All of us together

UK Health Data Research Alliance Symposium

Tuesday 1 December 2020
09:30 - 17:00

Trusted (and Productive) Research Environments for safe research

Chair: Susheel Varma, Health Data Research UK

Panellists: Matthew Howard, AWS; James Zwiers, NHS Digital; David Sibbald, Aridhia; Angela Wood, University of Cambridge

HDRUK
Health Data Research UK

UK Health Data Research Alliance

Health Data Research Innovation Gateway

→ Join the conversation:
hdruk.ac.uk/HDRAlliance20
@HDR_UK
#HDRAlliance20

aws machine learning

Trusted Research Environments

Dr. Matthew Howard
International Head of Public Sector Healthcare Data Science & AI
Amazon Web Services

Who Are We? & What is Cloud?

A broad and deep platform that helps customers build sophisticated, scalable, secure applications

Why do customers choose AWS?

Agility

Allows teams to experiment and innovate quickly and frequently

Cost Savings

Only pay for what you use, lower upfront expenses

Go Global in Minutes

Most extensive, reliable, and secure global cloud infrastructure

Innovate Faster

Ability to focus on business differentiators, not infrastructure

Elasticity

Stop guessing capacity, scale up and down with demand

Service Breadth & Depth

180+ fully featured services to support any cloud workload

aws

Architected for European Compliance Requirements

ISO 9001
Global Quality Standard

ISO 27001
Security Management Controls

ISO 27017
Cloud Specific Controls

ISO 27018
Personal Data Protection

GDPR
General Data Protection Regulation

SOC 1
Audit Controls Report

SOC 2
Security, Availability, & Confidentiality Report

SOC 3
General Controls Report

Data Security & Protection Toolkit
Standards Exceeded

C5 [Germany]
Operational Security Attestation

ASIP HDS [France]
Personal Health Data Protection

Cyber Essentials Plus [UK]
Cyber Threat Protection

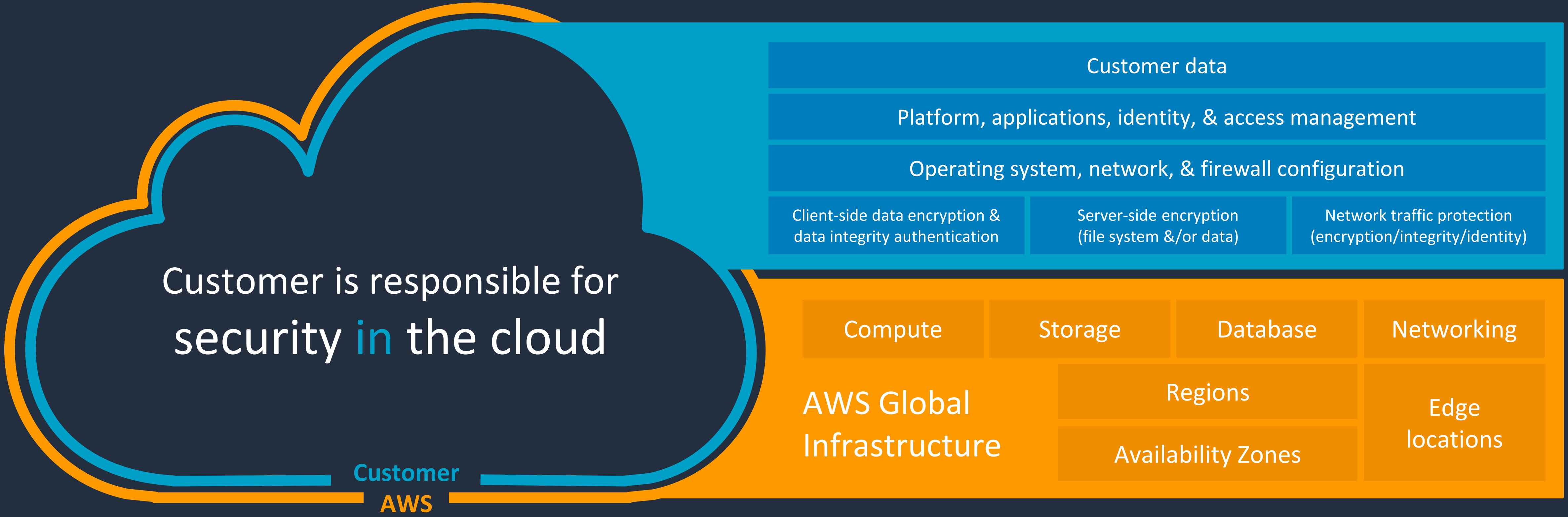
ENS High [Spain]
Spanish Government Standards

NCSC [UK]
Cloud security guidance

CSA
Cloud Security Alliance Controls

And many more . . . https://aws.amazon.com/compliance/

Share your security responsibility with AWS



Customer data

Platform, applications, identity, & access management

Operating system, network, & firewall configuration

Client-side data encryption & data integrity authentication	Server-side encryption (file system &/or data)	Network traffic protection (encryption/integrity/identity)

Compute | Storage | Database | Networking

AWS Global Infrastructure

Regions

Availability Zones

Edge locations

Customer is responsible for security in the cloud

AWS is responsible for security of the cloud

Example: Service Workbench on AWS

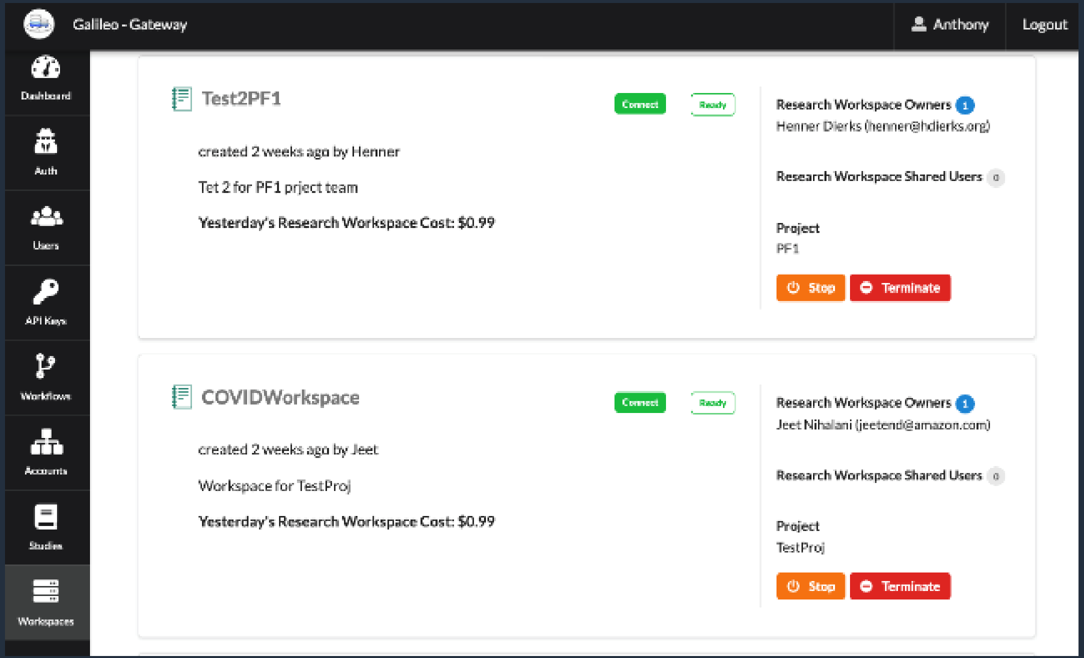
Web-based, Collaborative Research Framework

- An **open source** research web-application for you and your peers to collaboratively work with federated data, launch compute and tools within minutes

- Enables IT to provide you with secure, repeatable, pre-configured research environments that meet your institution's compliance needs (**HITRUST, HIPAA, ISO, FedRAMP Mod-eligible AWS services)**

- Provides you and IT with cost transparency and spend controls to help your projects stay within budget

- As an AWS open source solution, you only pay for the underlying AWS services consumed



Test2PF1

created 2 weeks ago by Henner

Tet 2 for PF1 prject team

Yesterday's Research Workspace Cost: $0.99

Research Workspace Owners 1
Henner Dierks (henner@hdierks.org)

