

USE CASE

Building a modern infrastructure for collaborative healthcare data science, machine learning and artificial intelligence



VISHNU CHANDRABALAN

Consultant Surgeon and Head of
Data Science and AI

Lancashire Teaching Hospitals

NHS Foundation Trust



Building a modern infrastructure for collaborative healthcare data science

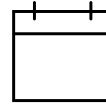
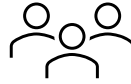
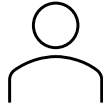
Vishnu V Chandrabalan MD FRCS

Consultant Surgeon
Head of Data Science

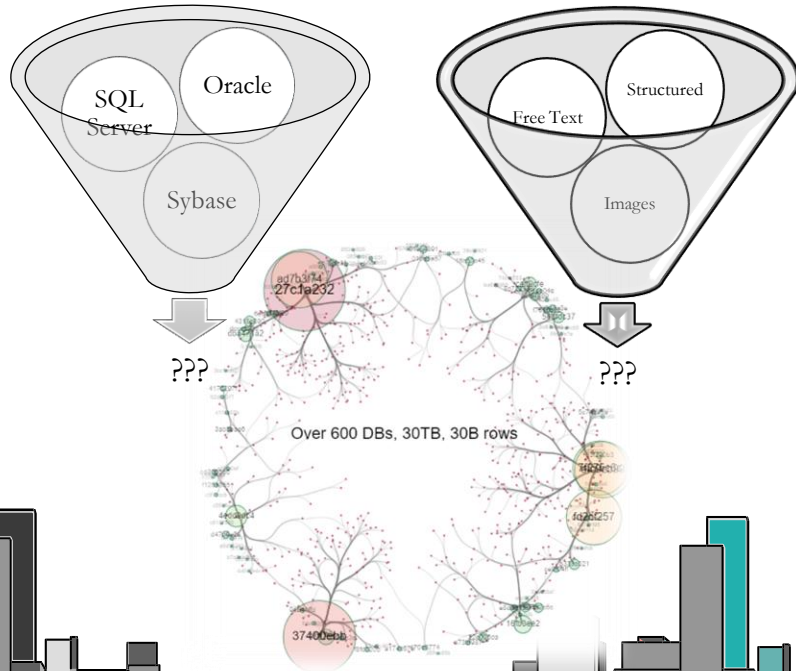


Lancashire Teaching
Hospitals
NHS Foundation Trust

ABOUT



Current Data Landscape



- EPR 15 years, nearly paperless
- Multiple technologies and data formats – nearly all on-prem
- Schema knowledge patchy and locked within a busy but excellent BI team
- All ETL/ELT is focussed on BI requirements
- Massive manual repetition, little documentation, no version control
- No “research-ready” datasets or dataflows

