Simon Thompson

Professor of Health Informatics

Director Secure e-Research Platform
Co-Director SAIL Databank
Associate-Director Dementia Platform UK
Co-Investigator MS Register
Member HDRUK & ADR Technical Leadership & ELIXIR UK Node member
Co-I DARE UK TREvolution
Co-I HDRUK FED-A

POPULATION DATA SCIENCE

AT SWANSEA UNIVERSITY MEDICAL SCHOOL

PIONEERING POPULATION DATA SCIENCE FOR PUBLIC GOOD

Swansea University Prifysgol Abertawe
Population Data Science
Faculty of Medicine, Health & Life Science
Gwyddor Data Poblogaeth
Y Gyfadran Meddygaeth, Gwyddor Iechyd a Bywyd

Contribution from

Anwar Gaungoo,
Centre for Health Informatics, Nottingham University

HDRUK FED-A

Trusted Research Environment

Protect **OUR** data/identity to ensure **YOU** get the benefits of world-leading research.

Beyond Data.

Implementations of SeRP

SeRP Canada supporting health services in the Department of Health for British Columbia (BC). Enabling projects across BC, through easier dataset querying, wider access to the Canadian Ministry of Health data (Health Canada) thereby streamlining the data access process. Sharing datasets across BC Health Boards in a safe, secure and governed environment within British Columbia.

Over 1,900 data users located in more than 150 organisations globally

SeRP UK is operated as a private research cloud and as a multi-tenancy model which means you control how and who uses it all under one pricing structure. It lets you build economies of scale and offers a cost-effective data management and governance environment for users whilst also enabling them to be part of the SeRP UK research community.

SeRP Australia supporting Australian Government Department of Health and Monash University for utilising data from hundreds of clinical trials, medical and population health studies and patient registries to offer approved researchers ground-breaking access to complex healthcare data to better understand a host of disease burdens faced by Australia including cancer, neurodegenerative disorders, heart disease and mental illness.

SeRP

SeRP Tenants & Users (selected)

HDRUK Health Data Research UK
ICODA International COVID-19 Data Alliance
COVID Symptom Tracker
Llywodraeth Cymru Welsh Government
GIG CYMRU NHS WALES Iechyd a Gofal Digidol Cymru Digital Health and Care Wales
NDR Adnodd Data Cenedlaethol National Data Resource
Partneriaeth Genomeg Cymru Genomics Partnership Wales
BREATHE Health Data Research Hub
ADR UK Data-driven change
GIG CYMRU NHS WALES Iechyd Cyhoeddus Cymru Public Health Wales
NHS Digital
Genomics england
biobank uk
Office for National Statistics
SAIL DATABANK
MS REGISTER
Department of Health & Social Care
Department for Business, Energy & Industrial Strategy
CHILDREN OF THE 90S
Dementias Platform UK Medical Research Council
DATA PORTAL
Ministry of Justice
Department for Education
Adolescent Mental Health DATA PLATFORM
Doeth am Iechyd Cymru HealthWise Wales
Department for Transport
Home Office
CLIMB BIG DATA
UK Longitudinal Linkage Collaboration
East London Genes & Health

Australian Government Department of Health
Curtin University
MONASH University
BRITISH COLUMBIA Ministry of Health
Health Canada
Santé Canada
IC/ES Data Discovery Better Health
UBC THE UNIVERSITY OF BRITISH COLUMBIA

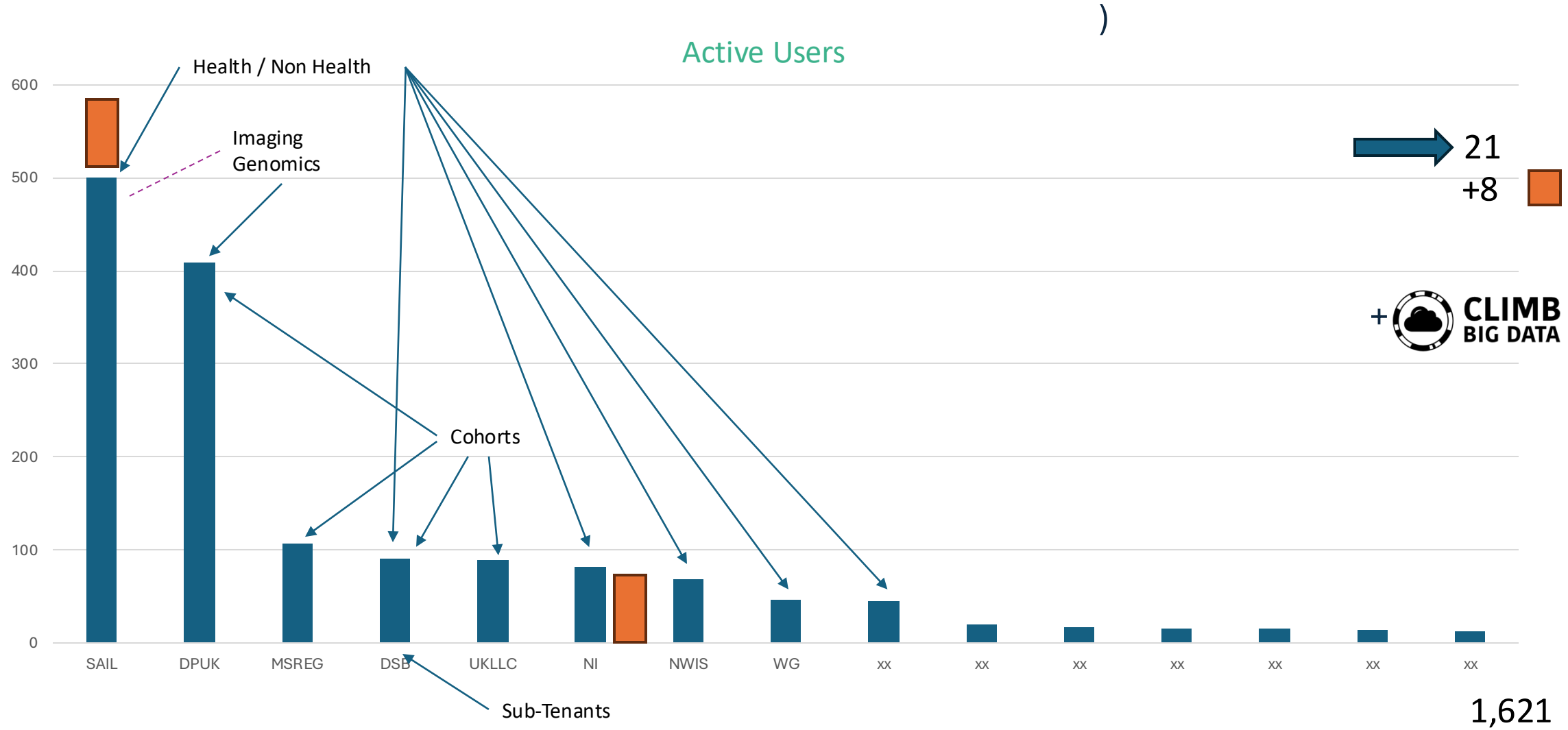
What is SeRP UK?

Provider of TRE services to research programmes

1 TRE = n users + x project + y datasets

SeRP

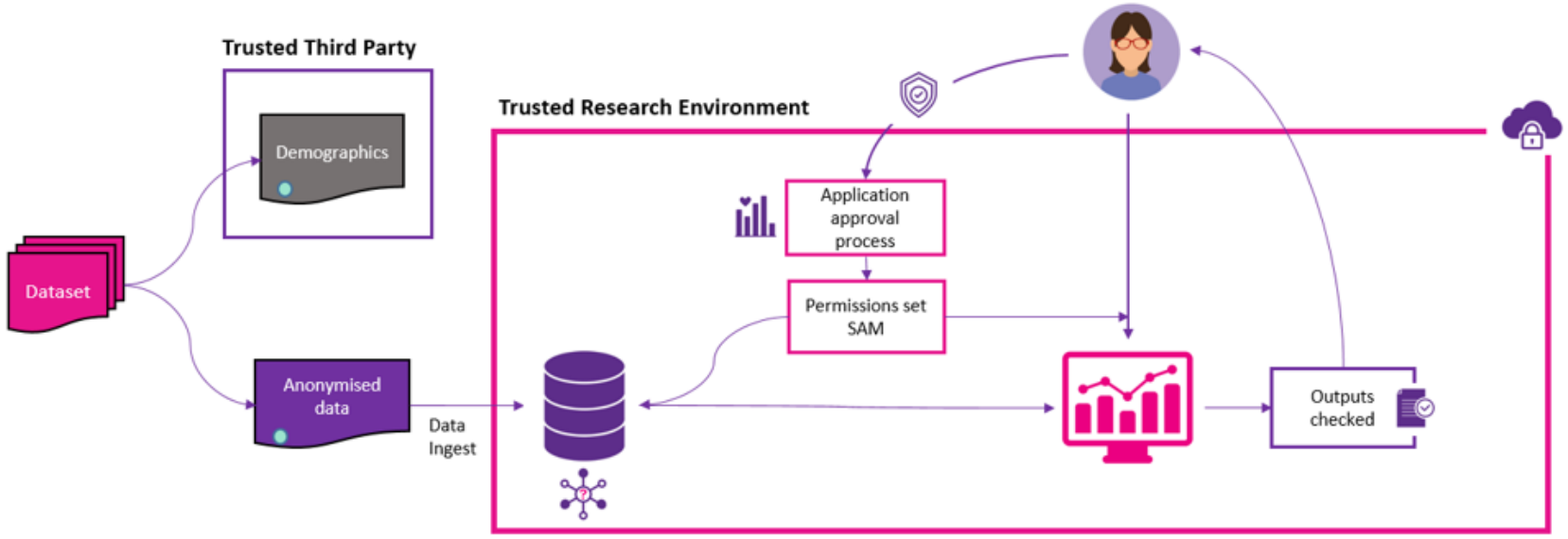
SeRP: Public Funded TRE Provider



Australia & Canada (VC) & NI & Ireland coming (this year)

Trusted Research Environment

High level Overview



Safe People
Only trained and specifically accredited researchers can access the data

Safe Projects
Data is only used for ethical, approved research with the potential for clear public benefit

Safe Settings
Access to data is only possible using secure technology systems – the data never leaves the TRE

Safe Data
Researchers only use data that have been de-identifed to protect privacy

Safe Outputs
All research outputs are checked to ensure they cannot be used to identify subjects

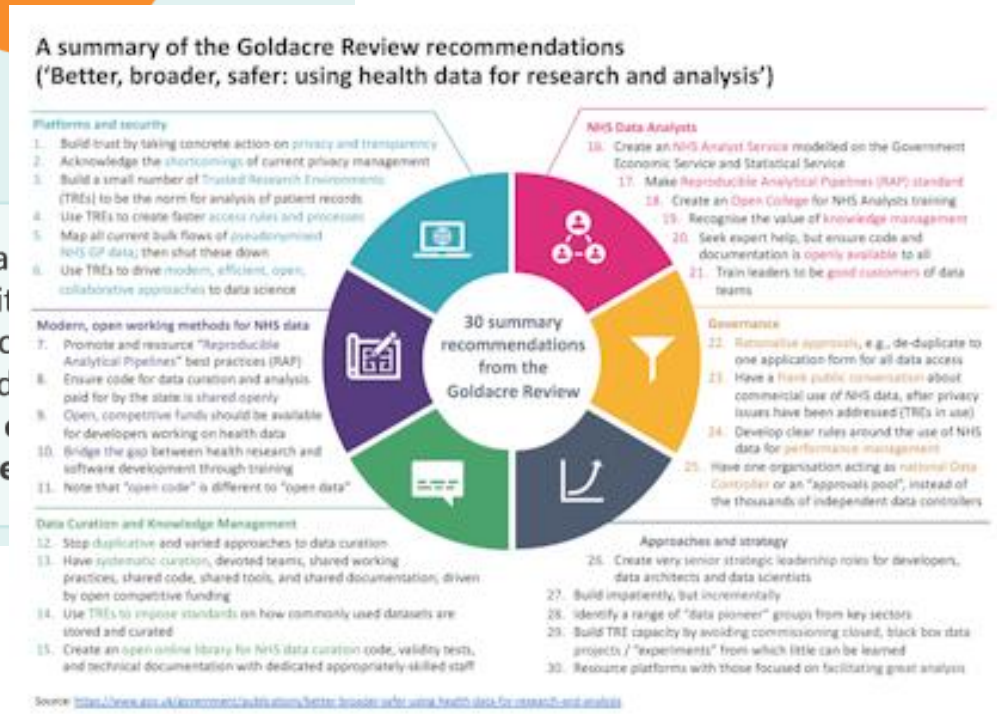
Transition to Trusted Research Environments

Why are they **important?**

TREs make research safer. Making data available through a TRE means that people can be **confident** that their personal health data is accessed **securely** and their **privacy protected.**

TREs help make **research efficient, collaborative** and **cost effective,** providing rich data that enables **deep insights** which will go on to improve healthcare and **save lives.**

TREs provide a
researchers wit
location to acc
datasets. The d
tools are all in
like a **secure re**



More recently: Sudlow Report


The Future going to be beautiful

Everybody building a TRE


SDE
VS
TRE

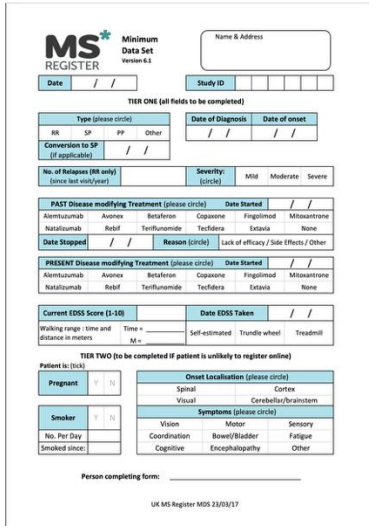
Starting to see Standard Definitions of TRE

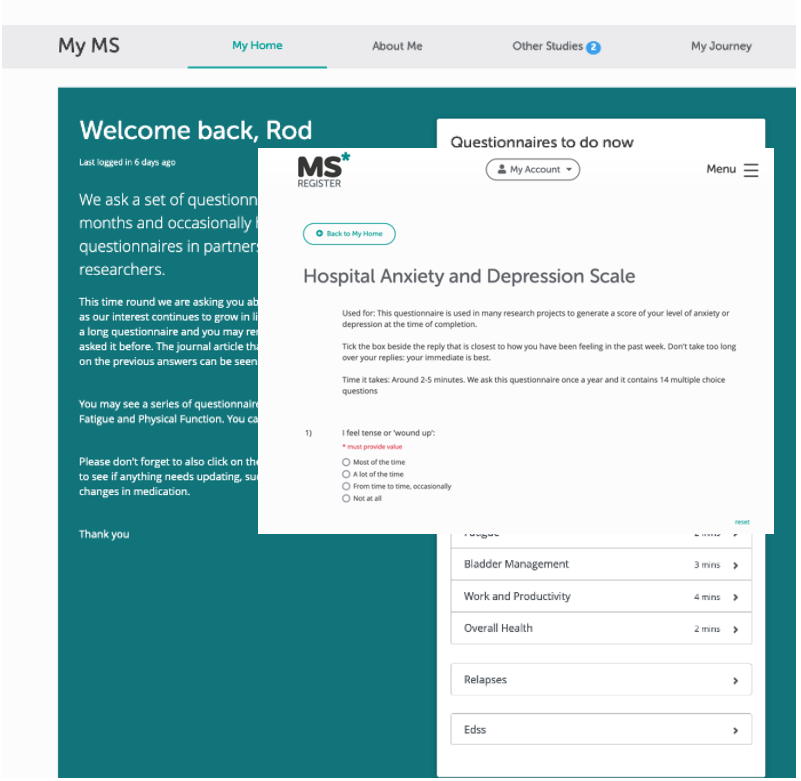
eosc | ENTRUST
European Network of Trusted Research Environments

EUROPEAN OPEN SCIENCE CLOUD

Data Collection and Registers

- MS Register – Clinical
- BHF Vanguard Studies
- Brainwaves
- *DPUK
- *TBI
- *DATAMIND
- Trials Delivery Framework
- MS Register
- Nurture (Kidney)
- Great Minds

