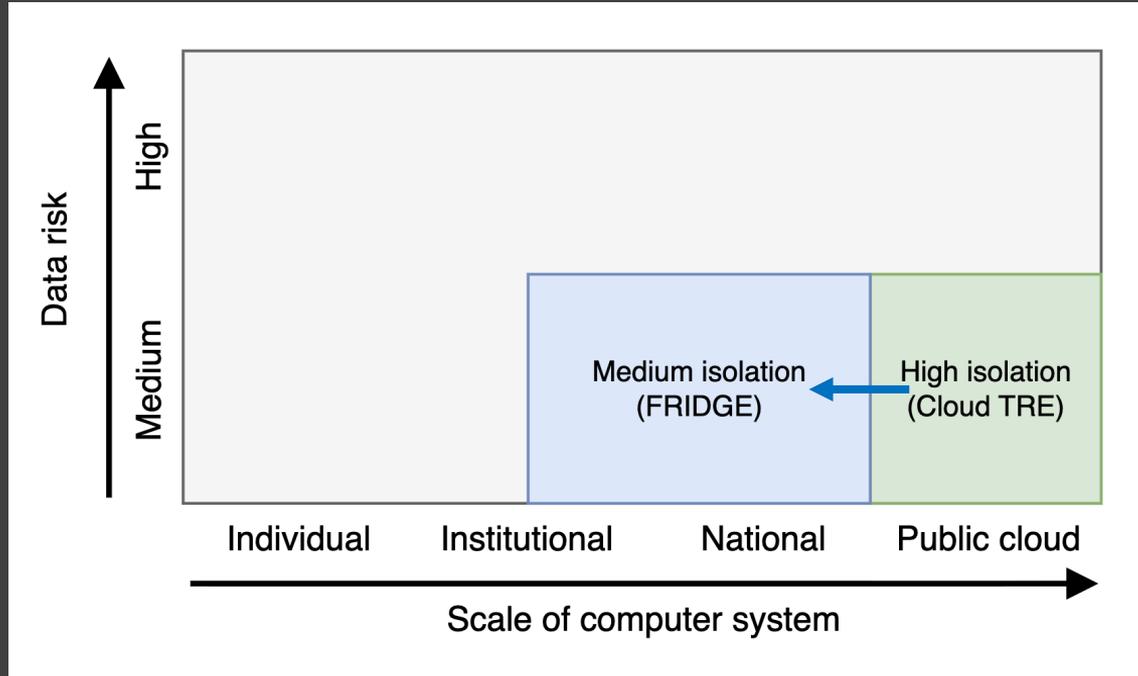
# Cloud TREs: Shared responsibility model

- Cloud service provides secure building blocks
- TRE operator securely “wires together” these building blocks into a secure system
- Cloud provider promises not to look at data stored and processed on their systems
- TRE operator **responsible for user access and data management** (information governance)

# Cloud TREs: Shared responsibility model



# FRIDGE: TREs on supercomputers



---

# How FRIDGE works

- Provides **technical protections** to **strongly isolate groups of computers** within a supercomputer

---

# How FRIDGE works

- Provides technical protections to strongly isolate groups of computers within a supercomputer
- Allows these isolated groups of computers to be controlled and configured by existing TRE owners

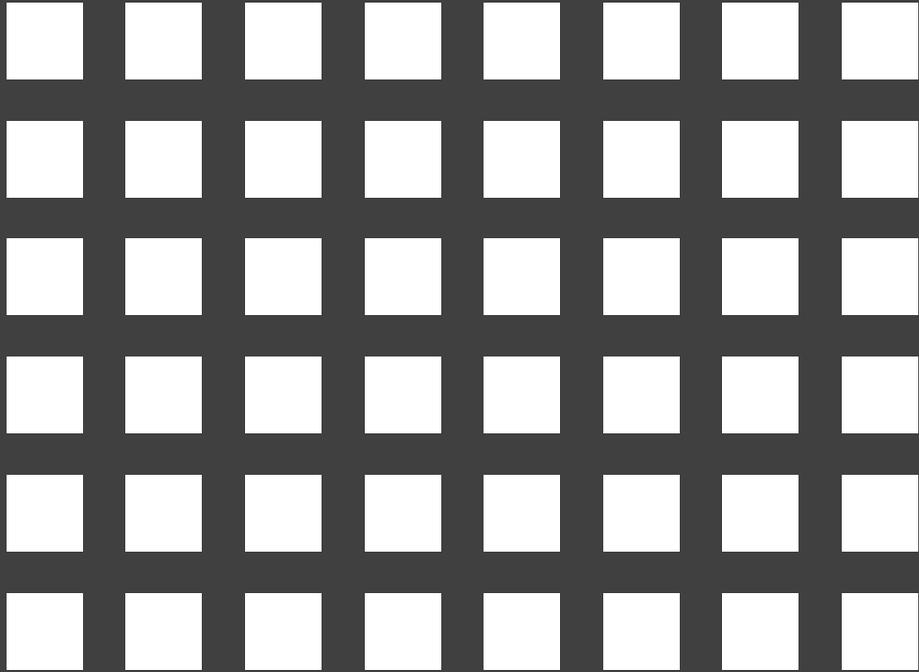
---

# How FRIDGE works

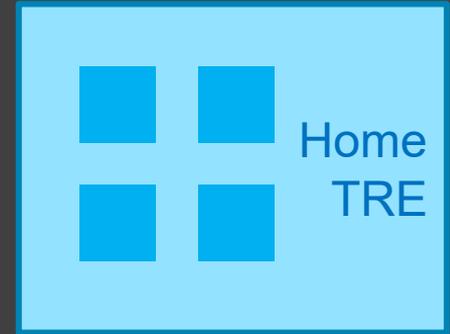
- Provides technical protections to strongly isolate groups of computers within a supercomputer
- Allows these isolated groups of computers to be controlled and configured by existing TRE owners
- Tells the **supercomputer** owner what they need to do to **ensure security**