Information Silos

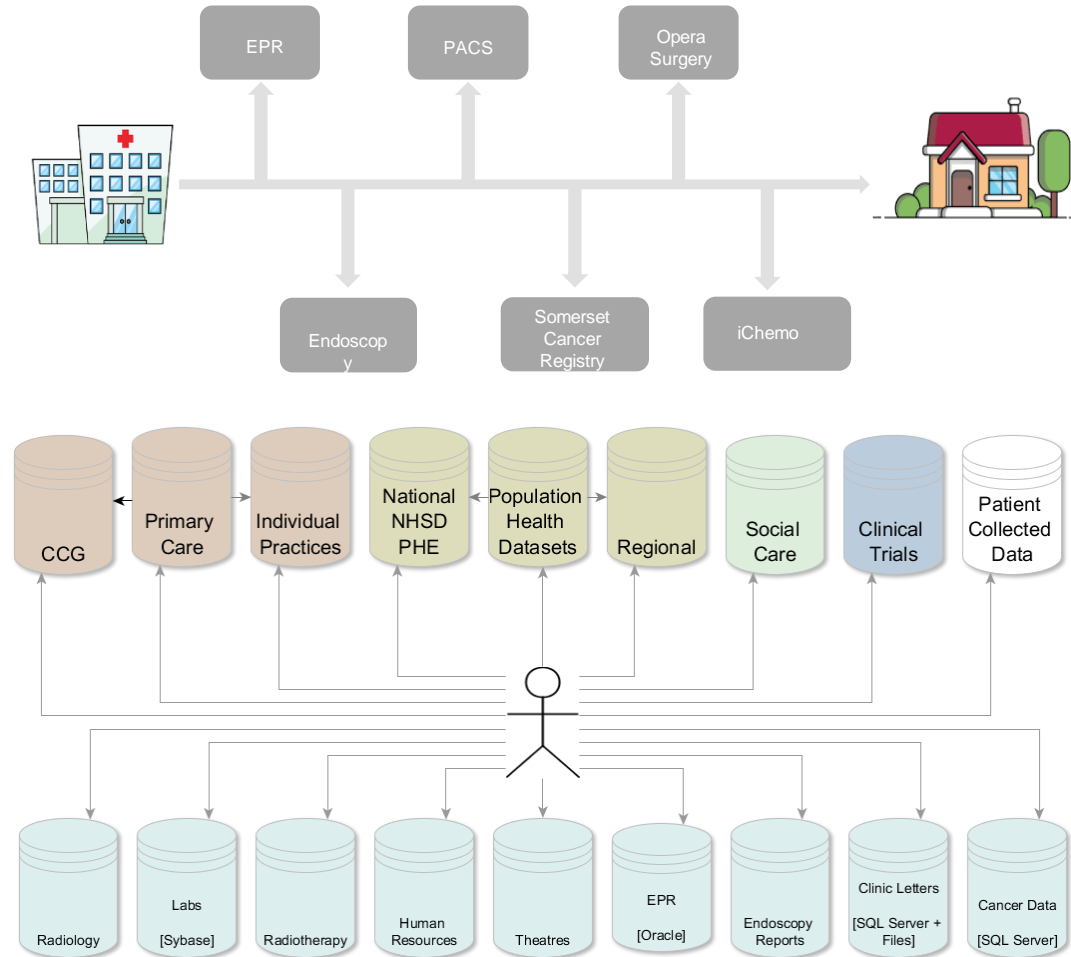
Imagine a patient ...

Referred from primary care

On the urgent 2-week pathway

With suspected bowel cancer

Now imagine ... asking any useful questions of this entire patient journey!



Common Data Models and Data Catalogues

OMOP - Common Data Model

Systematic analysis of disparate observational databases

Transform data into a common data model (CDM)

Person-centric rather than application-centric

Athena - Standardised Vocabulary

Data is mapped to SNOMED, ICD10, LOINC or other ontology of choice making it interoperable with other similarly mapped data

OHDSI Analytics Tools

ATLAS, HADES, ACHILLES

Open-source R packages for large scale analytics

Federated analytics without need to share data

OHDSI International Community

Global Collaborators

EHDEN.EU

22+ UK OHDSI Data partners

Funding

We are funded by:

- EHDEN – HDRUK 7th Data Partner Call
- NIHR Clinical Research Network

Data Sources

EPR Data Warehouse (Visits, Labs, Vitals, Drugs, ...)

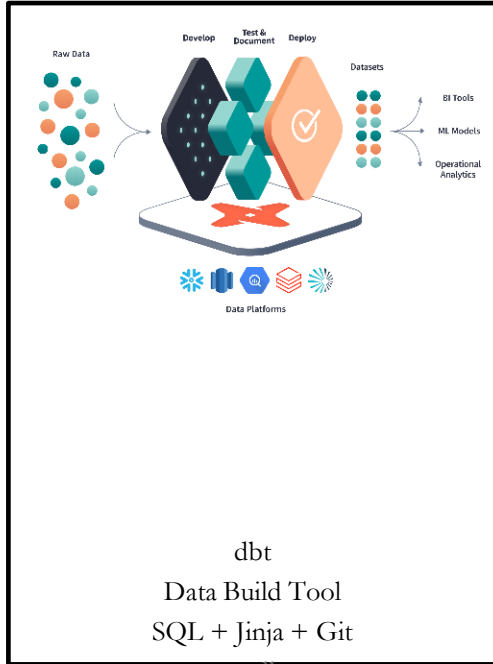
Somerset Cancer Registry, iChemo, SwissLab

