
# DATA SCIENCE @ ED

EDINBURGH DATA SCIENCE AND MANAGING NATIONAL DATA SERVICES AT EDINBURGH

PROF MARK PARSONS

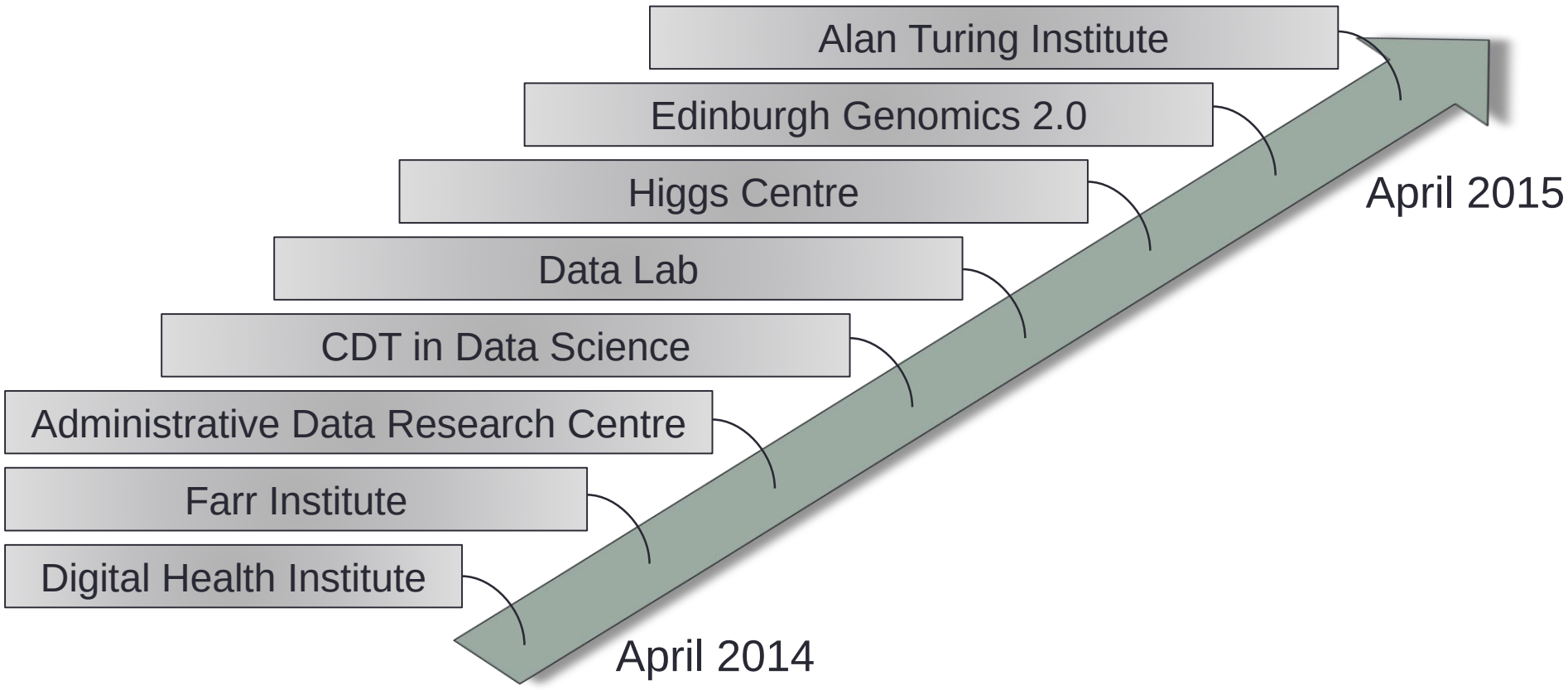
EPCC Executive Director
Associate Dean for e-Research

epcc

# Edinburgh Data Science

- An initiative spanning all of the University to bring together all of the world-class research on, and using, Data Science techniques
- Goals for next 3 years
  - Set the pace for strategic activities
  - Be at the hub of the UK Data Science network
  - Develop a safe haven for unique data assets
  - Be deeply integrated in pursuit of bold goals
  - Be recognised as a world centre