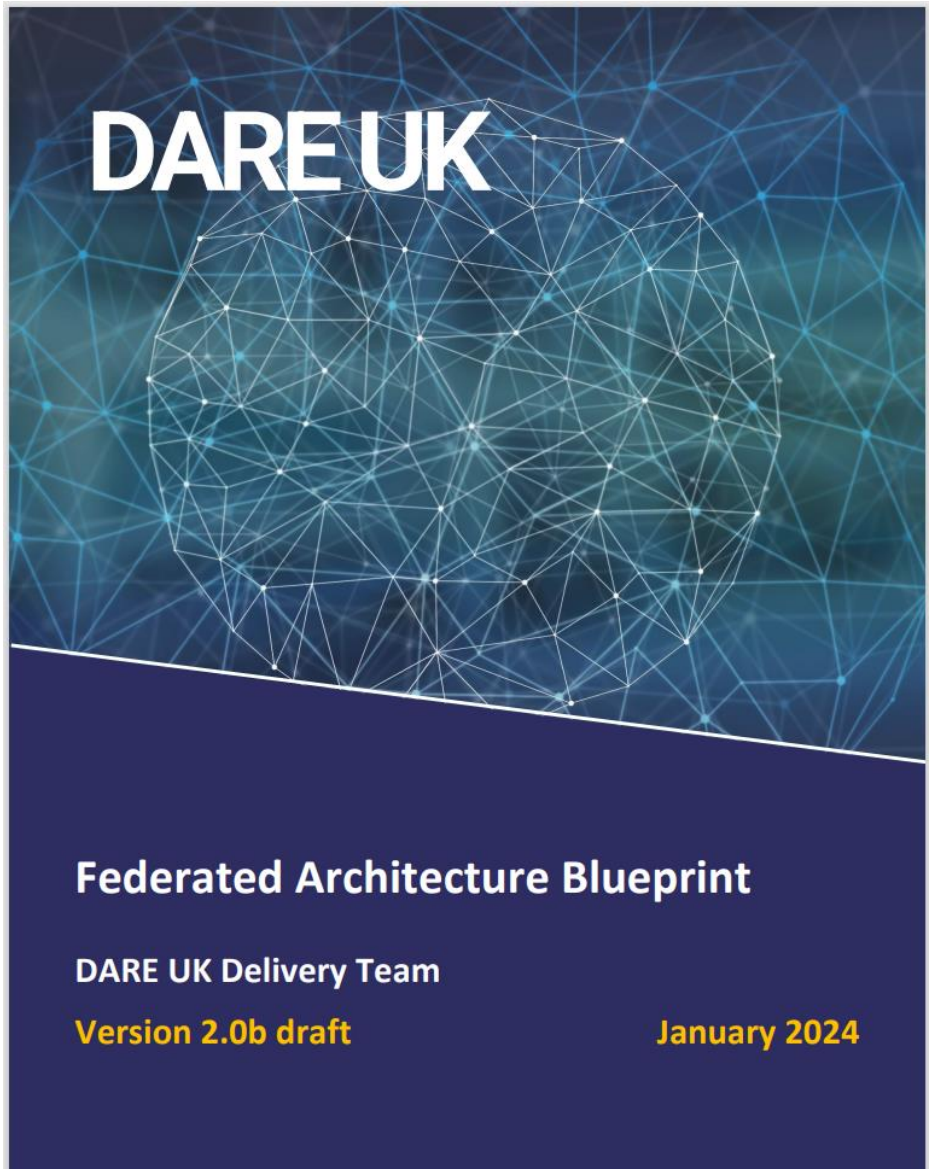




What do we mean by Federated Analysis

Formal Definition Coming

120 pages of indulgent fun



Lens

Useful to consider “sensitive data” is the focus

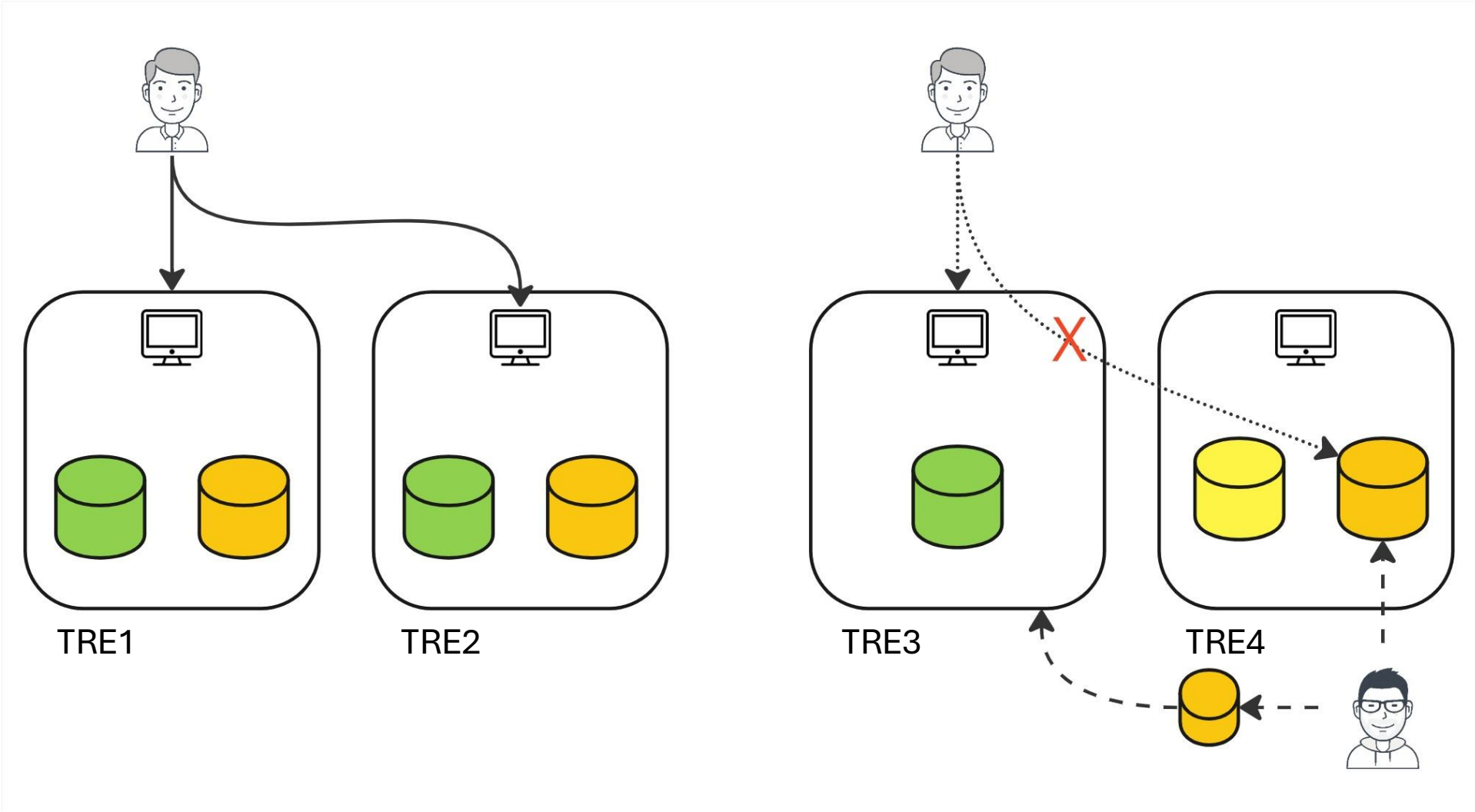
Open data can be included but does not need most of the same controls

Working Definition

Federation is the joining or making available data from more than one Trusted Research Environment
[for the purpose of doing science]

The mechanics of a federated infrastructure must still implement **ALL** the “5 Safe’s” but in a network sense, while fulfilling the requirements of the connected endpoints

Reason for federation

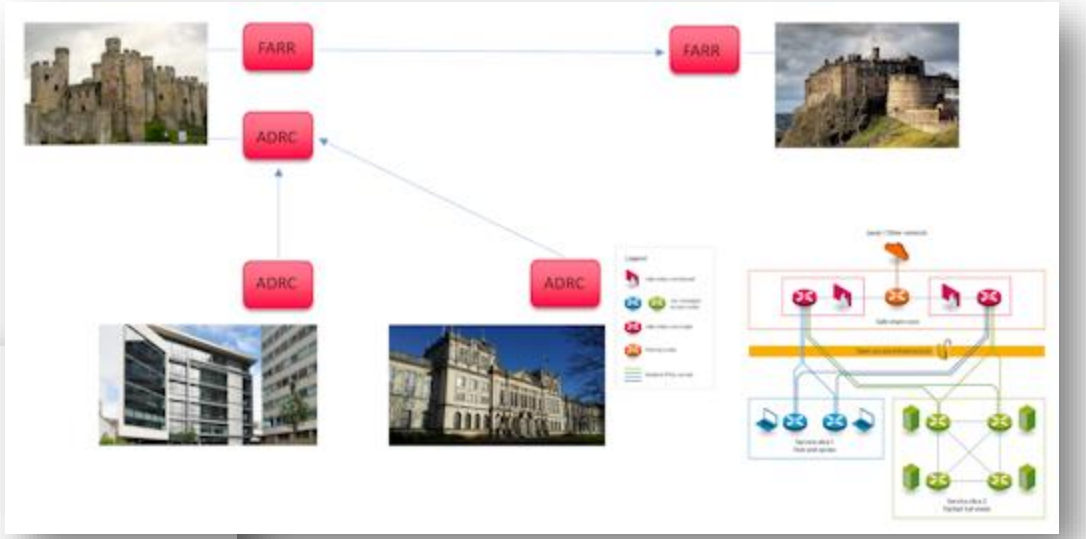


Been a potential problem for a while (2014)

Everything outside is vulnerable

- End users / access – have to let people in
- External systems / data suppliers

JISC “SafeShare” Project
before its time



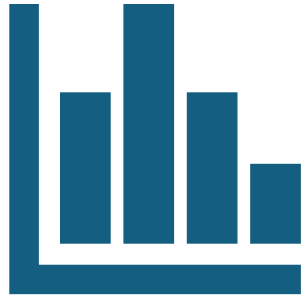
Creating Silos

- How do you work together ?
- How do you share data if you never data out ?

We need to trust each other and create tunnels for safe passage of data/access/people

Two Approaches

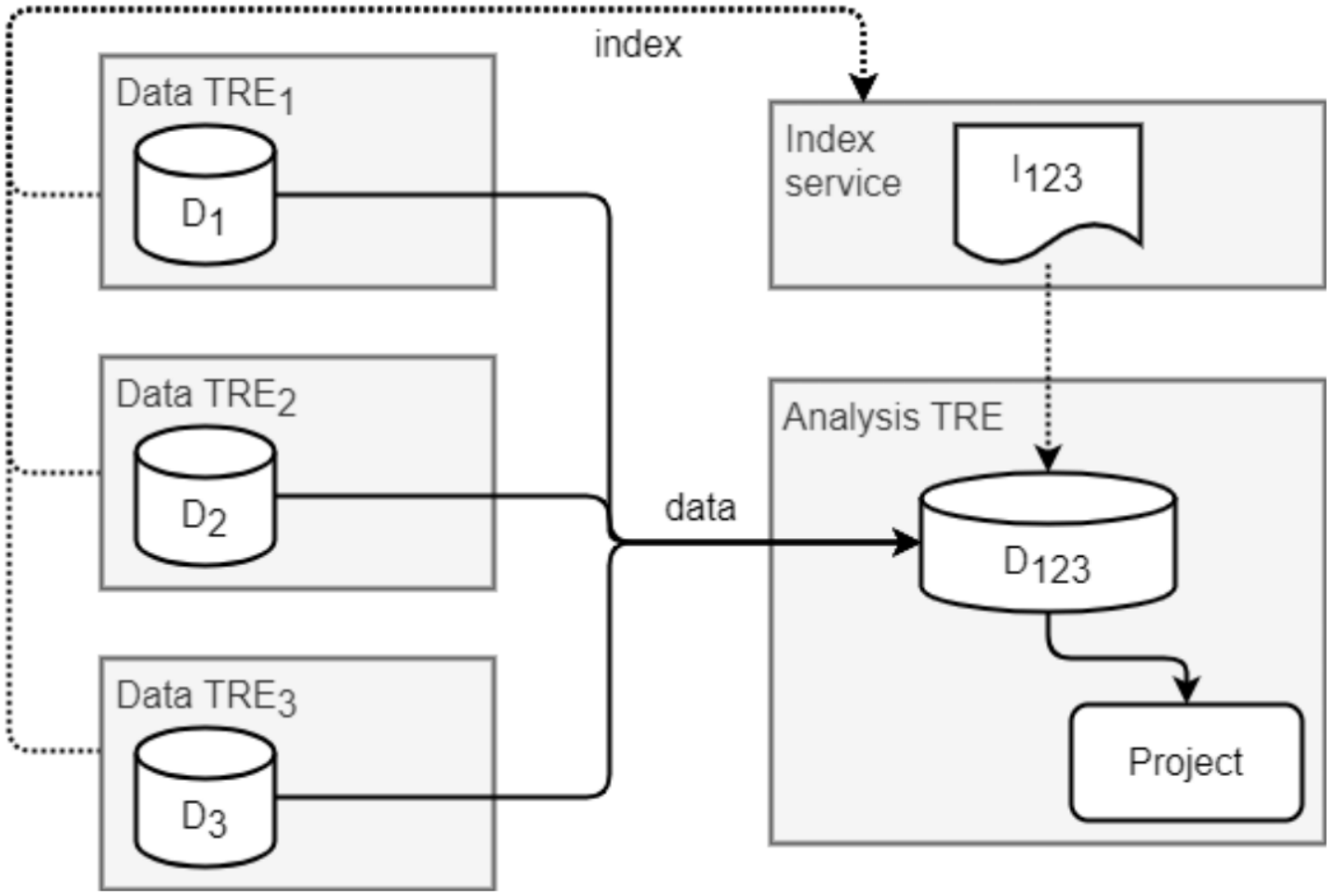
Data Pooling



Federated Analytics
+learning

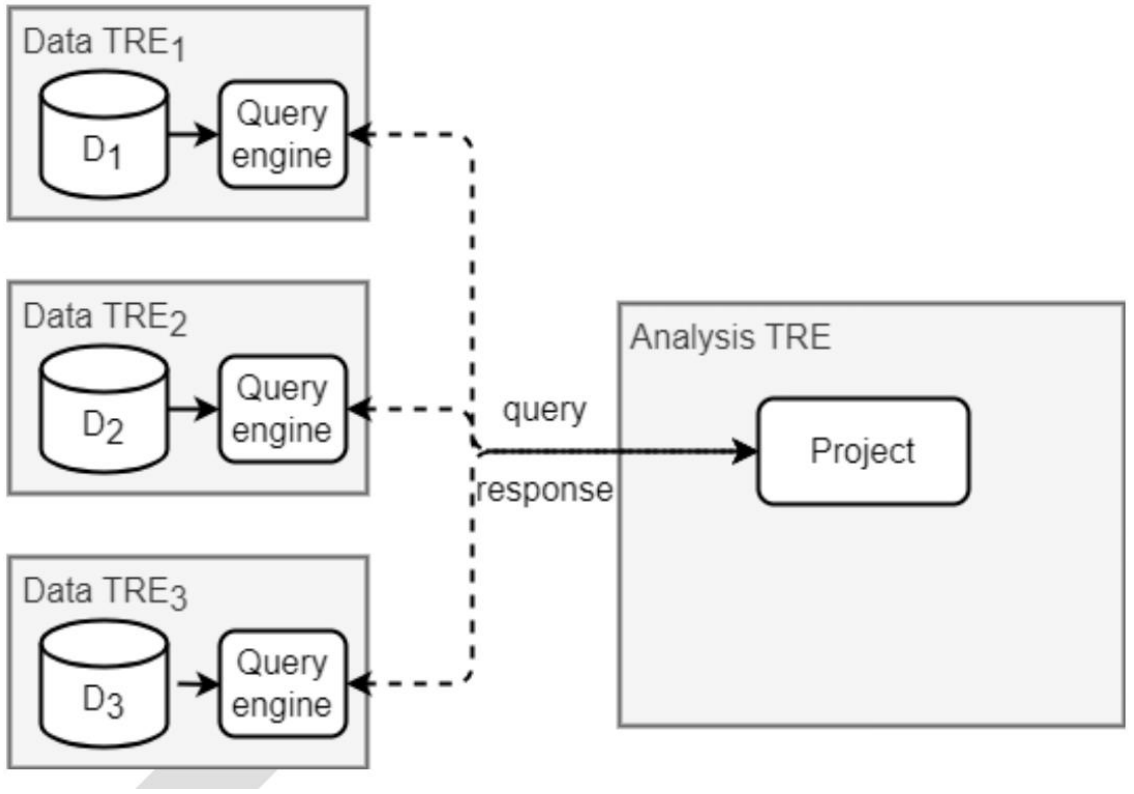
Data Pooling

Data pooling, where approved datasets or data extracts are moved between TREs, pooled in a single location and optionally linked, before being provided to a research team as a project. Analysis tools and resources are provided at the pooling location to support the project.