Will become core component of NW SDE

Feeds TriNetX Platform

The screenshot shows the Athena website interface. At the top, there is a navigation bar with the Athena logo, a search bar, and links for 'SEARCH', 'DOWNLOAD', and 'LOGIN'. Below the navigation bar, there is a 'Search' section with a search input field containing 'aspirin' and a 'Search' button. A small text box below the search bar provides instructions: '1. Usage of quotation marks forces an exact-match search' and '2. In case of a typo, or if there is a similar spelling of the word, the most similar result will be presented'. The 'Explore domains' section features six tiles: 'Drugs' (5,276,001), 'Conditions' (76,826), 'Procedures' (726,000), 'Devices' (461,001), 'Observations' (267,246), and 'Measurements' (266,041). Below this, there are several 'Method' cards: 'Cohort Method' (New-user cohort studies using large-scale regression for propensity and outcome models), 'Self-Controlled Case Series' (Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality), 'Self-Controlled Cohort' (A self-controlled cohort design, where time preceding exposure is used as control), 'Patient Level Prediction' (Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms), 'Case-control' (Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort), 'Case-crossover' (Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control)), 'Empirical Calibration' (Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design), 'Method Evaluation' (Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods), and 'Evidence Synthesis' (Combining study diagnostics and results across multiple sites). At the bottom, there is a 'Map of Collaborators' section with a world map showing the locations of OHDSI collaborators. A text box next to the map states: 'The OHDSI community brings together data from across the world to make better use of the data already being generated in health care. We are now adding new partners and generating reports of our activities.' Below the map, there is a 'OHDSI By The Numbers' section with the following statistics: 2,367 collaborators, 74 countries, 21 time zones, 8 continents, and 1 community.

OMOP Extract Load Transform



dbt

Search for models...

Overview

Project Database

OHDSI/OMOP Data Harmonisation project

Data Science Team, Lancashire Teaching Hospitals NHS Foundation Trust

Introduction

Lancashire Teaching Hospitals NHS Foundation Trust (LTH) is a digitally mature secondary care provider, major trauma centre and multi-specialty tertiary referral centre in Lancashire and South Cumbria (LSC). LTH is developing a cloud-native, secure, data science platform on Microsoft Azure that has proven measurable by enabling data scientists from regional, national, and international organisations to undertake advanced analytics without transferring data out. This led to LTH being a partner in a successful bid for £11 million to build a north-west Secure Data Environment (SDE).

OHDSI/OMOP

LTH have access to routinely collected healthcare data for over 2.25 million patients spanning 15 years, covering most aspects of secondary care. This data is stored in multiple disparate databases.

We have invested in a multi-year, large-scale data harmonisation program with the [Observational Medical Outcomes Partnership \(OMOP\) Common Data Model \(CDM\)](#) as the target model. We have secured additional external funding from OHDSI/OMOP further validating our strategy.

OMOP is supported by the [Observational Health Data Science and Informatics \(OHDSI\)](#) program, a multi-stakeholder, global collaborative that aims to deliver value out of health data through large-scale analytics, harmonising to OMOP makes our data immediately valuable using standardised, open-source, analytics software maintained by a global community of researchers.

LTH is a member of the [HDSUK Alliance](#) and will also become member of the global OHDSI federation collaborating on international research studies - both observational as well as clinical trials.

Benefits

Federated observational studies

Aggregated analysis across multiple organisations can be made without sharing any patient-level data or requiring complex data sharing agreements allowing rapid translational data science.

Clinical Trials

Cohort definitions for UK/International clinical trials prepared by any lead site using OMOP can be executed on our database to rapidly establish study feasibility and identify eligible patients, creating opportunities for a range of portfolio studies and build links with international academic and clinical collaborators. This is especially important for Lancashire and South Cumbria where participation in clinical trials has been historically low. It will also allow us to strategically target patient groups for all real-world clinical trials that require mapping with requirements.

(LTH are working with OHDSI, a global healthcare data platform, to enhance our capabilities for clinical trial feasibility assessments, cohort discovery and evidence generation using real-world data. The OMOP/OMOP data mapping will be a critical enabler for further automated data transformation into SDE).