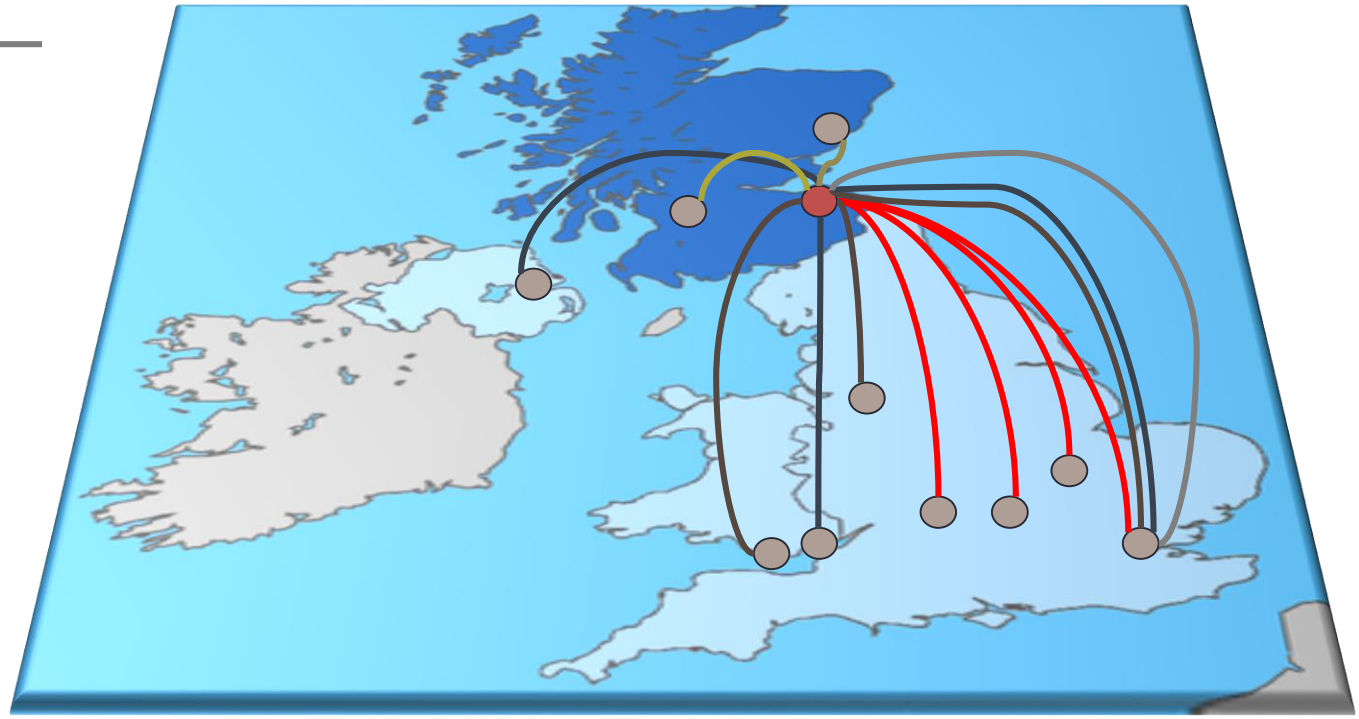
# Accumulating success …



# EDS networks … and I don't mean wiring

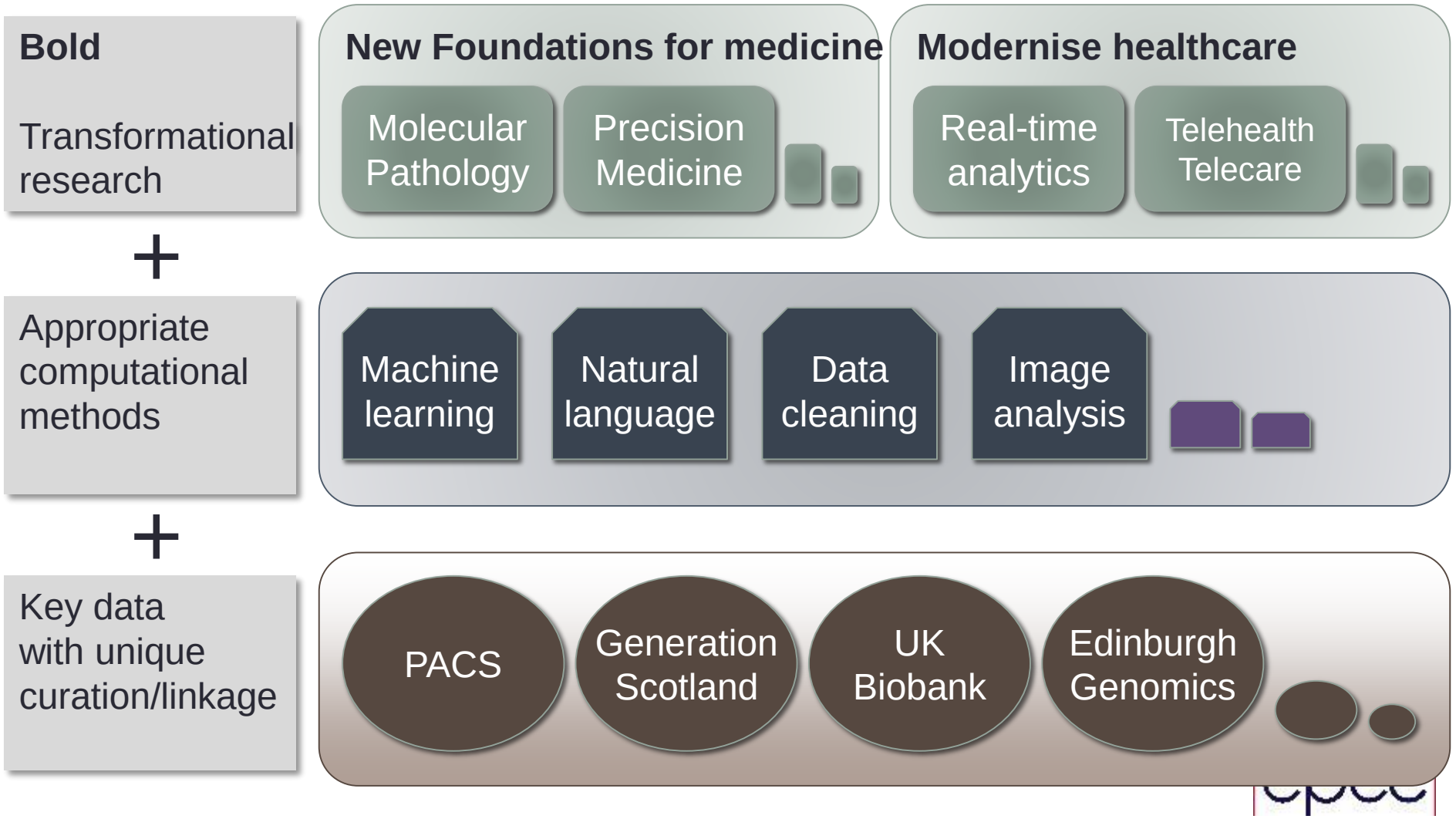
Turing
Farr
ADRC
ICs
ICT Labs

Networks hold > £150M UK funding in next 4 years
Will influence distribution of further funding



epcc

# Example of EDS stimulating research integration



# Physical Presence

Building 9 BioQuarter
Healthcare/administrative data

Higgs Innovation Centre
Physics/engineering data

Data Technology Institute

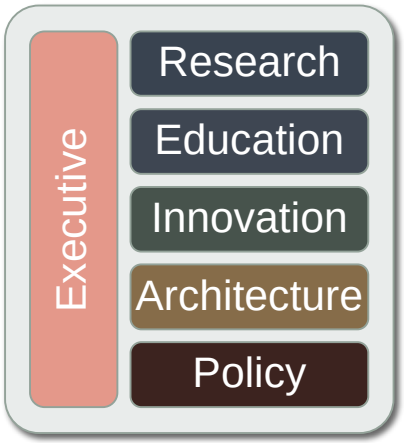
Advanced Computing Facility
HPC and Big Data

Easter Bush Innovation Centre
Bio/genomics data

# What EDS is *and* isn't

## EDS is

- An amplifier
- An integrator between centres
- A promoter of Data Science
- Quick to act
- Lightweight




## EDS isn't

- A source of direct funding
- An integrator within centres
- A replacement for existing efforts
- A source of frequent committees
- Rich!