---

# FRIDGE: Extending governance control

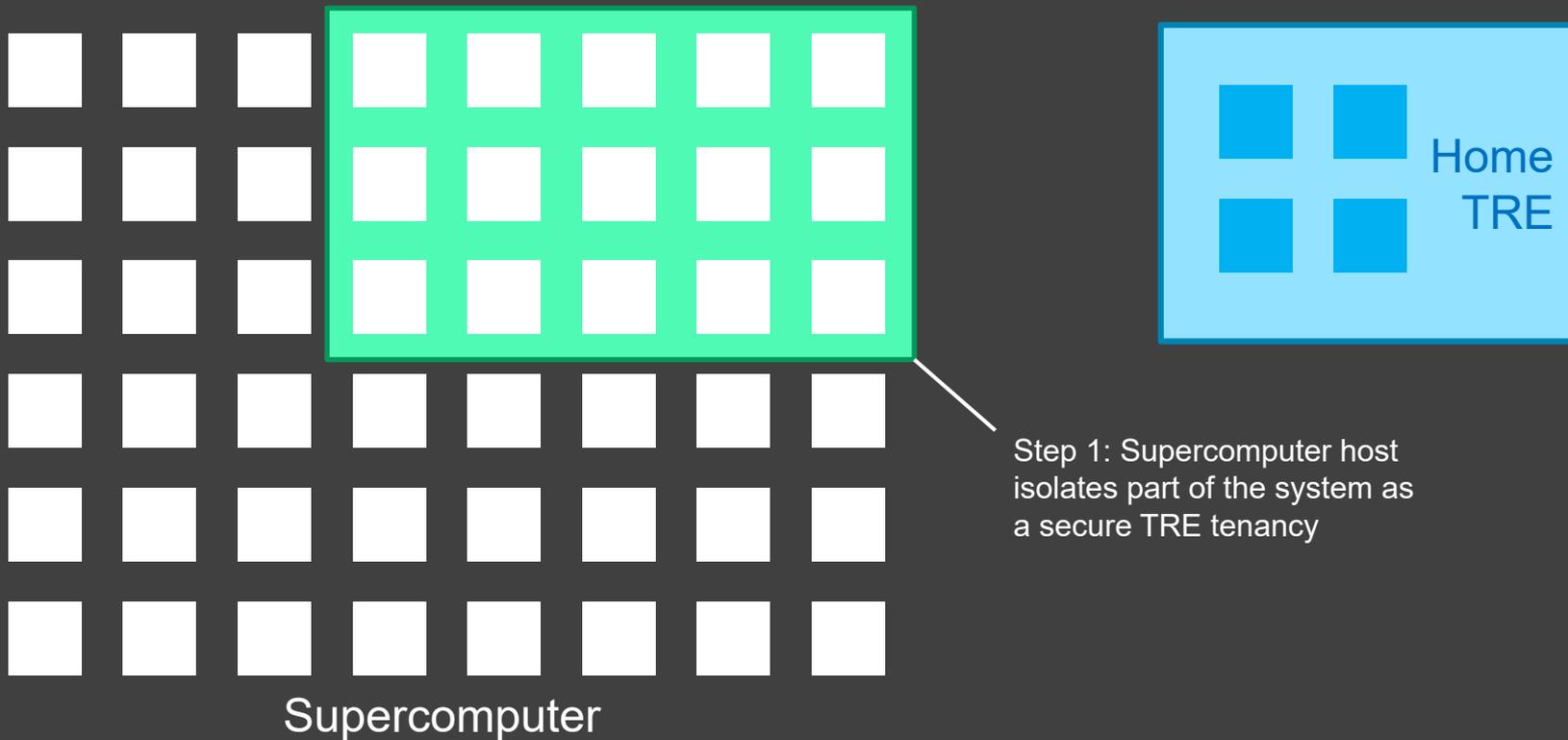


Supercomputer



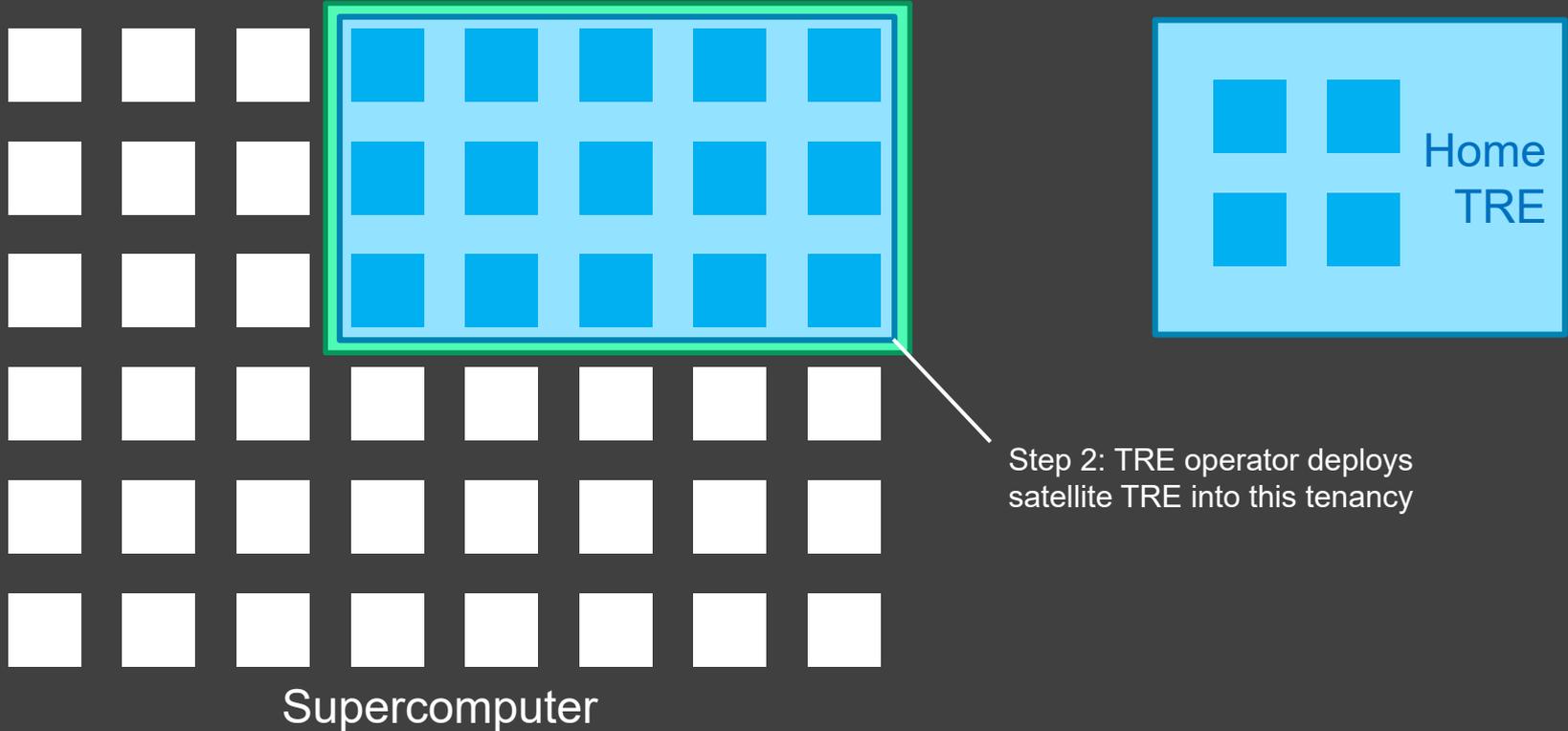
---

# FRIDGE: Extending governance control



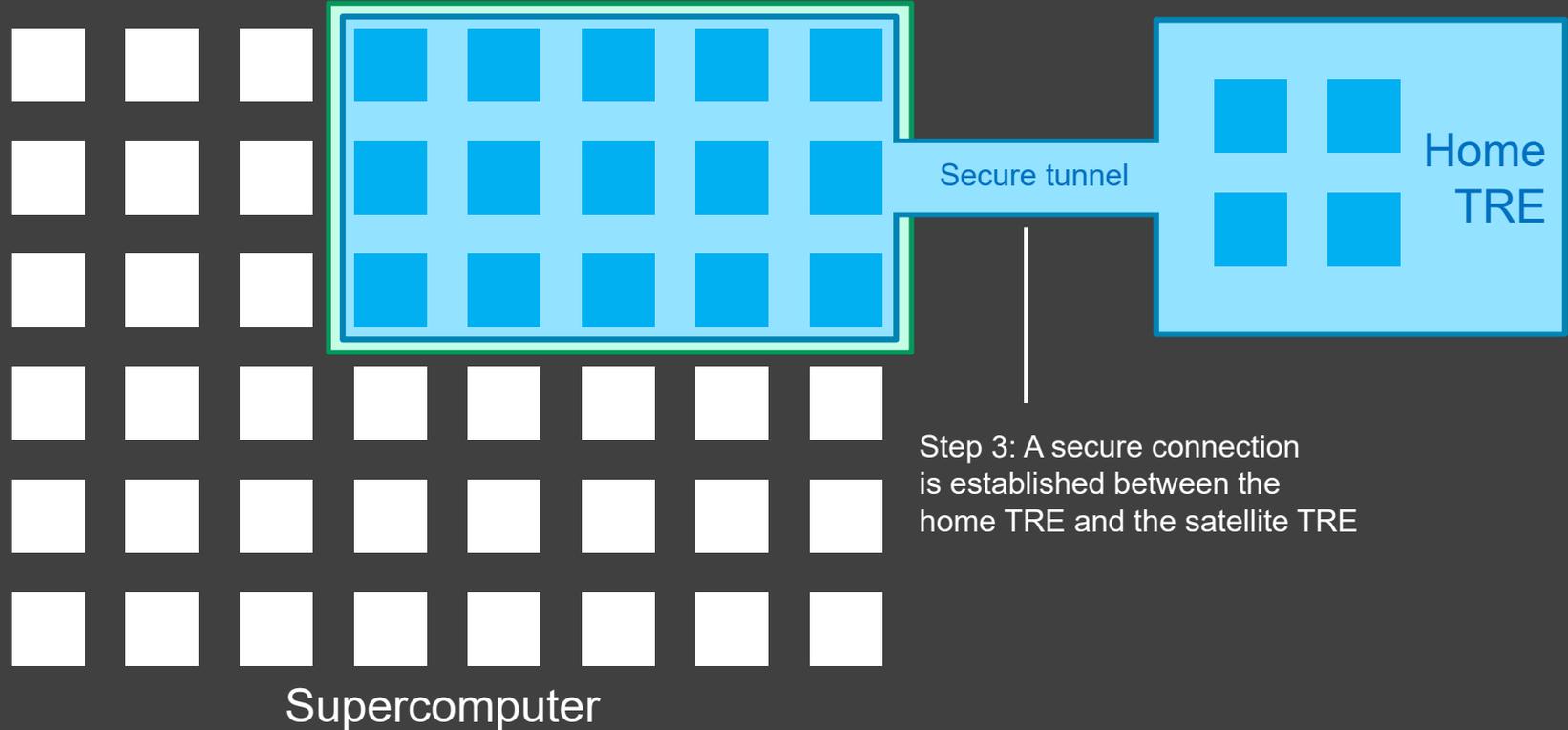
---

# FRIDGE: Extending governance control



---

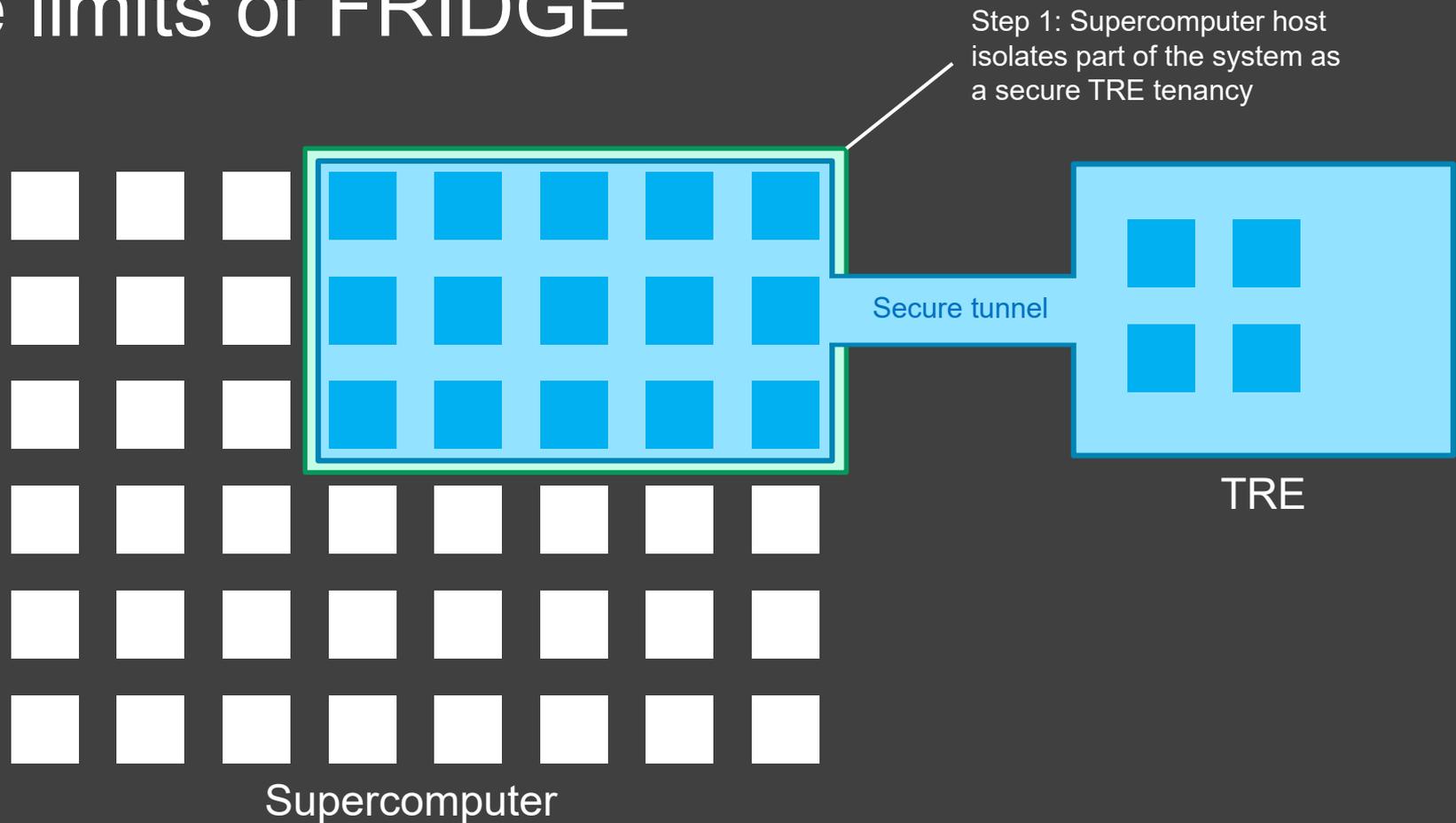
# FRIDGE: Extending governance control



---

# The limits of FRIDGE

# The limits of FRIDGE



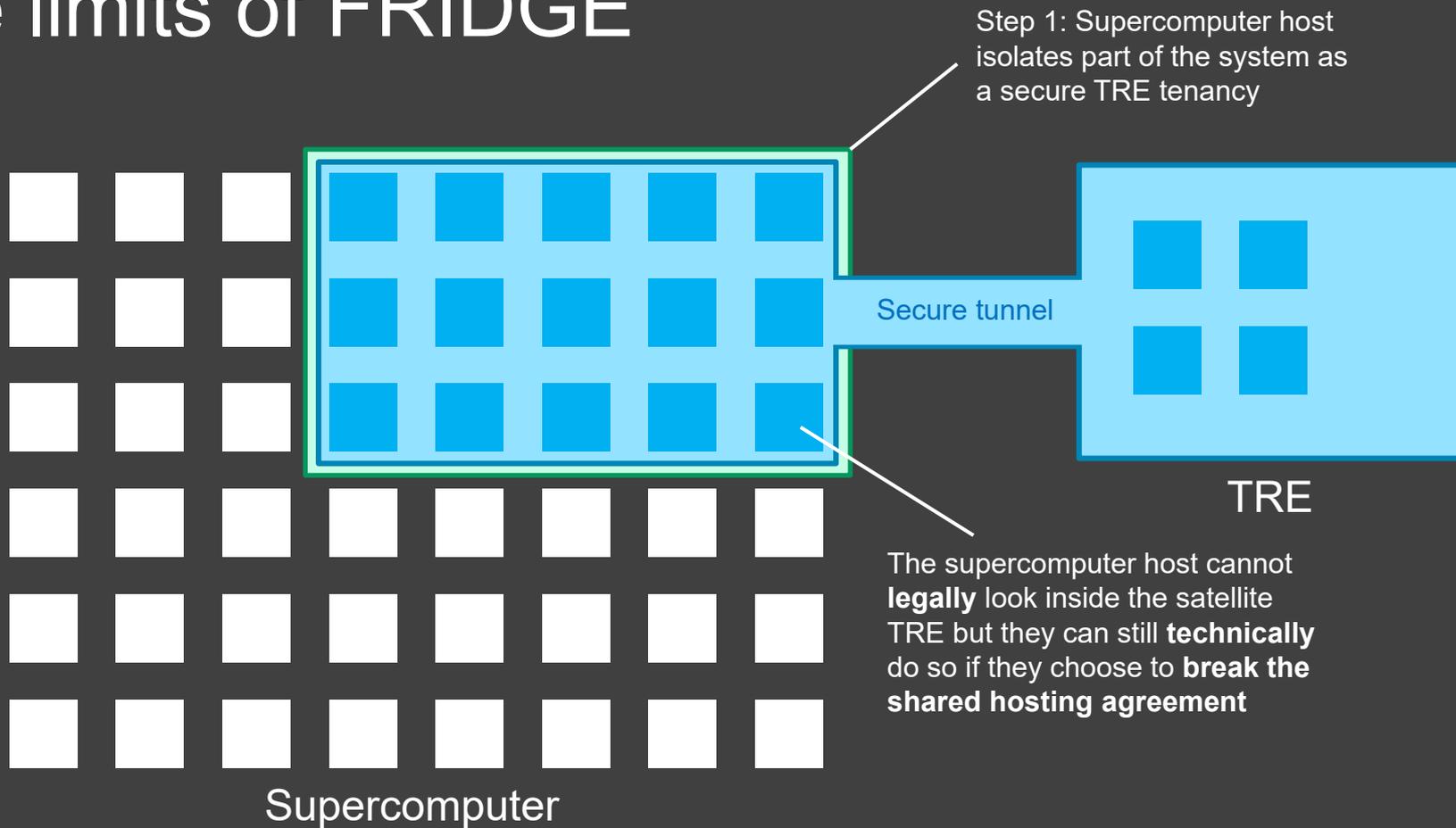
Step 1: Supercomputer host isolates part of the system as a secure TRE tenancy