Data Lineage, Self-Documenting, Metadata Generation

Diverse source and target architectures

Shallow Learning Curve

stash On vc/main: !!GitHub_Desktop<1

index on vc/main: ac50207 Addition of r

qa/main -> remotes/origin/qa/main

- Change of workflow
- Created of vocab layer to condition occi
- Added in cons code to flex provider id
- Change of colours in lineage graph
- remotes/origin/th/main Correcting da
- Creation of new drug era tables
- removal of duplicate field name
- Addition of new seed file
- Merge branch 'qa/main' of https://github
- Update README.md
- Addition of drug route mappings
- Adding drugs seed to yml file
- formatting
- New workflow for DRUG_EXPOSUR
- Workarounds for drug exposure
- Changes to include visit merges
- PROCEDURE_OCCURRENCE
- Removal of old proc_occurrence que
- Addition of full seed files

Git + GitHub

Version Control

CI/CD

LANDER Architecture Overview

What is a TRE?

A TRE is a **Trusted Research Environment**. Also known as 'Data Safe Havens', TREs are highly secure computing environments that provide remote access to health data for approved researchers to use in research that can save and improve lives.

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 approved researchers with a **single location** to access valuable datasets. The data and analytical tools are all in **one place**, a bit like a **secure reference library**.

Learn more about TREs and discover examples of how TREs are being used to enable life-saving health research.

How is my data safeguarded?

Health data should always be kept safe and secure, and used responsibly to ensure privacy. Health Data Research UK ensures these high standards are met by promoting the use of the 'Five Safes' model across all TREs.

- 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-identified to protect privacy
- Safe Outputs**
All research outputs are checked to ensure they cannot be used to identify subjects

Analytics Infrastructure

A 3-year story: April 2020 – March 2023

April 2020

On-prem only infrastructure
BI tools only – SSMS, Qlik
No data science / collaboration tools
USB drives, Emails, Dropbox(!)
Excel, MS Access, 1 Python user

Oct 2020 – Sept 2022

Microsoft Azure Landing Zone
Private VNETs ↔ On-prem over ExpressRoute
Auto-scaling Kubernetes-based TRE
Docker, Helm, Git, Python, R, Julia

Regional, national and international collaborations
Stronger ties with regional universities
Frugal Innovation