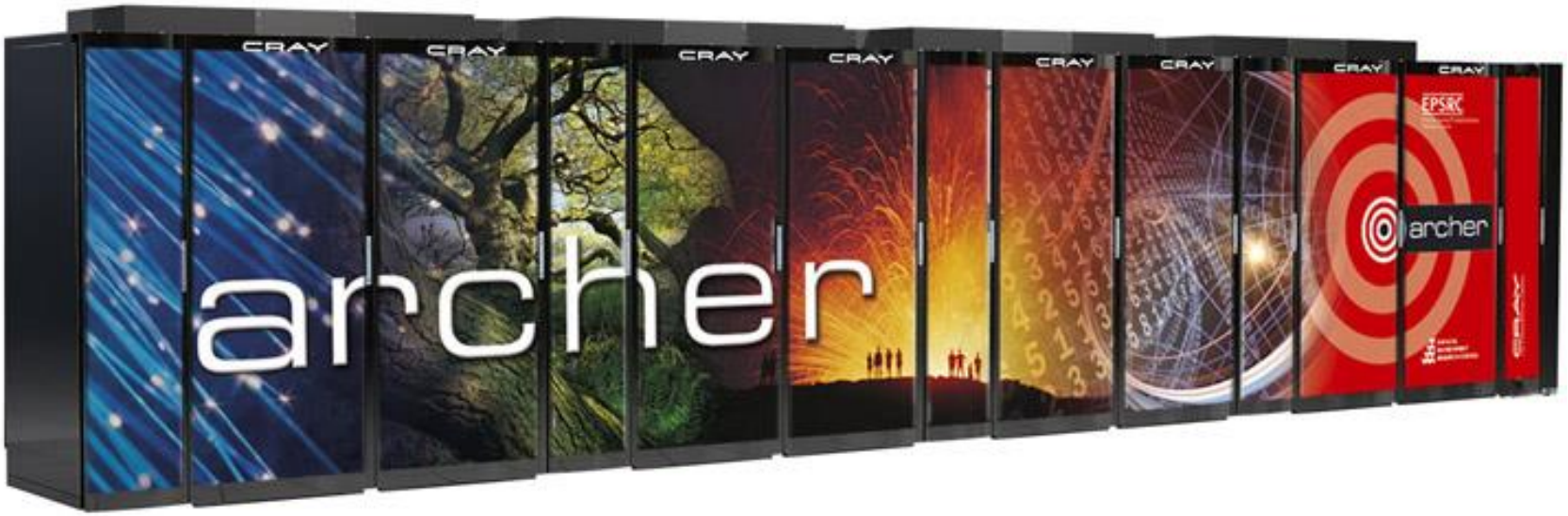
epcc

# Advanced Computing Facility

- The ‘ACF’
- Opened 2005
- Purpose built, secure, world-class facility
- Houses wide variety of leading edge systems and infrastructures
  - National services
    - ARCHER 118,080 cores (Cray XC30)
    - DiRAC 98,304 cores (IBM BlueGene/Q)
    - RDF (25Pb Disk / 50Pb Tape)
  - Local services
    - INDY – industry machine
    - ULTRA – SGI UV2000
    - HYDRA – research system
- £12m expansion in 2013: 6MW, 850m$^2$ plant room, 550m$^2$ machine


epcc

# ARCHER




- Cray XC30 – National HPC service contract managed by EPSRC
- 26 frames – 118,080 cores
- Comes with large 5Pb work filesystem
- Directly linked to 25Pb Research Data Facility for storage of simulation results
- EPCC won Service Provision contract in August 2013 – service opened in November 2013 for ~3,800 users