Secure tunnel

TRE

Supercomputer

# The limits of FRIDGE



---

# The limits of FRIDGE

- Relies on a **shared responsibility** model

---

# The limits of FRIDGE

- Relies on a shared responsibility model
- May not be sufficient for **highly sensitive** data

---

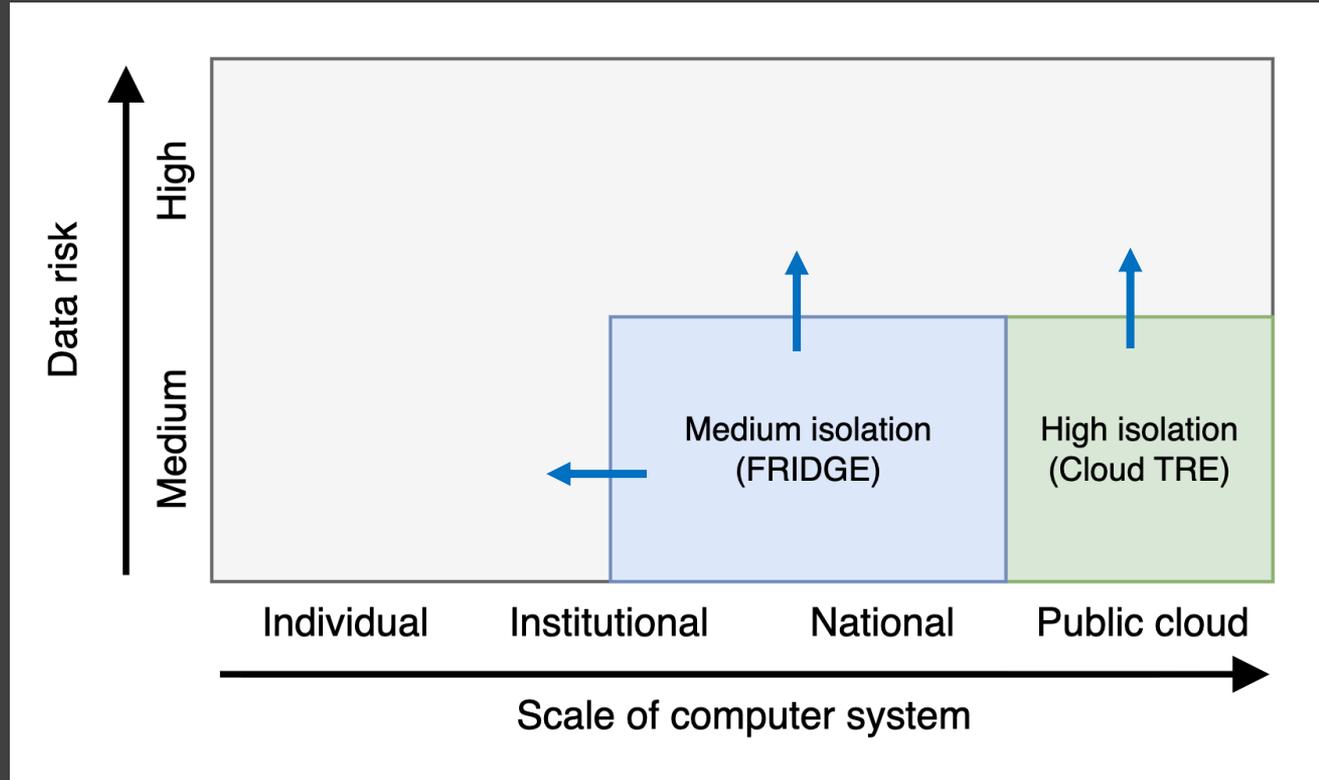
# The limits of FRIDGE

- Relies on a shared responsibility model
- May not be sufficient for highly sensitive data
- Doesn't support the “long tail” of **individual AI workstations** run by individual labs or researchers

---

Where we want to be

# Where we want to be



---

TRUSTEE: Extending technical control

---

# Confidential Computing

## Current encryption technologies



Data at rest

Encrypted on  
storage and  
databases



Data in transit

Encrypted when  
sent over the  
network

# Confidential Computing

## Current encryption technologies



Data at rest

Encrypted on  
storage and  
databases



Data in transit

Encrypted when  
sent over the  
network

## Confidential computing



Protect data in use

System owner  
cannot access data

---

# Confidential Computing

- **Technical** capabilities **built into** modern computer processors

---

# Confidential Computing

- Technical capabilities built into modern computer processors
- Uses a **secure mode** that makes the processor **inaccessible** to the **system administrators**

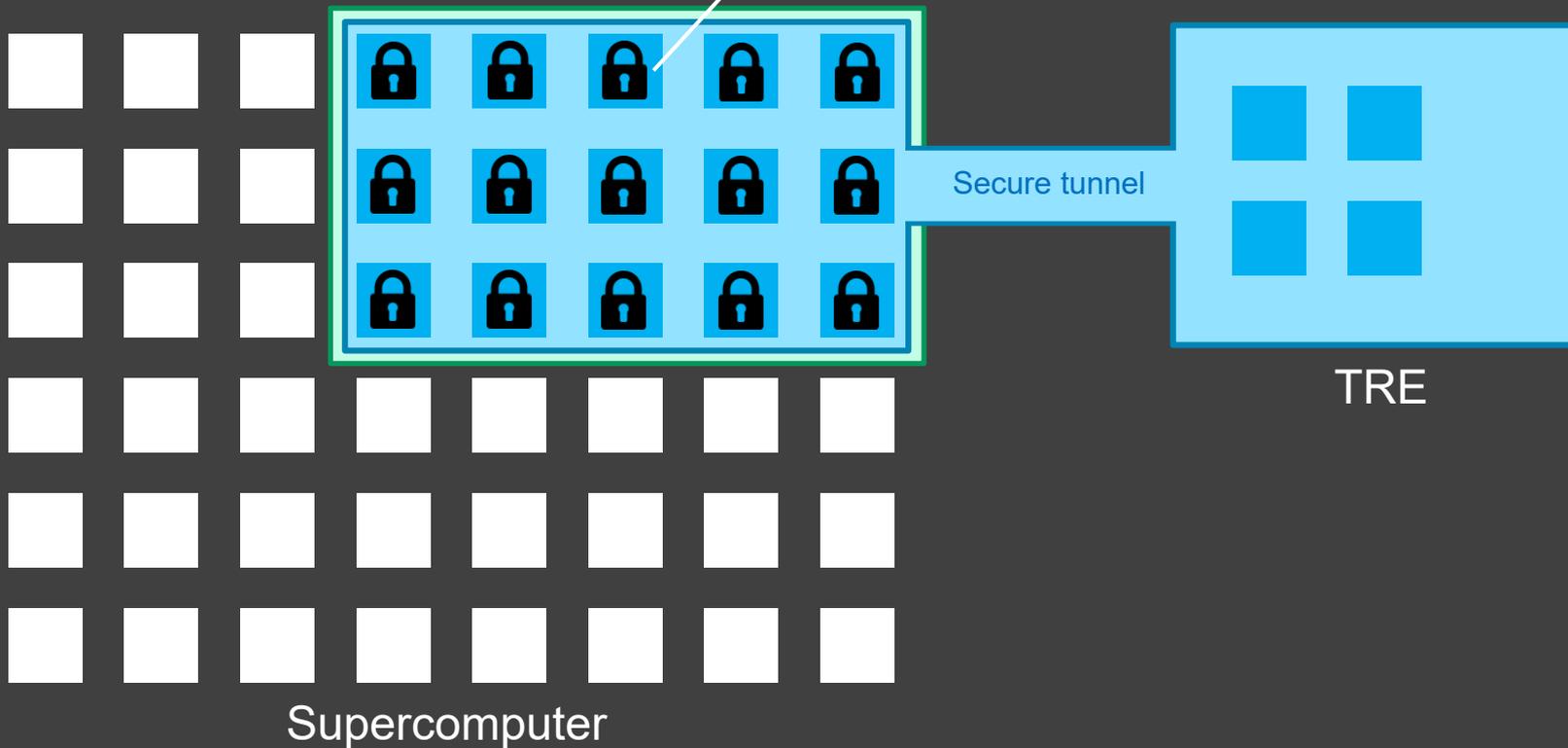
---

# Confidential Computing

- Technical capabilities built into modern computer processors
- Uses a secure mode that makes the processor inaccessible to the system administrators
- Provides a way for the user to **verify processor security** and **encrypt data and code**

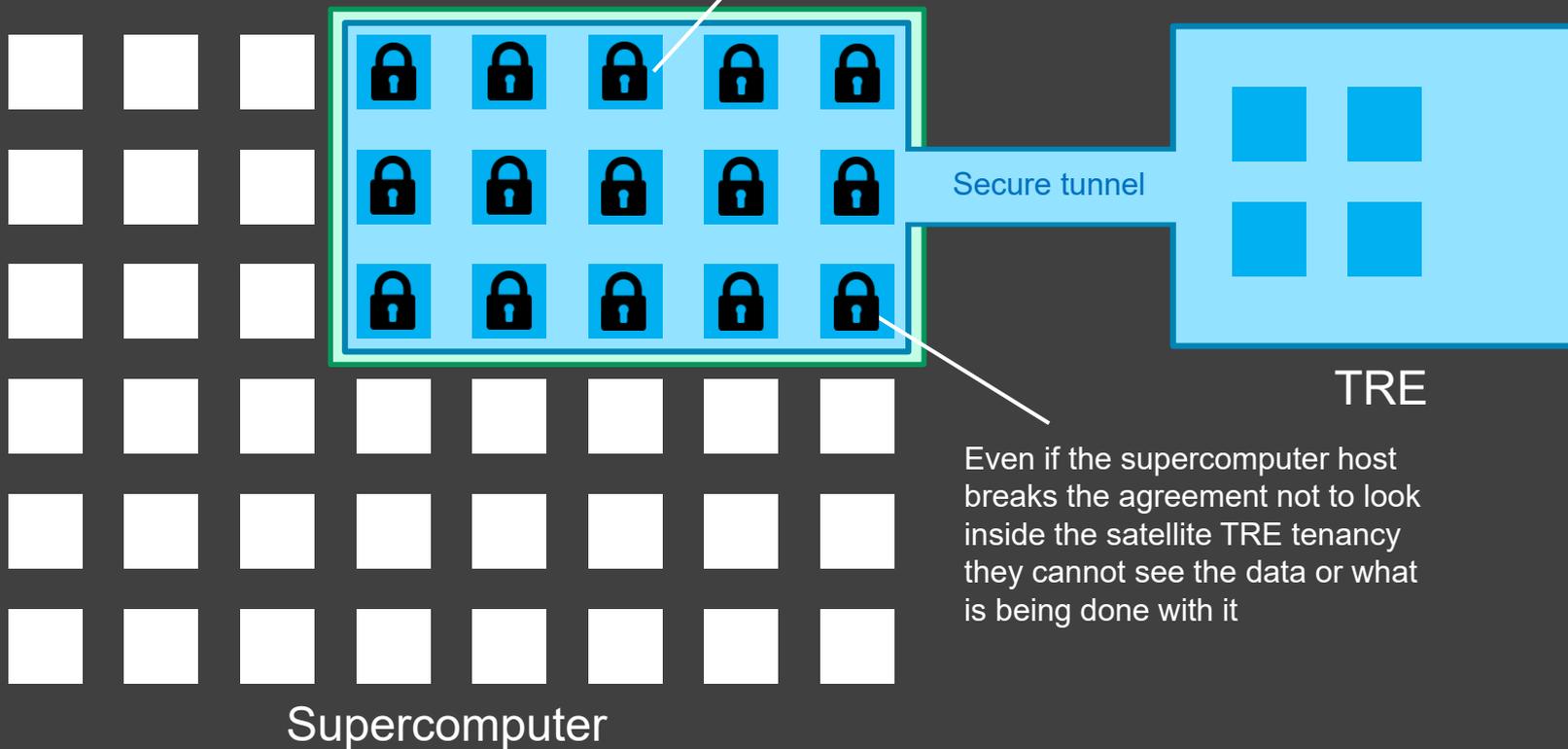
# Beyond FRIDGE (TRUSTEE)