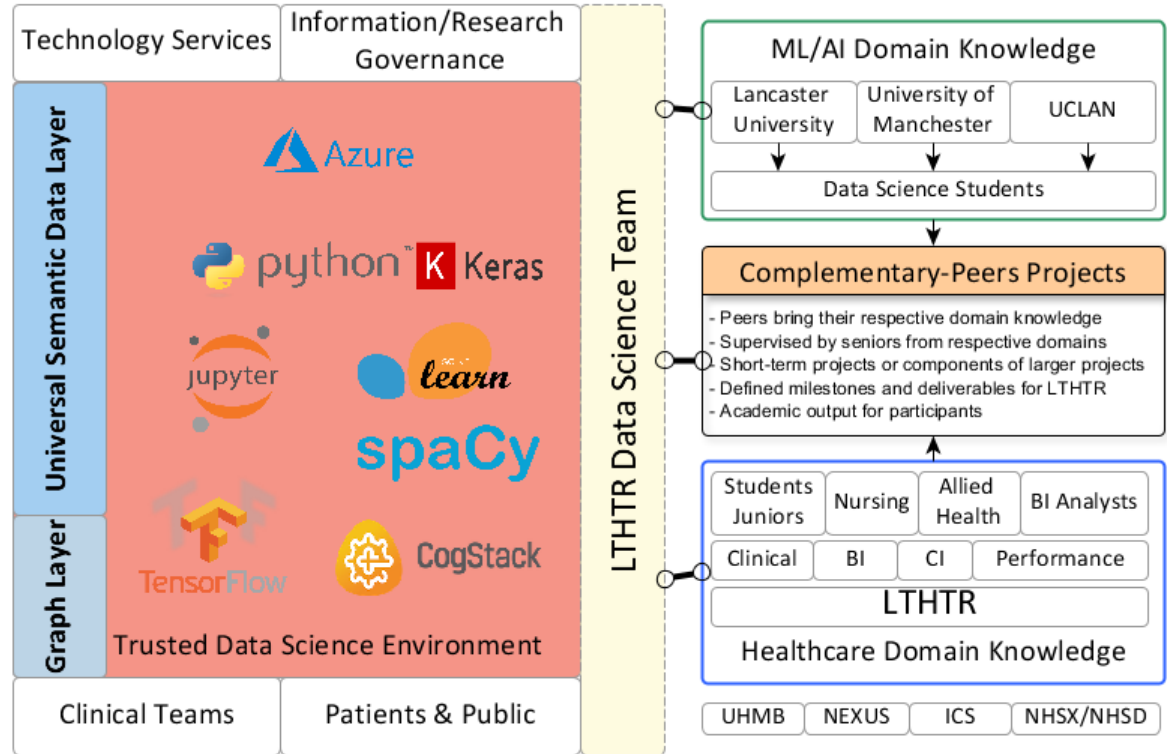
Sept 2022 - March 2023

Part of NW SDE
Building LSC ICB Azure Infrastructure
Data Lake + OMOP + TRE
NLP and Computer Vision on TRE
2 Data Scientists, 2 PHM fellows
Research Software Engineers
GitHub Enterprise

Lancashire Data
Environment for Research

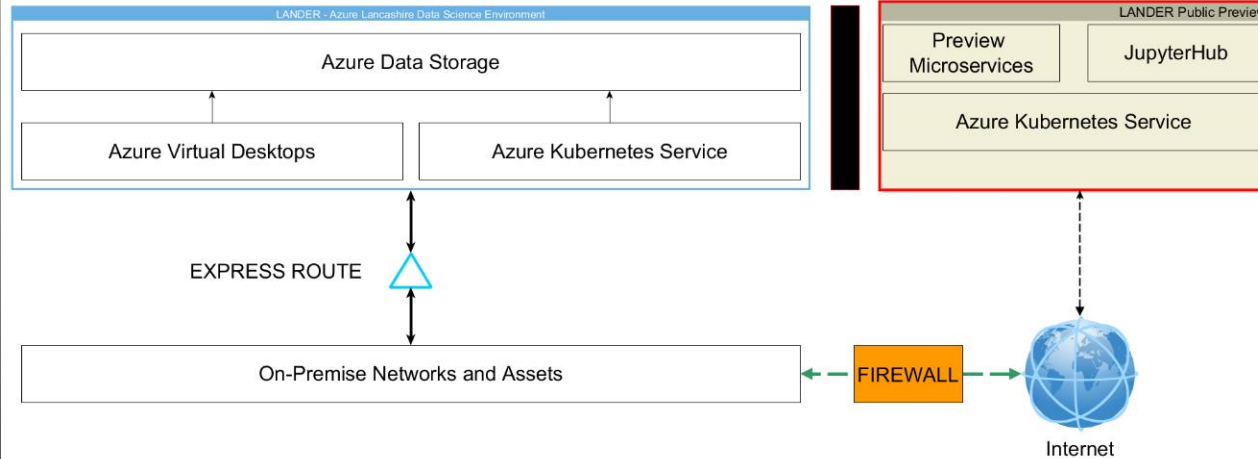
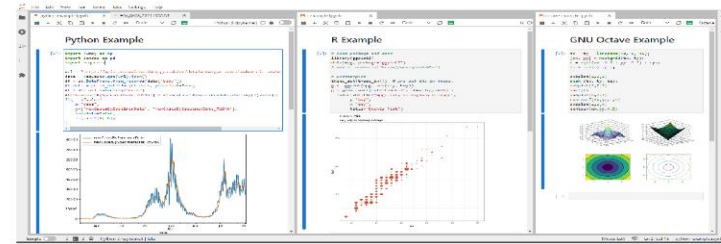
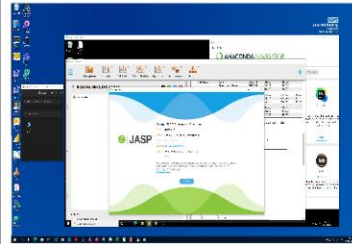
LANDER Architecture Overview

Slide from first LTH data
science meeting
– circa May 2020



Lancashire Data Environment for Research

LANDER Architecture Overview



LANDER Users and Projects

Users

Lancaster University
University of Manchester
Manchester Metropolitan University
University of Central Lancashire
Imperial College London
NHSX Analytics Unit
Beamtree Australia

Internal Users – CIO (!), Data
Scientists, BI, R&D

DataScience@clhr.lanc.ac.uk for further information.' At the bottom, it says 'Lancashire STAR2020 Platform'." data-bbox="262 226 723 680"/>

Welcome to Lancashire Data Science Environment

LANDER

Collaborative, auto-scaling, hyper-secure, cloud-based computing cluster for advanced health care data analytics, machine learning and artificial intelligence.

