

Connecting castles securely with safe share

Simon Thompson

Director of IT for Informatics

@ Swansea University Medical School

How are we ?

Swansea University medical School – 3rd Best in UK

Somewhere called Oxford / Cambridge are better apparently ©

<http://www.walesonline.co.uk/news/health/swansea-universitys-medical-school-uks-12986635>



Wales is the best part of the united kingdom sitting proudly on the west side of the country.

Devolved government in which Health is devolved

Population 3.5m humans and 10m sheep (with excellent healthcare)

SAIL Databank (<https://saildatabank.com/>)

- Over 9 billion records for >5 million people
- Much data goes back 10-20 years
- All pre-linked data
- 300+ feeder systems across Wales
- >£5 million investment in high performance IT
- Industrial strength, reusable infrastructure.
- >300 users,
- >£90m projects from UKRCs
- 140+ approved SAIL projects, with 79 active today
- 100 staff in Swansea working on Health Informatics-related projects

SAIL Databank Reach



OVER
370
USERS

78
ORGANISATIONS
WORLD-WIDE






- MRC-funded institute of health informatics
 - £9.3 million investment at Swansea
 - Four Centres across the UK
 - Centre for Improvement in Population Health through E-records Research (**CIPHER**)
 - (Swansea, Cardiff, Bristol, Uni of W Australia, Curtin, Ottawa)
 - Focus on large scale studies
- **Aim:** provide the physical and electronic infrastructure to facilitate collaboration across the four nodes
 - **UK Secure e-Research Platform (UKSeRP)**
 - **National Research Data Appliances (NRDA)**
 - New methods, public engagement, innovative governance, capacity building

- ADRN is a UK-wide partnership:
 - Universities
 - Government
 - National statistics authorities
 - Third sector
 - Funders
 - Researchers
 - ADRC-W one of four centres
 - £8m investment from ESRC
 - Part of the focus on governmental data sharing
 - Information assurance and privacy protection
- 
- Secure environment for research
 - Using SAIL infrastructure in Wales, with UK SeRP
 - Aimed at UK social researchers
 - Help accredited researchers carry out social and economic research
 - Help to using linked, de-identified administrative data – information which is routinely collected by government organisations.

The logo for the Medical Research Council (MRC), consisting of the letters 'MRC' in a white, sans-serif font on a dark brown rectangular background.

Cloud Infrastructure
for Microbial
Bioinformatics

- 
- A large, blue, downward-pointing arrow that spans the width of the slide, pointing from the top towards the two columns of text.
- £4m investment from MRC
 - Swansea, Cardiff, Birmingham, Warwick Universities
 - Large in CPU or Memory servers as host servers
 - OpenStack VM Stack
 - Capacity >1000 virtual research servers
 - Compute cloud for academics
 - 2,880 CPU cores
 - 4PB storage (2.8PB usable)

UK Secure e-Research Platform (UKSeRP)

Funded by: MRC & ESRC

AIM: to provide best in class informatics research platform for national research programmes

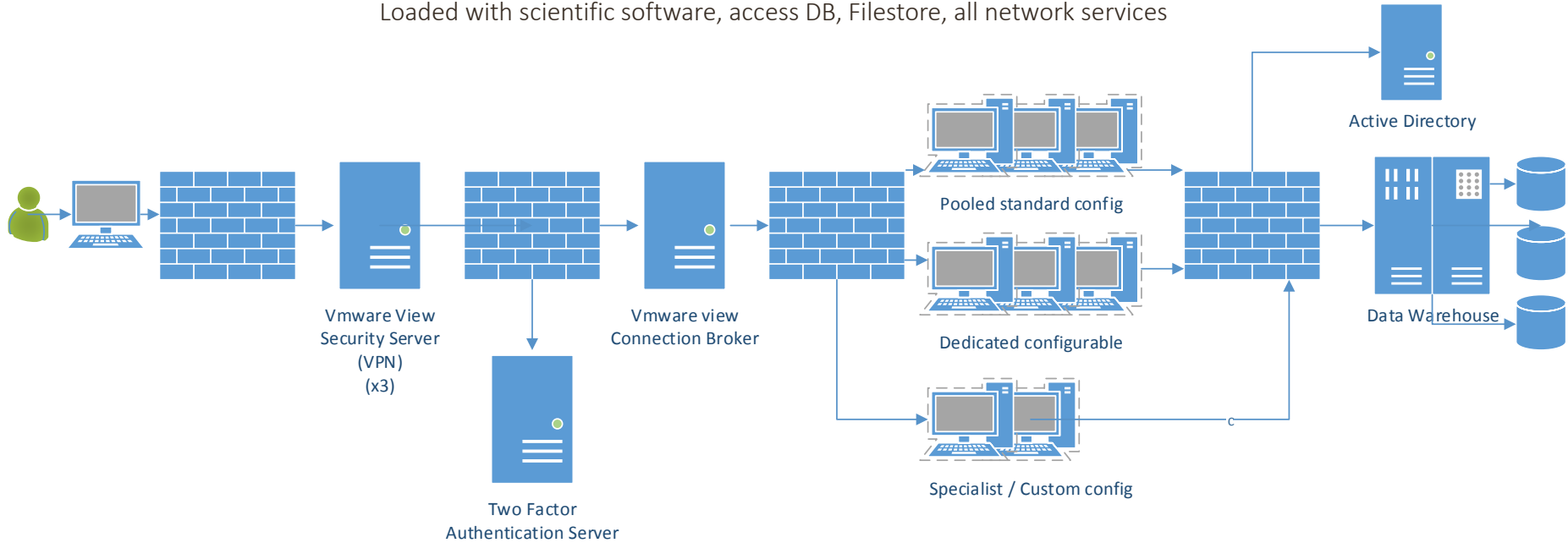
Just gone live: Medical Imaging

Next to go live: Include Biomedical / Genomics

Research Access Portal

- End User Experience – Remote desktop to Windows 10 / Ubuntu shared/dedicated system

Loaded with scientific software, access DB, Filestore, all network services

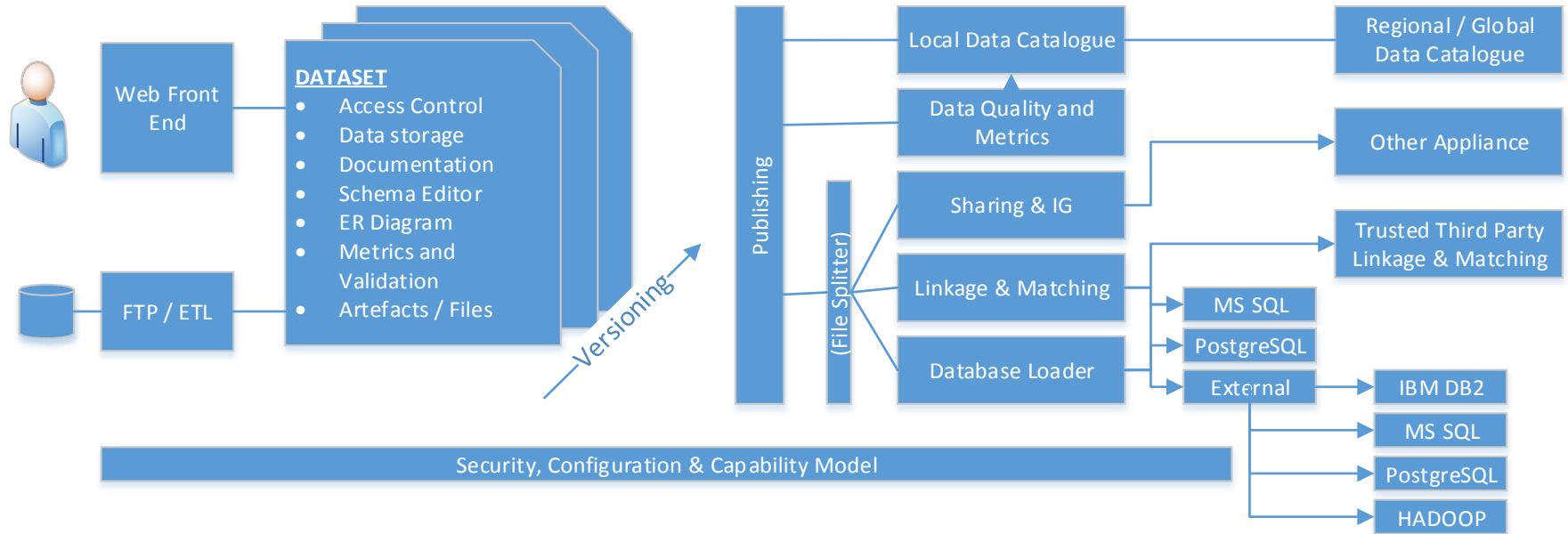


New Super-Size Desktops

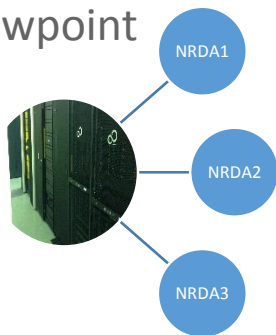
64 Cores, 4TB memory
6TB local SSD raid array

Window (UKSeRP) / Linux (CLIMB)

UKSeRP / National Research Data Appliance (NRDA)



Simplistic Viewpoint



User interface for dataset management
 Matching and Linkage
 Data Loader
 Data Quality
 Data Catalogue
 Pluggable architecture



A Dataset

Logomark

James Good 30 Notifications Sign Out Help

Home Programme 1 Programme 2 External Data Catalogues

Projects & Datasets Local Data Catalogue Users & Permissions Data Out

My Datasets (5) Favourite Datasets (10) Search

Congenital Anomaly Register and Information Service (CARIS)

Version 5 Published 21/03/2013

Administrative Contact

Dr. Jones
ABMU Health Board
01792 123456
a.jones@rersfongdomain.co.uk

Request Subscription to Data

Print Dataset Page

Overview

Description

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

[Data File](#) [Another Data File](#) [Another Data File](#)
[Another Data File](#)

View all Files

Purpose

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam architecto beatae vitae dicta explicabo.

[Another Data File](#)

View all Files

Data Quality Report 2.6MB

Theme Health

Date Type Clynical System Data

Dataset Level Individual Person

Tags Health, Wales, CARIS, ABMU, Swansea, Congenital Anomoly Register

Specific version & Date

Contact

Request

All section attach files

VIMO

Theme / Type / Level

Tags

A Dataset (cont.)

Coverage

Data Coverage

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione sequi nesciunt.

- [Data File](#)
- [Another Data File](#)
- [Another Data File](#)
- [Another Data File](#)

[View all Files](#)

Dataset Period

March 1989 – December 2001

[Data File](#)

[View all Files](#)

Inclusion / Exclusion Criteria

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni qui ratione voluptatem sequi nesciunt.

[Data File](#)

[View all Files](#)

Collection

Data Collection Method

This information is not available.

[Data File](#)

[View all Files](#)

Refresh Frequency

Data is refreshed every month w processing.

[Data File](#)

[View all Files](#)

Highlights & Know Issues

Data Highlights

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui sequi nesciunt.

[Data File](#)

[View all Files](#)

Known Issues

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

[Data File](#)

[View all Files](#)

[↑ Back to top](#)

Data Files

Data Tables (2/2)

[View all](#)



Name of Data Table

At vero eos et accusamus et iusto odio dignissimos ducimus qui blandit praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias.



Name of Data Table

At vero eos et accusamus et iusto odio dignissimos ducimus qui blandit praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga.



DDI, SPSS, SAS, STATA

SAIL Specific

Dataset Category
Core

Access Requirements

As a core SAIL dataset available in accordance with our standard Information Governance procedure.

Supporting Files

Supporting Files (5/20)

[View all](#)



Entity Relationship Diagram

orig-file-name.png (345 kb)

At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...



Friendly Name Here

orig-file-name.doc (4.2 mb)

At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...



Friendly Name Here

orig-file-name.pdf (1.6 mb)

At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...



DDI File

orig-file-name.pdf (3.6 mb)

At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...