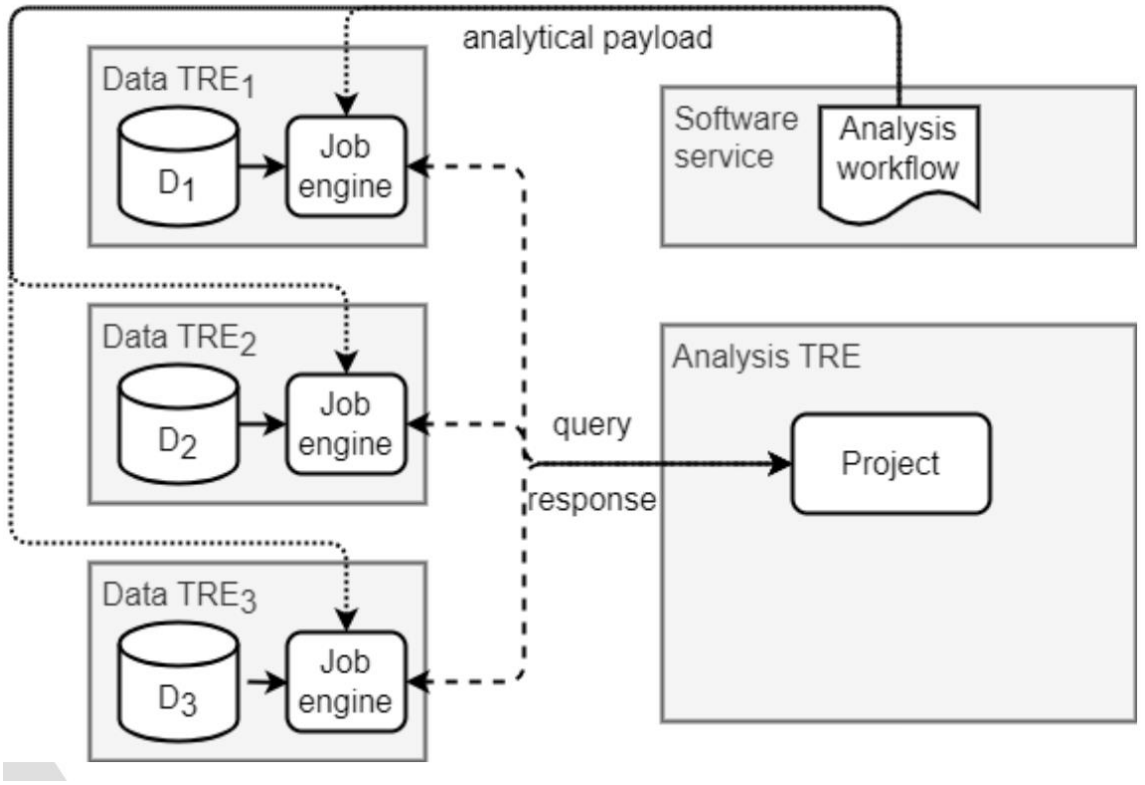
Federated Analysis: Direct Querry

The “query” here is fully encapsulated in the request from the analysis TRE; no additional information or external software is needed by the data TREs to execute the query. The actual query may be simple (e.g., an SQL COUNT) or it may be a complex object containing partial training results from a machine learning model

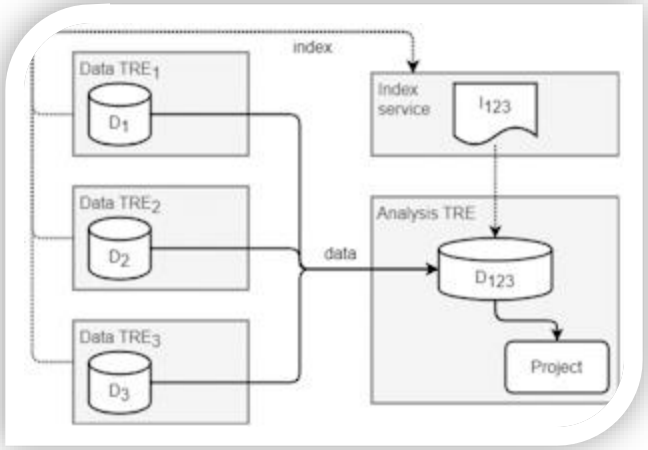


Federated Analysis: Indirect Query

Federated analytics using job submission: a job request is created by researchers on a project and sent to participating “data TREs”. Again, the datasets (D1, D2 and D3) remain within their provider organisations. To execute the job query, the TREs must download the actual “analytical payload” (a workflow, for example) from another source, run it, and return the response to the originating service.

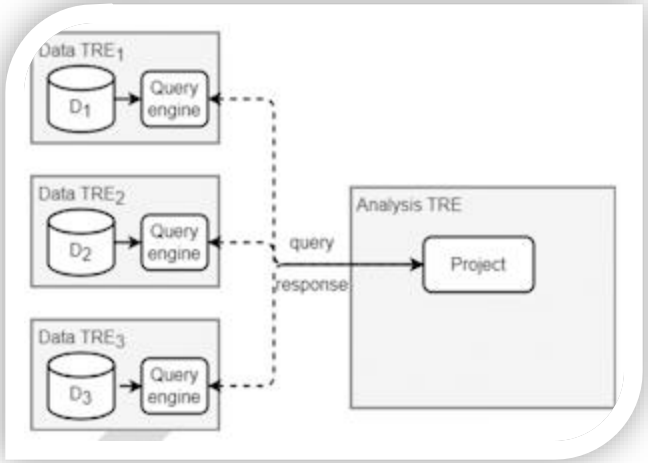


UK Proof of concept: Phase 2 in place



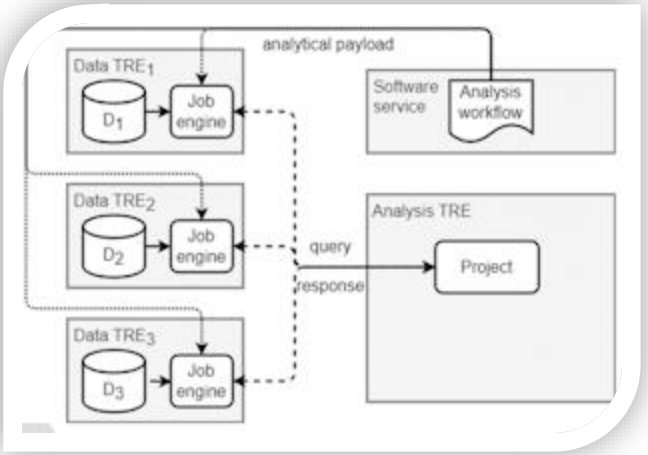
Data Pooling

DARE: TELEPORT (virtual)



Direct Query

DARE: **TELEPORT**



Indirect Query

DARE: **TREFX**

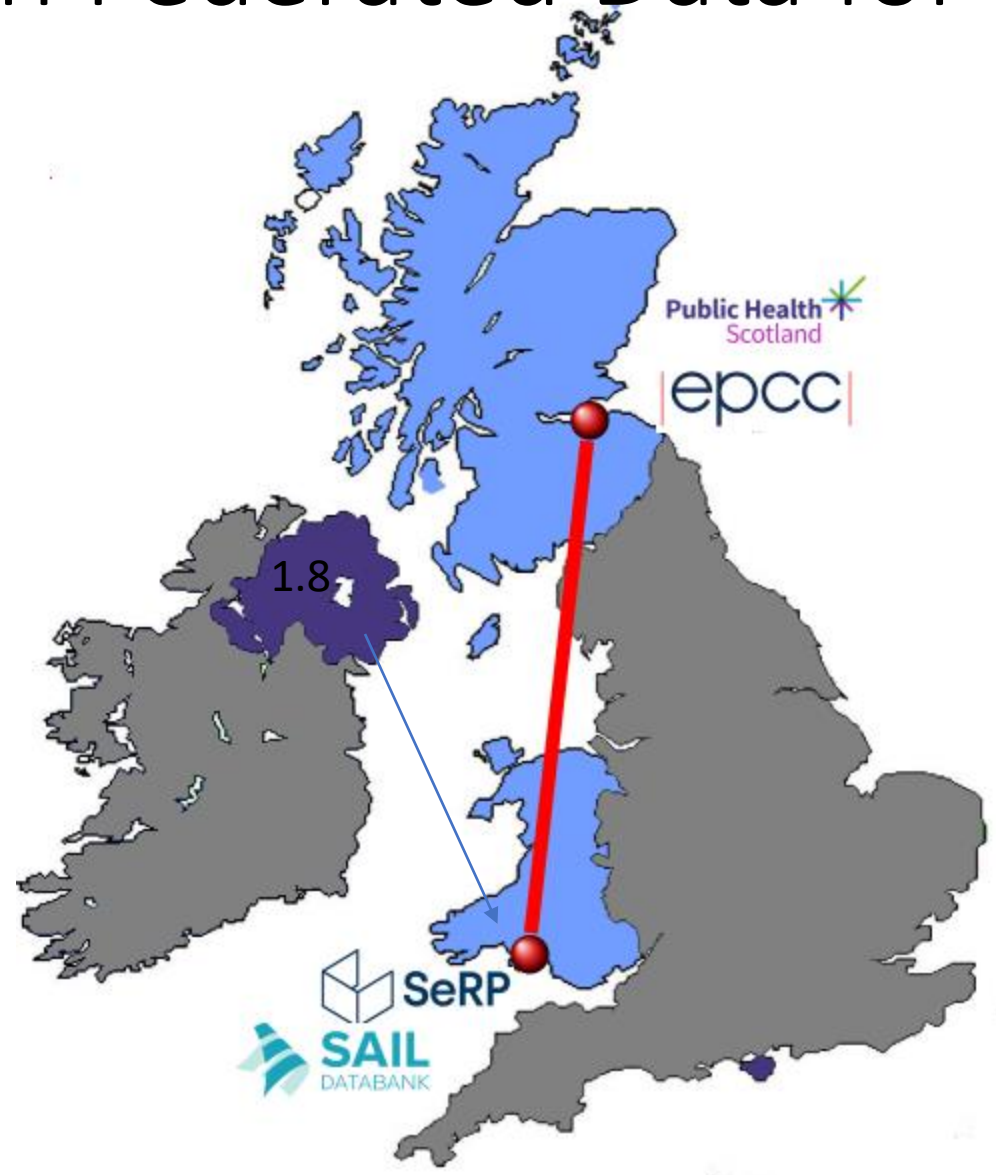
DARE: Federation Phase 2

Federated Learning

DARE-TELEPORT
“eyes on”

Connecting two Major National TRE's : Establish Federated Data for research

National Safe Haven
For Wales (3.3m)



National Safe Haven
For Scotland (5.4m)

Inter TRE plumbing

To allow data to be visible to each other and enable inter operability

“POP-UP TRE”

Ephemeral Project Specific TRE

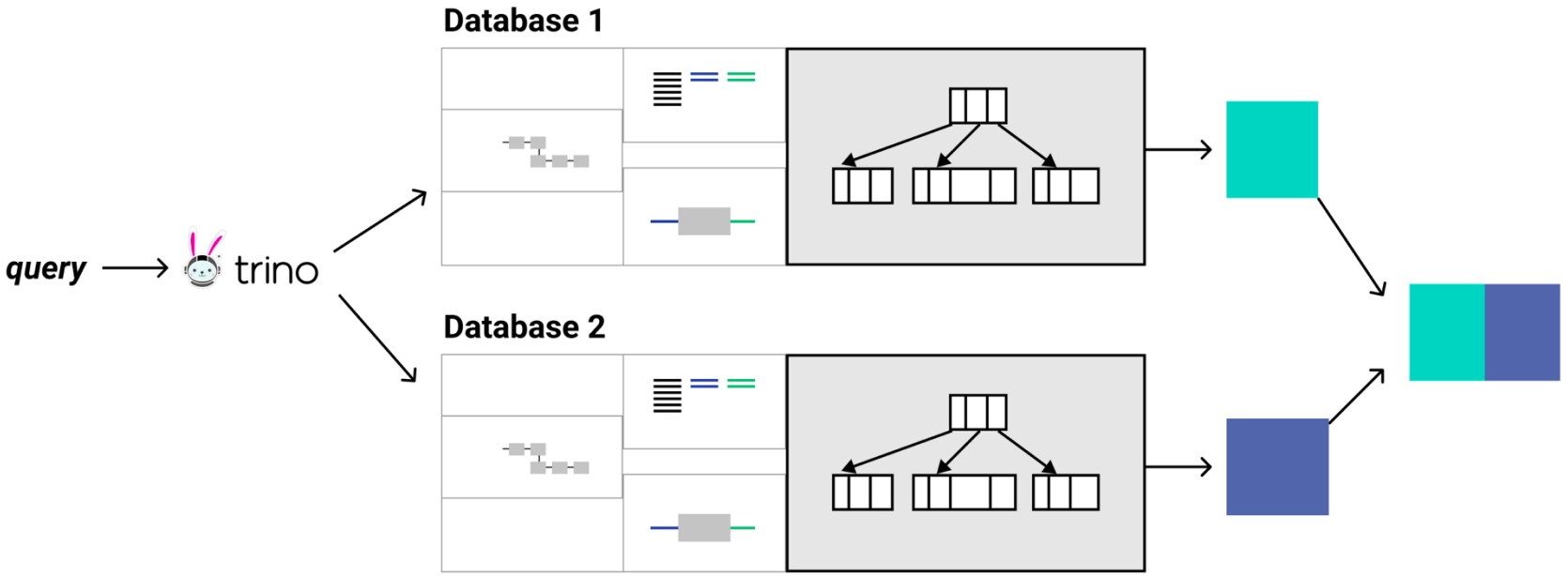
Socialising the idea of “popping-up” in either national TRE and being able to access the data from either

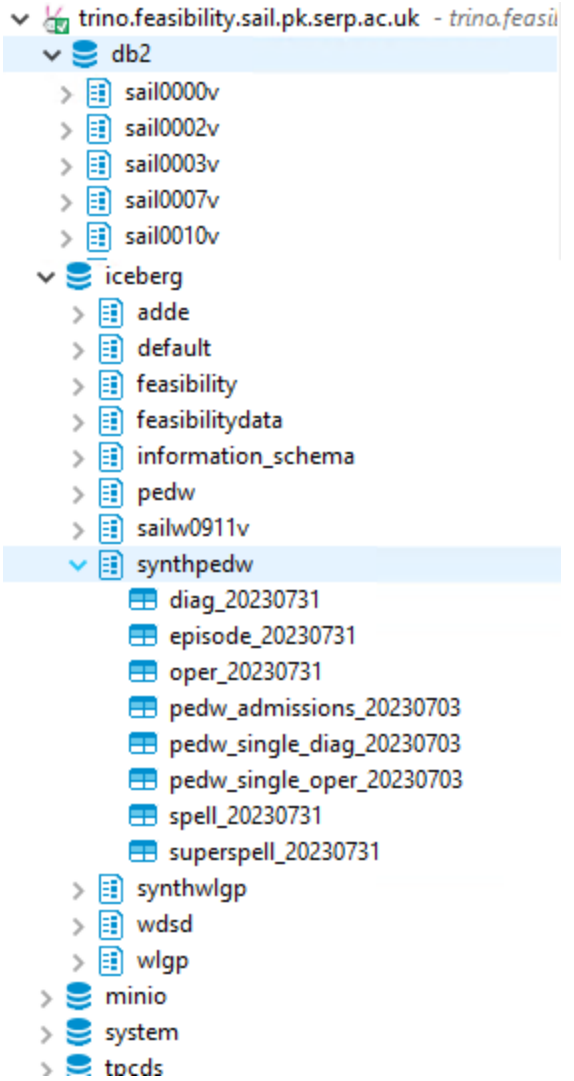
- Consistent capability
- Consistent governance
- Consistent user experience