Email DataScience@clhr.lanc.ac.uk for further information.

Lancashire STAR2020 Platform

Core architecture based around JupyterHub, Kubernetes, Docker
Familiar interfaces: JupyterLab, VS Code, Desktop and soon Matlab
Choice of languages: Python, R, Julia, Octave
Auto-scaling, workspace-specific compute including GPU instances

Selected Projects

Structured data

Neurology Informatics
Risk Stratified Clinical Harm Reviews
Demand Forecasting

Free Text

NLP & NER+L using MedCAT
Patient feedback sentiment analysis

Computer Vision

Diabetic Foot Ulcer Images
MRI white matter hyperintensities (in the pipeline)

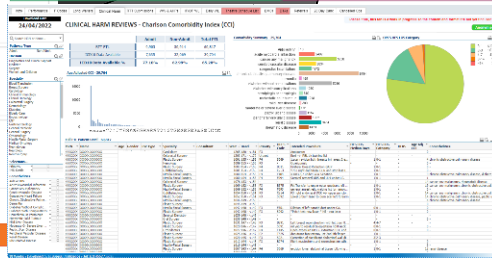
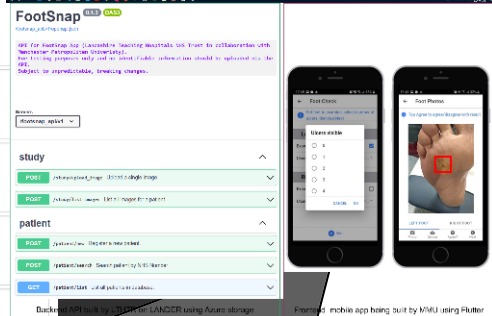
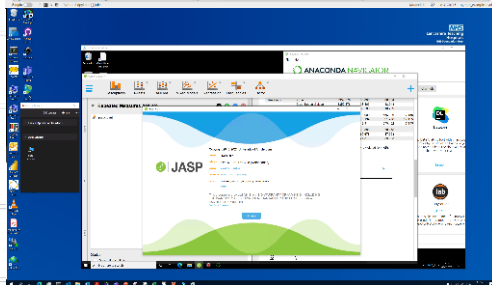
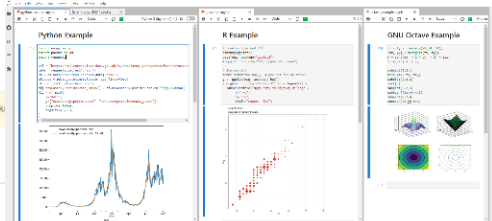
LANDER - Private Cluster

Lancashire Data Science Environment

Preview only. Subject to rapid, breaking changes. Contact Vishnu Chandrabalan@lth-rhls.uk with feedback

Analytics Workspaces

- Advanced Generic Workspace**
Advanced environment for testing with Python R and Julia.
Image: `datascience-notebook-2023-01-10`
Your access expires in : 2770 days
- VSCode and Desktop Server Workspace**
JupyterLab with additional options of Code Server and XCFE Desktop
Image: `datascience-code-server/latest`
Your access expires in : 2770 days
- Beamtree Ainsoff Deterioration Index Analysis**
RStudio environment for Service Evaluation of ADI for identifying deteriorating patients by Beamtree Australia. All work must be stored in `~/home/joyan/beamtree_adi`. Any work saved outside of this folder or its children will be lost between sessions.
Image: `beamtree-rstudio-0.2.0`
Your access expires in : 0 days.
- Colorectal ZWW Analytics**
Colorectal ZWW pathway analysis
Image: `datascience-notebook-2023-01-10`
Your access expires in : 2770 days.
- FFT Sentiment Analysis**
FFT 1 Sentiment Analysis NLP collaboration with Imperial College London
Image: `fft-notebook-0.1.0`
Your access expires in : 2770 days.
- GPU Workspace**
GPU enabled workspace
Image: `gpu-jupyter-v1.4_cuda-11.2_ubuntu-20.04`
Your access expires in : 953 days.
- LTH Data Science Team Workspace**
Shared workspace for LTH DST
Image: `datascience-notebook-2023-01-10`
Your access expires in : 953 days.
- Neurology Informatics Workspace**
Common workspace for all neurology informatics workloads including NLP
Image: `scipy-code-server-py-38-2023-02-23`
Your access expires in : 191 days.
- NHSX NLP Workspace**
JupyterLab, Code-Server, XCFE with MedCAT and related tools.
Image: `scipy-code-server-2023-01-26`
Your access expires in : 2770 days.
- NHSX NLP Workspace (Python 3.8)**
JupyterLab, Code-Server, XCFE with MedCAT and related tools
Image: `scipy-code-server-py-38-2023-02-23`
Your access expires in : 227 days.
- VC-SD Workspace**
Shared workspace for VC and SD
Image: `datascience-notebook-2023-01-10`
Your access expires in : 953 days