Another File

orig-file-name.pdf (3.6 mb)

At vero eos et accusamus et iusto odio dignissimos ducimus qui voluptatum...

- Data “schema” automatically computed based on data contained in uploaded file

	All Fields ▾	A-Z ▾	⚠ Errors (3) ▾	Bookmarks (5) ▾	PID (3) ▾
Field Name	PROV_UNIT_CD ⚠	SPELL_NUM_E	EPI_NUM		
Friendly Name	Organisation Code (Unit Code of Provider)	Hospital Provider Spell Number	Episode Number		
Field Description	This is the organisation code of the health care provider. The provider code identifies the ...	This is the organisation code of the health care provider. The provider code identifies the ...	This is the organisation code of the health care provider. The provider code identifies the ...		
Personal Identifiable Data (PID) Type	N/A	NHS Number	N/A		
Field Type	Scale: 5 Char 3	Precision: 15 Scale: 5 Float	Char 2		
Validation	Please Specify... View/Edit	Range View/Edit	Local Lookup View/Edit		
Toggle All	<input checked="" type="checkbox"/> Primary Key	<input type="checkbox"/> Primary Key	<input type="checkbox"/> Primary Key		
Toggle All	<input checked="" type="checkbox"/> Show in Data Quality Report	<input type="checkbox"/> Show in Data Quality Report	<input checked="" type="checkbox"/> Show in Data Quality Report		
Encryption	Encryption: Key	Decryption: Key 50	Choose Encryption Type		
Decryption	Choose Decryption Type	Choose Decryption Type	Choose Decryption Type		
	<input type="checkbox"/> Bookmark	<input type="checkbox"/> Bookmark	<input type="checkbox"/> Bookmark		

Data Catalogue – a specific table

Friendly Name of Data Table

[Download](#) Modified 3 months ago

Records
55,160,095
Fields
9

Description

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore ventitatis et quasi architecto beatae vitae dicta sunt explicabo.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Validation (VIMO)

Category	Percentage
Valid	80%
Invalid	10%
Missing	7%
Outliers	3%

Connection String

server.database.table.12345

Heading?

All Fields | A-Z

Code Name	PROV_UNIT_CD	SPELL_NUM_E	EPI_NUM	OPER_DT
Friendly Name	Organisation Code (Unit Code of Provider)	Hospital Provider Spell Number	Episode Number	Operatio dte
Description	This is the organisation code of the health care provider. The provider code identifies...	A number (alphanumeric) to provide a unique identifier for each hospital provider...	A number used to identify episodes uniquely, and is a sequence number for each...	A numbe position to a pati
Field Type	char Size 3	int Size 4	char Size 2	smallint :
VIMO	100% Valid View	100% Valid View	91.43% Valid View	100% View
Metrics	Top 10 Values View Min, Max, Mean	Top 10 Values View Min, Max, Mean	Top 10 Values View Min, Max, Mean	Date



Records
55,160,095
Fields
9

VIMO

Valid

Invalid |

Missing |

Outliers |

Top 10 Values

Rank	Value	Percentage
1	93.98%	
2	4.39%	
3	1.02%	
4	0.32%	
5	0.11%	
6	0.05%	
7	0.02%	
8	0.01%	
9	0%	
10	0%	