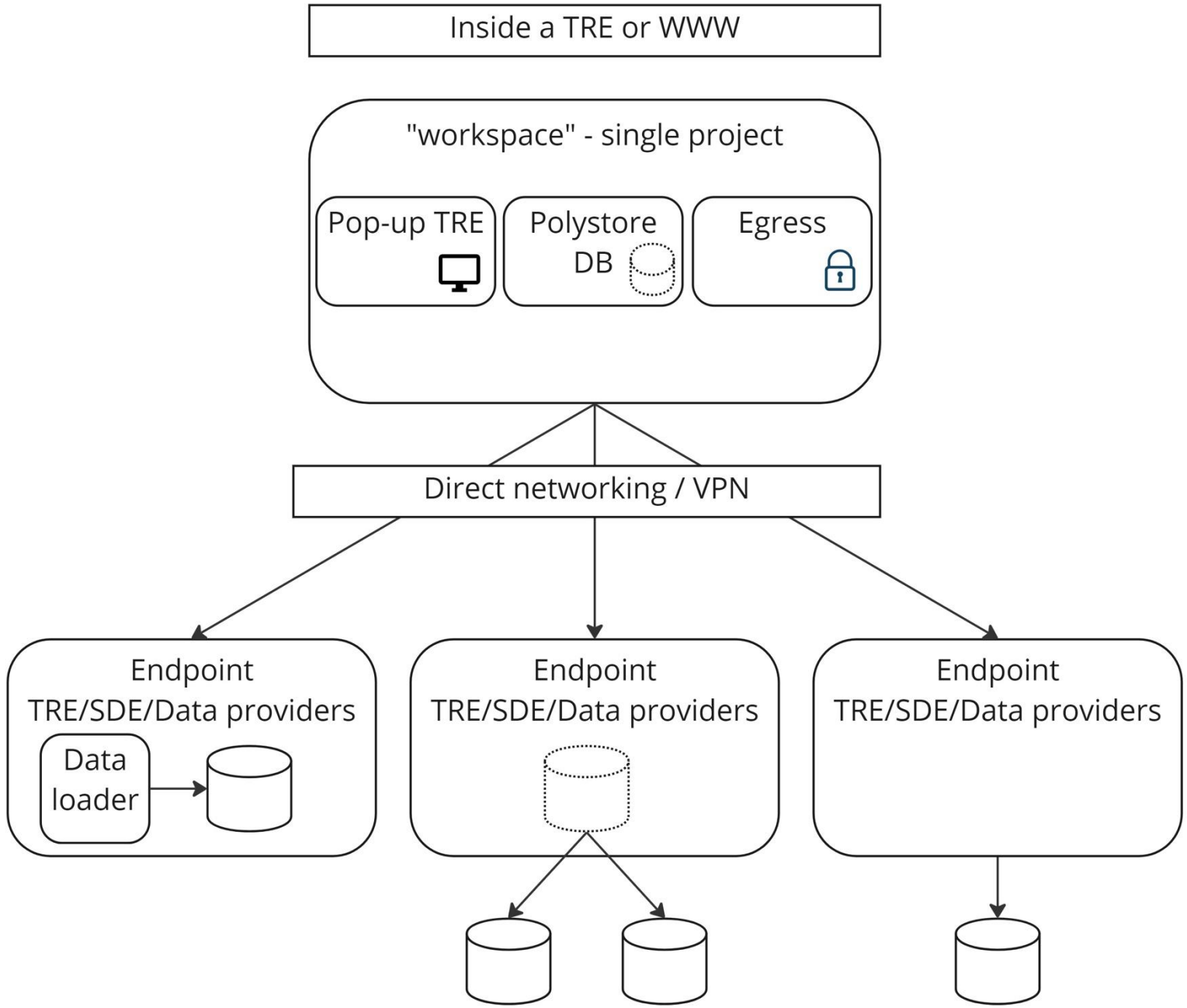
- Keeping existing TRE approaches (no changes)

Multi TRE: Single Pane of Glass

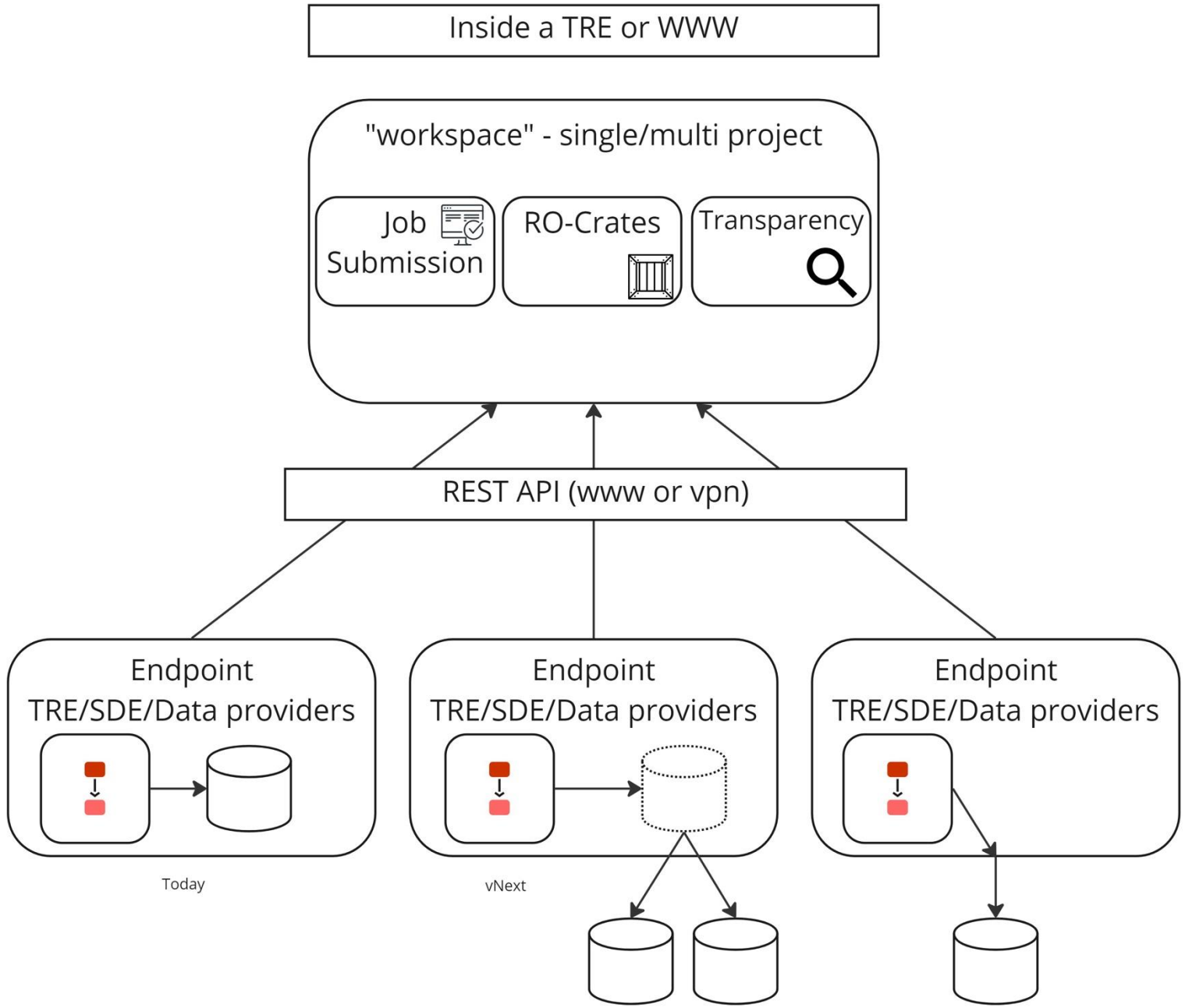
- Presented as Single database
- See each TRE as a database in the same system and be able to query both and across these systems



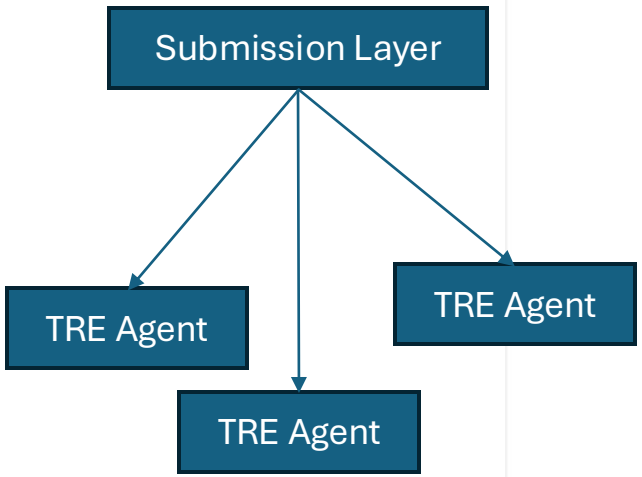


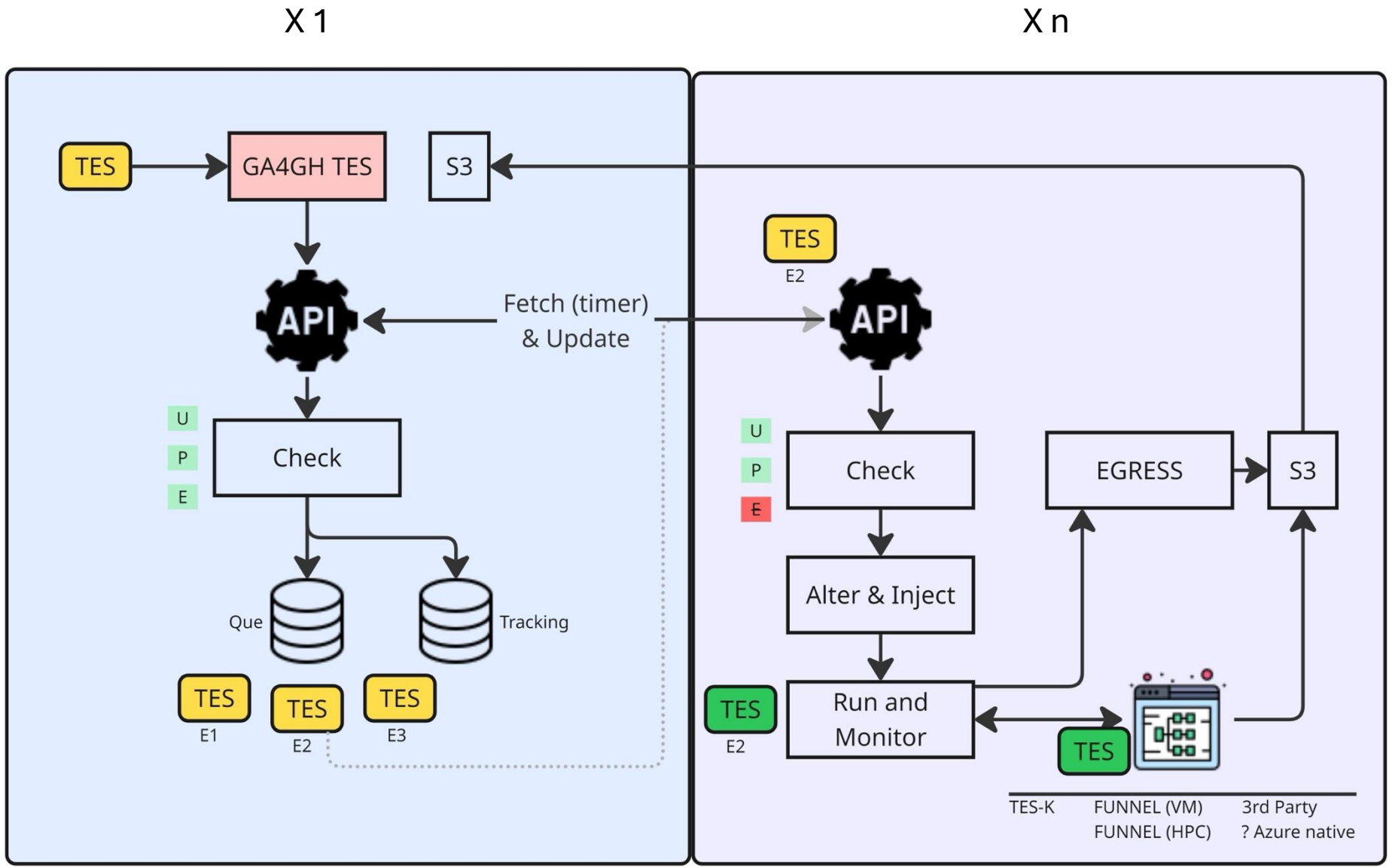

Inside a TRE or WWW
"workspace" - single project
Pop-up TRE
Polystore DB
Egress
Direct networking / VPN
Endpoint TRE/SDE/Data providers
Data loader
Endpoint TRE/SDE/Data providers
Endpoint TRE/SDE/Data providers

DARE-TREFX
“eyes off”

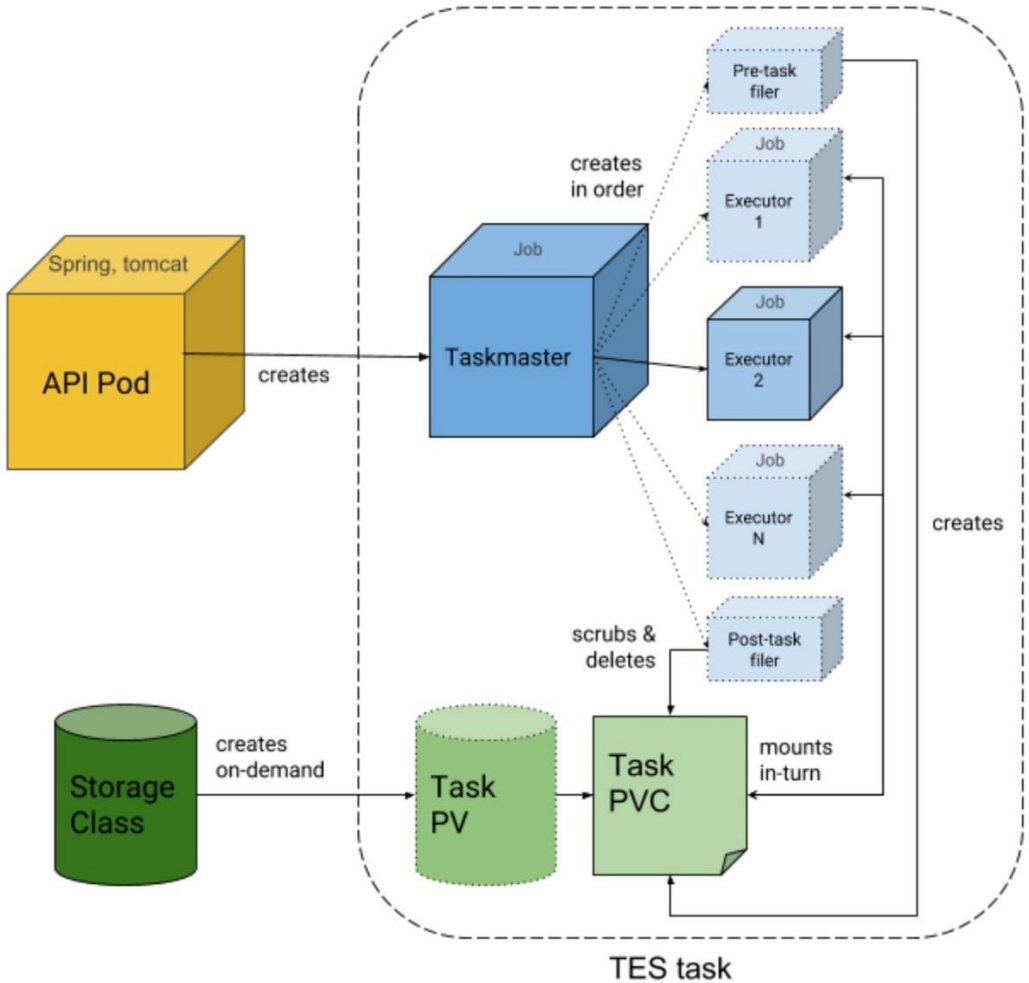

Inside a TRE or WWW
"workspace" - single/multi project
Job Submission
RO-Crates
Transparency
REST API (www or vpn)
Endpoint
TRE/SDE/Data providers
Endpoint
TRE/SDE/Data providers
Endpoint
TRE/SDE/Data providers
Today
vNext