LANDER Capabilities

Secure Remote Collaboration

Data stays within LTH; All existing security and network policies apply; No direct access to TRE from public internet; Robust processes for on-boarding researchers

Cloud-native and Hybrid

Kubernetes, microservices architecture for deploying multiple diverse workloads

COST SAVINGS

Frugal; Pay for usage; No custom-configured laptops, USB keys, etc.

EASY TO USE

Familiar interfaces and most common data science tooling cater to >90% of expected use cases.

Digital Workforce Strategy

Infrastructure

No dedicated support (yet)
Microsoft and certified partners
Azure cloud engineers (0.2 FTE x 2 since Dec 2022)
Research Software Engineers (2023 -)

Data Science

OMOP Analytics Engineer (NIHR CRN funded)

0.6 FTE Data Scientist

Data science student placements

- 2021: Summer 3, Autumn 12
- 2022: Summer 4
- 2023: Summer 2

Population Health

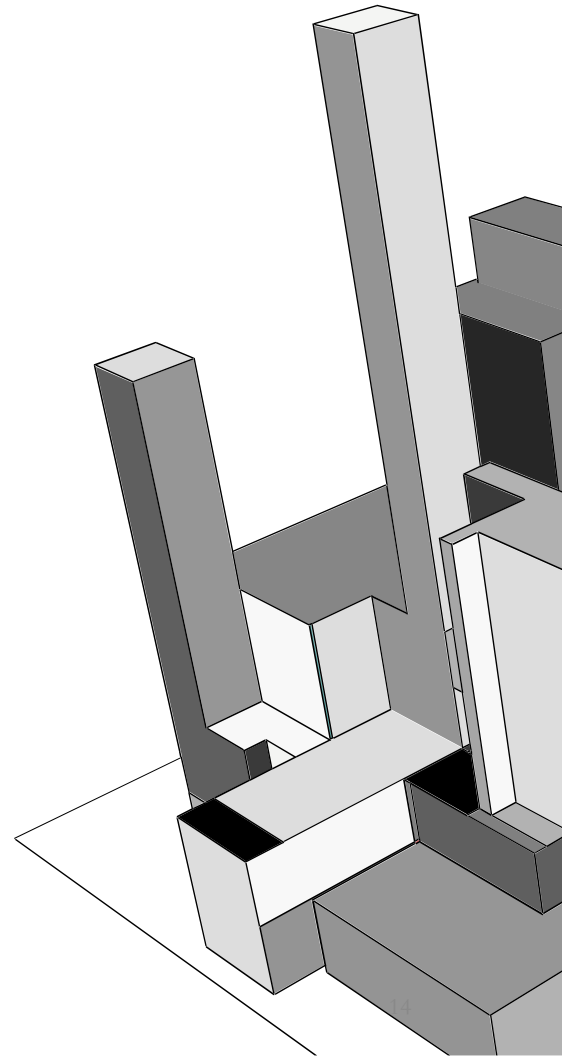
2 x HEE funded PHM fellows

- Establish PPIE for LSC
- Diabetic Retinopathy Screening Inequalities

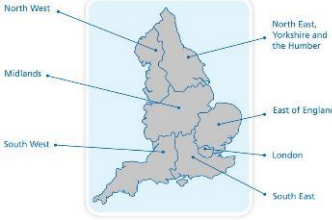
Career Development

MSc in Healthcare data science
New Skills: Python, R, dbt, git + GitHub,
Docker, NLP, Solr, OMOP