Step 4: Lock down individual computer processors within the satellite TRE tenancy with confidential computing

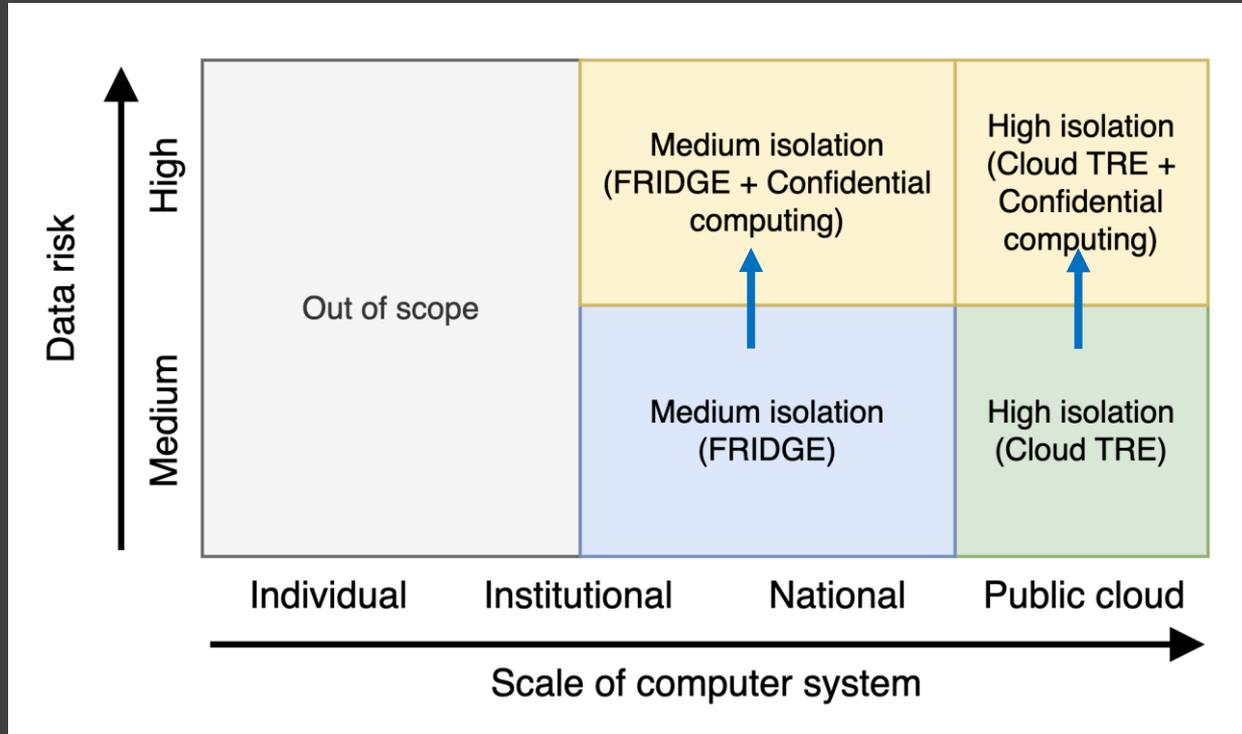


# Beyond FRIDGE (TRUSTEE)

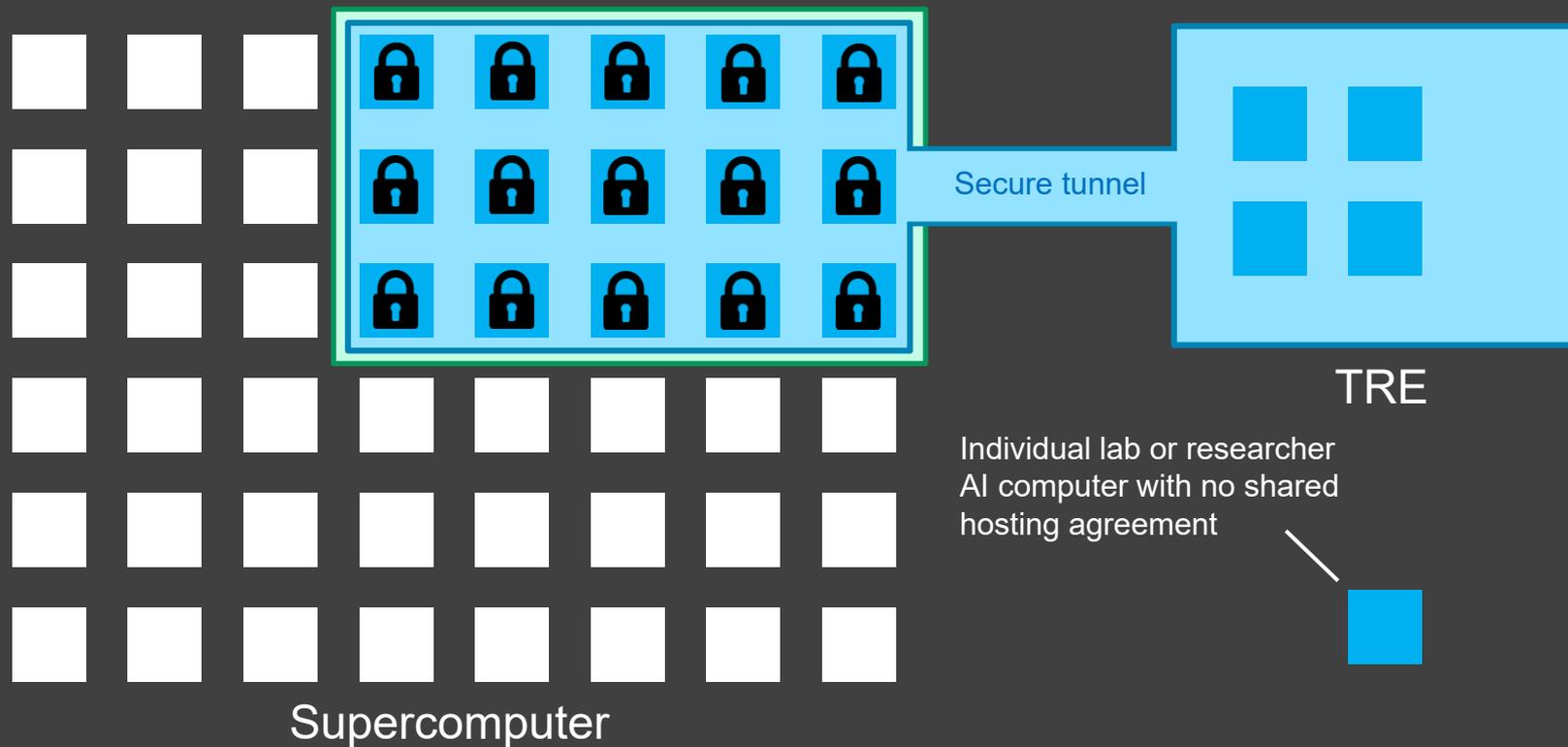
Step 4: Lock down individual computer processors within the satellite TRE tenancy with confidential computing



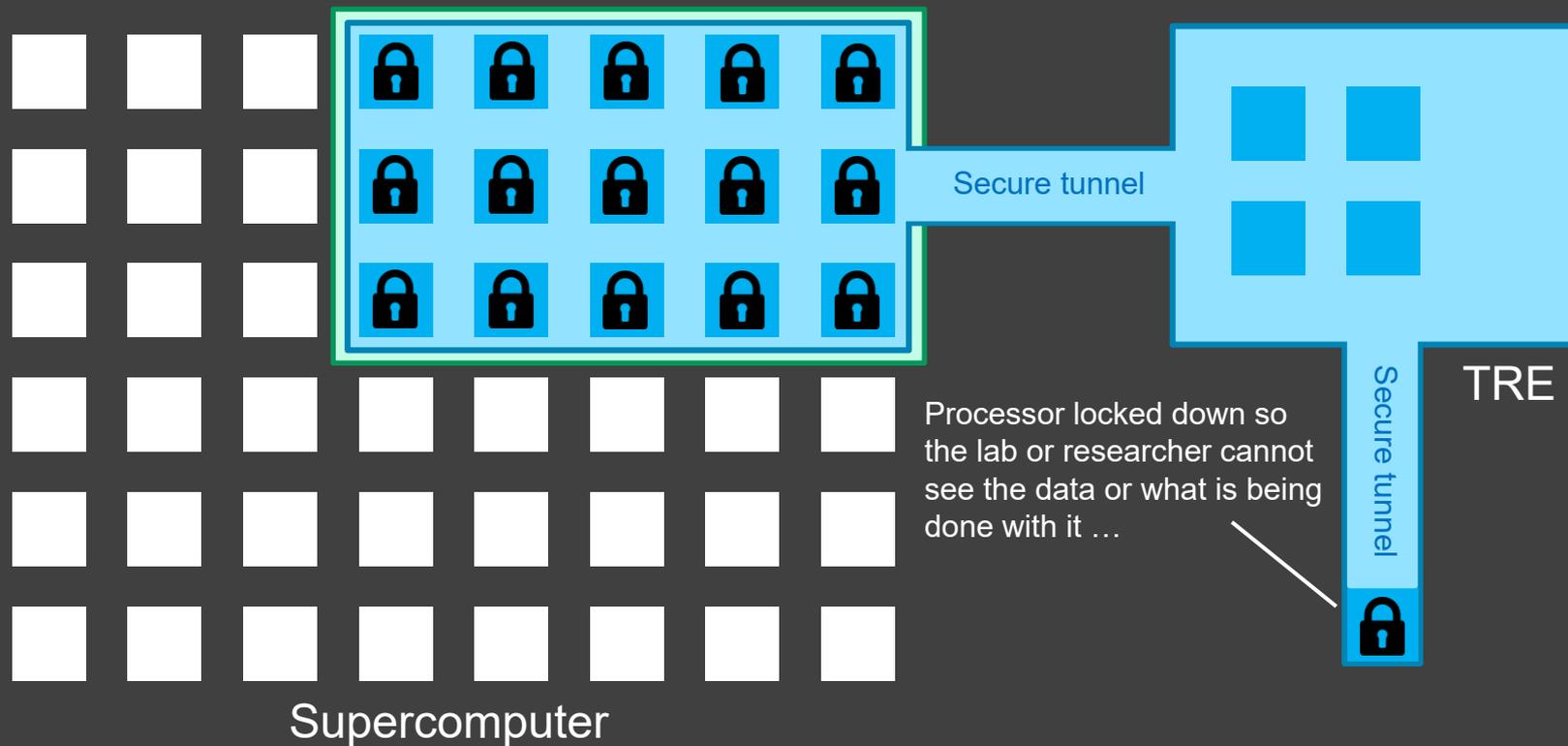
# Beyond FRIDGE (TRUSTEE)