Datetime Range

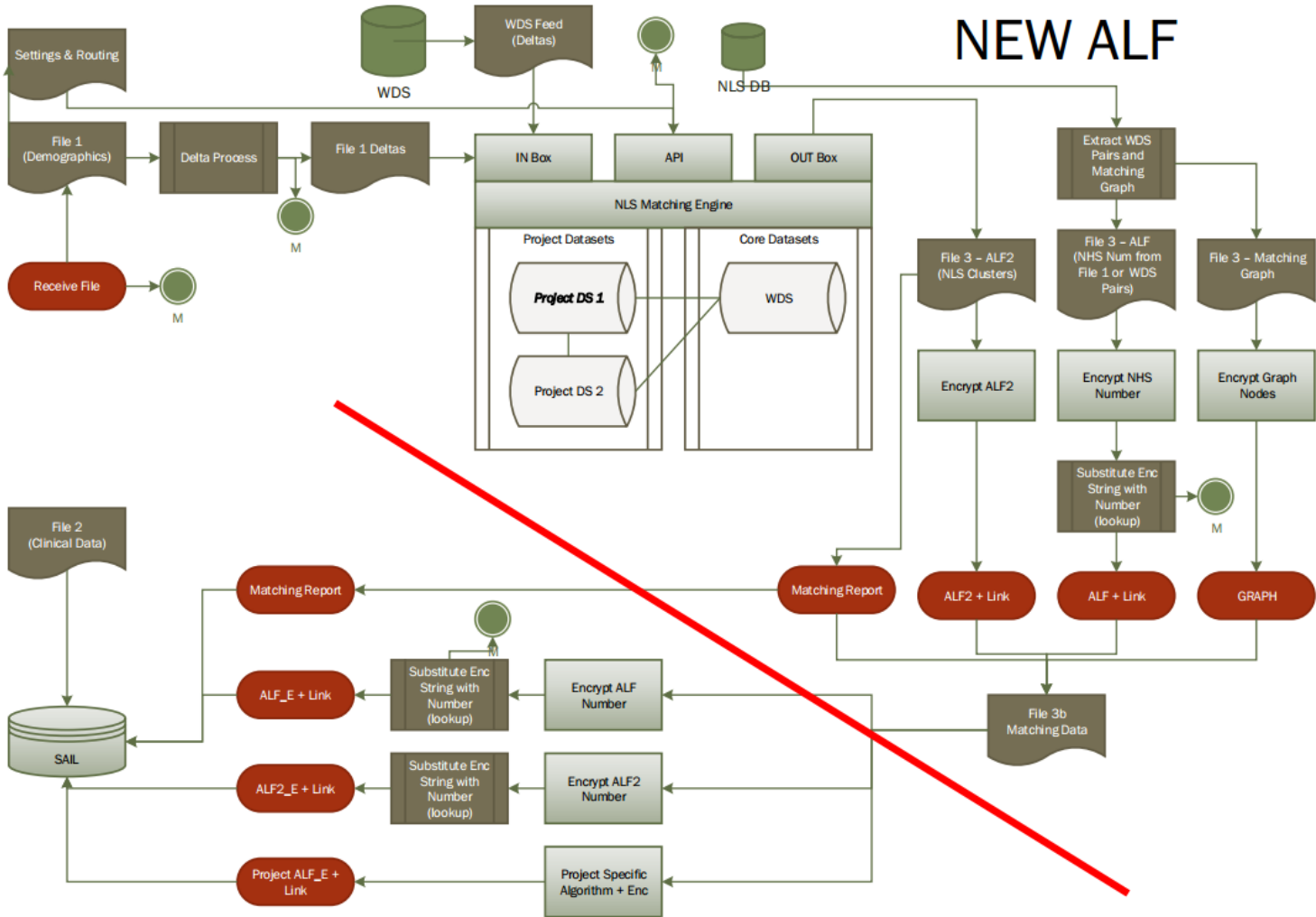
Start Datetime: 1900-01-01 00:00:00

End Datetime: 2020-12-28 00:00:00

View Min, Max, Mean

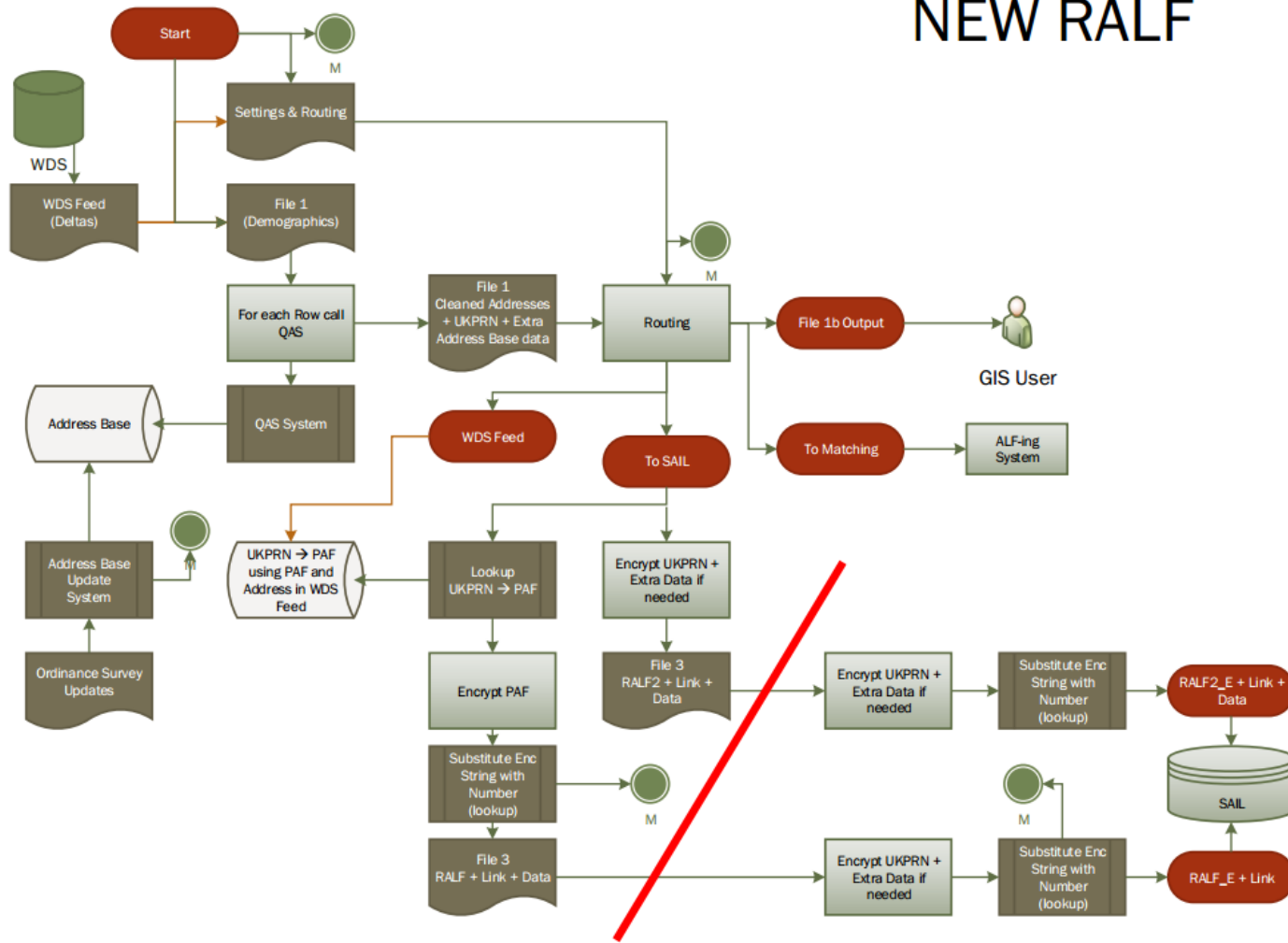
Min: 1, Max: 12, Mean: 2.18225767767603

New ALF Process



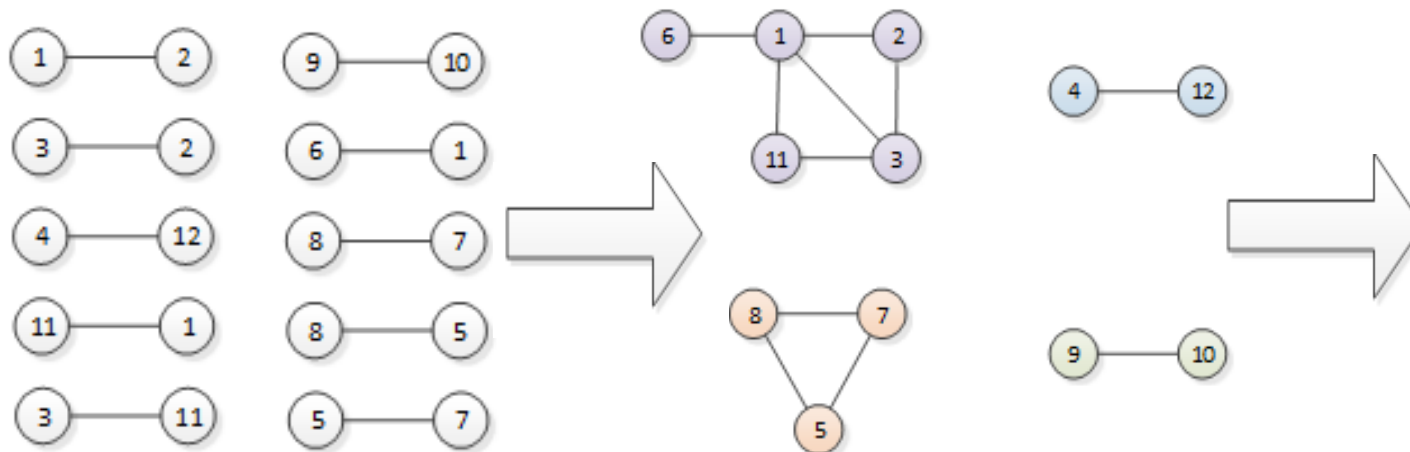
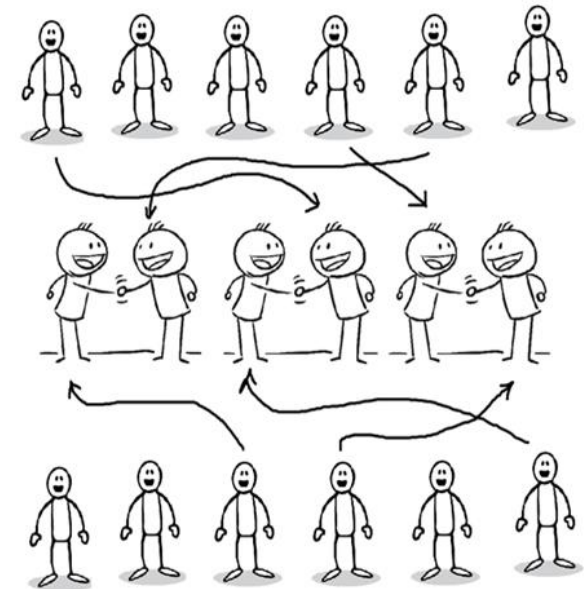
RALF – Residential ALF. Relevant as now we can pipeline address cleaning to improve matching

NEW RALF



NRDA brings world leading linkage

- Assessment of all pairs to decide if they belong to the same person
- Identify all pairs of records for each individual
- Combine 'true positive' pairs together into Groups
- Group output provides the linkage map



Person A:
Records 1, 2, 3, 6, 11

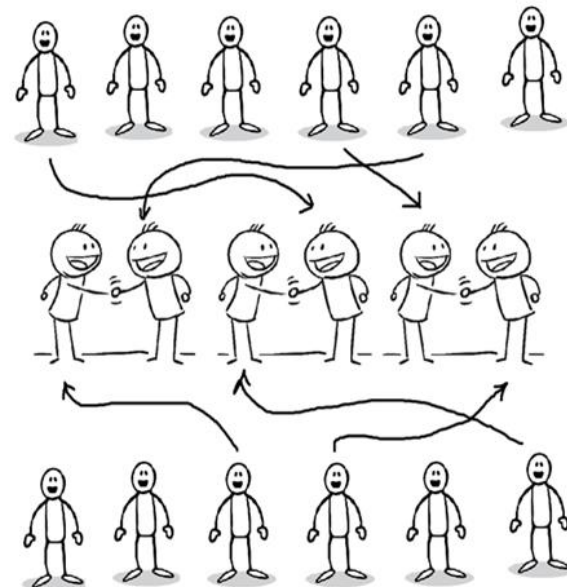
Person B:
Records 5, 7, 8

Person C:
Records 4, 12

Person D:
Records 9, 10

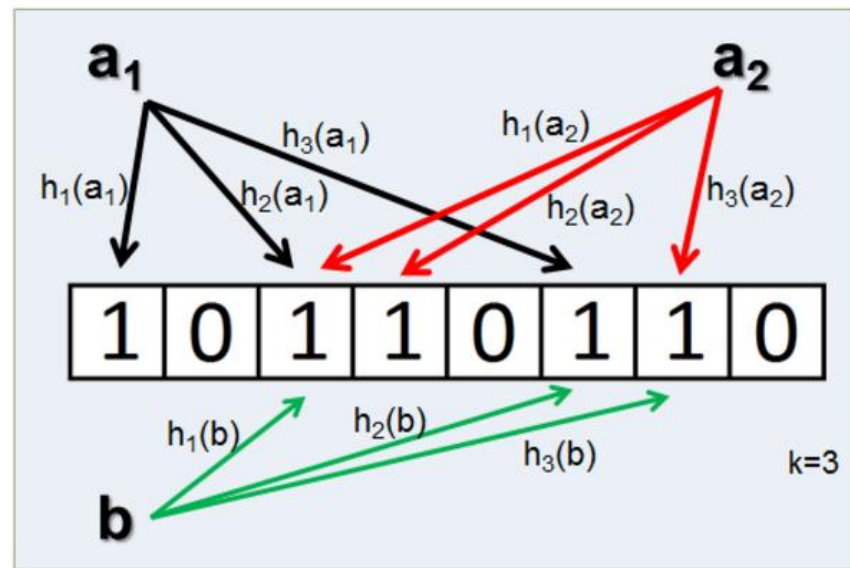
NRDA brings world leading linkage

- Hashed and Bloom Filtered Linkage Capability
- Demographics Hashed/Encrypted at source
- Deterministic and Probabilistic matching strategies
- Very small % drop in matching viability

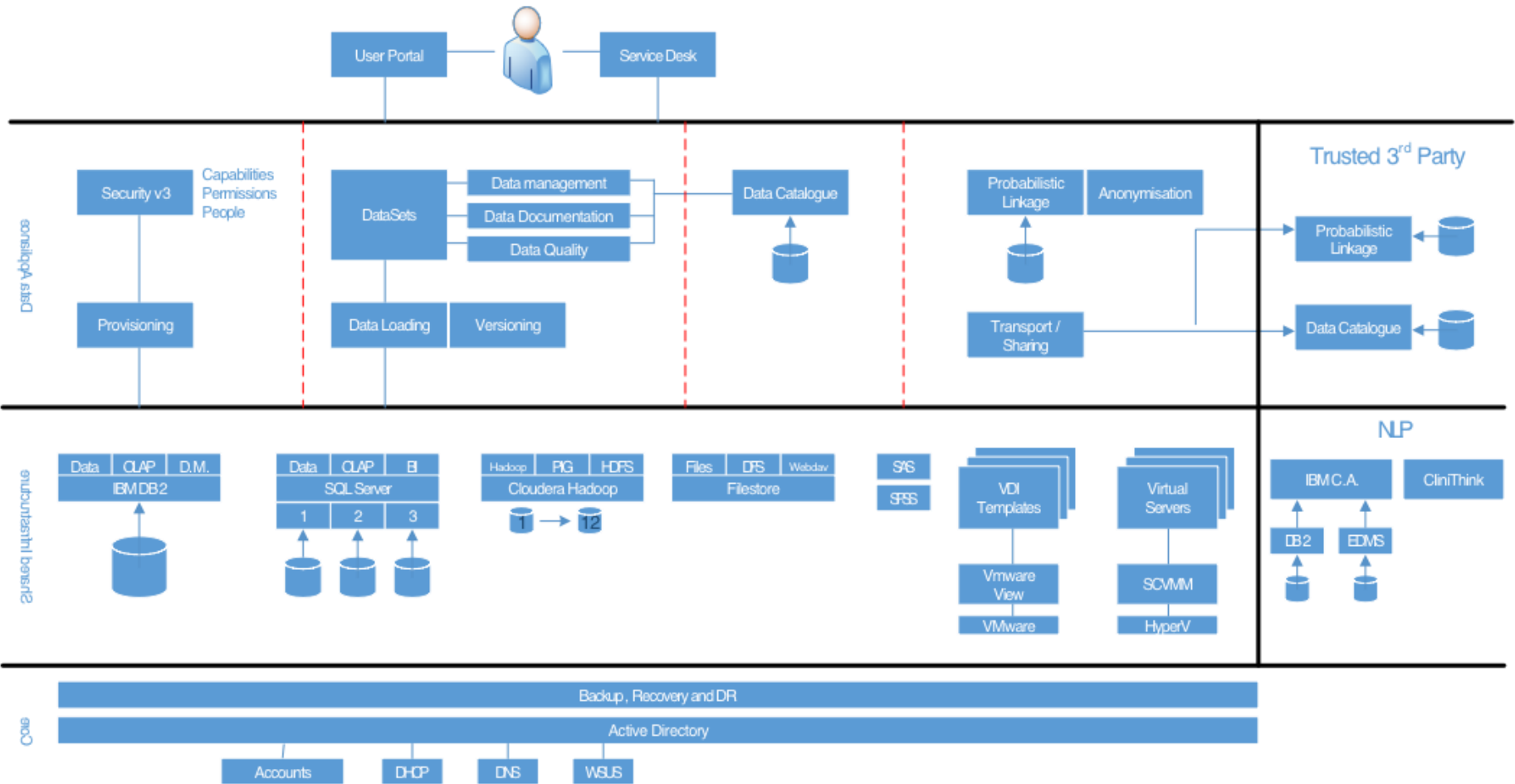


Technology productised,
integrated and available
now

Wales – Australia –
Germany – Canada
collaboration.