Research Workspace Shared Users 0

Project
PF1

COVIDWorkspace

created 2 weeks ago by Jeet

Workspace for TestProj

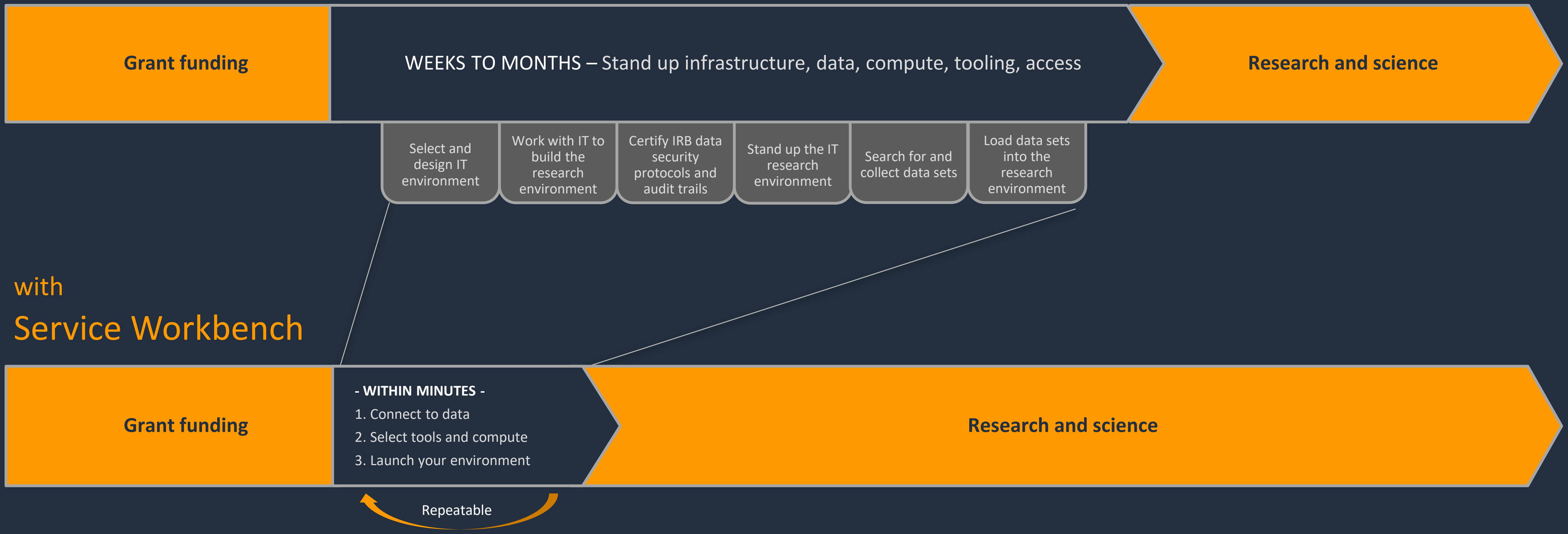
Yesterday's Research Workspace Cost: $0.99

Research Workspace Owners 1
Jeet Nihalani (jeetend@amazon.com)

Research Workspace Shared Users 0

Project
TestProj

Use Cloud to reduce time to science



Existing

Grant funding

WEEKS TO MONTHS – Stand up infrastructure, data, compute, tooling, access

Research and science

- Select and design IT environment
- Work with IT to build the research environment
- Certify IRB data security protocols and audit trails
- Stand up the IT research environment
- Search for and collect data sets
- Load data sets into the research environment

with Service Workbench

Grant funding

- WITHIN MINUTES -

1. Connect to data
2. Select tools and compute
3. Launch your environment

Research and science

Repeatable

Promotes

- ✓ Repeatable and configurable
- ✓ Secure infrastructure and environments
- ✓ Cost estimation, tracking, and controls

Why researchers are choosing Service Workbench on AWS

Reduce time to science
Access research environments in minutes

Controlled access to data
Controlled access to datasets at scale

Conduct research securely
Maintain consistent security, compliance, and governance

Globally accessible
Collaborate with researchers around the world

Scale and agility to grow
Virtually limitless tooling via AWS Service Catalog

Spend controls
Cost visibility, centralized budgeting and chargeback management

Broad and Deep Functionality

TECHNICAL & BUSINESS SUPPORT

- Support
- Professional Services
- Optimization Guidance
- Partner Ecosystem
- Training & Certification
- Solutions Management
- Account Management
- Security & Billing Reports
- Personalized Dashboard

MARKETPLACE

- Business Apps
- Business Intelligence
- DevOps Tools
- Security
- Networking
- Databases
- Storage

ANALYTICS

- Data Warehousing
- Business Intelligence
- Hadoop/Spark
- Streaming Data Analysis
- Streaming Data Collection
- Elasticsearch
- Data Pipelines
- Interactive SQL Queries
- ETL

APP SERVICES

- Queuing & Notifications
- Workflow
- Search
- Email
- Transcoding

DEV OPS

- One-click App Deployment
- Resource Templates
- Build & Test
- Application Lifecycle Management
- DevOps Resource Management
- Triggers
- Containers
- Analyze & Debug
- Patching

MOBILE SERVICES

- API Gateway
- Single Integrated Console
- Identity
- Sync
- Mobile Analytics
- Mobile App Testing
- Targeted Push Notifications

IoT

- Rules Engine
- Device Shadows
- Device SDKs
- Device Gateway
- Registry
- Local Compute

MACHINE LEARNING

- Custom Model Training & Hosting
- Image & Scene Recognition
- Facial Recognition & Analysis
- Facial Search
- Text to Speech
- Conversational Chatbots
- Deep Learning (Apache MXNet, TensorFlow, & others)

ENTERPRISE APPS

- Virtual Desktops
- Sharing & Collaboration
- Corporate Email
- App Streaming
- Communications
- Contact Center

HYBRID ARCHITECTURE

- Data Integration
- Integrated Networking
- Integrated Identity & Access
- Integrated Resource & Deployment Management
- Integrated Devices & Edge Systems

MIGRATION

- Schema Conversion
- Exabyte-Scale Data Migration
- Application Migration
- Database Migration
- Server Migration

INFRASTRUCTURE

- Regions
- Availability Zones
- Points of Presence

CORE SERVICES

- Compute
  VMs, Auto-scaling, Load Balancing, Containers, Virtual Private Servers, Batch Computing, Cloud Functions, Elastic GPUs, Edge Computing
- Storage
  Object, Blocks, File, Archivals, Import/Export, Exabyte-scale data transfer
- Databases
  Relational, NoSQL, Caching, Migration, PostgreSQL compatible
- Networking
  VPC, DX, DNS
- CDN

SECURITY & COMPLIANCE

- Identity Management
- Configuration Compliance
- Access Control
- Key Management & Storage
- Monitoring & Logs
- Account Grouping
- Assessment & Reporting
- Resource & Usage Auditing
- Web Application Firewall
- DDOS Protection

MANAGEMENT TOOLS

- Manage Resources
- Monitoring
- Service Catalogue
- Server Management
- Configuration Tracking
- Resource Templates

aws

Service Workbench Core Capabilities

- Comes preconfigured with 5 simple research environments for
  - EC2 VM for Linux
  - EC2 VM for Windows
  - RStudio on EC2
  - SageMaker on EC2 with Jupyter Notebook
  - Hail on EMR

- Federated access to S3 storage whereby datasets are emulated as local file folders with writable storage

- Requires that your IT has resources familiar with AWS services, including AWS Service Catalog, AWS CloudFormation templates

- https://github.com/awslabs/service-workbench-on-aws

Thank You!

mjhoward@amazon.com

aws.amazon.com/health/

github.com/awslabs/service-workbench-on-aws

NHS Digital

NHS Digital Trusted Research Environments

Enabling Repeatable, Reliable Population Scale Research

HDRUK Alliance Symposium

1st December 2020

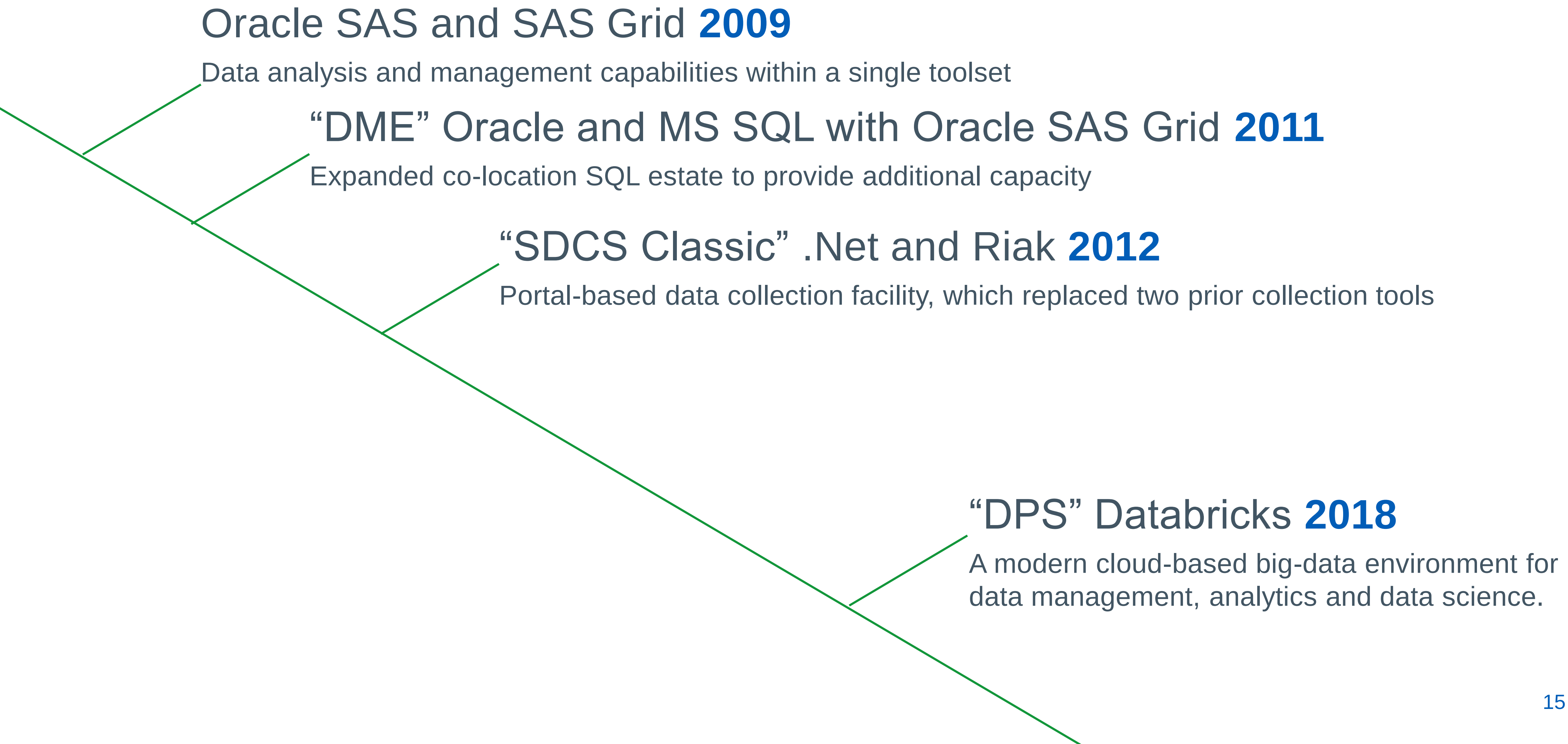
Information and technology
for better health and care