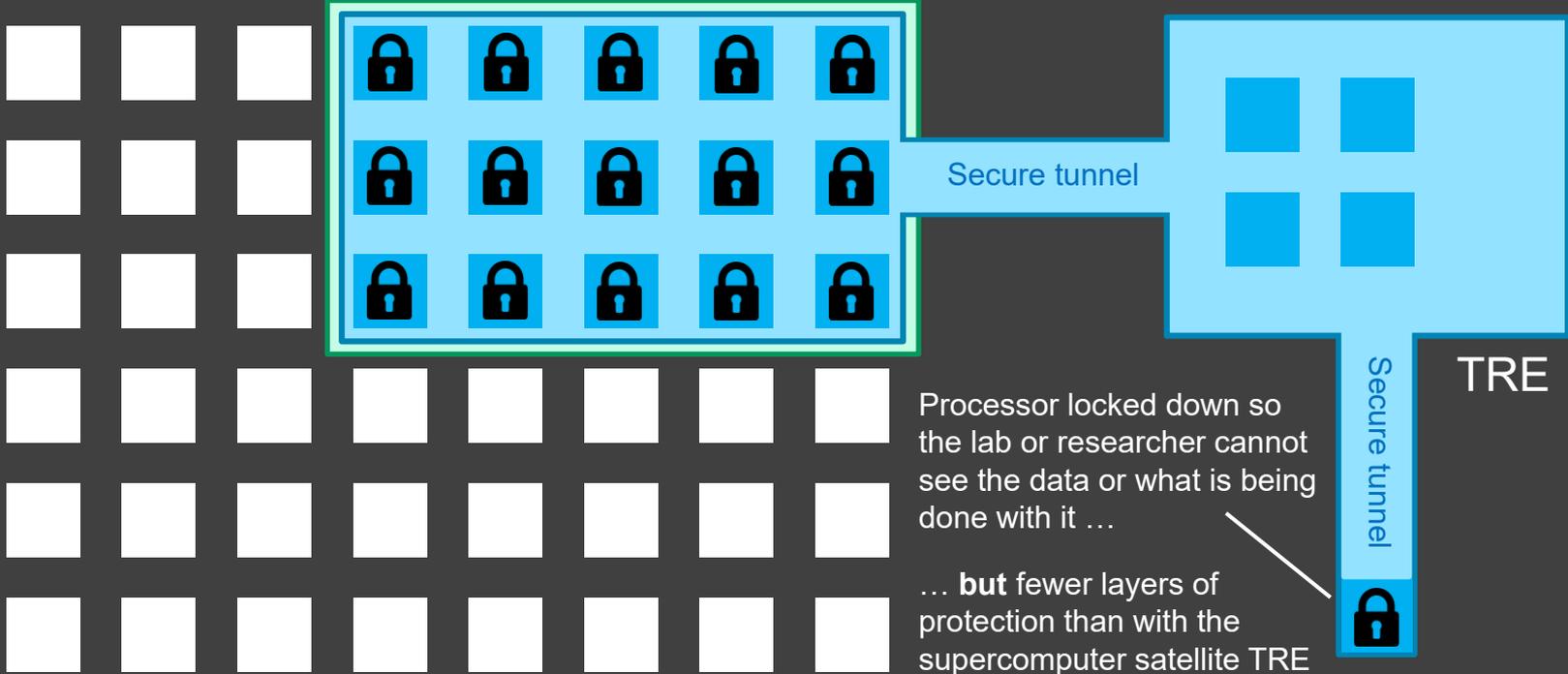
# Beyond FRIDGE (TRUSTEE)



# Beyond FRIDGE (TRUSTEE)



# Beyond FRIDGE (TRUSTEE)



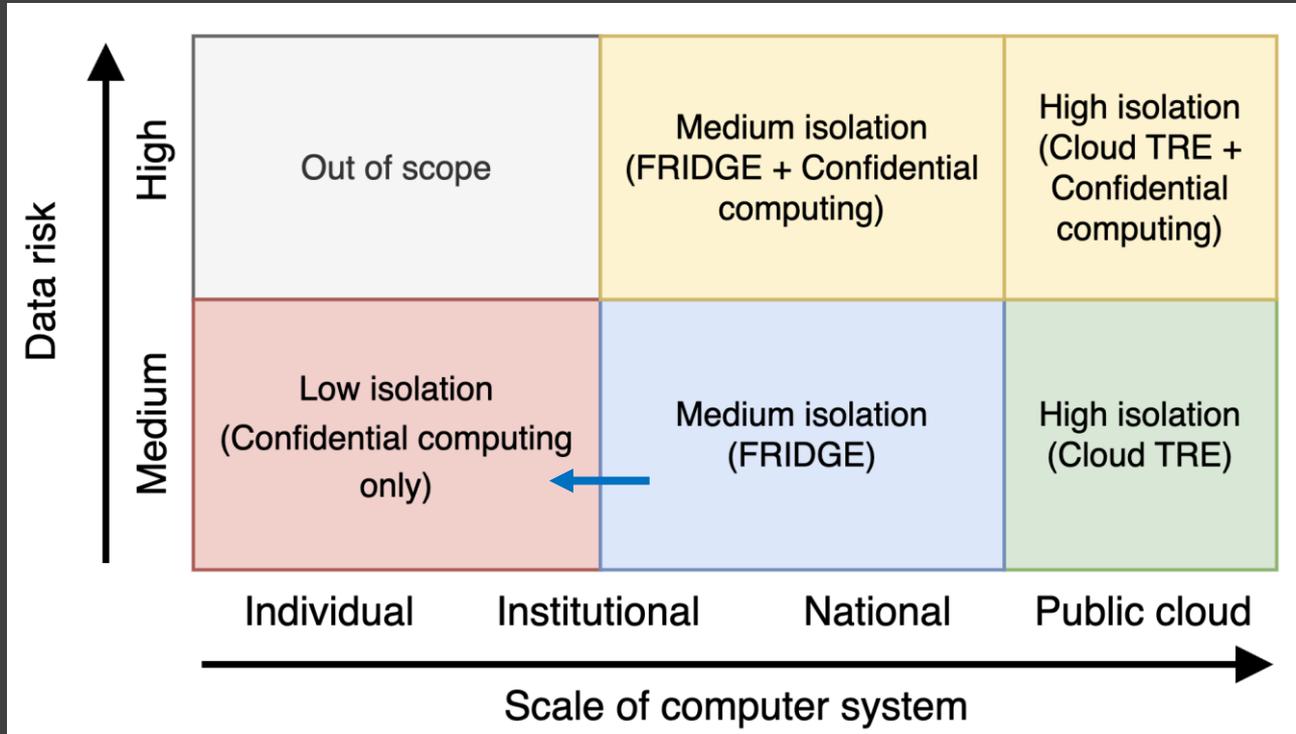
Supercomputer

Processor locked down so the lab or researcher cannot see the data or what is being done with it ...

... but fewer layers of protection than with the supercomputer satellite TRE

TRE

# Beyond FRIDGE (TRUSTEE)



---

# Our approach

---

# PIE programme

- Dedicated PIE lead as part of core project leadership team

---

# PIE programme

- Dedicated PIE lead as part of core project leadership team
- Public advisory group embedded in project as public voice with regular co-working workshops

---

# PIE programme

- Dedicated PIE lead as part of core project leadership team
- Public advisory group embedded in project as public voice with regular co-working workshops
- Integration with DARE UK TREvolution PIE workstream

---

# PIE programme

- Dedicated PIE lead as part of core project leadership team
- Public advisory group embedded in project as public voice with regular co-working workshops
- Integration with DARE UK TREvolution PIE workstream
- Sessions at public TRE Community Collaboration Cafes

---

# Community co-design

- Targeted co-design workshops with DARE UK TREvolution programme

---

# Community co-design

- Targeted co-design workshops with DARE UK TREVolution programme
- Wider stakeholder engagement workshops through virtual and in-person events

---

# Community co-design

- Targeted co-design workshops with DARE UK TREVolution programme
- Wider stakeholder engagement workshops through virtual and in-person events
- Attending workshops and seminar series held by other stakeholder communities

---

# Community co-design

- Targeted co-design workshops with DARE UK TREVolution programme
- Wider stakeholder engagement workshops through virtual and in-person events
- Attending workshops and seminar series held by other stakeholder communities
- Embedding public voice in project from the start with our embedded public advisory group

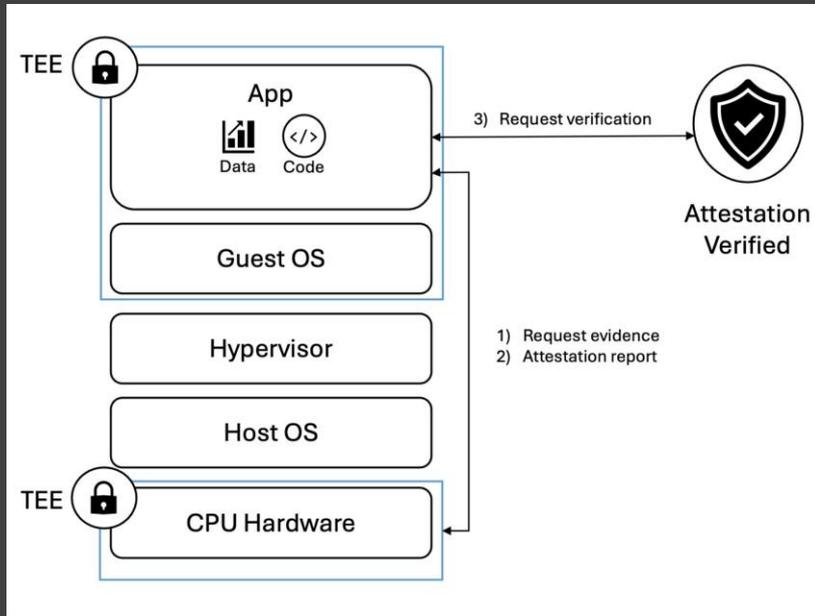


---

Potential Q&A slides

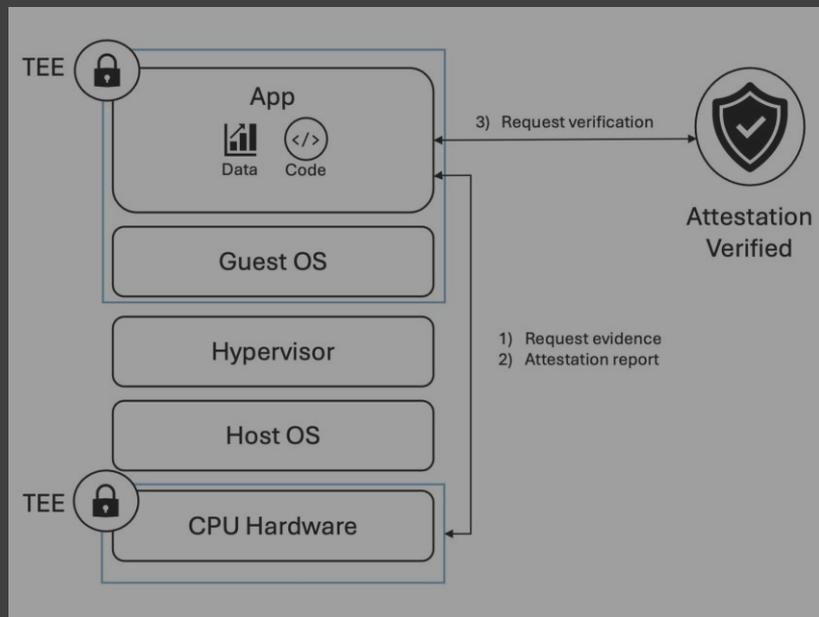
# Trusted Execution Environments (TEEs)