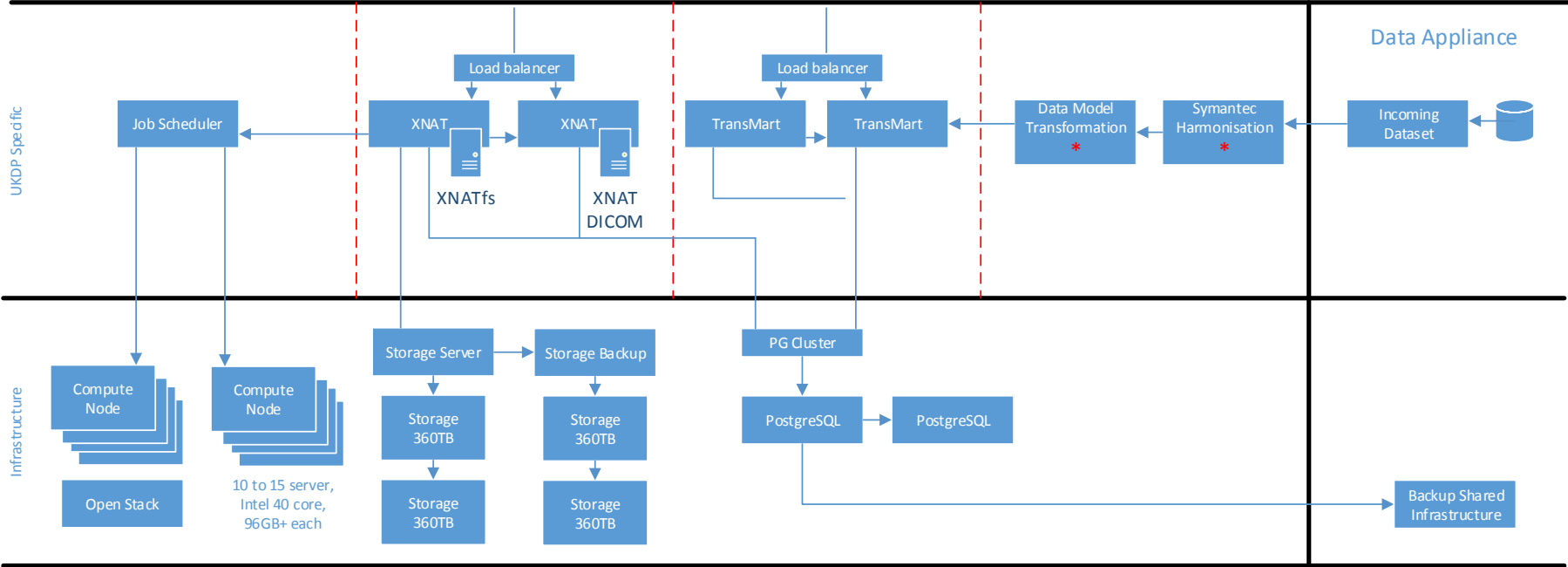
UKSeRP uses NRDA

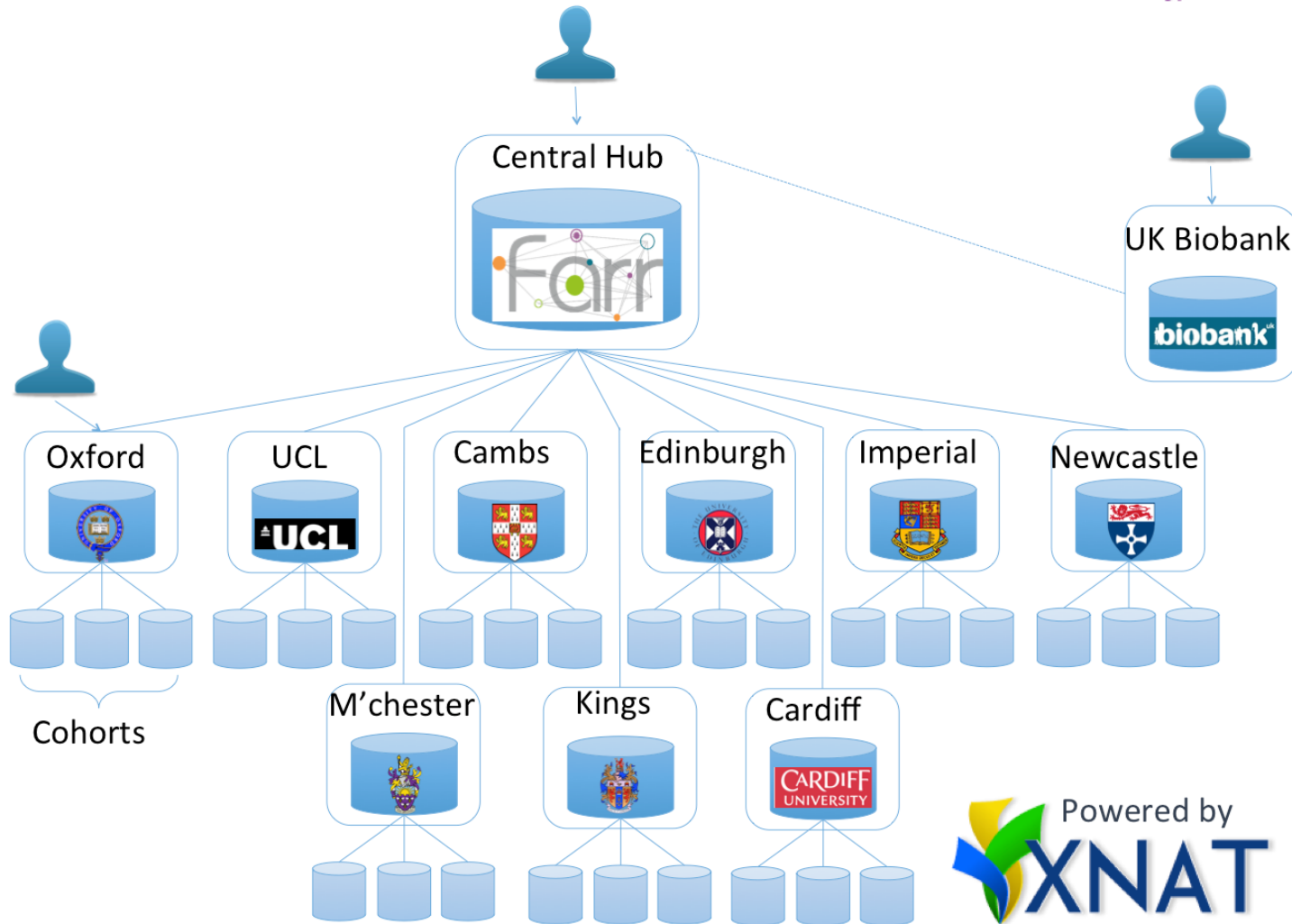


★ +Postgres, +MySQL, +MongoDB, +ElasticSearch

Recent Expansion

- Image storage, HPC Cluster, (Transmart), EMIF tooling

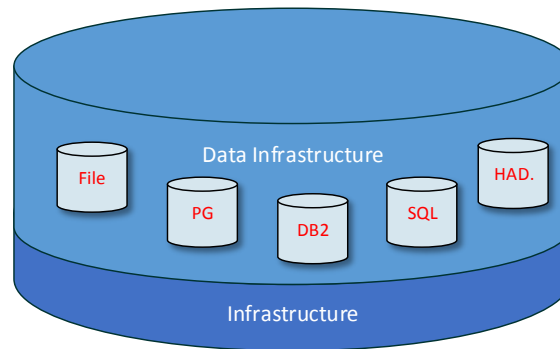
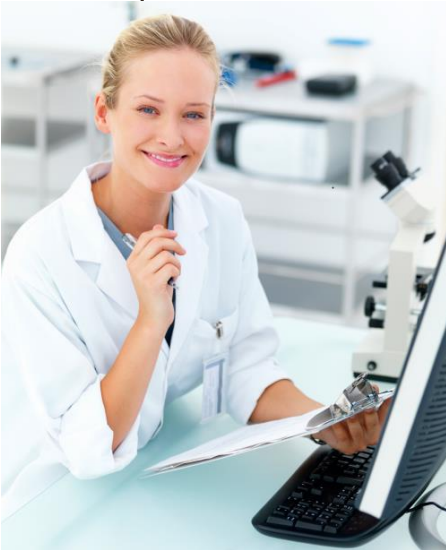


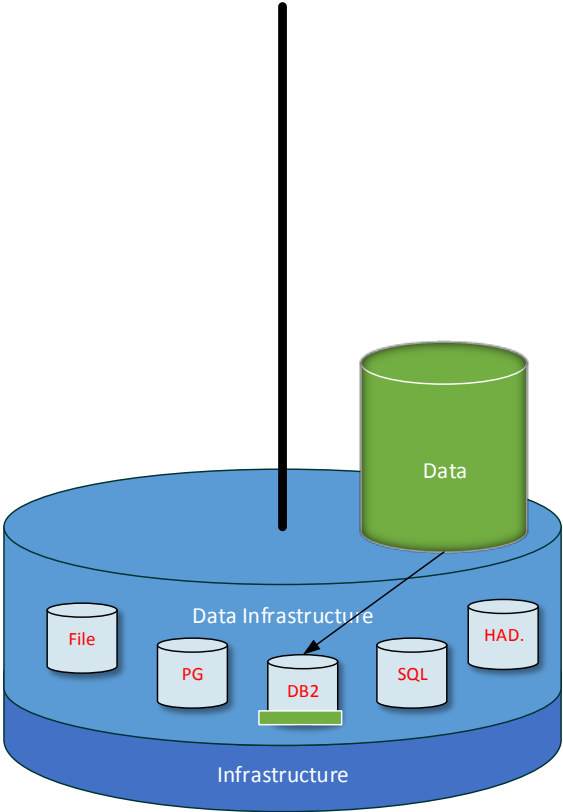


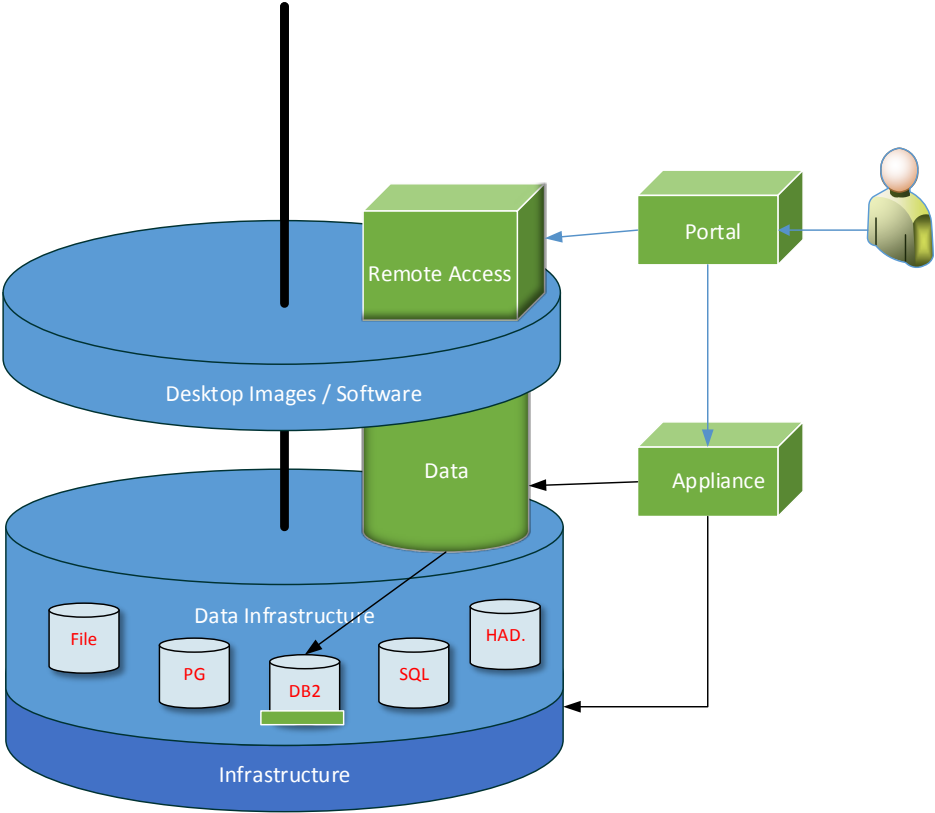
Multi centred model – central hub providing central location of cross site analysis and external data processing/contributions.