Presented by
James Zwiers

Data Platform

History & Capabilities

History



Oracle SAS and SAS Grid **2009**

Data analysis and management capabilities within a single toolset

“DME” Oracle and MS SQL with Oracle SAS Grid **2011**

Expanded co-location SQL estate to provide additional capacity

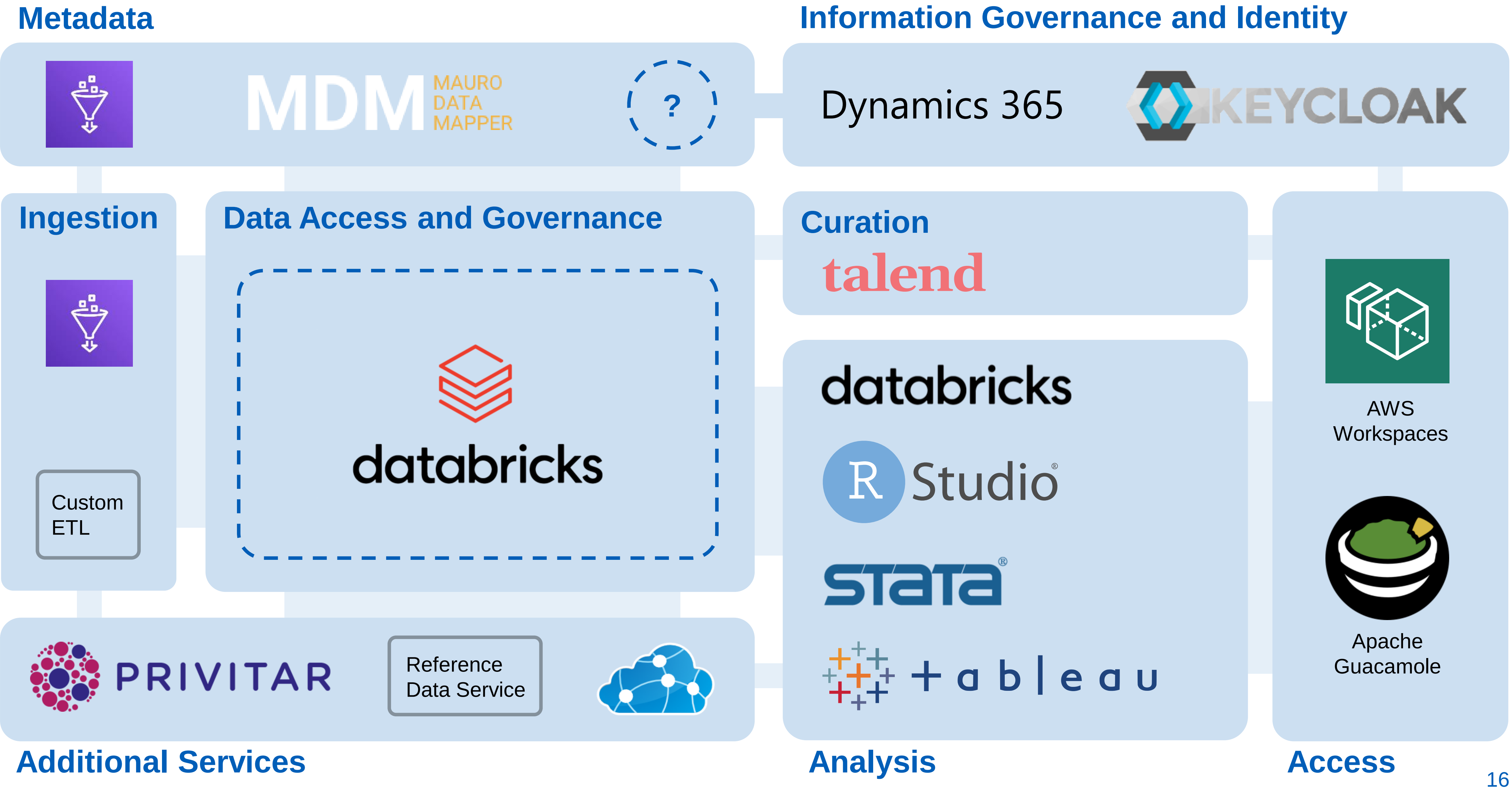
“SDCS Classic” .Net and Riak **2012**

Portal-based data collection facility, which replaced two prior collection tools

“DPS” Databricks **2018**

A modern cloud-based big-data environment for data management, analytics and data science.

Architectural Overview



Metadata

MDM MAURO DATA MAPPER

?