RCEM pre-doctoral fellowship
? NIHR Research Scholar Program




What Next



NW SN SDE
Northwest Sub-National
Secure Data Environment



Architect cloud-based
integrated, intelligence
architecture combining data
lake and TRE



Strong PPIE
Widen and strengthen
collaborations
Diverse, Linked Data Sets

“Stuff” not covered today

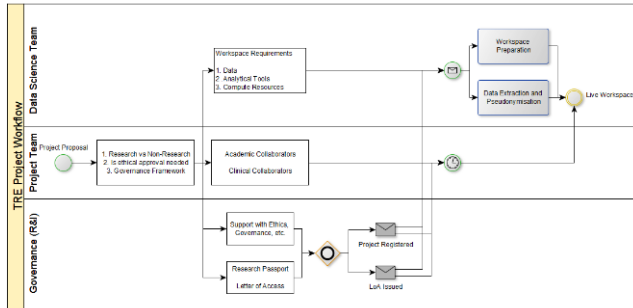
Kubernetes and miscellany

- KEDA and autoscaling microservices
- Complex networking, proxies, CIDR, SSL, etc.
- Identity management, AVD, etc.
- Terraform, API deployments, CI/CD
- And a whole pile of other cool-stuff

AzureTRE, data lakes, etc.

AzureTRE test deployment in separate tenant
LSC ICB data lake – Synapse vs Snowflake
OpenSafely

Governance and 5-Safes

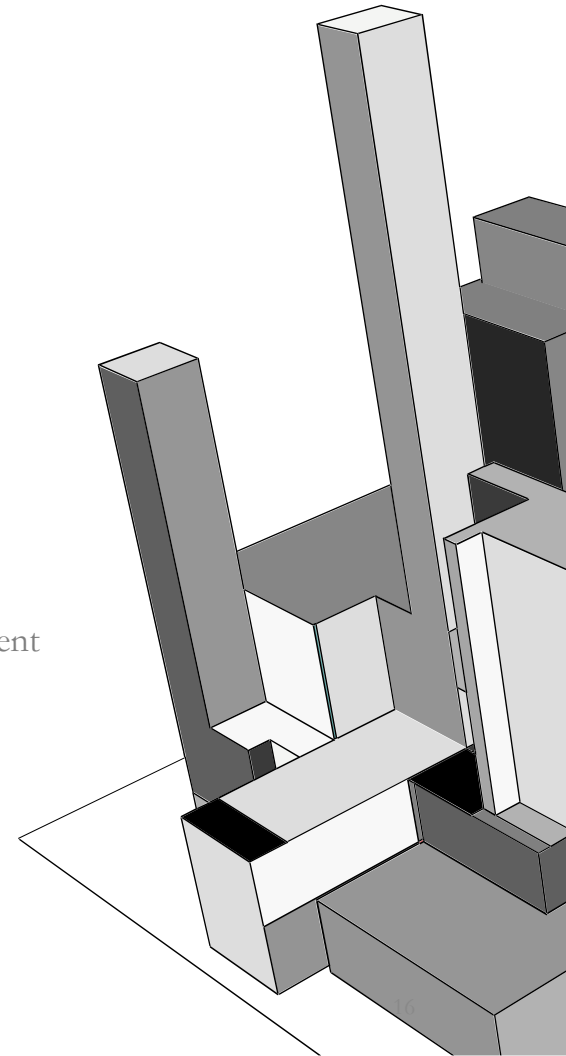


NW SDE and Workforce Development

LSC + GM + C&M

Challenges around equity and levelling-up
Role of in-house workforce vs third-party development

Research Software Engineering Hub – RSE
Society membership



The Team

- **Stephen Dobson**, Chief Information Officer
- **Saeed Ummar**, Head of Technical Services, **Paul Woodhouse**, Senior Technical Specialist
- **Tim Howcroft**, **Quinta Ashcroft** – Data Scientists
- **Paul Brown**, **Kina Bennett** – Research & Innovation
- **Dale Kirkwood**, HEE PHM Fellow and PPIE Lead
- Academic partners – Lancaster University, UCLAN, UoM, MMU
- Regional NHS partners and Clinical Colleagues
- Commercial partners – Microsoft and certified partners

INTELLIGENT HEALTH UK 2023

Breaking down the barriers
between tech and healthcare