TRE-FX


Submission Layer
TRE Agent
TRE Agent
TRE Agent

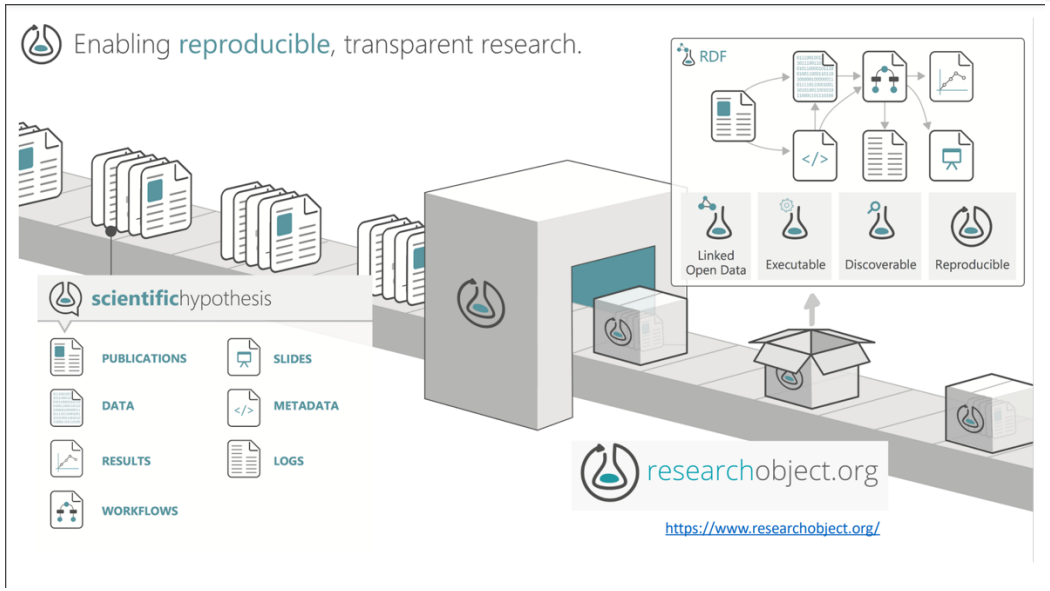

X 1
X n
TES
GA4GH TES
S3
API
Fetch (timer)
& Update
U
P
E
Check
Que
Tracking
TES
E1
TES
E2
TES
E3
TES
E2
API
U
P
E
Check
Alter & Inject
Run and
Monitor
TES
E2
EGRESS
S3
TES
TES-K
FUNNEL (VM)
FUNNEL (HPC)
3rd Party
? Azure native

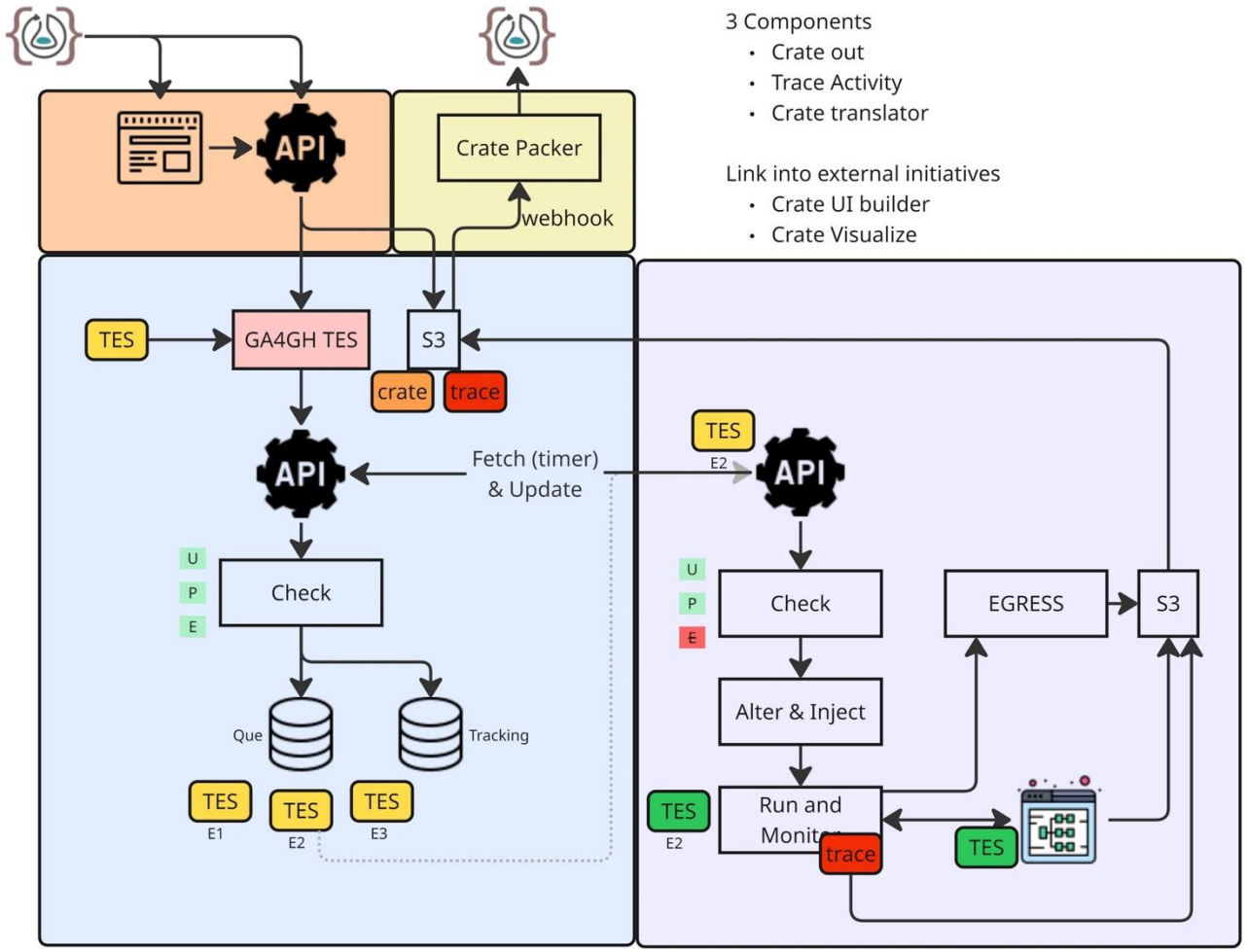
ELIXIR TESK (documentation)




```
{
  "name": "MD5 example",
  "description": "Task which runs md5sum on the input file.",
  "tags": {
      "custom-tag": "tag-value"
  },
  "inputs": [
      {
          "name": "infile",
          "description": "md5sum input file",
          "url": "/path/to/input_file",
          "path": "/container/input",
          "type": "FILE"
      }
  ],
  "outputs": [
      {
          "name": "outfile",
          "url": "/path/to/output_file",
          "path": "/container/output"
      }
  ],
  "resources": {
      "cpuCores": 1,
      "ramGb": 1,
      "diskGb": 100,
      "preemptible": false
  },
  "executors": [
      {
          "image": "ubuntu",
          "command": [
              "md5sum",
              "/container/input"
          ],
          "stdout": "/container/output",
          "stderr": "/container/stderr",
          "workdir": "/tmp"
      }
  ]
}
```

Even when extending: its still GA4GH TES





How to build analysis when eyes off

SANDBOX of TRE

- Direct Data access
- Subset locally
- TELEPORT to all sandbox endpoints

OpenSAFELY – provides sample data, a key part of the solution (GP only datasets currently)

+++ support training

DARE-TREFX

Technically difficult for end users

- Send “Query” to Endpoints
- See only metadata
- Normal Governance
- Egress / Disclosure control outputs
- Combine “Results”

Extremely similar to OpenSafely

DARE-TELEPORT

Technically difficult for TRE’s

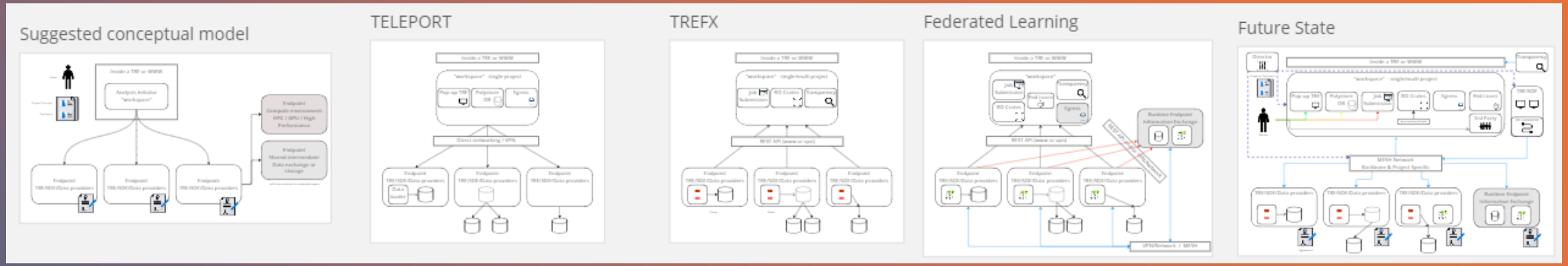
- Access “Data” from Endpoints
- See all data
- Normal Governance
- Combine / Join “Results”
- Egress Combined “Results”