Information Governance and Identity

Dynamics 365

KEYCLOAK

Ingestion

Custom ETL

Data Access and Governance

databricks

Curation

talend

databricks

R Studio

stata

tableau

AWS Workspaces

Apache Guacamole

PRIVITAR

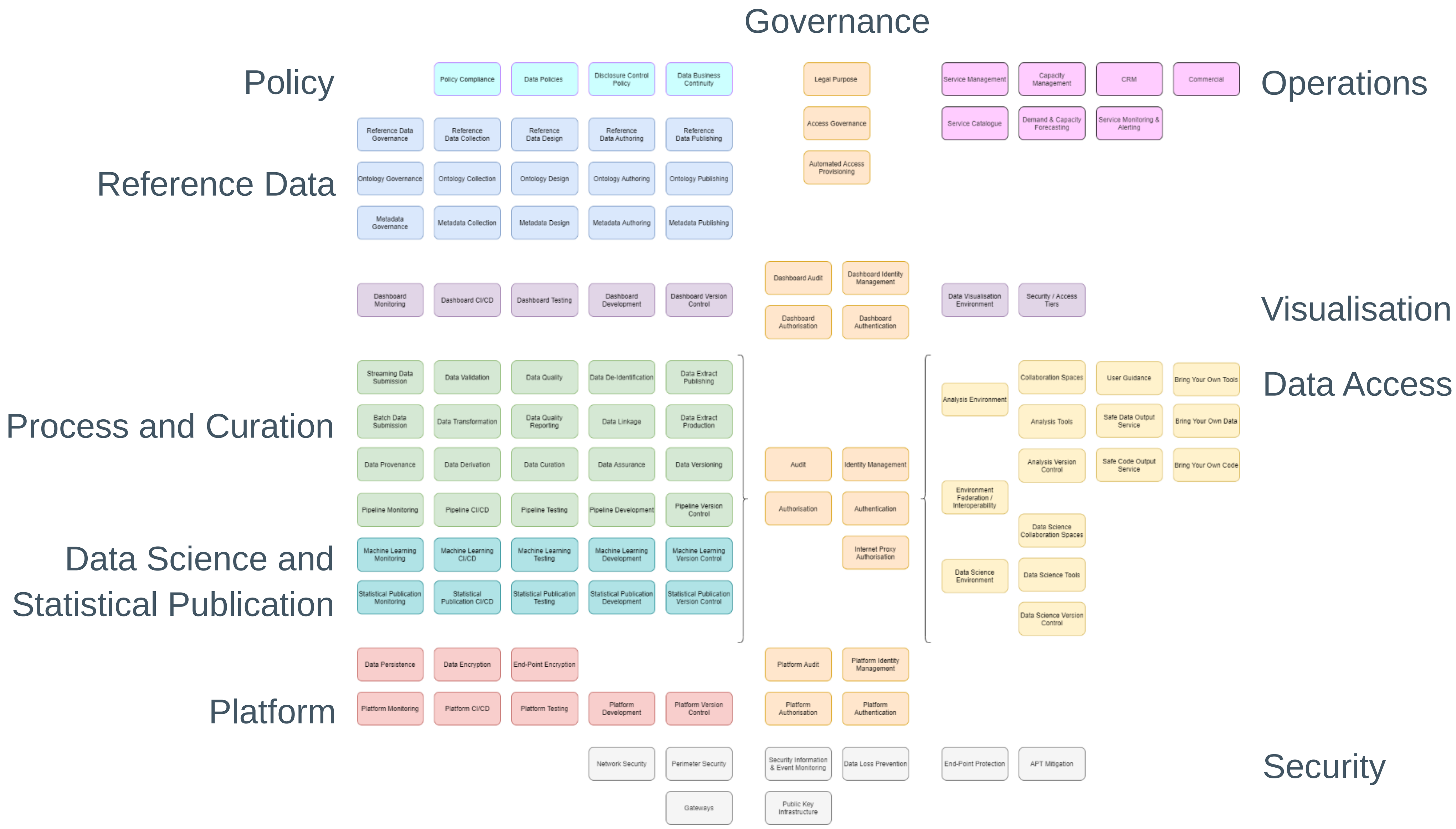
Reference Data Service

Additional Services

Analysis

Access

Platform Capabilities



Governance

Policy

Policy Compliance
Data Policies
Disclosure Control Policy
Data Business Continuity

Legal Purpose
Access Governance
Automated Access Provisioning

Operations

Service Management
Capacity Management
CRM
Commercial
Service Catalogue
Demand & Capacity Forecasting
Service Monitoring & Alerting

Reference Data

Reference Data Governance
Reference Data Collection
Reference Data Design
Reference Data Authoring
Reference Data Publishing
Ontology Governance
Ontology Collection
Ontology Design
Ontology Authoring
Ontology Publishing
Metadata Governance
Metadata Collection
Metadata Design
Metadata Authoring
Metadata Publishing

Visualisation

Dashboard Monitoring
Dashboard CI/CD
Dashboard Testing
Dashboard Development
Dashboard Version Control
Dashboard Audit
Dashboard Identity Management
Dashboard Authorisation
Dashboard Authentication
Data Visualisation Environment
Security / Access Tiers

Process and Curation

Streaming Data Submission
Data Validation
Data Quality
Data De-Identification
Data Extract Publishing
Batch Data Submission
Data Transformation
Data Quality Reporting
Data Linkage
Data Extract Production
Data Provenance
Data Derivation
Data Curation
Data Assurance
Data Versioning
Pipeline Monitoring
Pipeline CI/CD
Pipeline Testing
Pipeline Development
Pipeline Version Control

Data Science and Statistical Publication

Machine Learning Monitoring
Machine Learning CI/CD
Machine Learning Testing
Machine Learning Development
Machine Learning Version Control
Statistical Publication Monitoring
Statistical Publication CI/CD
Statistical Publication Testing
Statistical Publication Development
Statistical Publication Version Control

Audit
Identity Management
Authorisation
Authentication
Internet Proxy Authorisation

Data Access

Analysis Environment
Collaboration Spaces
User Guidance
Bring Your Own Tools
Analysis Tools
Safe Data Output Service
Bring Your Own Data
Analysis Version Control
Safe Code Output Service
Bring Your Own Code
Environment Federation / Interoperability
Data Science Collaboration Spaces
Data Science Environment
Data Science Tools
Data Science Version Control

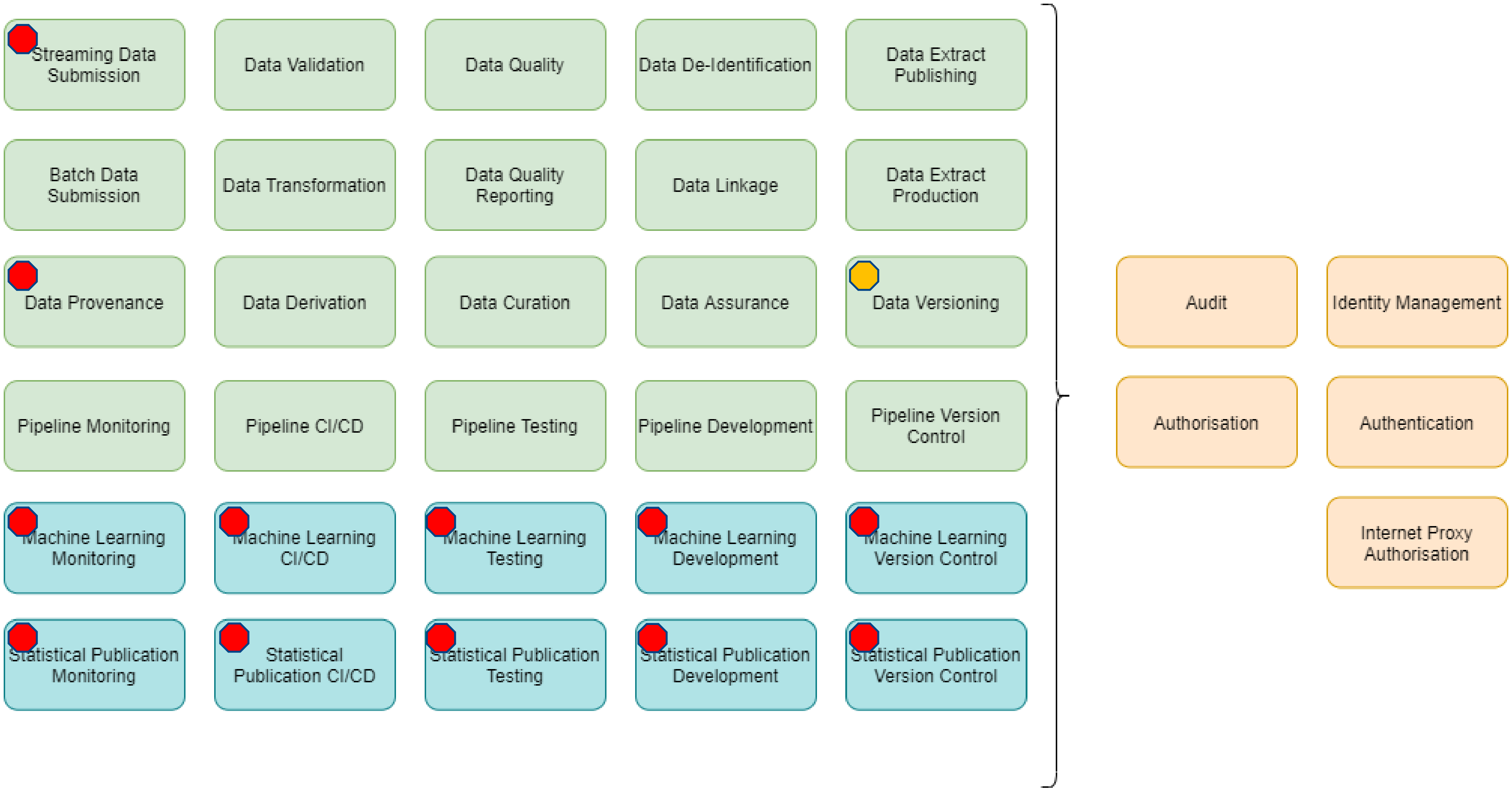
Platform

Data Persistence
Data Encryption
End-Point Encryption
Platform Monitoring
Platform CI/CD
Platform Testing
Platform Development
Platform Version Control
Platform Audit
Platform Identity Management
Platform Authorisation
Platform Authentication

Security

Network Security
Perimeter Security
Gateways
Security Information & Event Monitoring
Data Loss Prevention
Public Key Infrastructure
End-Point Protection
APT Mitigation