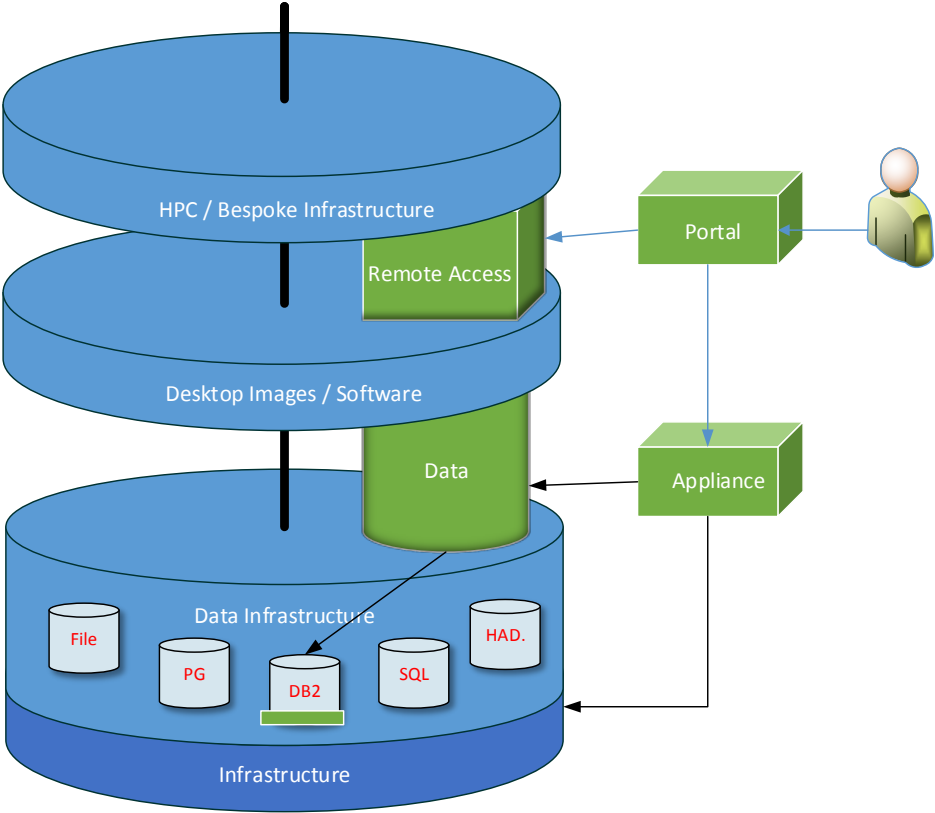
UKSeRP

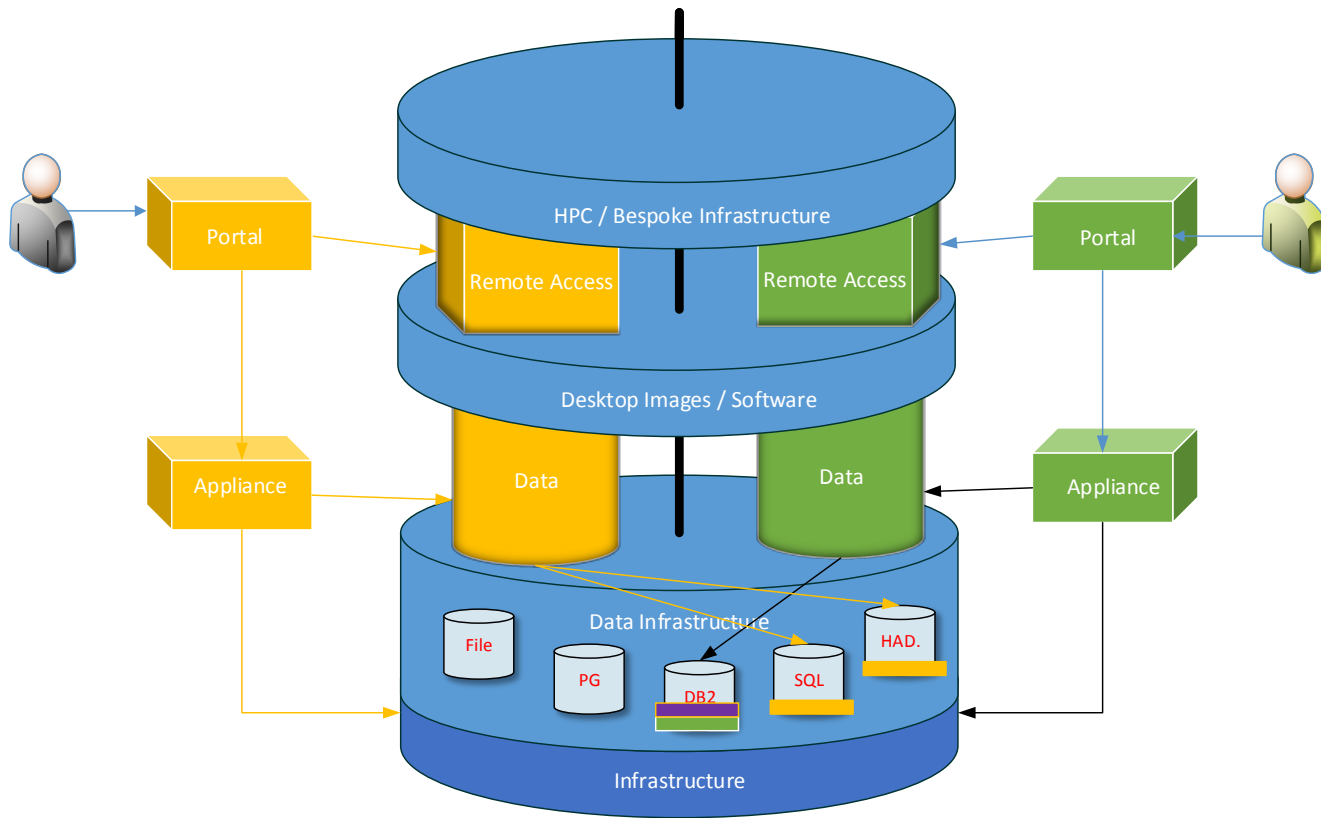
- Better ROI for funders
- Researchers can focus on doing research
- IT specialist run the IT
- Better performance from combined infrastructure







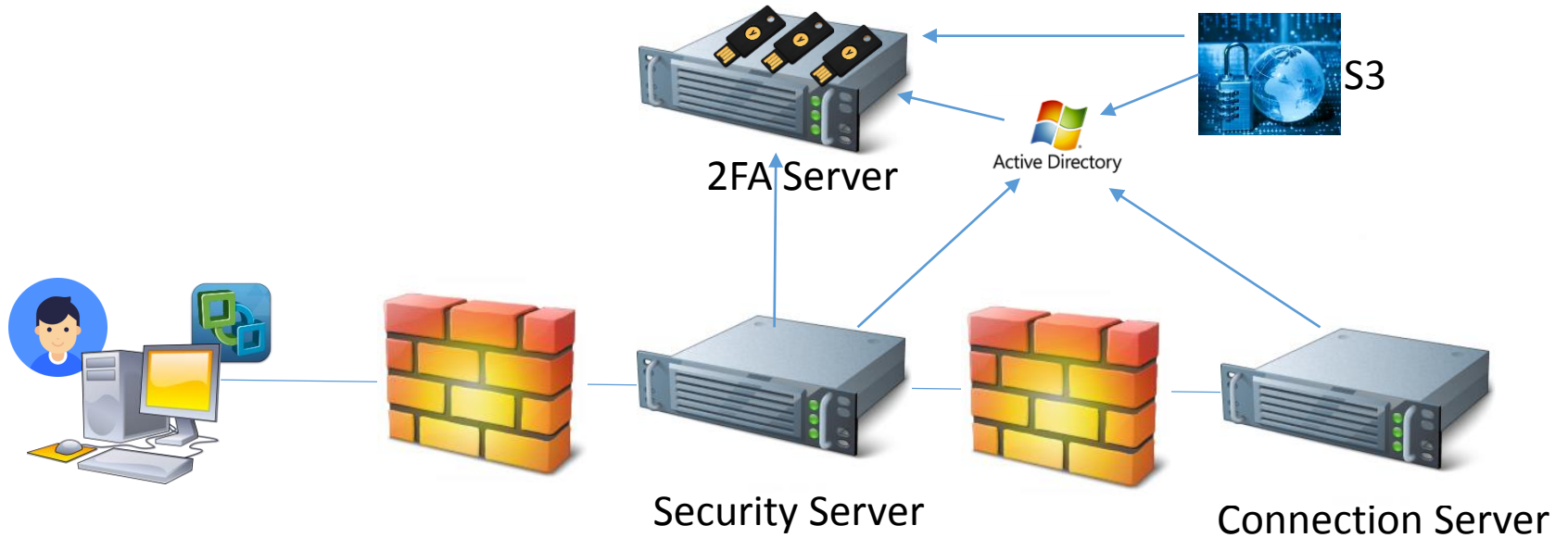




UK Secure e-Research Platform (UKSeRP)

- Large scale data and compute platform
 - Performance and scale
- A remote access analytics platform
 - Best practice: data management, security, governance
 - Suite of standard and bespoke data analytical tools
 - Accessible across UK and internationally
- Leaves data ownership with the cohorts/programme
 - Each 'controls' slice of UKSeRP
 - Devolved account and access control
 - Information governance remains with cohorts/programme
 - Brings together data for DPUK across cohorts
- Enables researchers to focus on the science

Reduce Costs, Reduce Risk



Last 2 days 21st and 22nd March 2017

Maxing at 16



Float Pool



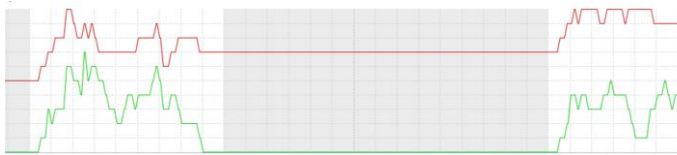
Maxing at 33



Dedicated Pool

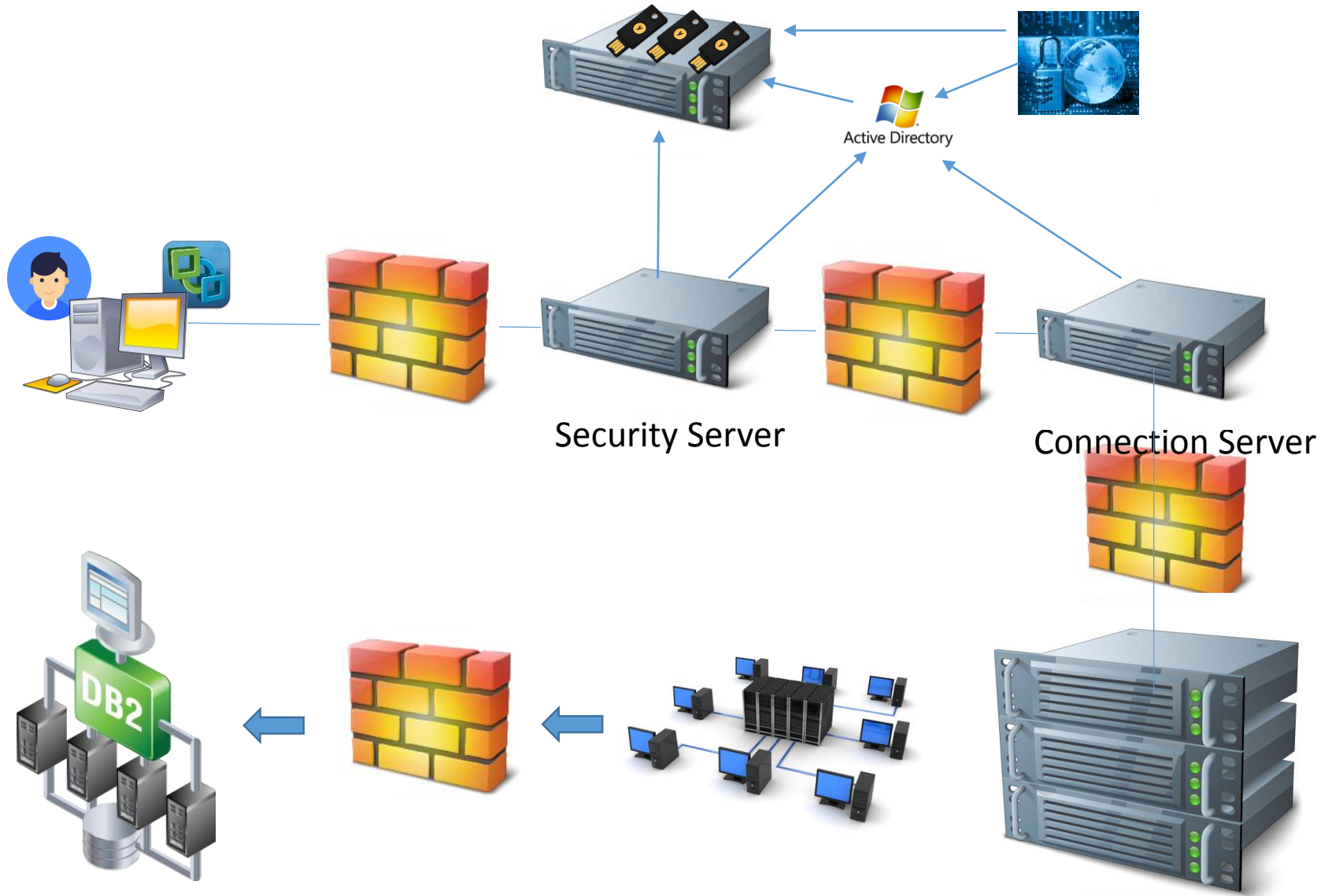


Maxing at 10



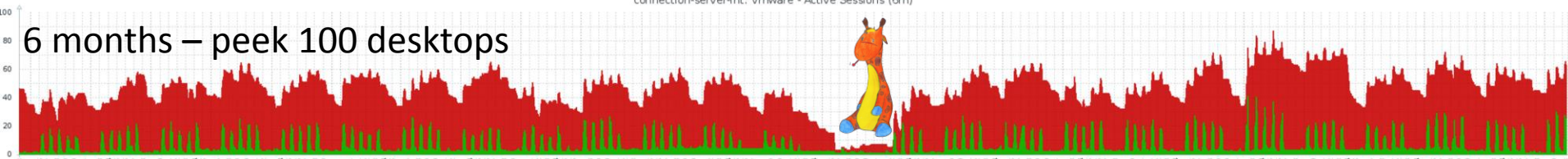
STATA Pool





connection-server-int: Vmware - Active Sessions (6m)