## Establishing trust

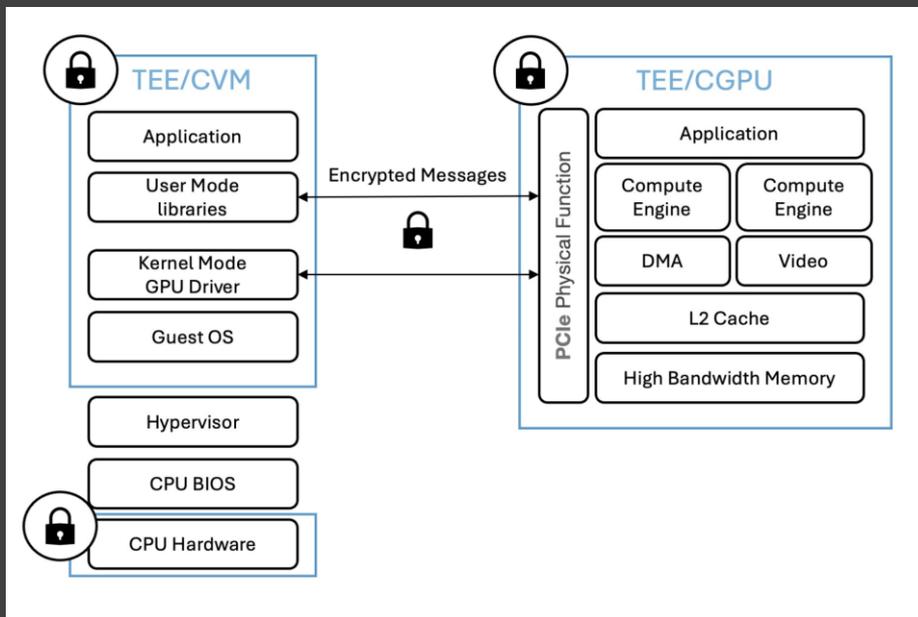


# Trusted Execution Environments (TEEs)

## Establishing trust



## Extending trust





# Synthetic data & TREs

*Enabling and Connecting TREs*  
*Transparency for the **public**, efficiency for the **researcher***

*Steve Harris*  
*NHS Consultant and Researcher*





**Sensitive data  
TRES (Level 3+)**

Area	Tier 0/1	Tier 2	Tier 3	Tier 4
<i>Package mirrors</i>	External internet package servers	Delayed package mirrors	Whitelisted package mirrors	Whitelisted package mirrors
<i>Inbound network</i>	Open internet	Institutional network only	Restricted network only	Restricted network only
<i>Outbound network</i>	Internet is accessible	Virtual network isolated	Virtual network isolated	Virtual network isolated
<i>User devices</i>	Open	Open	Managed	Managed
<i>Physical security</i>	None	None	Medium (access limited to research offices)	High (access limited to a secure room)
<i>User addition signoff</i>	Project Manager	Project Manager	Project Manager + DPR	Project Manager + DPR
<i>Remote connections</i>	Can access through ssh	Only remote desktops	Only remote desktops	Only remote desktops
<i>Software ingress signoff</i>	None	Investigator	Investigator + Referee	Investigator + Referee
<i>Airlock on software install?</i>	No	Yes	Yes	Yes
<i>High-security data ingress?</i>	No	Yes	Yes	Yes
<i>Copy-paste</i>	Permitted	Forbidden by policy but not configuration	Disabled on the remote desktop	Disabled on the remote desktop
<i>Classification Referee?</i>	No	Yes	Yes	Yes

Safe-guarded data  
TRES (Level 1/2)



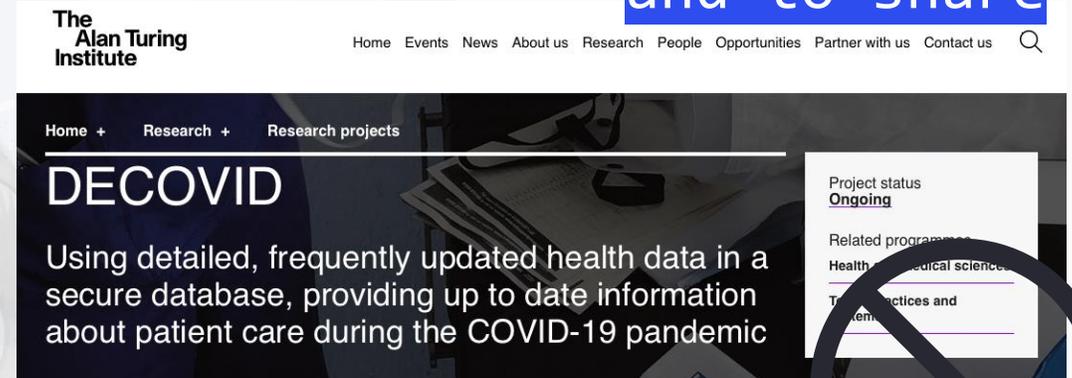


Synthetic data  
Public [TREs Level 0)]

Without synthetic data ...



Slow and difficult to start  
and to share



Without synthetic data ...

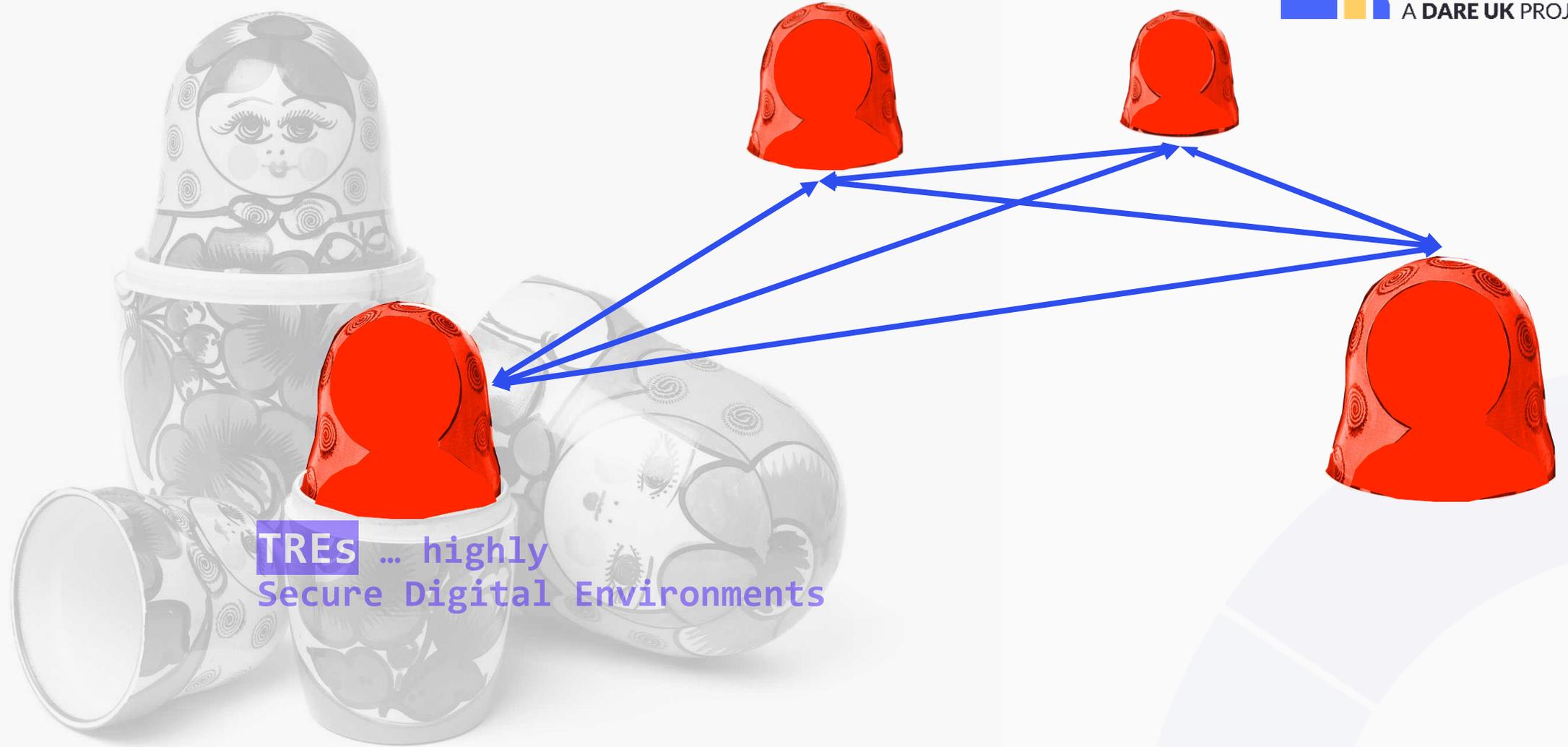


## Difficult to verify The “Reproducibility crisis”

Smith R. **Time to assume that health research is fraudulent until proven otherwise?** *The BMJ* (2021).  
Grey, A., Bolland, M. J., Avenell, A., Klein, A. A. & Gunsalus, C. K. **Check for publication integrity before misconduct.** *Nature* **577**, 167–169 (2020).



1. Outside  $\leftrightarrow$  Inside
2. TRE  $\leftrightarrow$  TRE



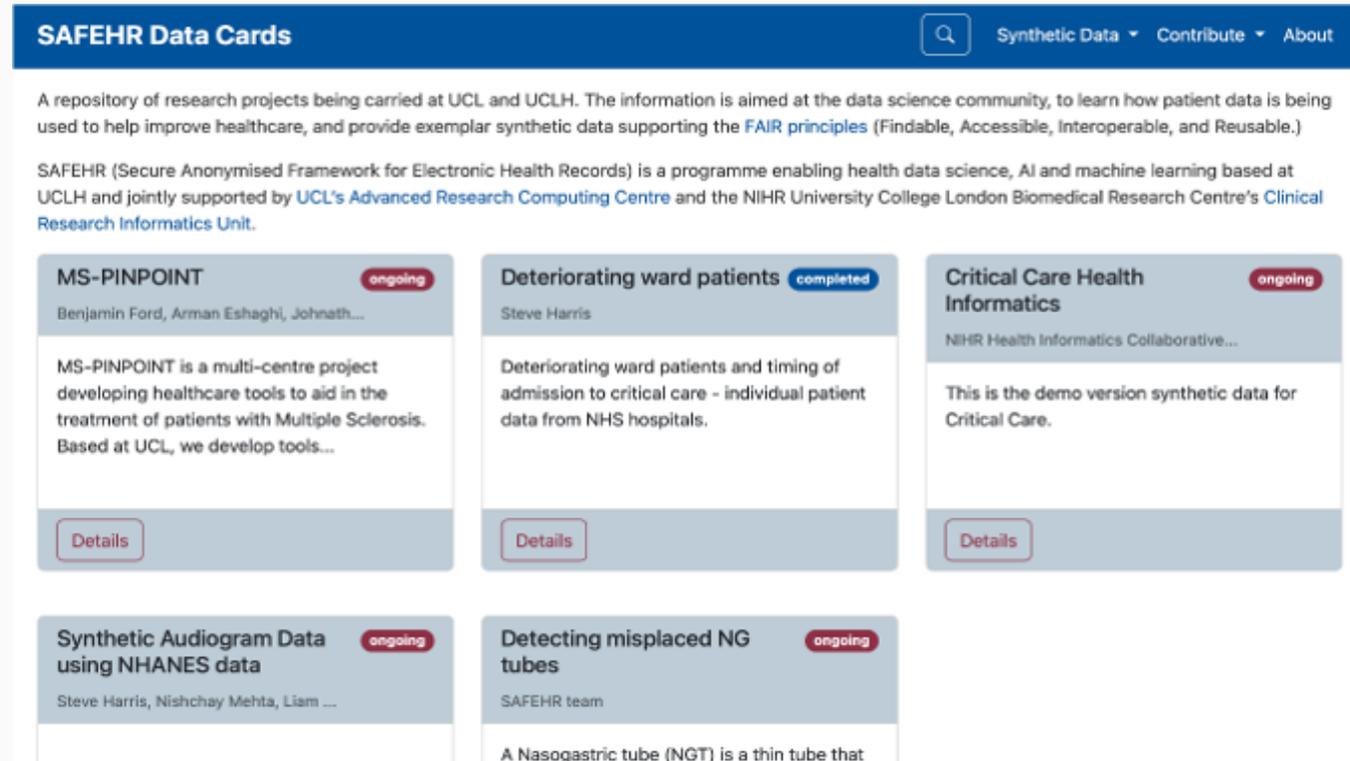
1. Outside  $\leftrightarrow$  Inside
2. TRE  $\leftrightarrow$  TRE

# Project aim

*Open synthetic data for  
collaboration and transparency*

# At University College Hospital London

## For every patient data study, a data card ...



**SAFEHR Data Cards** Search Synthetic Data Contribute About

A repository of research projects being carried at UCL and UCLH. The information is aimed at the data science community, to learn how patient data is being used to help improve healthcare, and provide exemplar synthetic data supporting the [FAIR principles](#) (Findable, Accessible, Interoperable, and Reusable.)