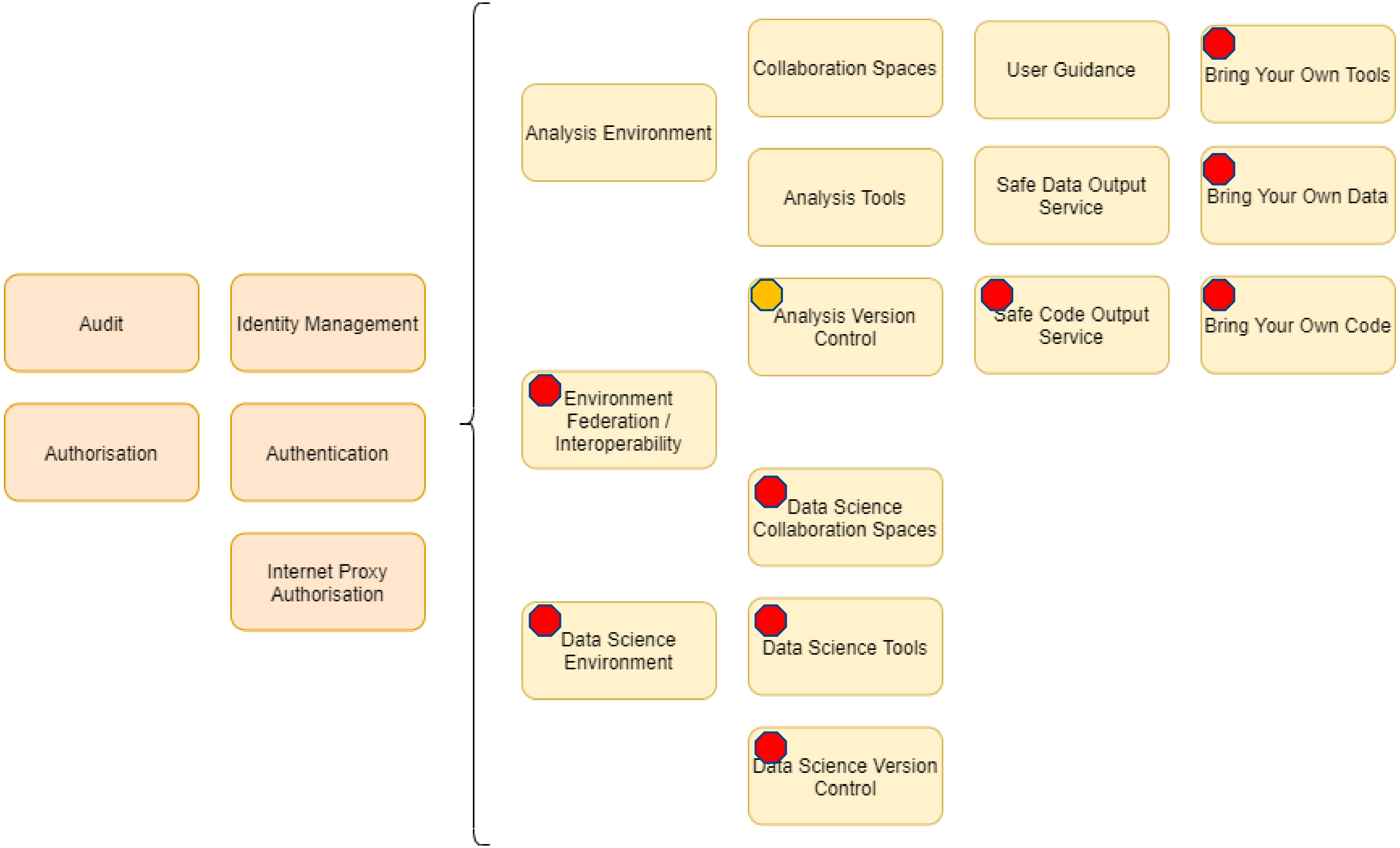
Platform Capabilities: Processing & Curation



🔴 Streaming Data Submission	Data Validation	Data Quality	Data De-Identification	Data Extract Publishing
Batch Data Submission	Data Transformation	Data Quality Reporting	Data Linkage	Data Extract Production
🔴 Data Provenance	Data Derivation	Data Curation	Data Assurance	🟡 Data Versioning
Pipeline Monitoring	Pipeline CI/CD	Pipeline Testing	Pipeline Development	Pipeline Version Control
🔴 Machine Learning Monitoring	🔴 Machine Learning CI/CD	🔴 Machine Learning Testing	🔴 Machine Learning Development	🔴 Machine Learning Version Control
🔴 Statistical Publication Monitoring	🔴 Statistical Publication CI/CD	🔴 Statistical Publication Testing	🔴 Statistical Publication Development	🔴 Statistical Publication Version Control

Audit	Identity Management
Authorisation	Authentication
	Internet Proxy Authorisation

Platform Capabilities: Data Access



Platform Forward View

Five Themes

Five Themes

DAE Usability

Adoption of a user-centric design for the Data Access Zone, providing modern, rich and performant analytical and data science environments.

Policy-Based Access Control

Replacing the current direct-grant access control with a highly-scalable policy-based approach will deliver substantive improvements to the data access processes.

Metadata-driven Models

The introduction of a robust, integrated metadata modelling layer, facilitating the exchange of business metadata with partner organisation and full-scale automation of the platform

Five Themes (cont)

Configuration-Based Ingestion

A configuration and orchestration approach to the ingest and pipelining of data allows for the rapid onboarding of novel data collections and reuse of standard pipeline components.

Data Manager Independence

Providing fully independent development environments and lifecycles for data managers and data scientists will enable organisation-wide adoption of the platform.

Five Themes: Impact on the TRE Service

Improving the analytical environment

We are in the process of developing a new analytical environment that will offer a vasty improved user experience, with more capabilities to deliver additional tooling choices. Further, we are expanding the collaborative features of the environment by adding version control that is shared amongst all of the members of a sharing agreement.

Automating the provision of data & access

We are investing in new tools that will automate the provision of data once a sharing agreement has been signed.

Rapid data pipelining and curation

New tools and methodologies are being adopted to radically increase the pace at which we can bring existing data collections onto the platform, as well as bring in new novel data collections.

Building a data science environment

We have begun gathering requirements for a data science environment, with the intention of enabling reliable, reproducible work to be undertaken safely.

NHS Digital

Questions?

Connect with us

@nhsdigital

company/nhs-digital

www.digital.nhs.uk

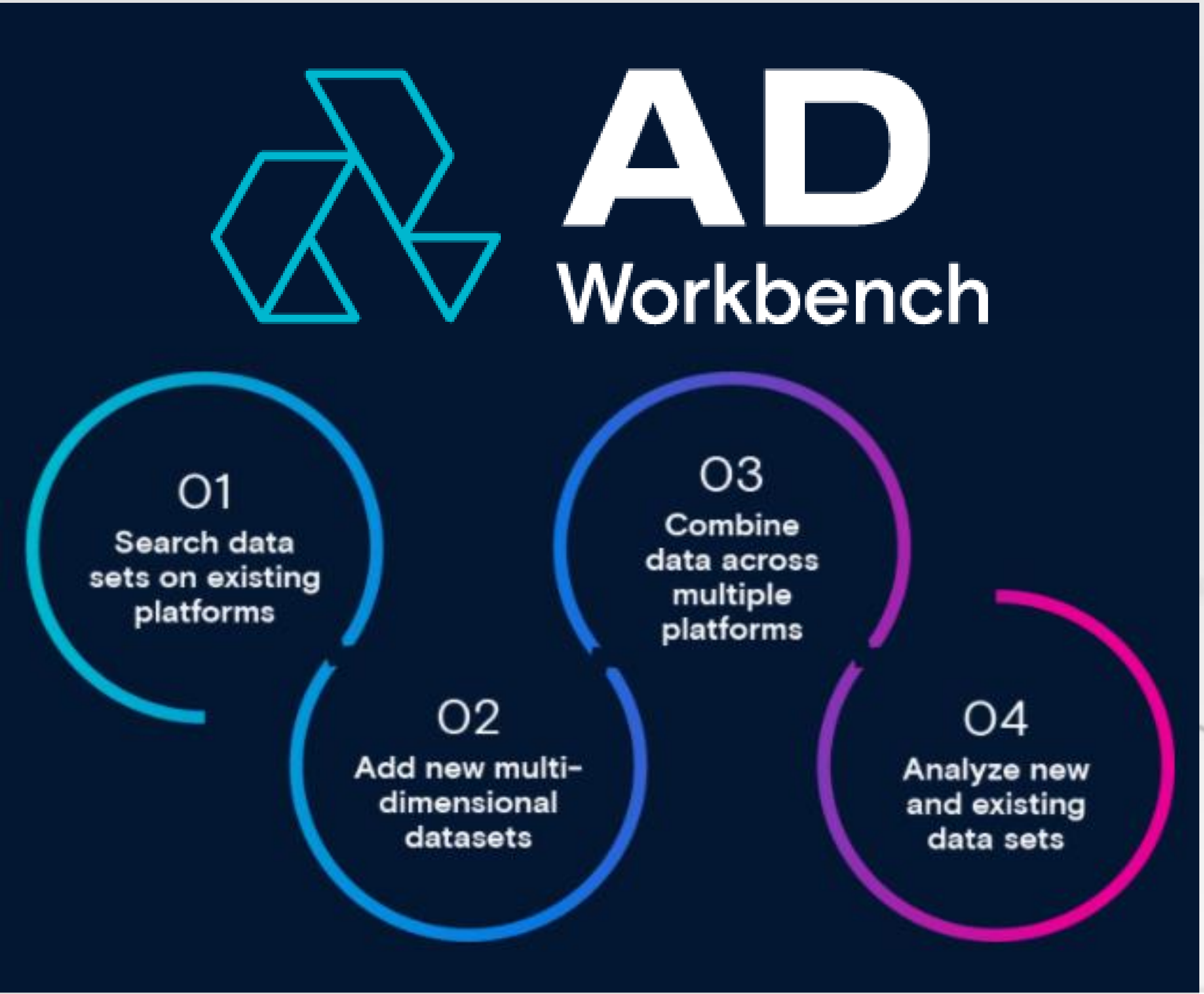
Information and technology
for better health and care

Creating an eco-system for research (of which TRE’s are one part)

- Alzheimers Disease Data Initiative (ADDI) – a newly formed independent MRO for AD
  - A global, networked approach to improving scientific discovery for AD
  - A two-year Pilot program now moving to global implementation
  - Philanthropic, public and private initiative at scale
- Solving for many to many interactions
  - Many data contributors (the Supply) and many researchers (the Demand)
  - How to discover data and how to work with data
  - Giving data contributors choice on data access approaches and reducing friction on data sharing
- Capacity building and working in the open
  - Open source programs
  - Freedom to operate for researchers
  - Serving the many researcher profiles that make up the community

aridhiaDRE
DIGITAL RESEARCH ENVIRONMENT

ADDI
Alzheimer's Disease
Data Initiative



AD Workbench

01 Search data sets on existing platforms

02 Add new multi-dimensional datasets

03 Combine data across multiple platforms

04 Analyze new and existing data sets

> "I'm optimistic that this will make a real difference in Alzheimer's research, because there are many examples where we've made progress on diseases after bringing together large amounts of data."
> **Bill Gates**

AD Workbench Video

Together, we can connect the dots and fight AD.