6 months – peek 100 desktops





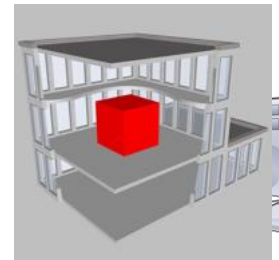
Security Server



Active Directory



Connection Server



Security Server



Active Directory



Connection Server



Security Server



Active Directory



Connection Server



Different Security requirements

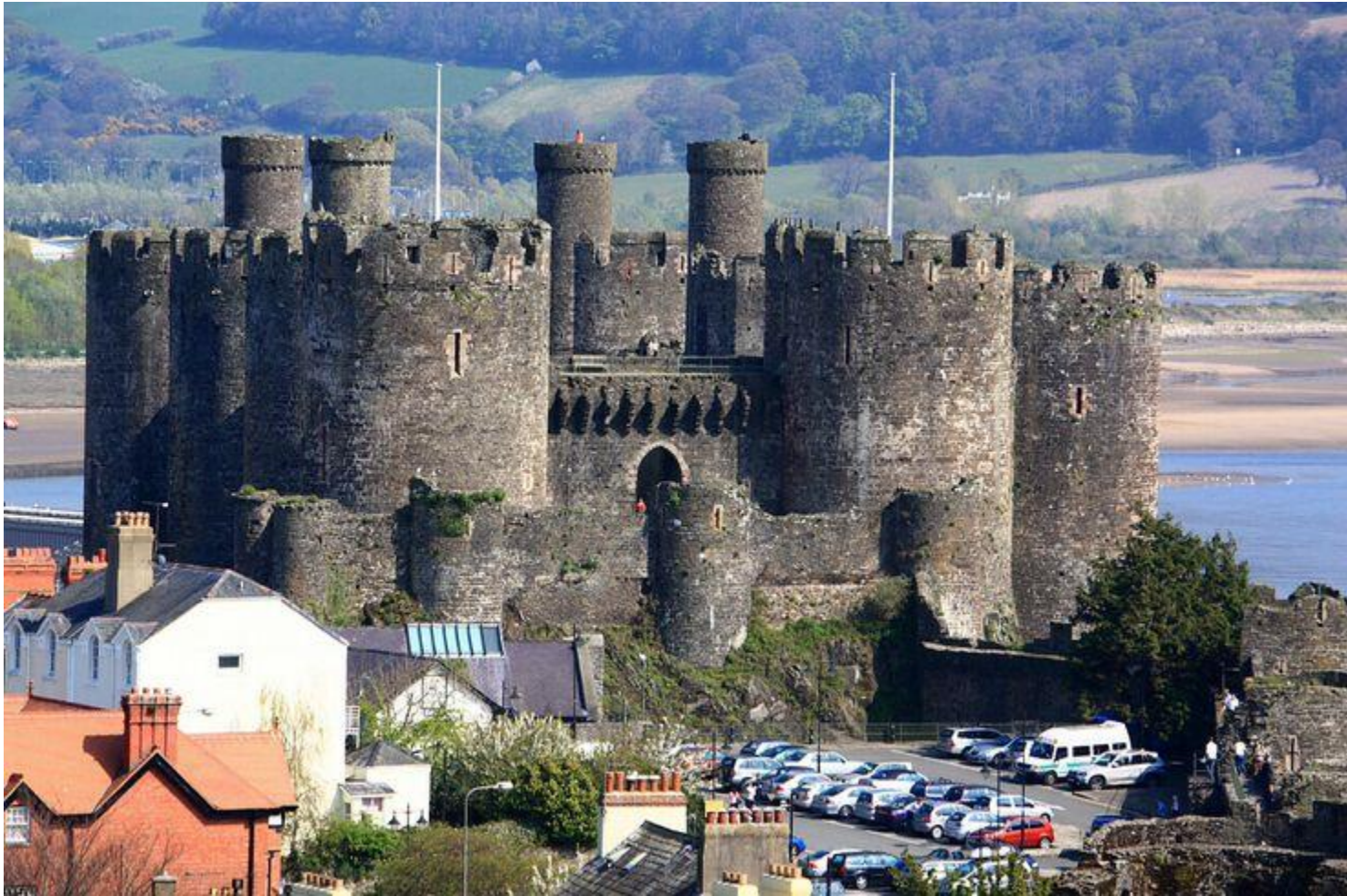
- SAIL Databank – No geo restrictions / NHS connections
- ADRC – Locked down environments / Safe-Rooms
- FARR – Also Identifiable datasets, vary by project
- DPUK – Multi model / integration, complex
- Biobank – Restricted to project members
- Alspac – Local secure direct access / Users UK
- (more not public yet) – like Perth, Australia

Protecting the data and ensuring only authorised access is key

Twice weekly vulnerability / pen testing – fully aware of our risk exposure (slept well this weekend)

Patching, compliance and Hardening core components

We built a castle !



Wales loves Castles – have 600 of them.

Conway Castle

Castles are Great — Big, Strong, Expensive

- Many Layers of security – restricted building, data centres, 2 inch steel door, independent alarms, devolved control of access control system, CCTV, facial detection, multi vender perimeter firewalls, internal firewalls 3rd vender, network segmentation and isolation.
- Garrison of solders to protect – team of DevOps staff, security officers, compliance and governance people
- Controlled Entry points

Everything outside is vulnerable

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



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

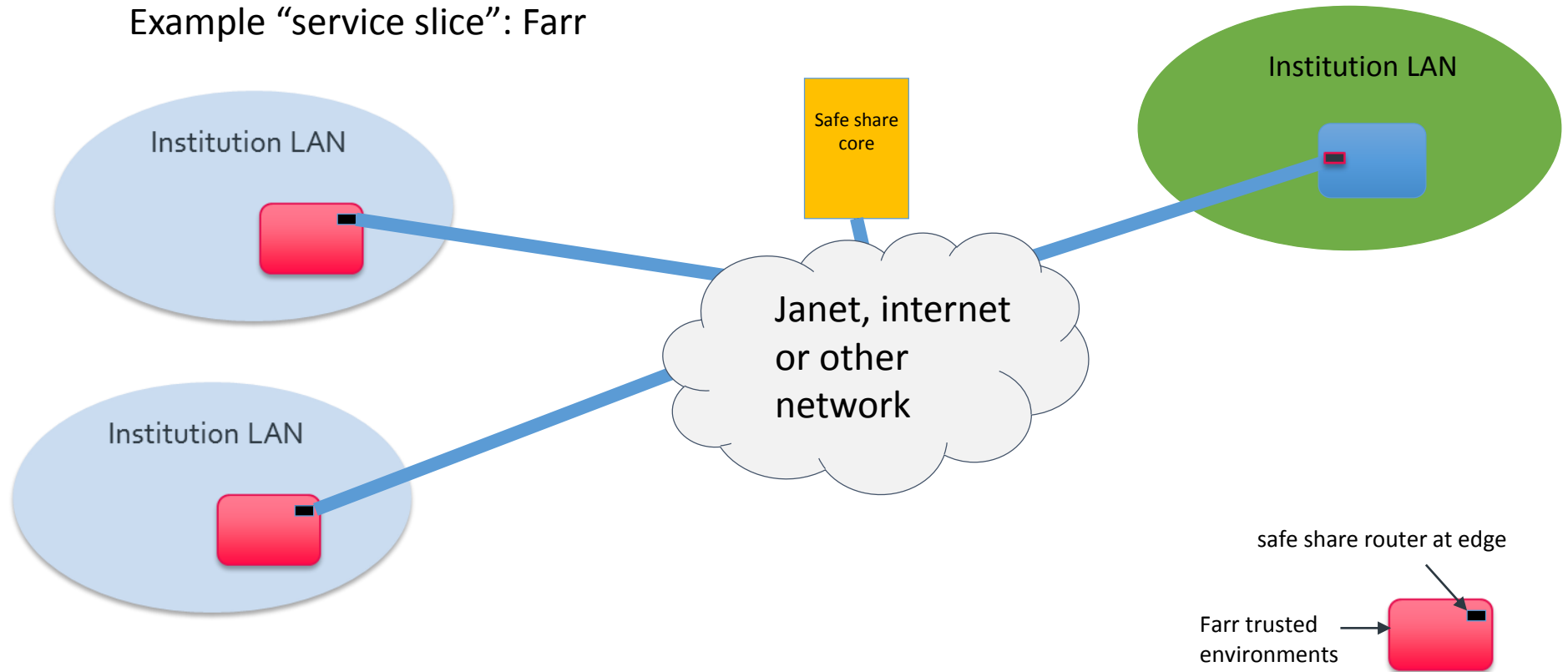
How / Who do you trust ?

- Equivalence, hard to measure so subjective
- Build a relationship up
- Standards and accreditation helpful.
ISO 27001 is great but does not cover governance
- Never fully trust, also put a portcullis at the end the tunnel just incase 😊



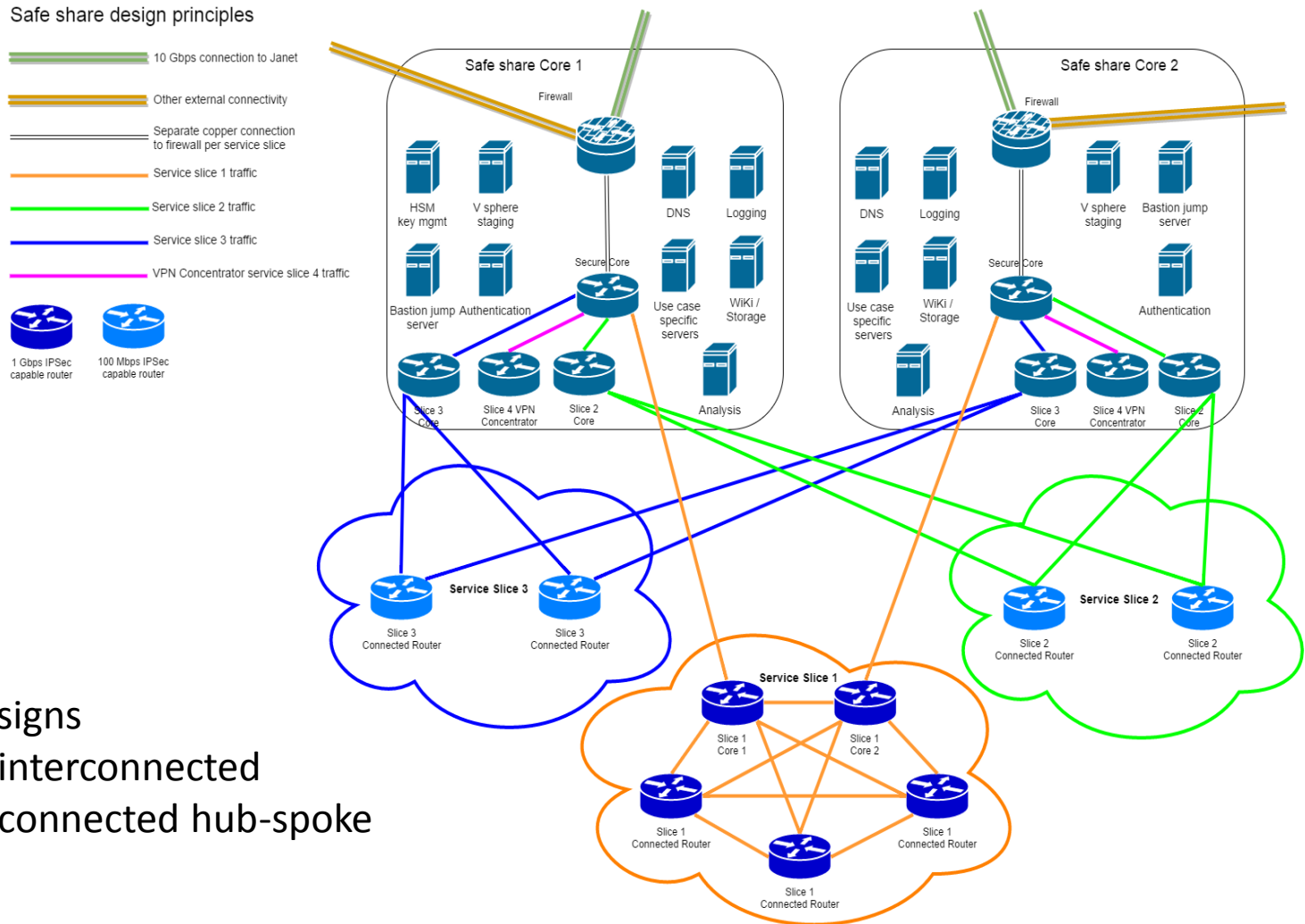
The safe share project

Example “service slice”: Farr



IT MAKES TUNNELS !!!