SAFEHR (Secure Anonymised Framework for Electronic Health Records) is a programme enabling health data science, AI and machine learning based at UCLH and jointly supported by [UCL's Advanced Research Computing Centre](#) and the NIHR University College London Biomedical Research Centre's [Clinical Research Informatics Unit](#).

### MS-PINPOINT

ongoing

Benjamin Ford, Arman Eshaghi, Johnath...

MS-PINPOINT is a multi-centre project developing healthcare tools to aid in the treatment of patients with Multiple Sclerosis. Based at UCL, we develop tools...

[Details](#)

### Deteriorating ward patients

completed

Steve Harris

Deteriorating ward patients and timing of admission to critical care - individual patient data from NHS hospitals.

[Details](#)

### Critical Care Health Informatics

ongoing

NIHR Health Informatics Collaborative...

This is the demo version synthetic data for Critical Care.

[Details](#)

### Synthetic Audiogram Data using NHANES data

ongoing

Steve Harris, Nishchay Mehta, Liam ...

[Details](#)

### Detecting misplaced NG tubes

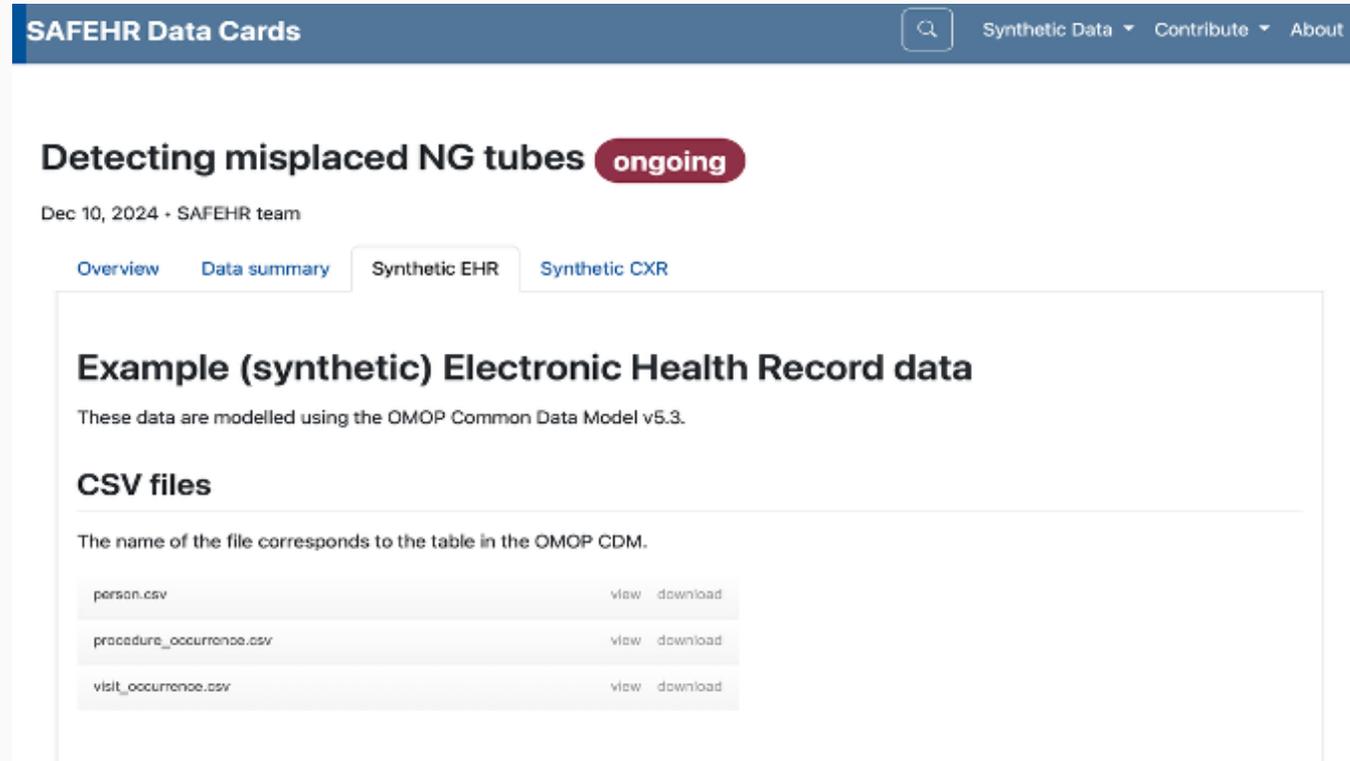
ongoing

SAFEHR team

A Nasogastric tube (NGT) is a thin tube that

[Details](#)

For every patient data study, a data card ...



**SAFEHR Data Cards** Search Synthetic Data Contribute About

## Detecting misplaced NG tubes ongoing

Dec 10, 2024 - SAFEHR team

[Overview](#) [Data summary](#) [Synthetic EHR](#) [Synthetic CXR](#)

### Example (synthetic) Electronic Health Record data

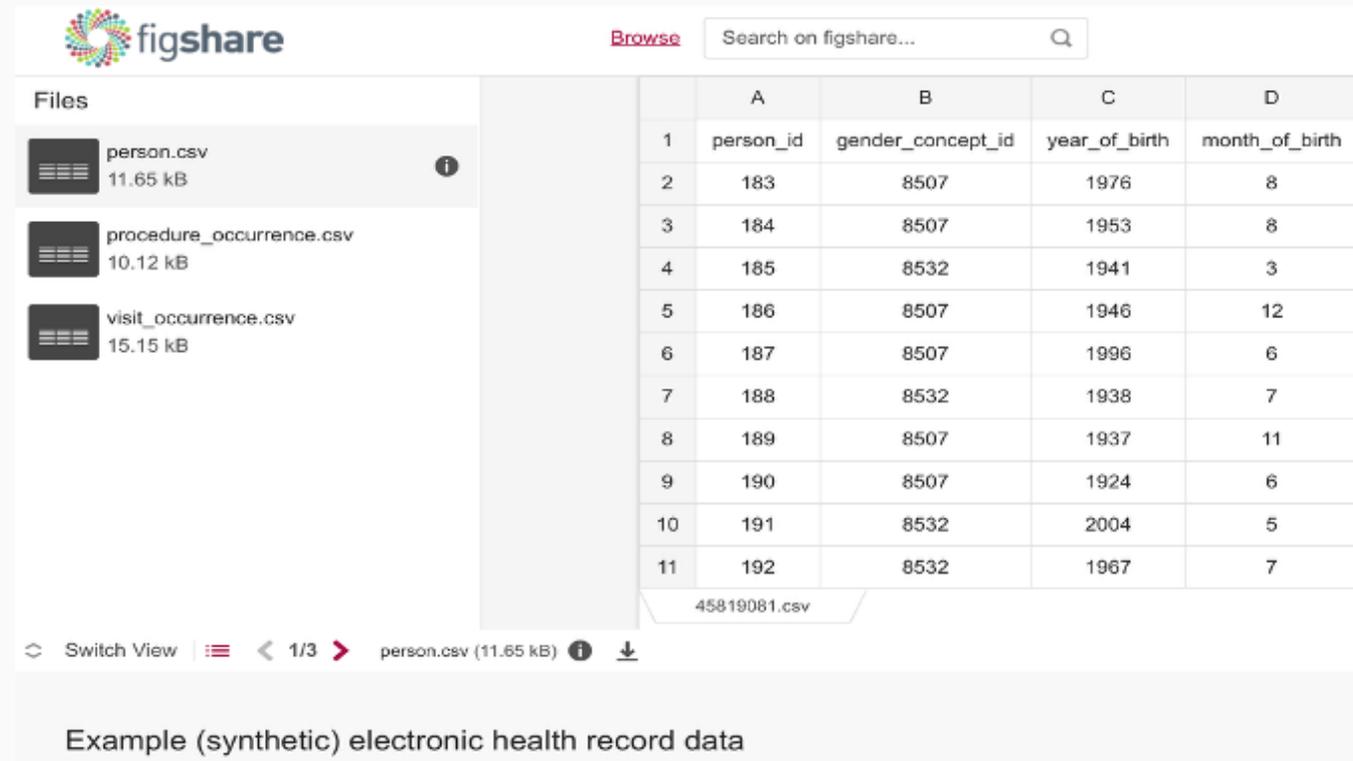
These data are modelled using the OMOP Common Data Model v5.3.

#### CSV files

The name of the file corresponds to the table in the OMOP CDM.

person.csv	<a href="#">view</a>	<a href="#">download</a>
procedure_occurrence.csv	<a href="#">view</a>	<a href="#">download</a>
visit_occurrence.csv	<a href="#">view</a>	<a href="#">download</a>

For every patient data study, a data card ...



The screenshot shows a Figshare interface with a file list on the left and a data preview table on the right. The file list includes 'person.csv' (11.65 kB), 'procedure\_occurrence.csv' (10.12 kB), and 'visit\_occurrence.csv' (15.15 kB). The preview table displays columns A, B, C, and D with corresponding data values for rows 1 through 11.

	A	B	C	D
1	person_id	gender_concept_id	year_of_birth	month_of_birth
2	183	8507	1976	8
3	184	8507	1953	8
4	185	8532	1941	3
5	186	8507	1946	12
6	187	8507	1996	6
7	188	8532	1938	7
8	189	8507	1937	11
9	190	8507	1924	6
10	191	8532	2004	5
11	192	8532	1967	7

Example (synthetic) electronic health record data

# Our tools

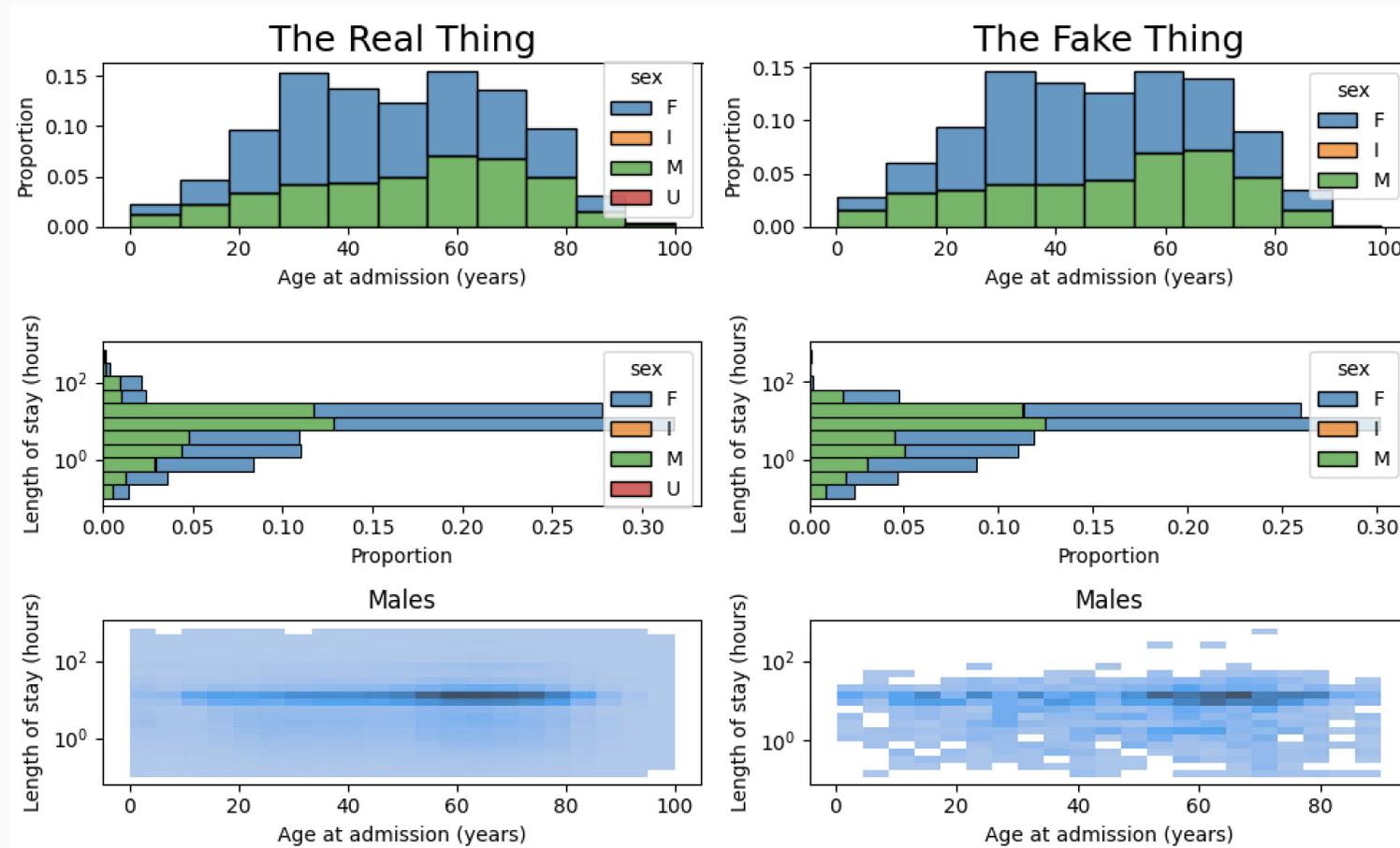
*White box* versus *black box*

# White box



Synthetic data  
Public [TREs Level 0)]