For Data Contributors – Federated Data Sharing

- Data sharing agreements are diverse and we need to reduce barriers for data sharing amongst data controllers
- **Our approach asks data controllers to self-select at what 'level' they can join the network**
- 3 levels of which federation is one (remember federation is not always the best choice, that's why choice is important!)
- Audit all use and transactions that run through the lifecycle of the project – transparency and productivity

- **For the Federated approach** –
  - Create an **open ecosystem** through a software development kit (SDK)
  - Partners can implement the API in a way that suits their local circumstances, join the ADDI network AND can re-purpose the implementation to join other networks
  - Researchers can implement their analysis plans or develop tools that suit their requirements
- Currently, the Federated API is an **open specification** shared by ADDI and other initiatives (HDR UK, ICODA)
  - https://github.com/federated-data-sharing/
- **Establish an open standard for data platforms to participate in open and closed data sharing networks**
  - Implement once, join multiple networks
  - Encourage convergence of existing (proprietary or niche) efforts
  - Encourage an ecosystem of tools & syndication
- **"Package" existing standards where they exist and fit**
  - OpenAPI
  - OAuth2 for API authentication
  - W3C DCAT standard for catalogues
  - GA4GH TES = Task Execution Service for compute tasks

For the researcher community – recognising the diversity of requirements

Different types of Users want to interact in different ways – there's a very **longtail** of use cases and skill sets

User with a Question

Some users have specific questions and very diverse skill sets

User with Code

Some bring in their code

User with Data

Some bring in new data

Standard

	Query	Visualise	Test	Model	Learn	Code Options
SQL console	✓					PostgreSQL
Data Table Analytics		✓	✓	✓		R code generation
R Console	✓	✓	✓	✓	✓	R
R-Shiny Apps	✓	✓	✓	✓	✓	R
Virtual Machine	✓	✓	✓	✓	✓	Multiple, optional GPU
Jupyter Notebooks - beta	✓	✓	✓	✓	✓	Python, R, Julia

Add on

	Query	Visualise	Test	Model	Learn
Containerised apps	✓	✓	✓	✓	✓
Azure ML Services					✓
Cromwell on Azure	✓				✓
Azure Cognitive services	✓				
3rd party Apps	✓	✓	✓	✓	✓
Cohort browsing	✓	✓			

How?

Query	Visualise	Test	Model	Learn
SQL Tidyverse, Pandas Bioinformatics pipelines, NLP Image analysis	Charts and reports, Interactive apps, Matplotlib, plotly Imaging tools Graph visualisation	CRAN packages R Studio Anaconda	CRAN packages, R Studio Anaconda	scikit-learn tensorflow keras pytorch Automated training

Working within a Trusted Research Environment (TRE) for access and analysis of population-wide patient-level healthcare data

My user experience in the NHS Digital TRE for England

Angela Wood
Reader in Health Data Science
Cardiovascular Epidemiology Unit
Department of Public Health and Primary Care
University of Cambridge

amw79@medschl.cam.ac.uk

British Heart Foundation
Data Science Centre

Led by Health Data Research UK

HDRUK
Health Data Research UK

Professor Cathie Sudlow,
Director of the
BHF Data Science Centre

Vision:

To improve the public's cardiovascular health using the power of large-scale data and advanced analytics across the UK

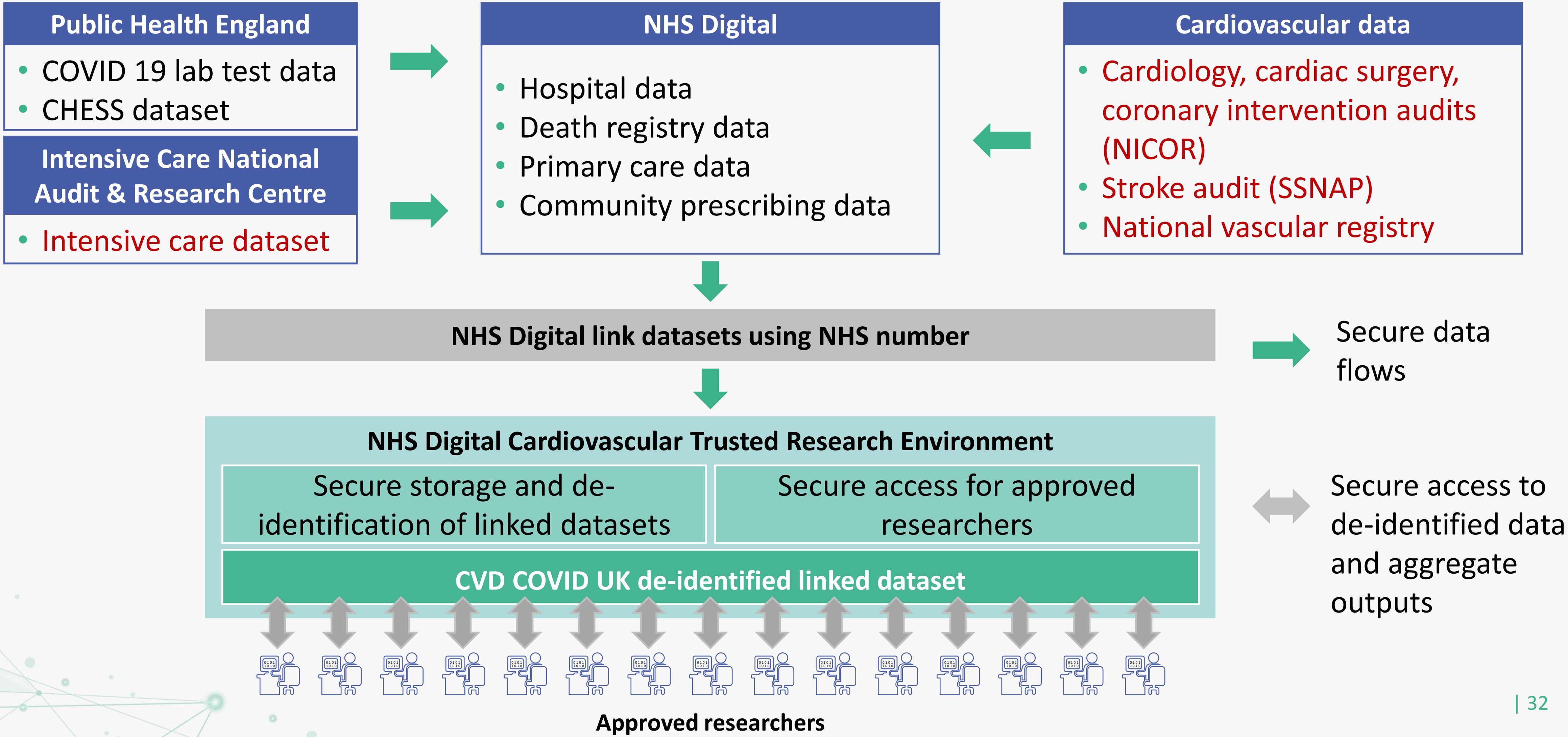
DRIVER PROJECT: The CVD-COVID-UK project

Aims to understand the relationship between COVID-19 and cardiovascular diseases such as heart attack, heart failure, stroke, and blood clots in the lungs through analyses of de-identified, linked, nationally collated healthcare datasets across the four nations of the UK

Infrastructure of the NHS Digital TRE

British Heart Foundation Data Science Centre
Led by Health Data Research UK

HDRUK Health Data Research UK



Public Health England

- COVID 19 lab test data
- CHESS dataset

Intensive Care National Audit & Research Centre

- Intensive care dataset

NHS Digital

- Hospital data
- Death registry data
- Primary care data
- Community prescribing data

Cardiovascular data

- Cardiology, cardiac surgery, coronary intervention audits (NICOR)
- Stroke audit (SSNAP)
- National vascular registry