TRE business as usual

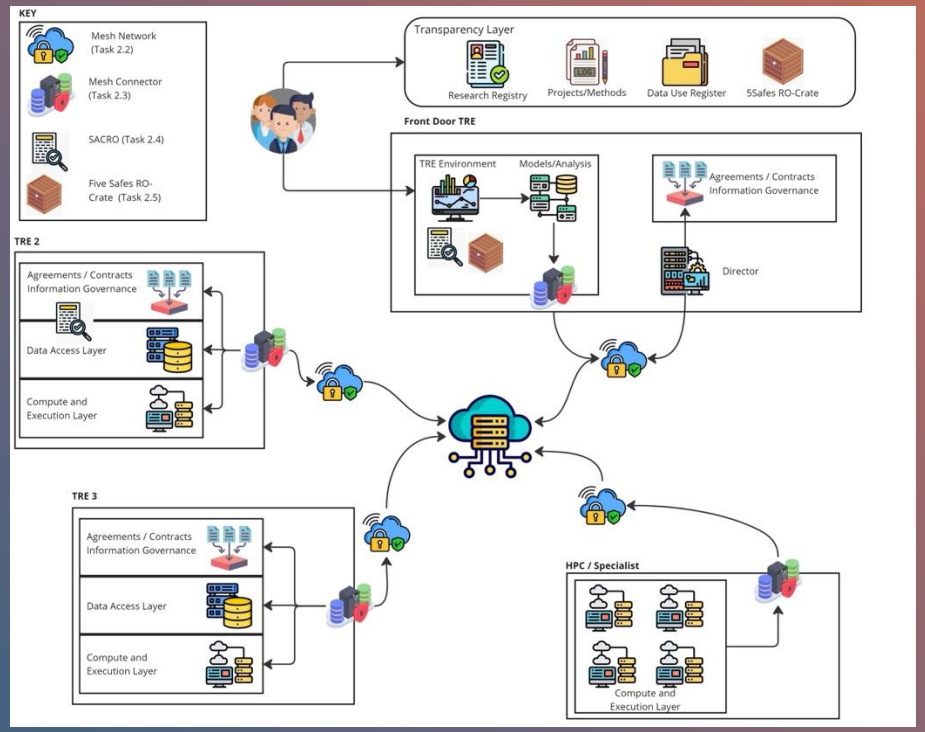
EGRESS: Safe Outputs

Jurisdiction, Control & Responsibility stop at the border

Federated Learning



Future state

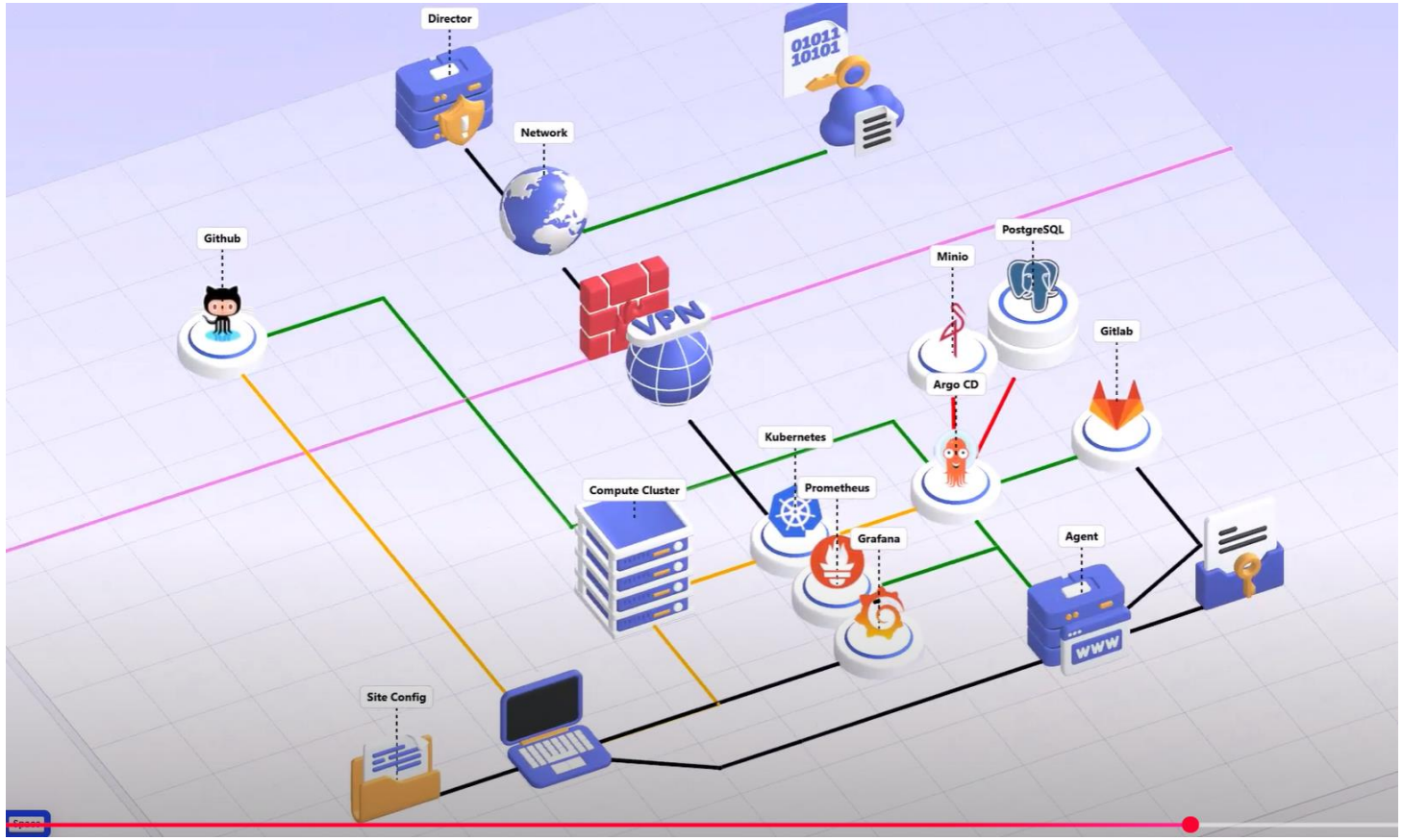


DIRECTOR

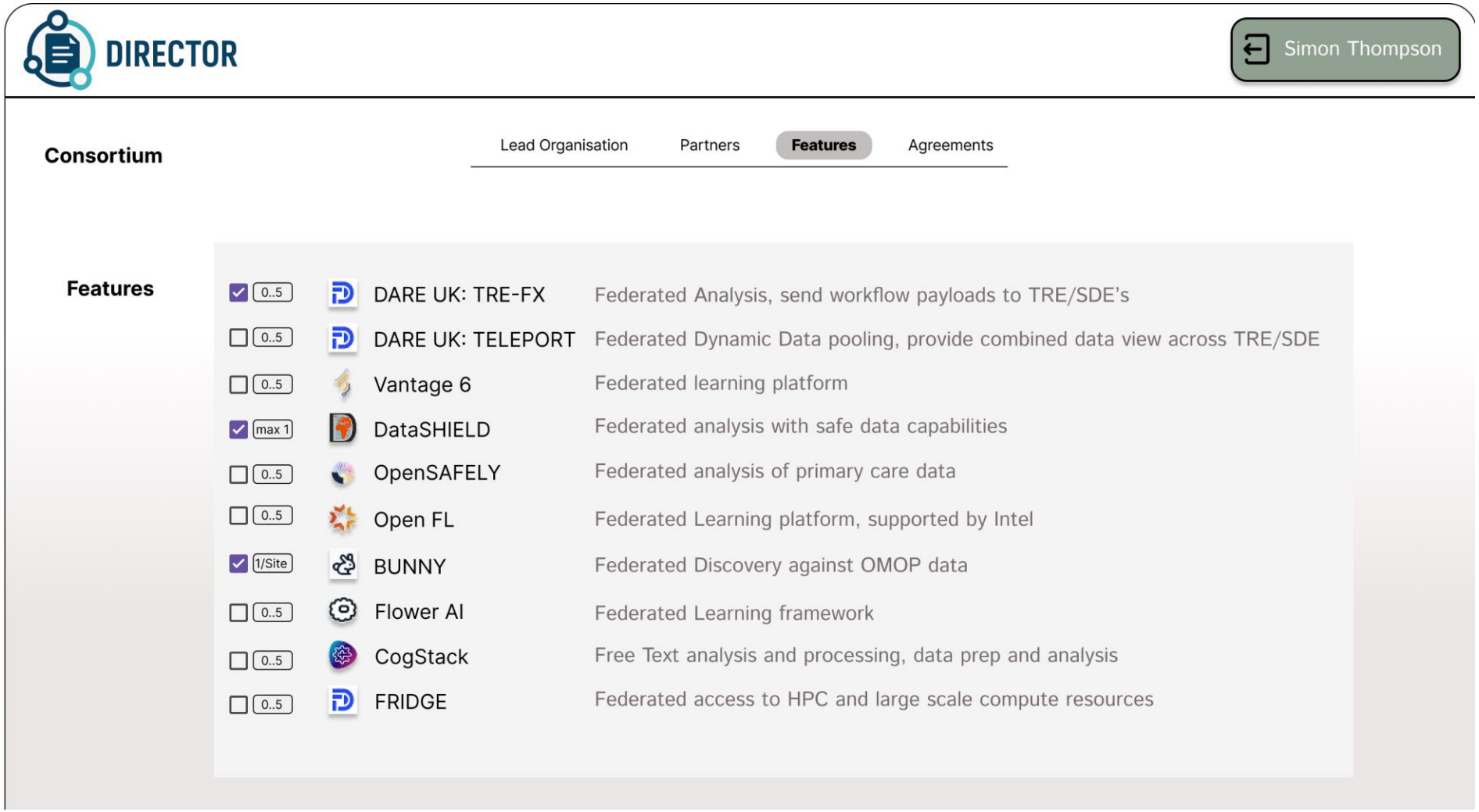
Ephemeral Compute space
Hub and spoke
Horizontal data connectivity
Federated learning and information exchange
Dynamic infrastructure – Pop-UP

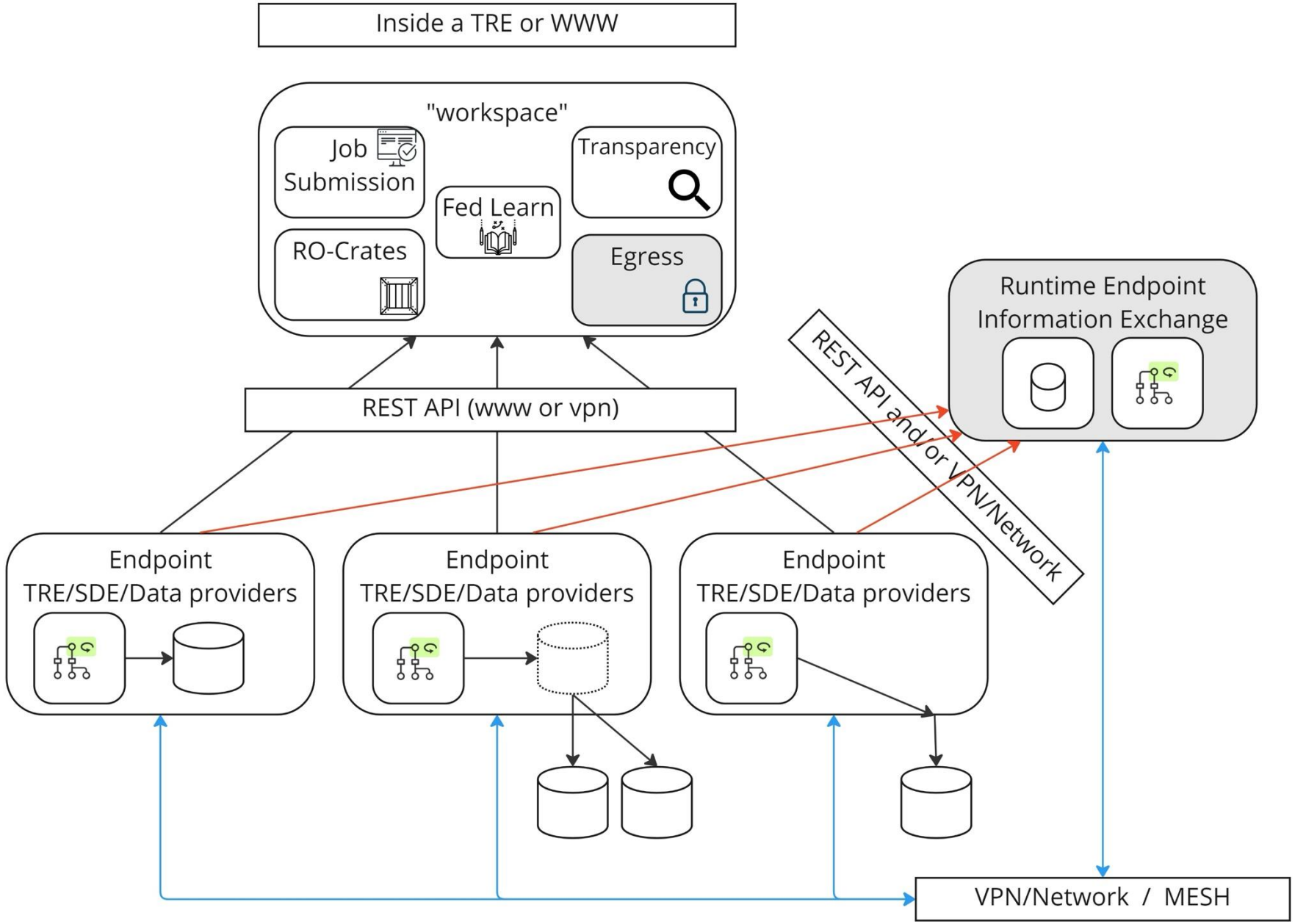
Director (future)

- Create interconnected sites
- Create data planes
- Install required software to data planes



Director (future)



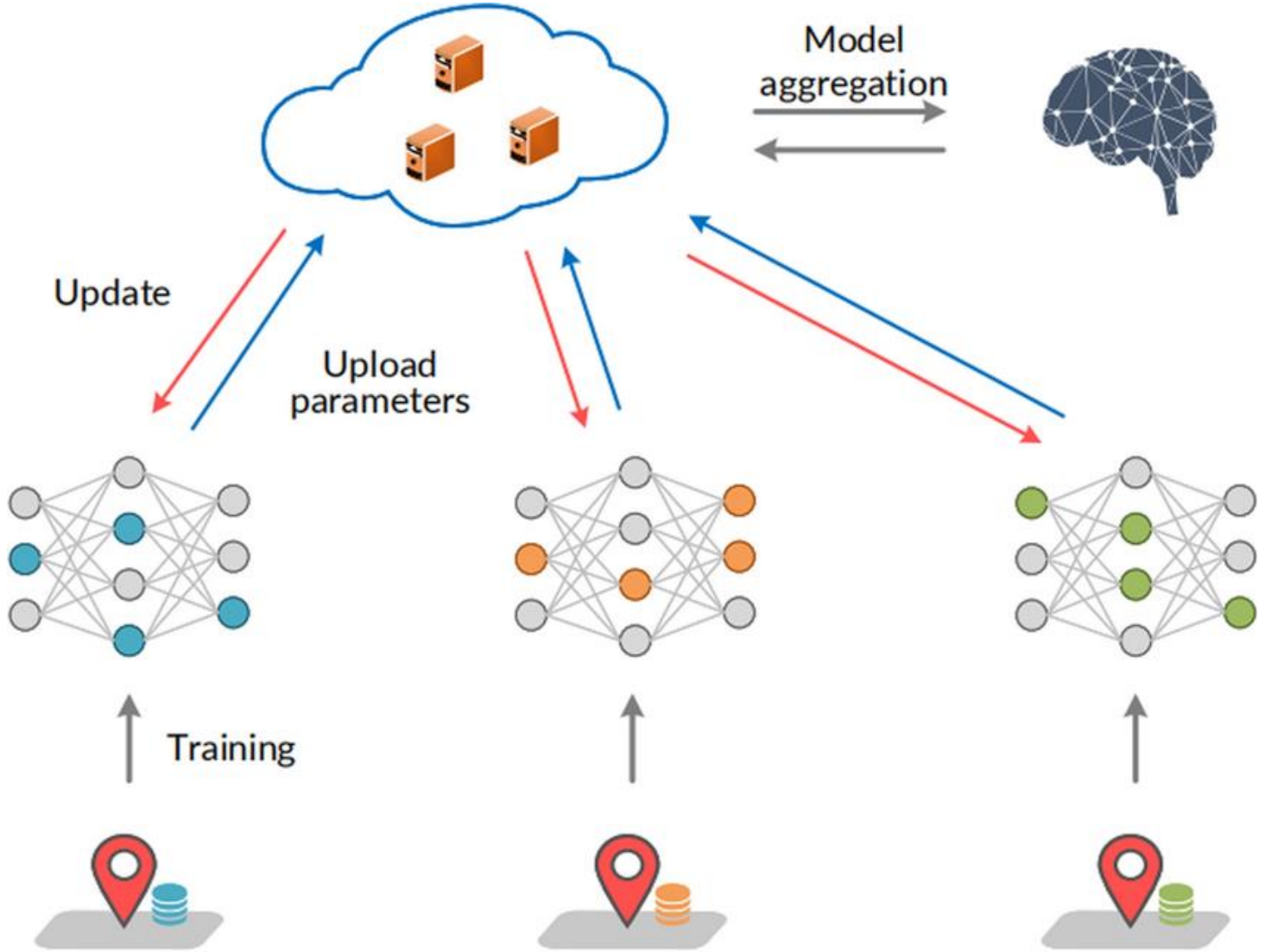

Inside a TRE or WWW
"workspace"
Job Submission
Transparency
Fed Learn
RO-Crates
Egress
Runtime Endpoint Information Exchange
REST API (www or vpn)
REST API and/or VPN/Network
Endpoint TRE/SDE/Data providers
Endpoint TRE/SDE/Data providers
Endpoint TRE/SDE/Data providers
VPN/Network / MESH

Federated learning

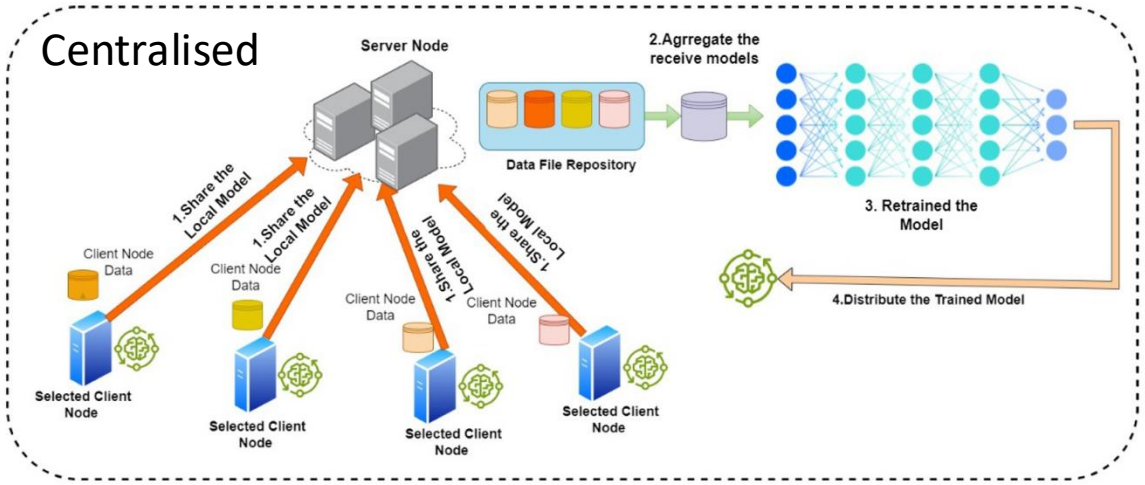
- Federated Learning is not the issue, it is **WHAT** data gets passed over the borders

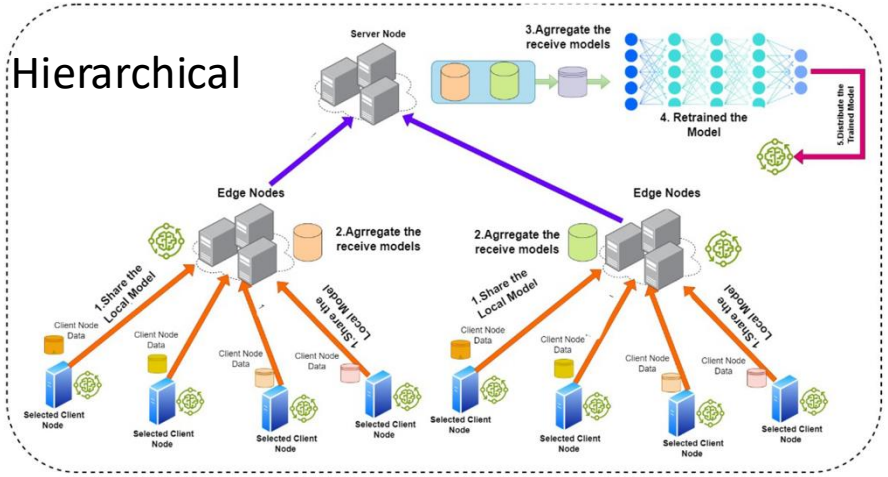
- This often depends on what analysis is being done and how it's approached

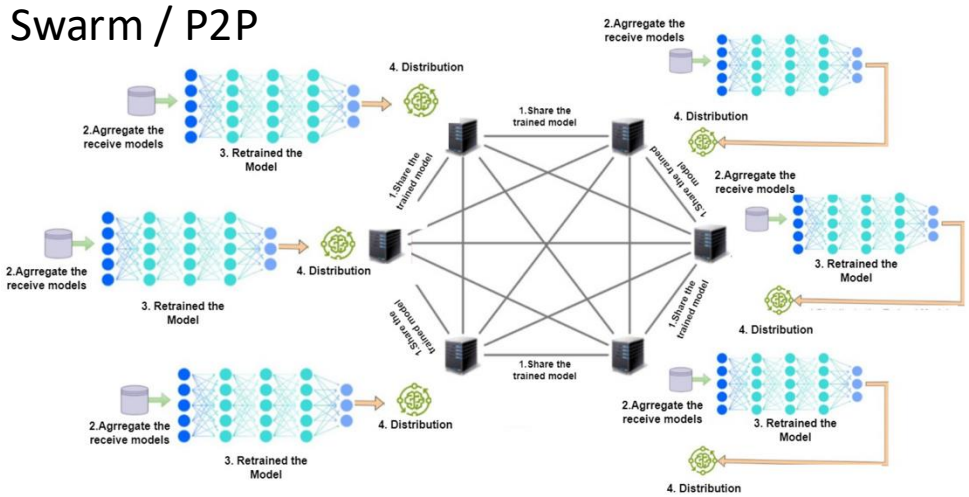
Which is often not defined till access to the data to explore what's possible