# White box: Data Faker\*



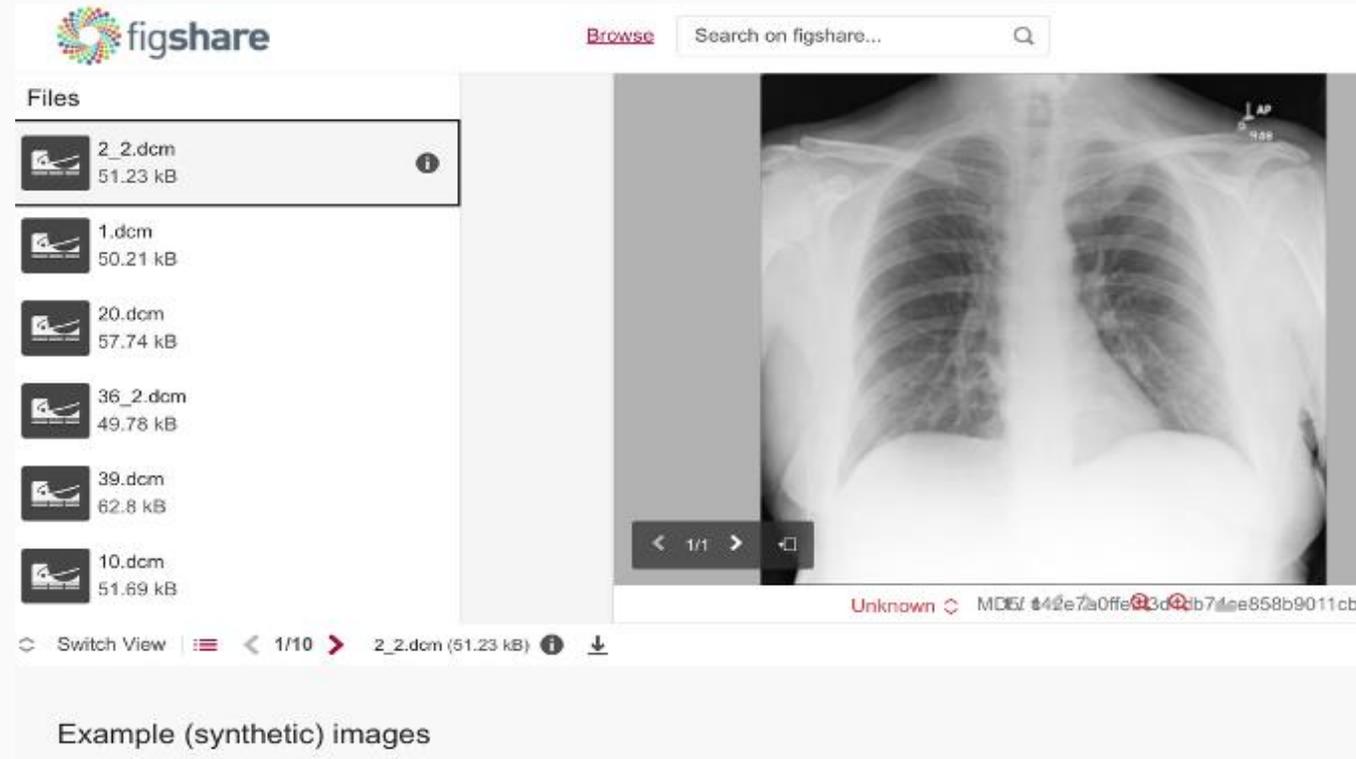
**The Alan Turing Institute**

# Black box



Safe-guarded data  
TRES (Level 1/2)

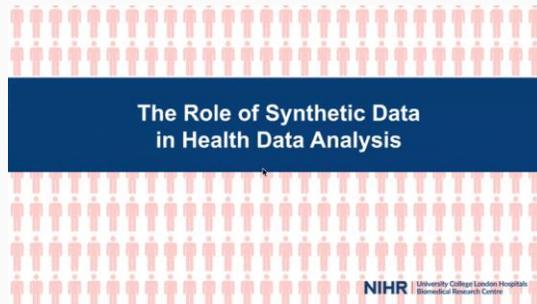
# Generative adversarial networks and more ...





# Our people

1. NHS Information governance teams
2. Patients and public



Anmol Arora



Eva Driskell



Martin O'Reilly



May Yong



## The Alan Turing Institute

Sarah Keating



Sulyman Abdulkareem



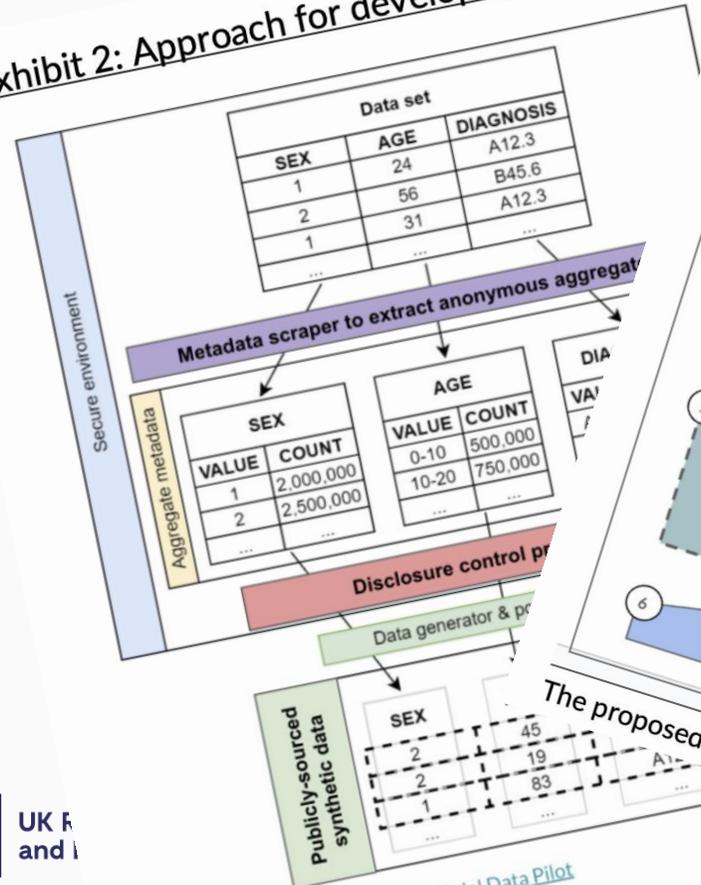
Steve Harris



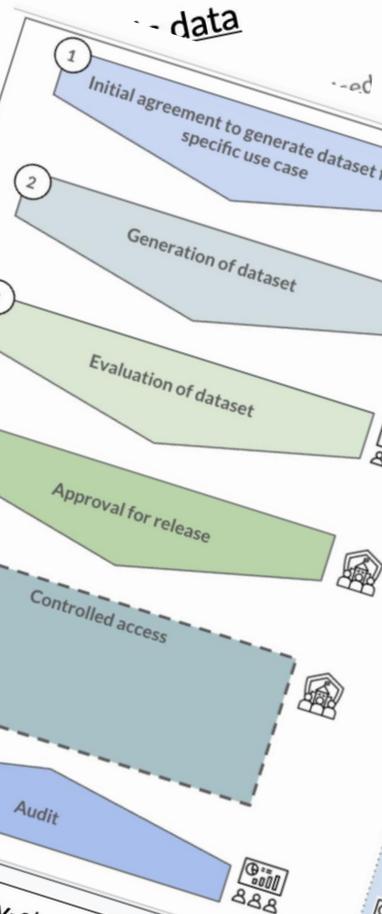
# Via policy ... for NHS trusts

## “Share information, not data”

Exhibit 2: Approach for developing publicly-sourced synthetic data



The proposed step-by-step procedure is as follows:

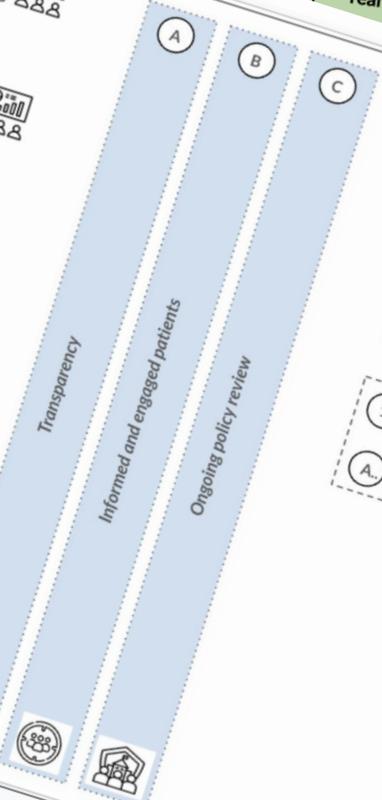
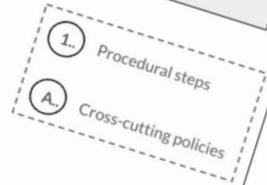
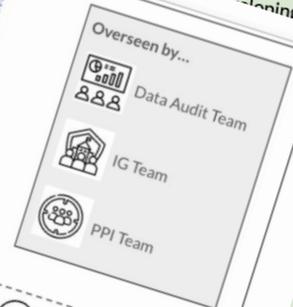


### 1.2 Summary of key proposals

Exhibit 1 below summarises this paper's key proposals (detailed further in Sections 6 and 7).

Exhibit 1: Summary of key proposals

Synthetic	Description	Access to underlying real data	Privacy risk	Example uses	Key proposed policies & procedures
					<ul style="list-style-type: none"> <li>Can always be publicly shared, provided no commercial/ IP concerns must be labelled as structural synthetic data</li> </ul>
					<ul style="list-style-type: none"> <li>Can always be publicly released, provided that the mechanism of its generation has been approved by IG and a standard disclosure control method has been applied to the statistical information used (and provided no commercial/IP concerns)</li> <li>The scope of internal use cases/applications must be defined and documented (for example, it cannot be used for direct clinical care purposes)</li> <li>Must be clearly labelled as publicly-sourced synthetic data and accompanied by documentation explaining its generation mechanism</li> <li>A record is kept of the uses of the synthetic data</li> </ul>
					<ul style="list-style-type: none"> <li>Can be released for specific use cases</li> </ul>



# Via training ... for users and approvers

## NHS data and information governance teams