NHS Digital link datasets using NHS number

NHS Digital Cardiovascular Trusted Research Environment

Secure storage and de-identification of linked datasets

Secure access for approved researchers

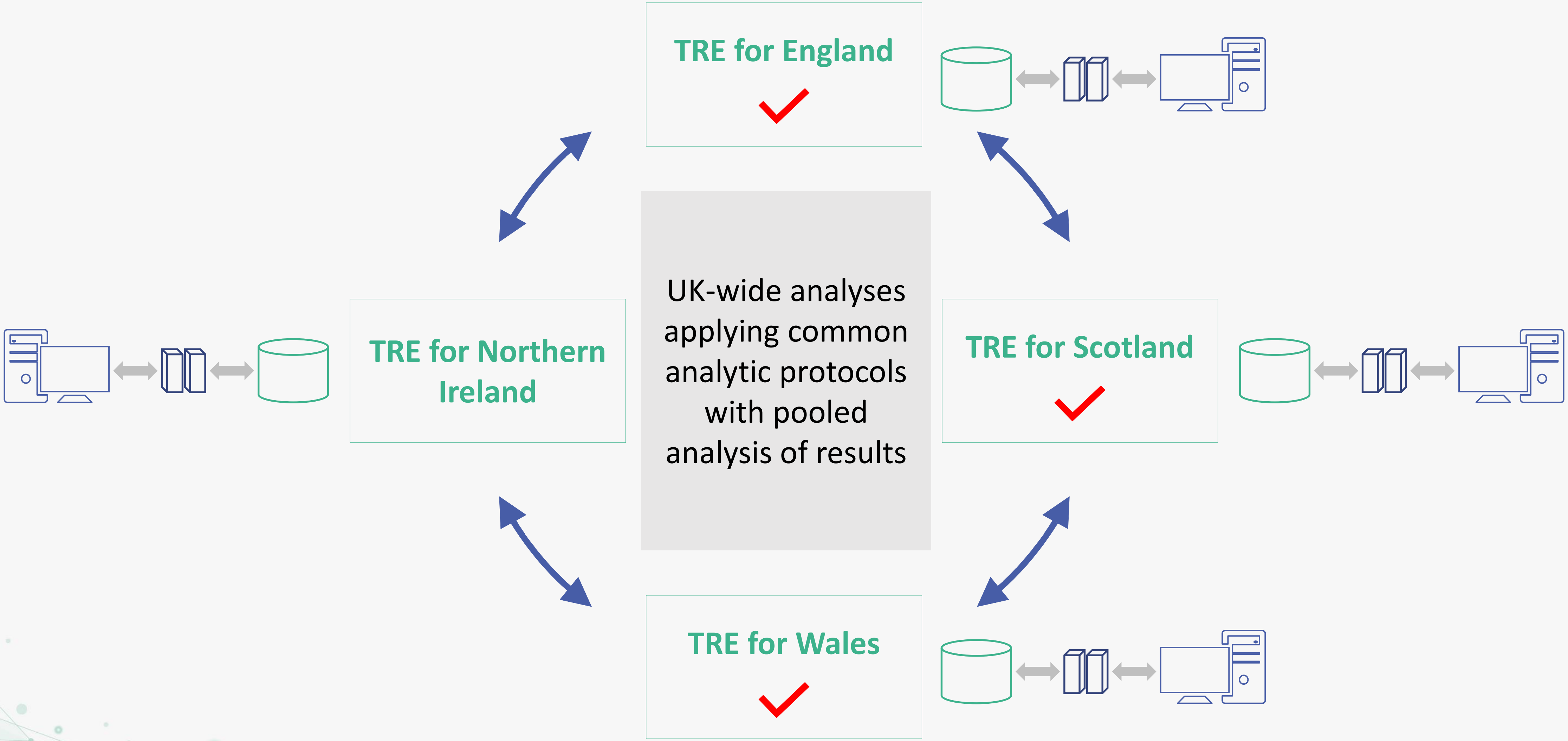
CVD COVID UK de-identified linked dataset

Approved researchers

Secure data flows

Secure access to de-identified data and aggregate outputs

CVD-COVID-UK: building UK-wide infrastructure



TRE for England ✓

TRE for Northern Ireland

UK-wide analyses applying common analytic protocols with pooled analysis of results

TRE for Scotland ✓

TRE for Wales ✓

My journey into the NHS Digital TRE for England

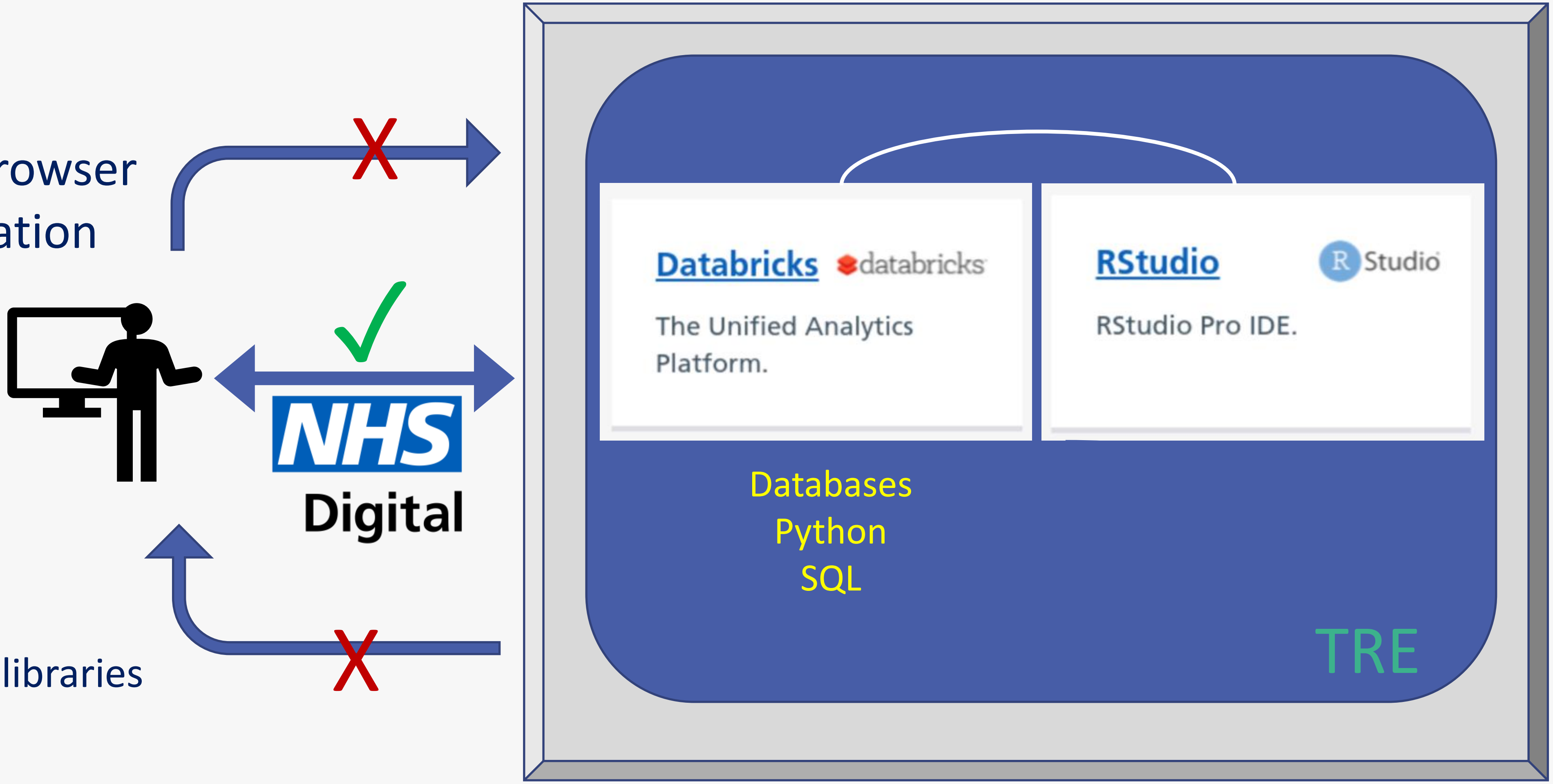
May 2020

Nov 2020

1. User input into design of platform
2. Structuring TRE collaboration folder spaces
3. Data quality checking, cleaning
4. Writing and refining analysis plans
5. Preparing datasets for planned analyses
6. Reference codelists to define variables of interest
7. "Approvals and Oversight board" with lay members
8. Descriptive analysis of available resources

Inside the TRE for England (my experience)

- Login in via web-browser
- 2-factor authentication



Databricks

The Unified Analytics Platform.

RStudio

RStudio Pro IDE.