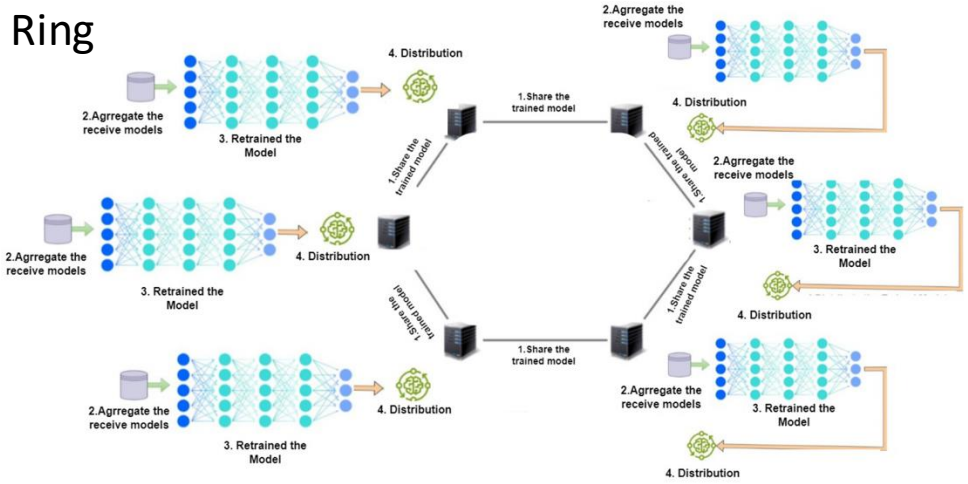


Federated Architectures



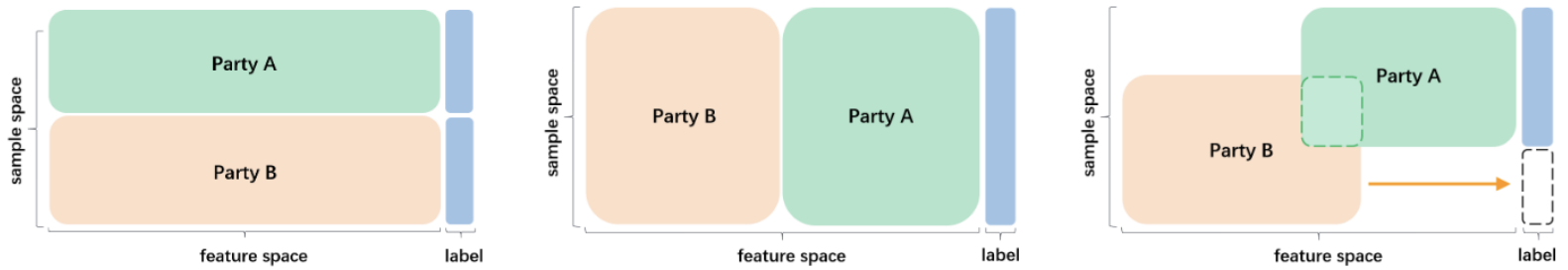






Different federated learning architectures [1].

Federated Architectures (vertical and horizontal)



(a) Horizontal Federated Learning (b) Vertical Federated Learning (c) Federated Transfer Learning

Diagram of vertical and horizontal federated learning paradigms [2].

We will focus on centralised, horizontal federated learning

- Simpler – less complexity compared to vertical federated learning or decentralized models
- Aggregation is simpler in both communication and coordination – algorithms and architecture
- Centralised/hierarchical is good for keeping federations separate and managed
- These architectures are supported by most established frameworks (FLARE, Flower, Vantage6, Datashield, Bitfount, etc.)
- Centralised can be scaled well using hierarchical architecture

Homomorphic Encryption

- Allows data to be operated on without the data itself being disclosed
- Increased privacy
- Allows for federated analytics and learning with less trust (e.g. less trusted data nodes, HPCs)
- May allow for certain analyses that would be otherwise disclosive

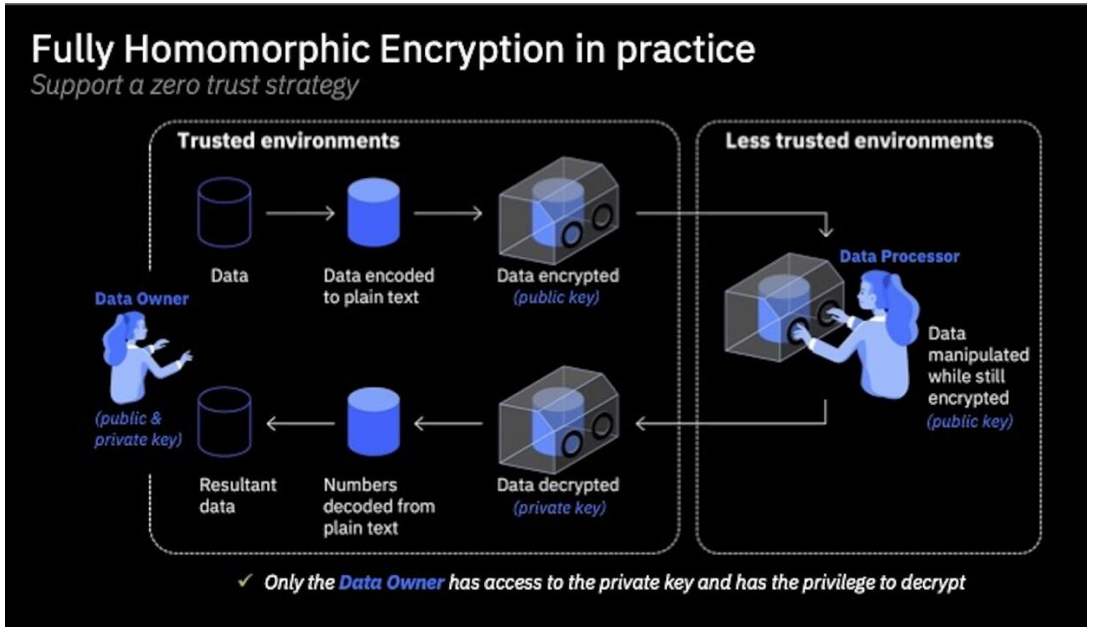


Illustration of homomorphic encryption [3]

Homomorphic Encryption

$$\text{enc}(a) + \text{enc}(b) = \text{enc}(a + b)$$

$$\text{enc}(a) \times \text{enc}(b) = \text{enc}(a \times b)$$

- Several different schemes, but a simple (somewhat) homomorphic scheme for encrypting ***bits*** is this [6]:

$$\text{enc}(m) = R \times p \ + r \times 2 \ + m$$

$$\text{dec}(c) = (c \mod p) \mod 2$$

$$\text{where } R > p \gg r$$

$$R, p, r \in \mathbb{Z}^+$$

p is the secret key (large prime), and R and r are random numbers.

- Adding or multiplying two encrypted numbers gives a result in the same form as an encrypted number
- When decrypted, the terms with p become 0 due to the mod p
- Terms with r become 0 due to the mod 2, because they are multiplied by 2

Homomorphic Encryption

- Inclusion of the error term makes the size of the encrypted result grow with every operation!
- If the error term gets as large as p, the scheme fails and the data can't be recovered.

$$\text{enc}(a) \times \text{enc}(b) = R_1 p_1 R_2 p_2 + R_1 p_1 m_2 + m_1 R_2 p_2 + m_1 m_2$$
$$+2m_1 r_2 + 2R_1 p_1 r_2 + 2r_1 R_2 p_2 + 2r_1 m_2$$
$$+4r_1 r_2$$

ERROR

- As the size approaches the limit, the result can be "bootstrapped"
- This is typically re-encrypting the already encrypted result
- It's like a repeater in a digital signal – a new signal is created from the degrading signal, but the new signal is clean
- It can be slow – really slow!

Homomorphic Encryption

- Levelled HE schemes exist – the idea is that we can avoid the highest computational cost by not bootstrapping, just avoid doing too many multiplications
- This can be supported for analytics and learning, when running inference – up to shallow neural nets.
- Deep learning inference will use too many multiplications.
- Training will use too many multiplications, but we can *usually* train in plaintext

Analytics and Inference with Homomorphic Encryption

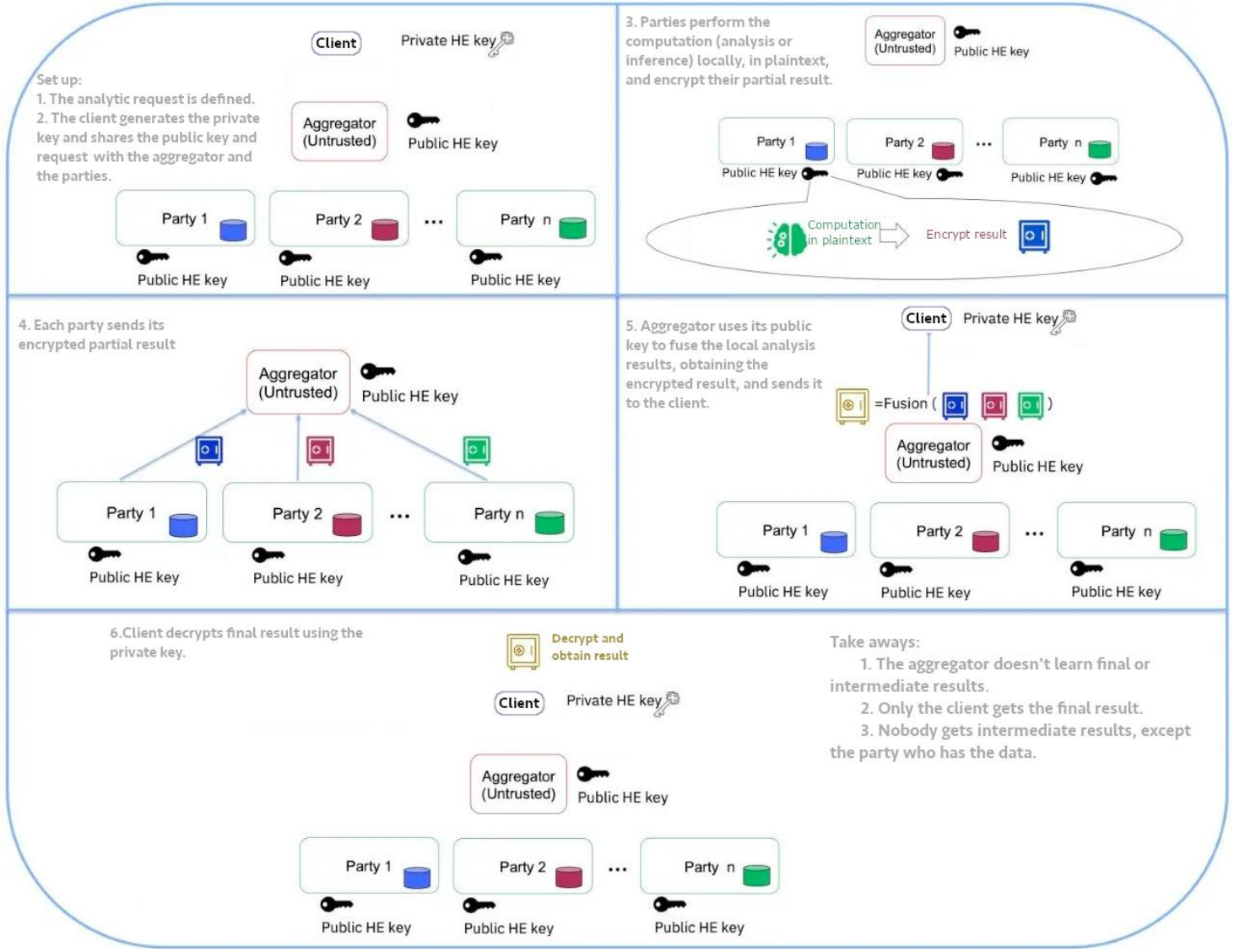


Diagram of analysis and inference scheme for machine learning models using homomorphic encryption, adapted from [4].