# Data management is very challenging

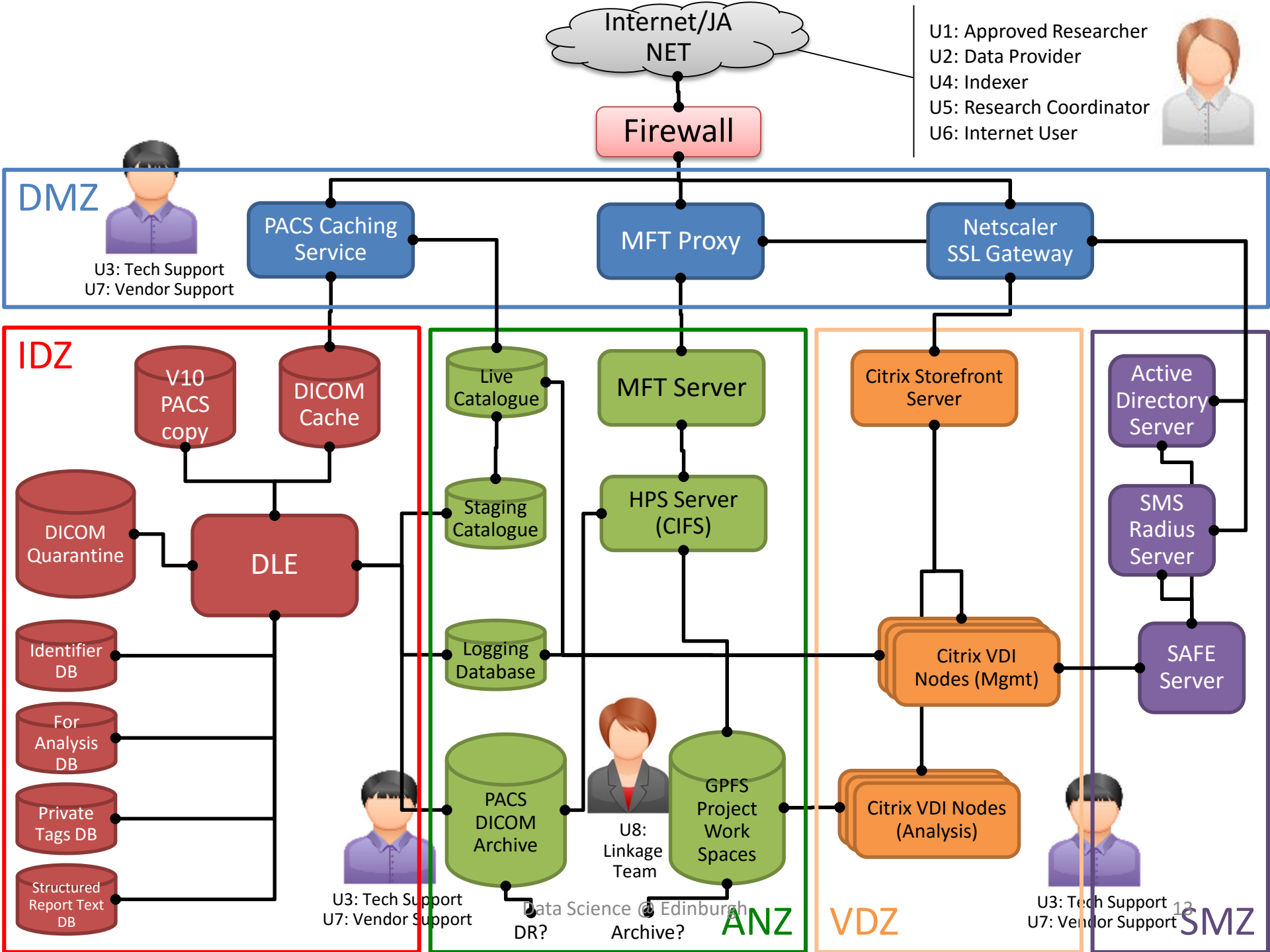
- Our ability to store data is out-stripping our ability to manage data
- Large supercomputers today are always slightly broken … but managing this is well understood
- It's less clear with regard to data …
- All EPCC's system challenges today are data related
  - Disk and enclosure failures
  - Unexpectedly poor performance
  - Difficulties with software tools
- But ...

# Data services

- On the RDF we guarantee to look after your data for a minimum of 10 years
- We have a 50Pb tape-based disaster recover system at KB
- But how should we provide data access?
  - Today the RDF is a large data store
  - Limited data services on top
- Big growth area will be proper – user focussed - data services
  - But everyone is struggling with this

# Example: Farr Institute

- Scotland has a nicely sized population ~5.5m
- Has a good history of archiving data
  - E.g. 25m image sets from PACS system
- Joining unique resources for medical and social research – data sets from
  - Farr Institute
  - Administrative Data Research Centre
  - Urban Big Data Centre
- Pseudonymised unconsented public data for research


Internet/JA NET
Firewall
U1: Approved Researcher
U2: Data Provider
U4: Indexer
U5: Research Coordinator
U6: Internet User
DMZ
U3: Tech Support
U7: Vendor Support
PACS Caching Service
MFT Proxy
Netscaler SSL Gateway
IDZ
V10 PACS copy
DICOM Cache
DICOM Quarantine
DLE
Identifier DB
For Analysis DB
Private Tags DB
Structured Report Text DB
Live Catalogue
Staging Catalogue
Logging Database
MFT Server
HPS Server (CIFS)
PACS DICOM Archive
GPFS Project Work Spaces
U8: Linkage Team
DR?
Archive?
ANZ
Citrix Storefront Server
Citrix VDI Nodes (Mgmt)
Citrix VDI Nodes (Analysis)
VDZ
Active Directory Server
SMS Radius Server
SAFE Server
SMZ

# Summary

- Edinburgh Data Science
- EPCC and its ACF data centre
- Data management is really hard
- Edinburgh has a unique set of resources **and** skills to work in the Big Data area

Questions?