Databases
Python
SQL

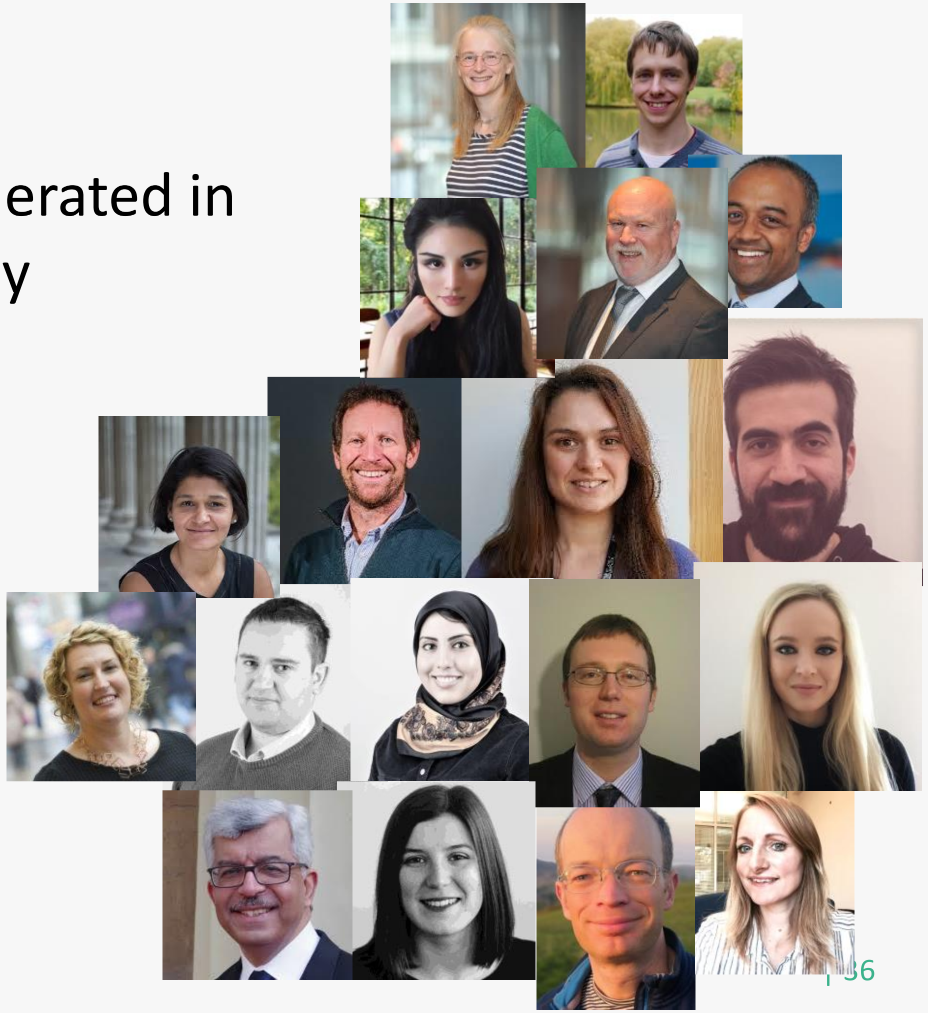
- Input / output

  software packages / libraries
  reference codelists
  analysis output

CVD-COVID-UK consortium – key ingredients

This ambitious project depends on:

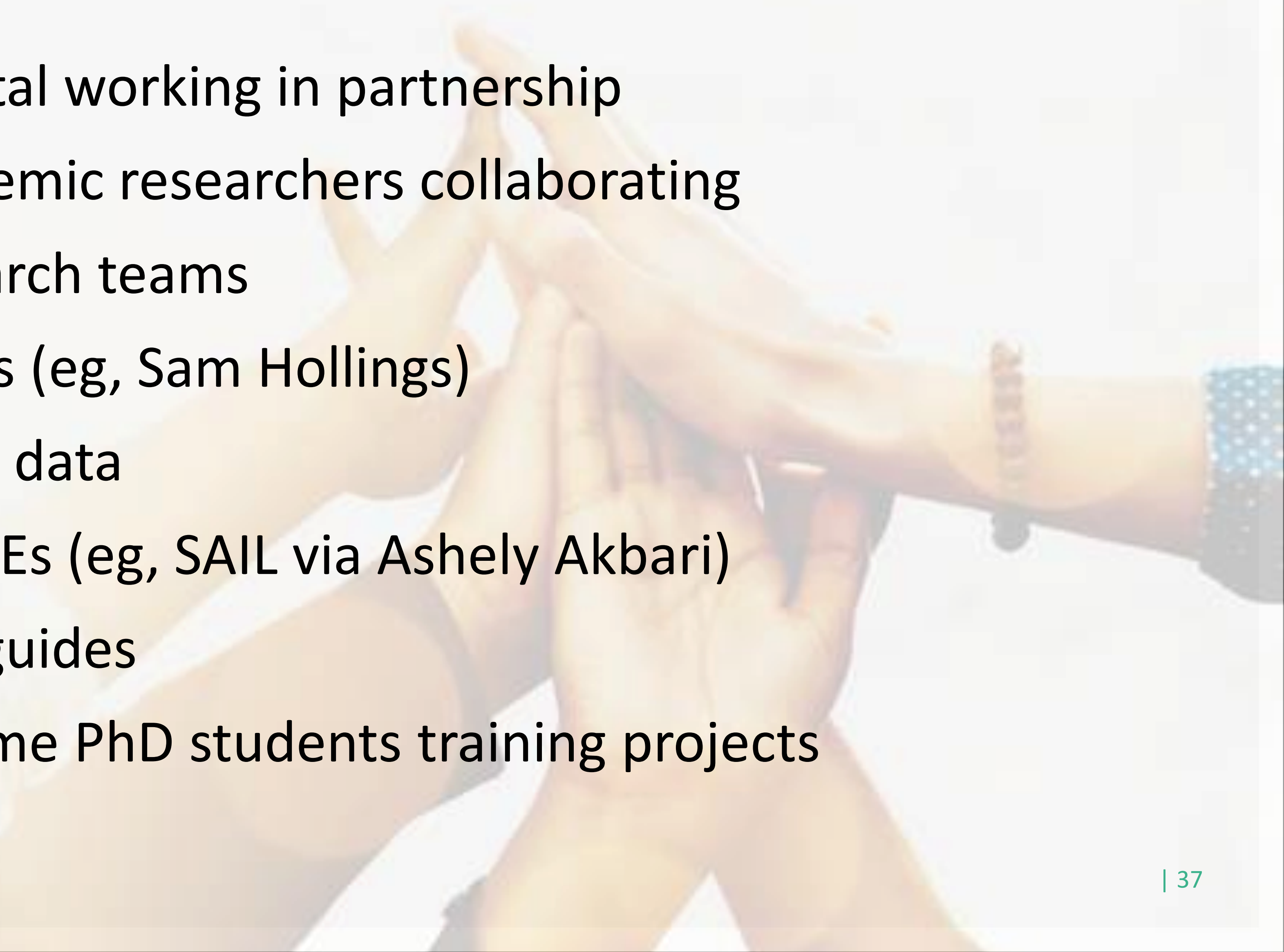
1. Availability of multiple different sources of data generated in NHS healthcare settings, brought together nationally
2. Researchers collaborating together
3. Data privacy, security and trustworthiness
4. Partnerships and transparency



An inclusive, open and transparent consortium

Committed to the 'Five Safes' (http://www.fivesafes.org/)

What is working well?



- ✓ Consortium + NHS Digital working in partnership
- ✓ Across institution academic researchers collaborating
- ✓ Multi-disciplinary research teams
- ✓ Dedicated data curators (eg, Sam Hollings)
- ✓ Hands-on access to the data
- ✓ Learning from other TREs (eg, SAIL via Ashely Akbari)
- ✓ NHS Digital manuals / guides
- ✓ HDR UK Turing/Wellcome PhD students training projects

Hold backs

- ✗ Delays in datasets availability and linkability
- ✗ Lack of user control (inputs/outputs)
- ✗ Users really need SQL / Python skills

Technology “must haves”

- ✗ Github / Gitlab for version control and collaborative work
- ✗ Need more memory on link between Databricks and RStudio
- ✗ GPUs / acceleration chips
- ✗ Known computing capacity

MUST HAVE!

To be determined…

- ? Success of monthly data updates & version controls
- ? Access to TREs for Wales, Scotland and N. Ireland
- ? Scaling up number of analysts / HPC requirements

Future: Increasing productivity

- Dedicated funded researchers
- Extension to support health-related research more broadly beyond COVID-19 & Cardiovascular
- Extending functionality so accessible to broader skillsets
- Establishing overarching analysis approaches across four national TREs

All of us together

UK Health Data Research Alliance Symposium

Tuesday 1 December 2020
09:30 - 17:00

Coming up next...

11:35-11:45 Break

11:45-12:35 Innovation Gateway Demo, Drop-in Session and Lightning Talks

Find the links to join in your brochure or at hdruk.ac.uk/symposium-links-20

HDRUK Health Data Research UK

UK Health Data Research Alliance

Health Data Research Innovation Gateway

→ Join the conversation:
hdruk.ac.uk/HDRAlliance20
@HDR_UK
#HDRAlliance20