Training with Homomorphic Encryption

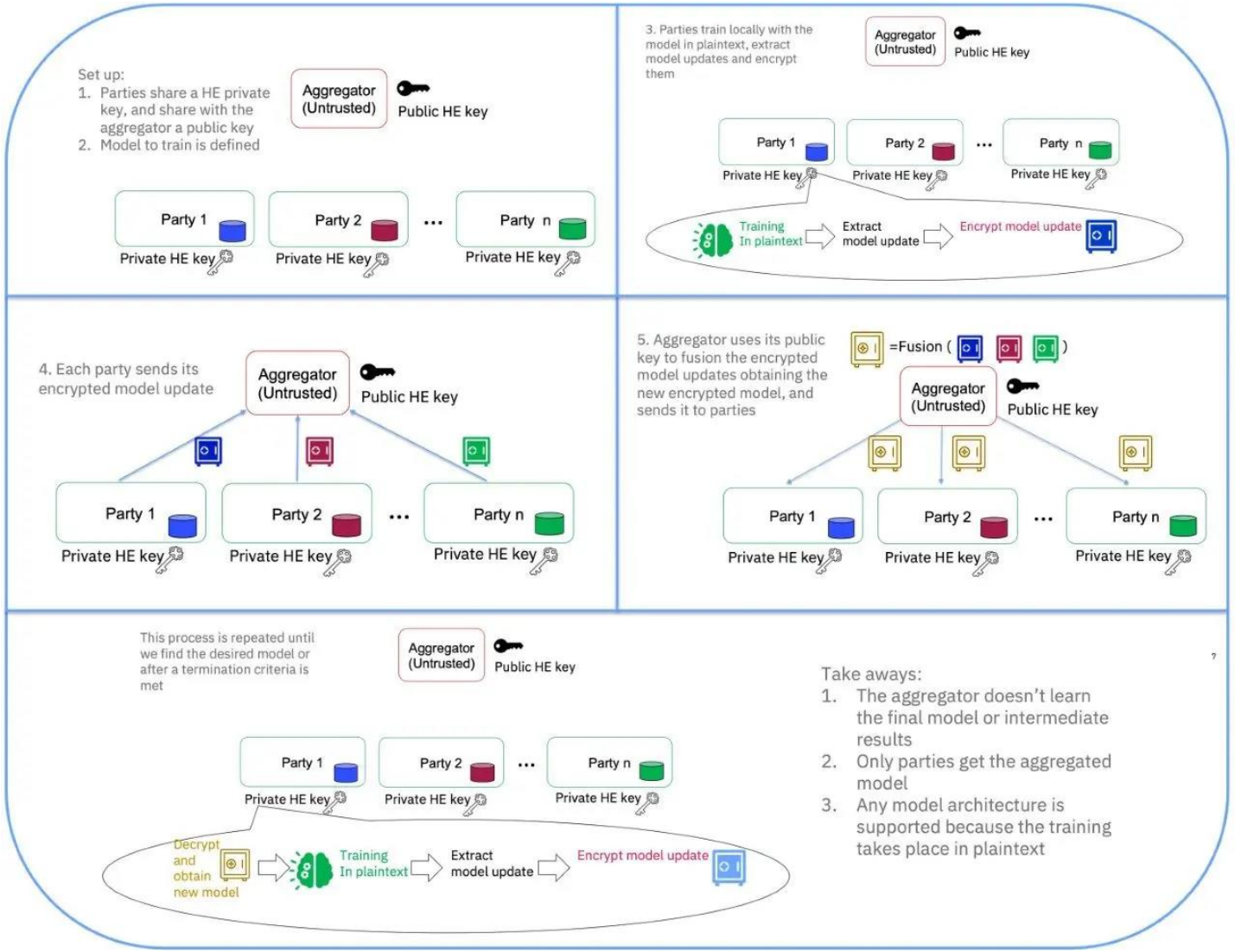
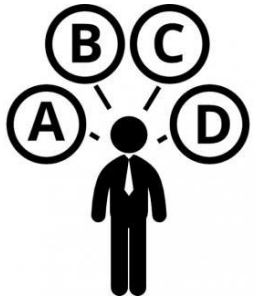
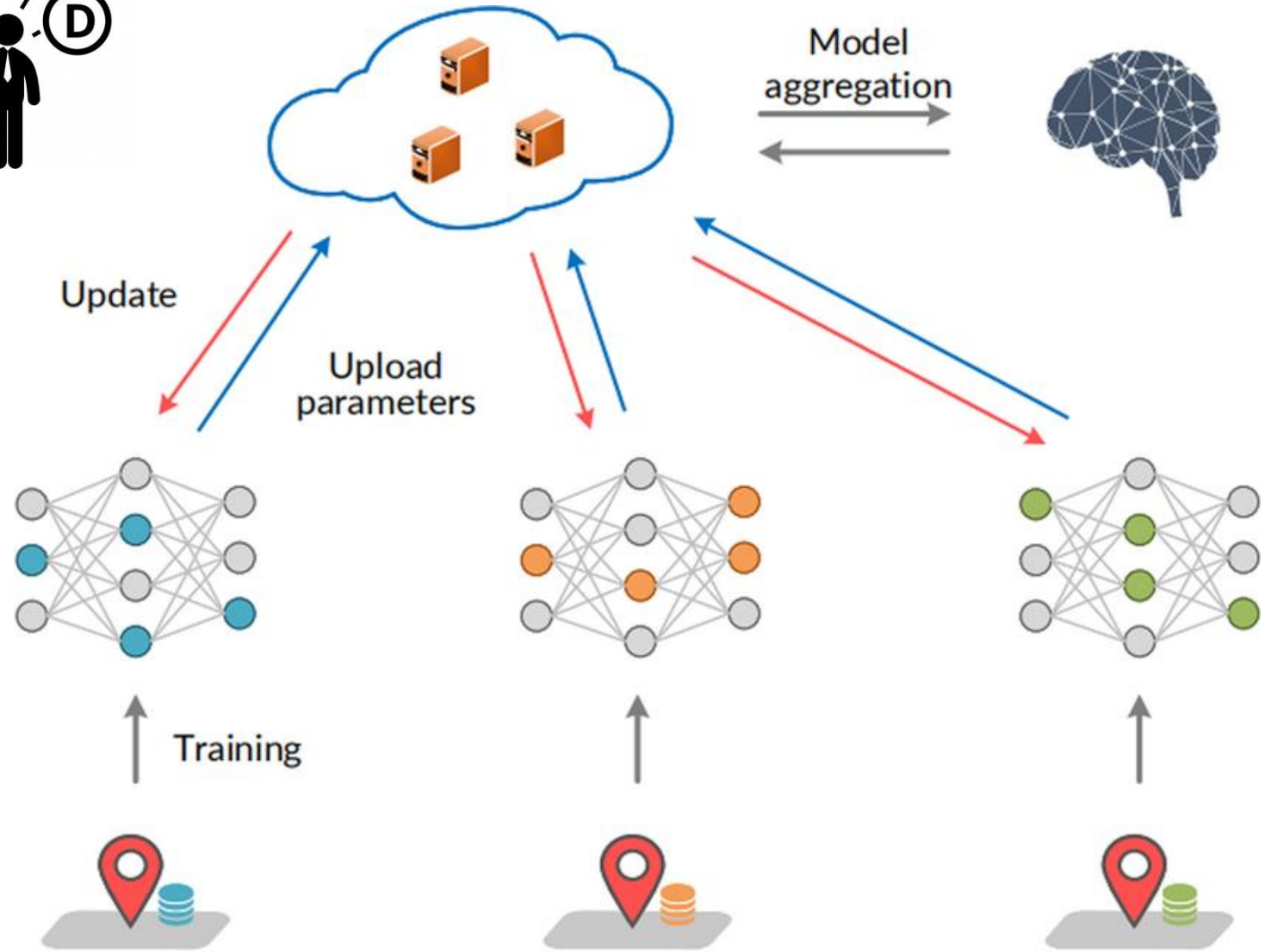


Diagram of training scheme for machine learning models using homomorphic encryption [4].

Federated learning





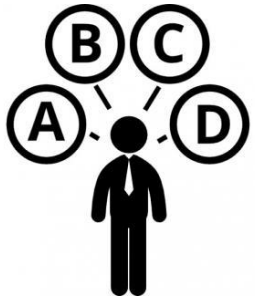
- Federated Learning is not the issue, it is **WHAT** data gets passed over the borders

- *This often depends on what analysis is being done and how it's approached*

Which is often not defined till access to the data to explore what's possible

- **Homomorphic Encryption approach (if possible)**

Federated learning

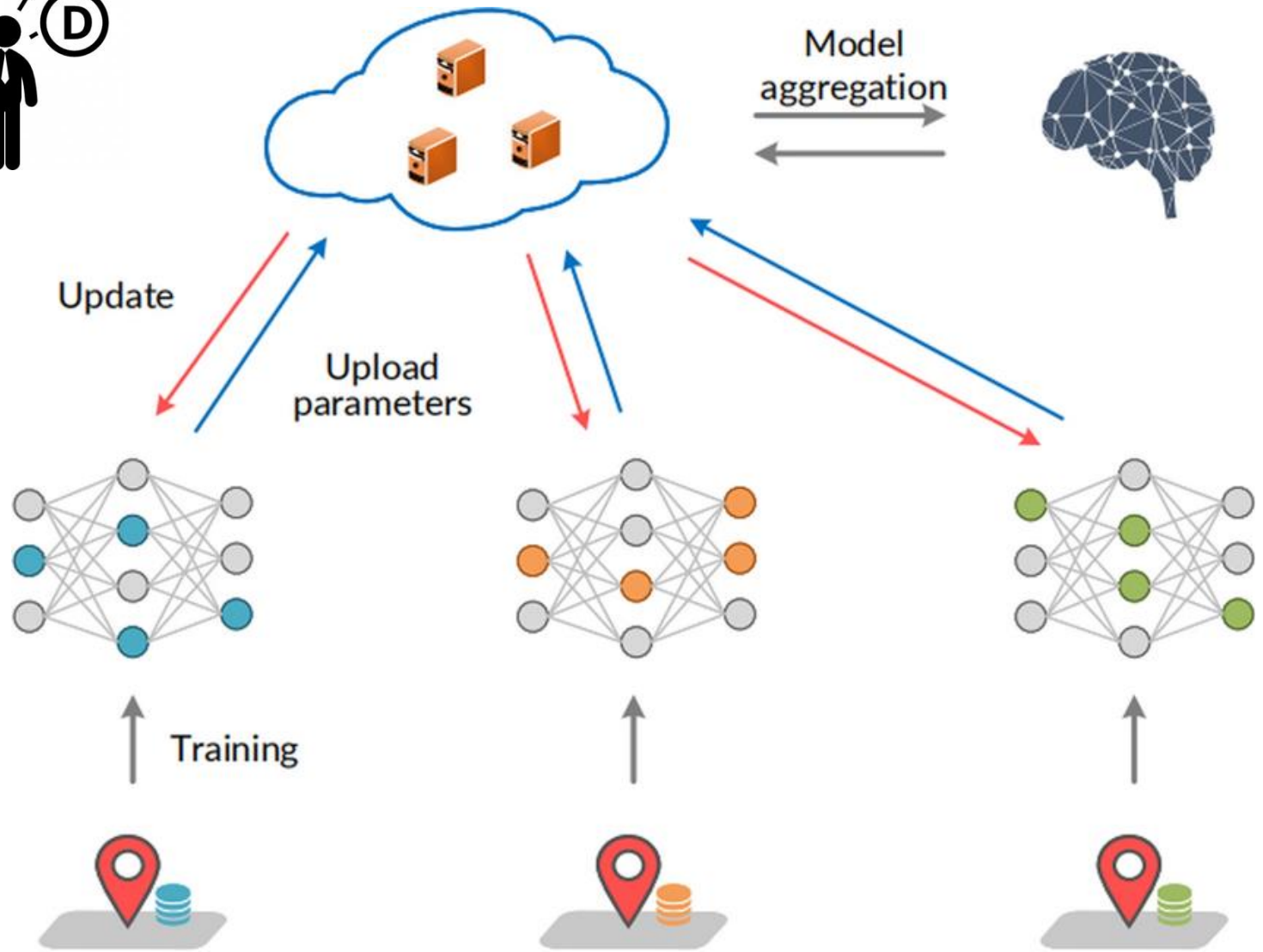


- Federated Learning is not the issue, it is **WHAT** data gets passed over the borders

- *This often depends on what analysis is being done and how it's approached*

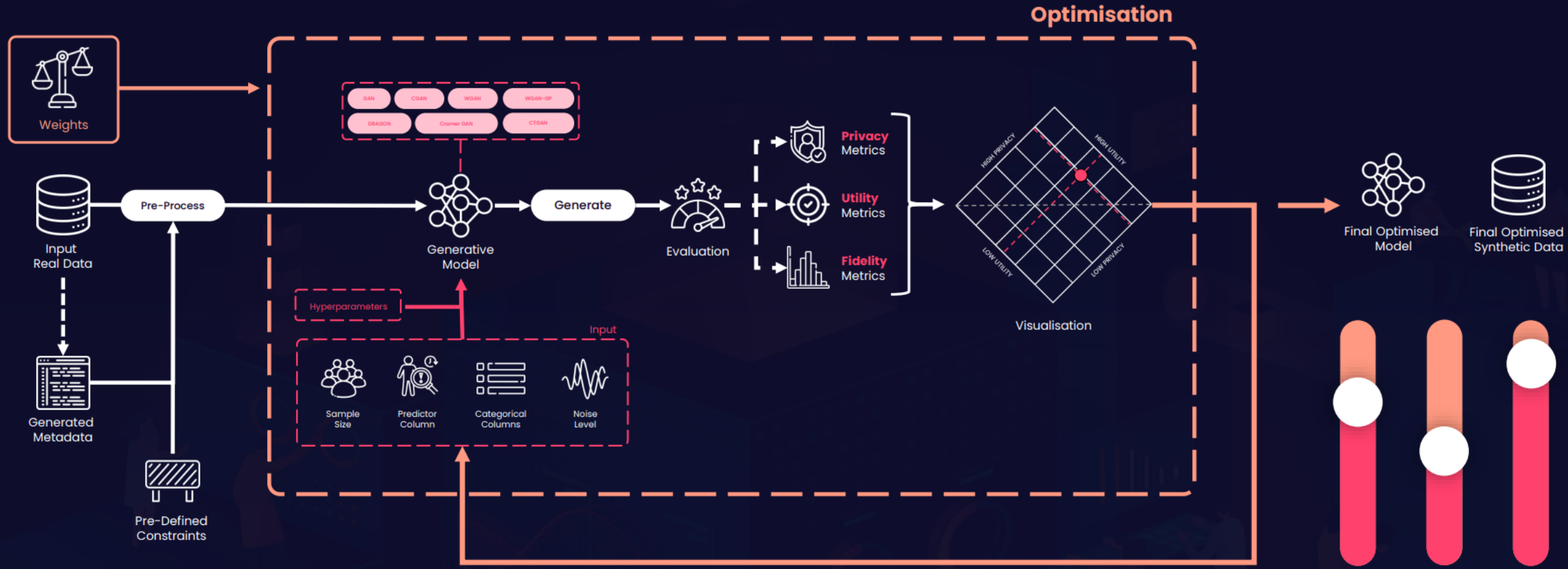
Which is often not defined till access to the data to explore what's possible

- **Synthetic data could allow this to be <u>defined</u> for a full approval and switch to "real" data**



SynthOpt
Synthetic Data Optimisation

Optimisation



Dementias Platform UK | Data Portal

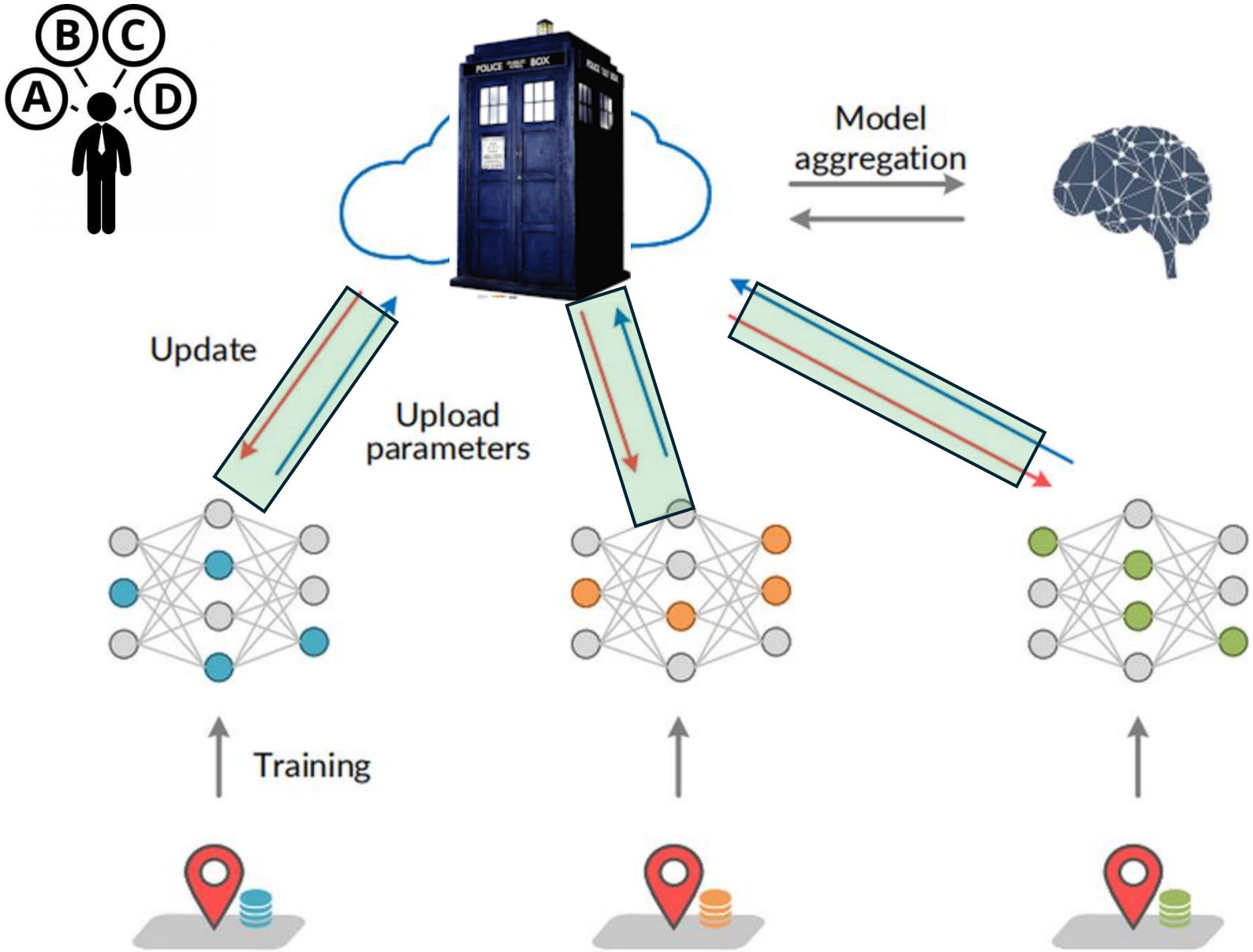
BRAIN Building Responsible Artificial Intelligence *in* Neuroscience

AI Risk Evaluation Community Group

Federated learning

- Federated Learning is not the issue, it is **WHAT** data gets passed over the borders

- **Remove the borders**
- **Delay Egress**
- **“Teleport” concept**



Future: Provide capability to all partners

- Get the plumbing sorted
- Develop the governance and operating models

- Let the science roll

Diolch yn fawr