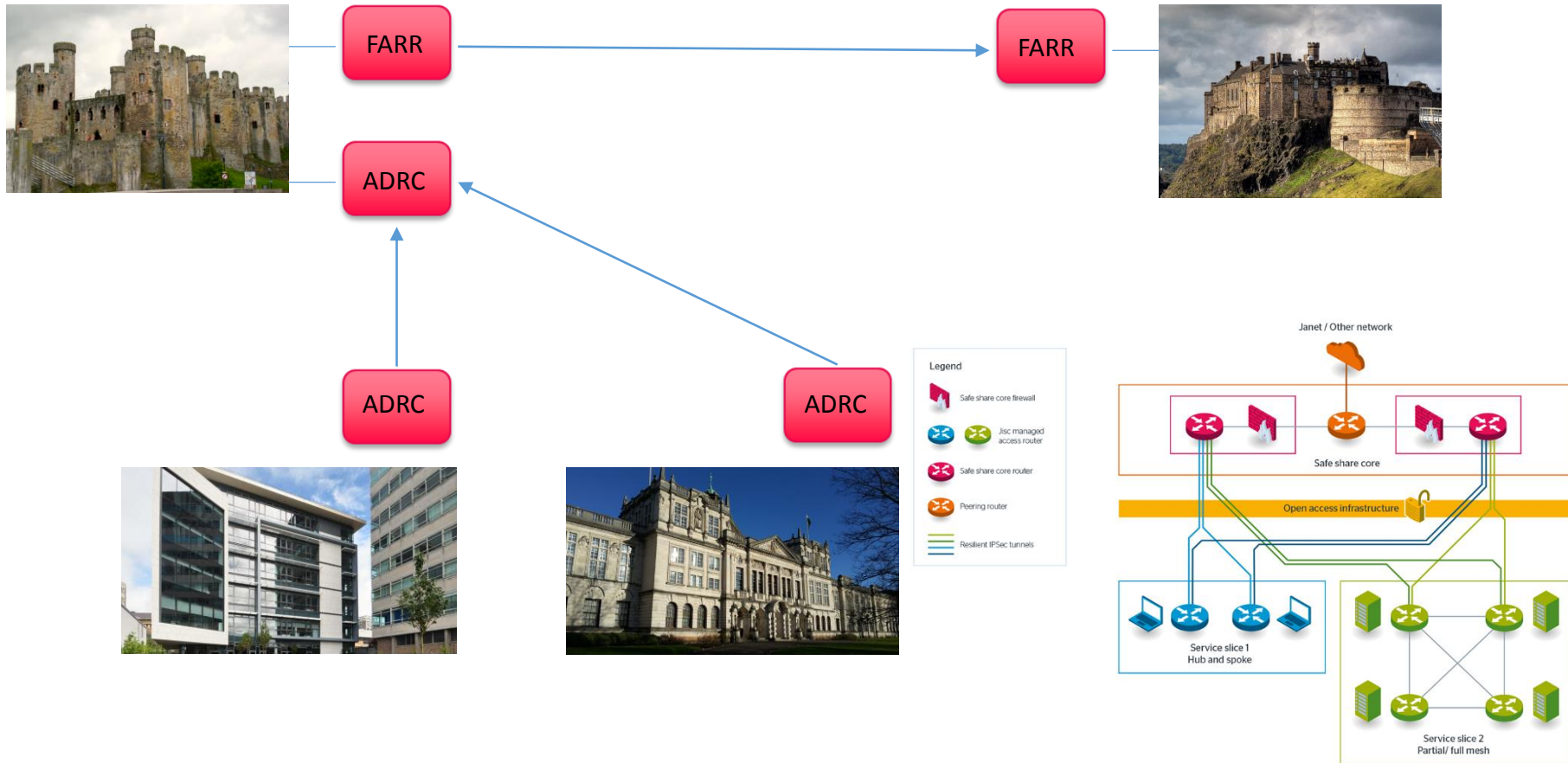
The safe share project: HAN design overview



TWO main designs

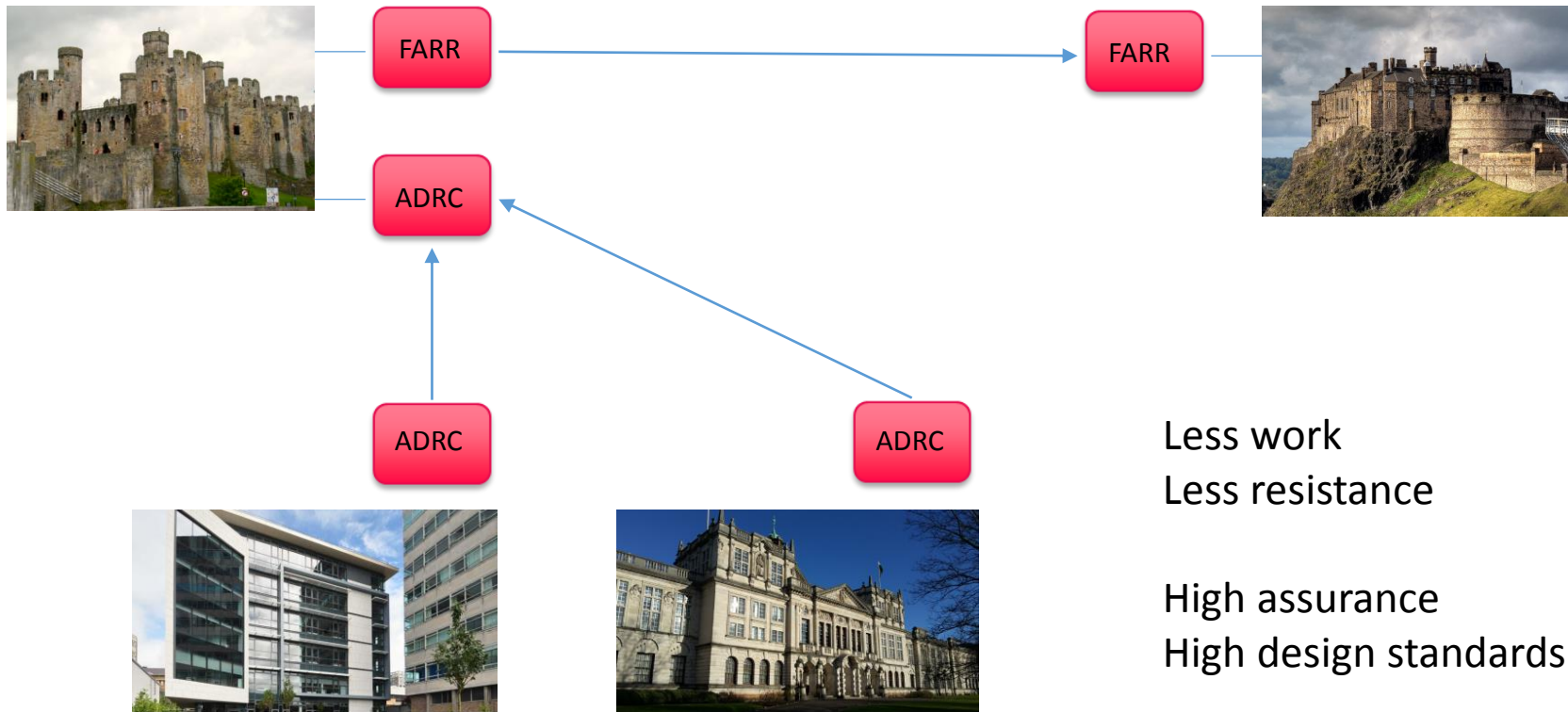
1. All Nodes interconnected
2. All Nodes connected hub-spoke

How we have used it



Why?, could do it ourselves ?

Already connecting 3 organisations – so independent trusted name like JISC advantage, accreditations, service management and governance strong “selling point”
Same strong standards, guaranteed interconnect / compatibility

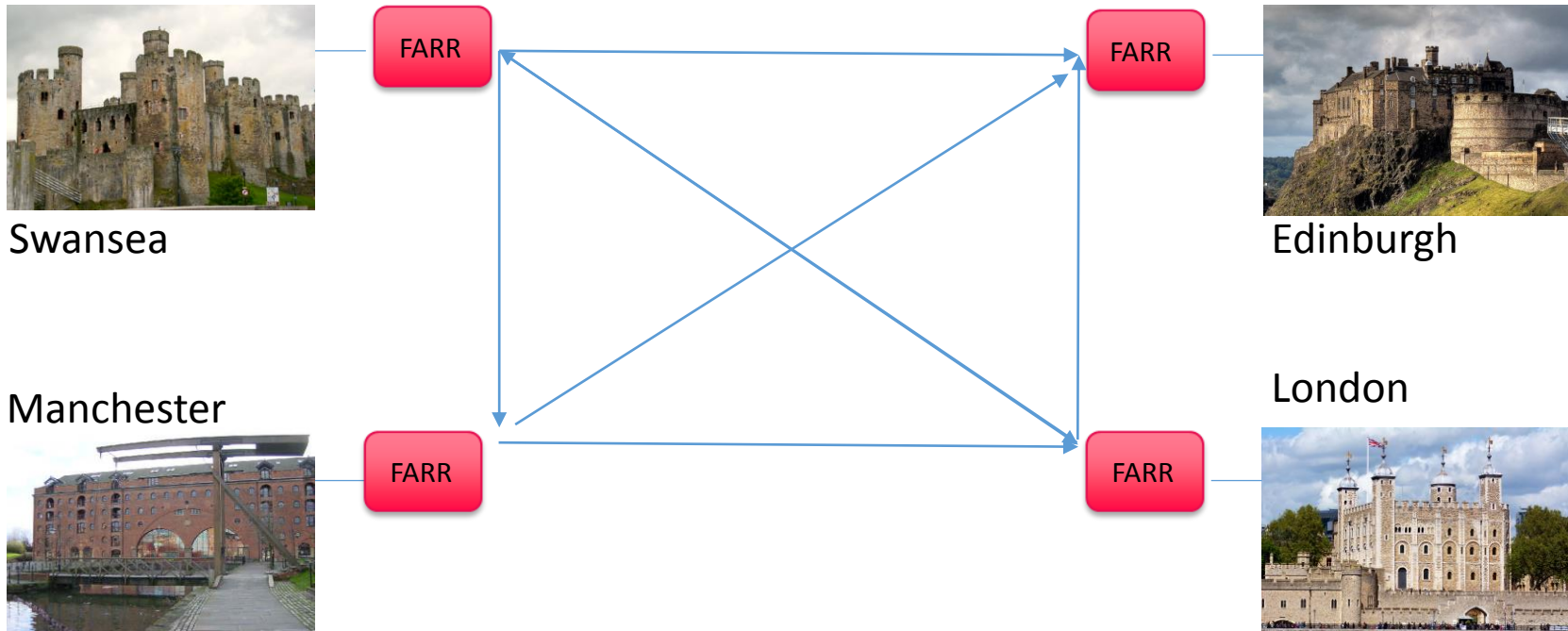


Less work
Less resistance

High assurance
High design standards

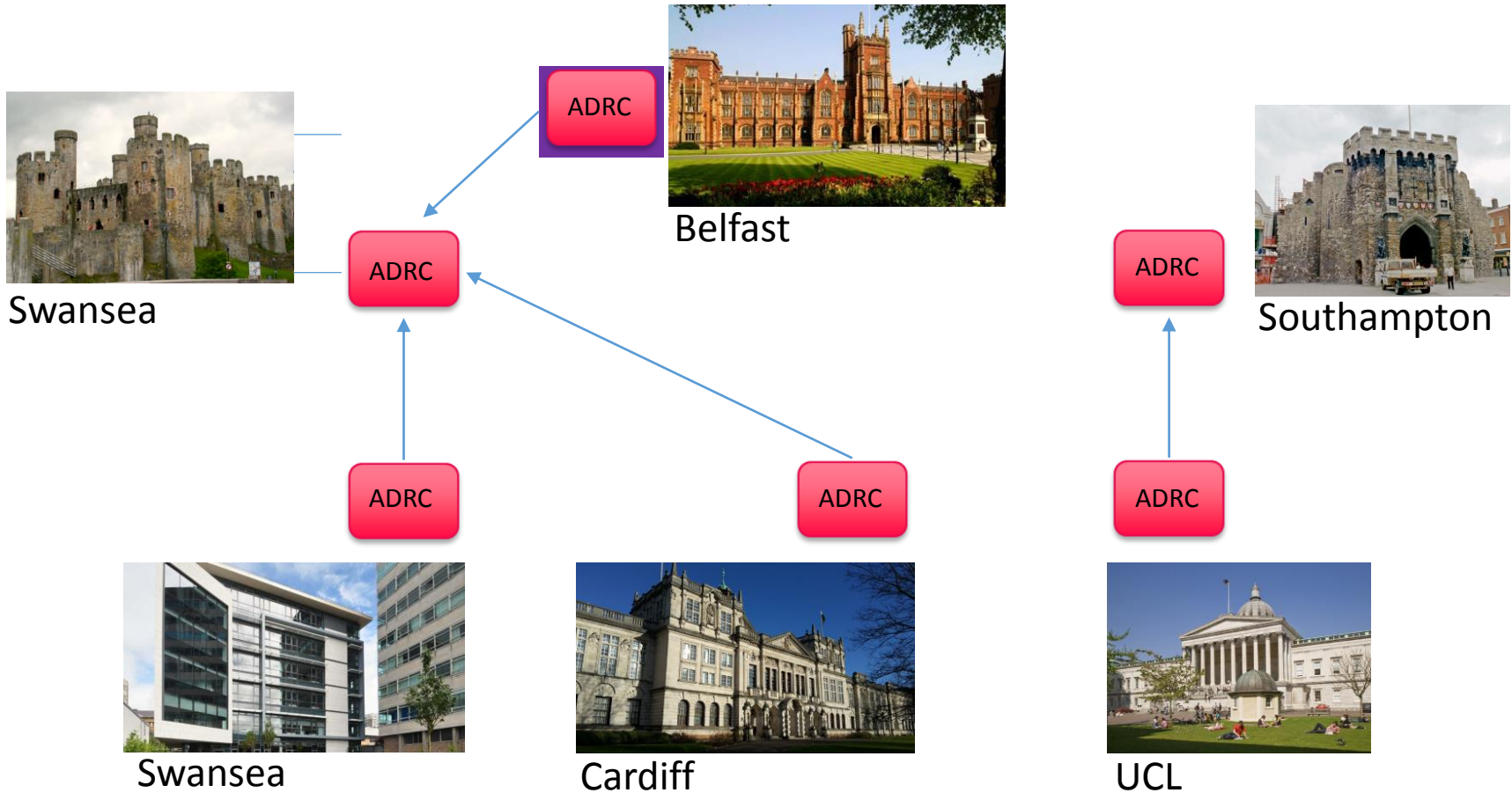
Could / going to - FARR

FARR v2 – 4 centres to 10 centres

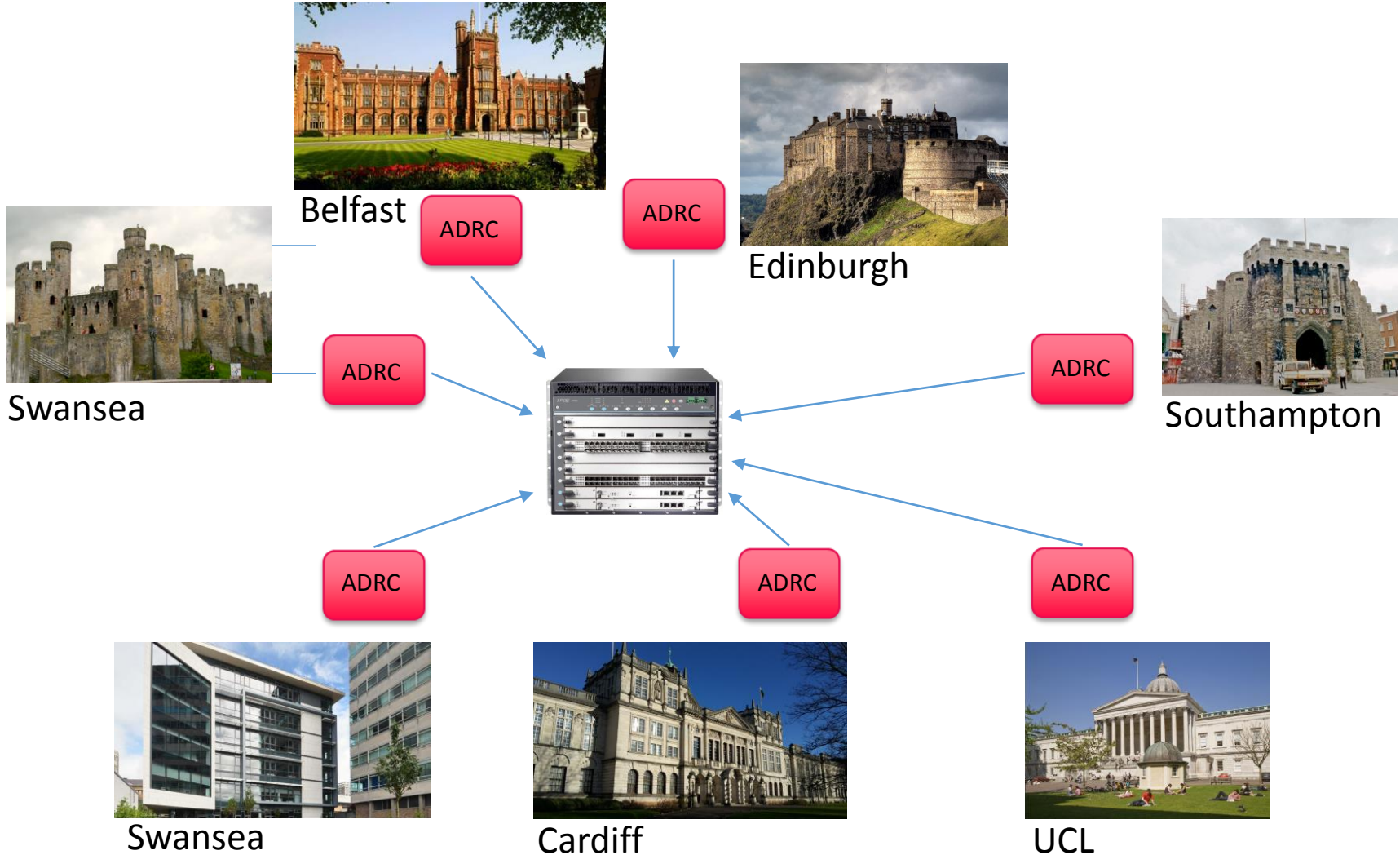


End users from each site could connect to any other infrastructure
Other infrastructure / HPC could securely access datasets in remote data centres

Could / going to - ADRC

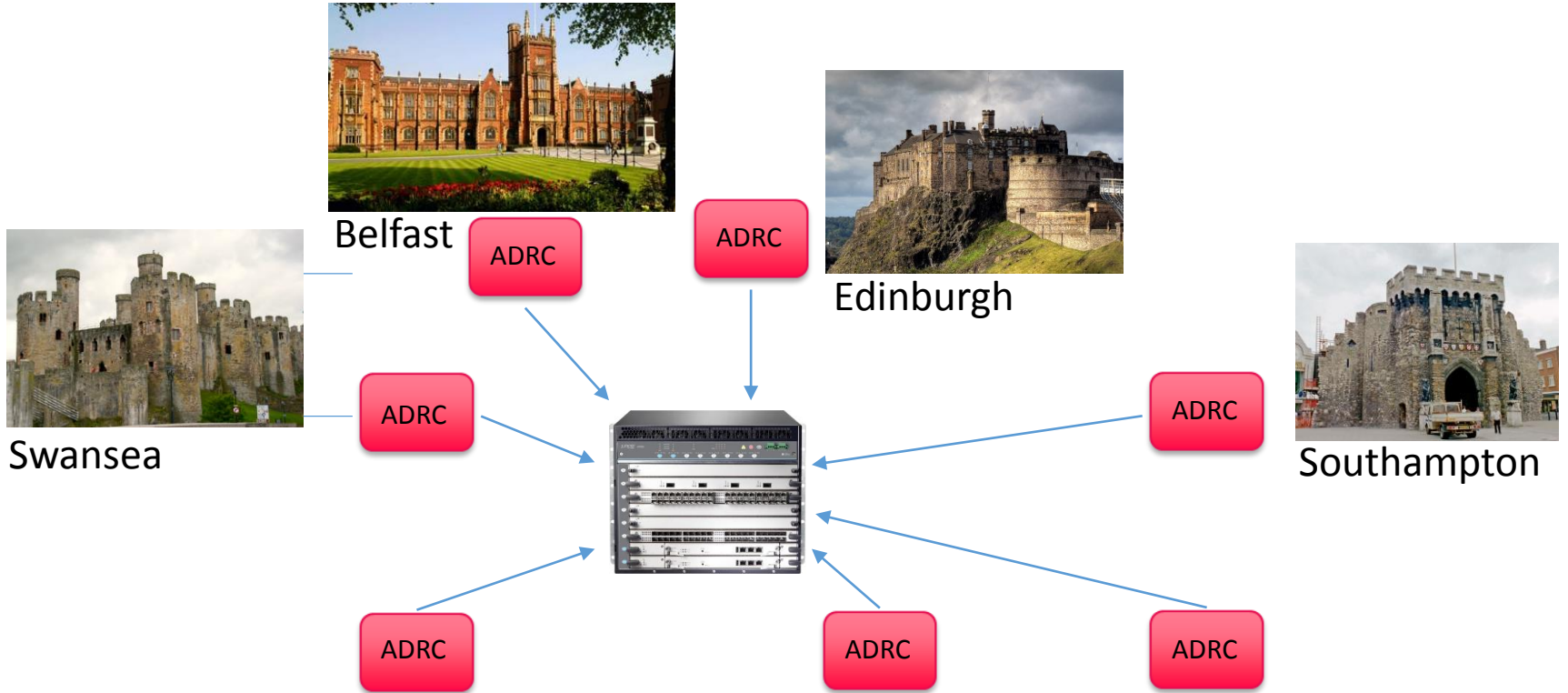


Could / going to - ADRC



All safe room can connect to all data centres – significantly increasing access

Could / going to - ADRC



Department
for Work &
Pensions



Office for
National Statistics

Secure transfer / access of sensitive large datasets

JISC Safe Share

- We need secure network inter connections
- High assurance / data governance
- Multiple over lays based on use case / requirements
- Trusted third party is the ideal solution

- The key will be for the academia community to buy into this vision, only at scale does it make sense.

- If we create silo's then in the end things will be worse its started....

Questions?



Simon Thompson

simon@chi.swan.ac.uk

