



Keystones for supporting collaborative research using multiple data sets in the medical and bio-sciences

David Fergusson


Head of Scientific Computing

The Francis Crick Institute



The Francis Crick Institute



- 
- The development of genomic and personalised medicine is having profound effect on the bio-medical research process.
 - Increasingly biomedical researchers have to link data relating to different aspects of a disease and representing a broad range of scales (ie. genomic, phenotypic, physiological, population).
 - This means accessing, synthesising and analysing disparate data sources, some of which may have significant access restrictions.
 - To do this a new paradigm of results as a service is required.
 - In order to support this demand networks and data source providers need to be able to present a seamless access model.



Biomedical data characteristics

How Much Physical Space?

Data Storage requirements for 100,000 Genomes


- 100,000 Genomes (3Gb per genome)
- 30 X coverage
- Minimal variant base calling (minimal overhead)
- Data replicated to protect over multiple years (2:1)

- 3Gbases per human genome = approx. 3Gbytes of data

3×30 (coverage) = 90 Gigabytes X 100,000 genomes = 9,000,000 Gbytes = 9,000 Terabytes or

9 Petabytes (PB) - single copy

Replicated (2:1) = 18 PB



Estimating operational capacity requirements - Crick

- Sequencing 1Pb pa
- High throughput screening 0.5Pb pa
- Proteomics (MS) 2.6Pb pa
- Microscopy 0.65Pb pa
- Electron microscopy 0.2Pb pa (rising to 10s Pb Pa)
- MRI 2.64 Pb pa
- Other imaging 0.1Pb pa
- All other sources 1Pb pa

Approximately 8 Pb pa in total



Increases in requirements

- Sequencing expected to increase rapidly
- EM – techniques to capture 60Pb per mouse on the near horizon
- Proteomics – expected increases could be up to 26Pb pa
- Expect other techniques to come along
- 100s of Pb pa expected within the next 5 years.

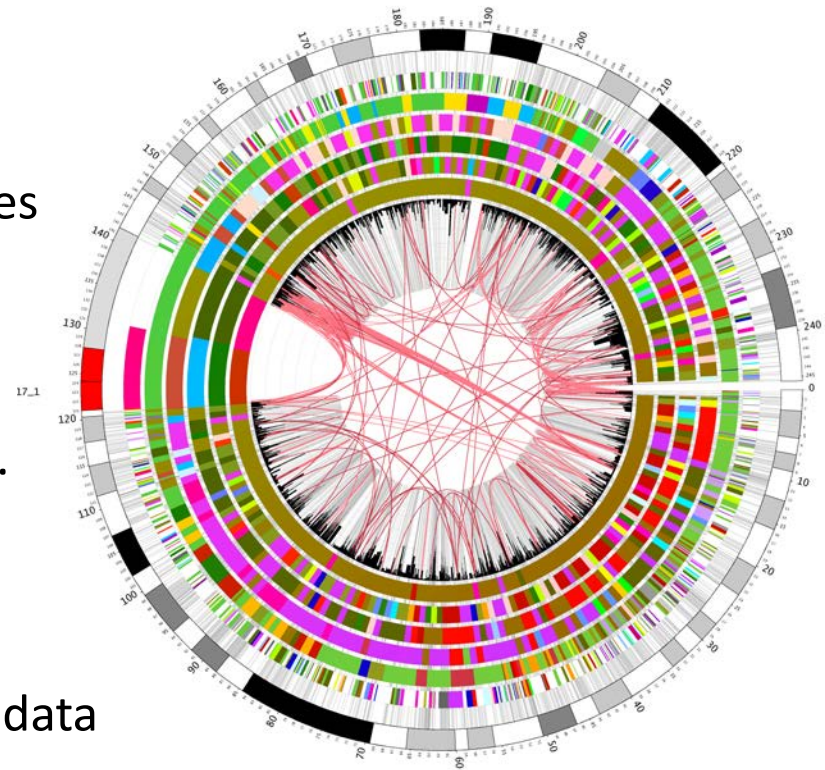


Sensitive and Identifiable Data

- Identifiable data
 - Allows identification of an individual
 - Note not always obvious : eg. for rare diseases (small population) this can be a collection of symptoms
 - Will sequence data become identifiable?
- Sensitive data
 - Data that, in conjunction with other data or knowledge could identify an individual

Complex Data

- Complex data / Complex analytics
- Distributed data in numerous data stores
- Clinical Data presents new challenges
- Legal, ethical, transmission security etc.
- Managing and tracking the data
- Securing and auditing access to clinical data
- Scale of the data involved



Challenge: To develop the tools/infrastructure/middleware in a common way as opposed to the many groups developing strategies independently and across the globe.



Changing the dynamic

- Data centric not compute centric.
- Data problems are harder to deal with than compute problems.
- Data is hard (expensive) to move.
- Data requires curation (provenance).
- Big data silos – trusted data suppliers
- Move the compute to the data
- Provide services around data (SaaS)
 - Improve speed
 - Streamline workflows
 - Support better data practice
 - (no opportunity to leave CDs on trains)



Global Alliance

- Initiative announced 5th June 2013
- <http://www.sanger.ac.uk/about/press/2013/130605.html>
- More than 60 leading health care, research and disease advocacy organisations from across the world are joining together to form an international alliance dedicated to enabling secure sharing of genomic and clinical data.
- Each of these organisations has signed a 'Letter of Intent', pledging to work together to create a not-for-profit, inclusive, public-private, international, non-governmental organisation (modelled on the World Wide Web Consortium, W3C) that will develop a common framework.
- Organisations from North and South America, Europe, Asia and Africa have joined together to form a non-profit global alliance which will work on a common framework of international standards designed to enable and oversee the sharing of **genomic and clinical data** in an effective, responsible, and interpretable manner.
- It will consider legal, ethical issues and PLC - Portable Legal Consent.



Global Alliance

- Data regulations differ across the globe – U.S., EU, Far East, Scandinavia, etc.
- The framework does not mandate open data but it does promote the concept of an open platform for data sharing i.e. common APIs, standards etc., that will be derived to promote conformity between global organisations and allow data to be shared, analysed in a secure and consistent manner.
- Individuals will have the right to revoke legal consent electronically which will remove tracked data or revoke access mechanisms according to local regulation.
- Many UK organisations have signed LOI including Wellcome Trust, CRUK, NIHR, Oxford Uni, Sanger, EBI and many more.



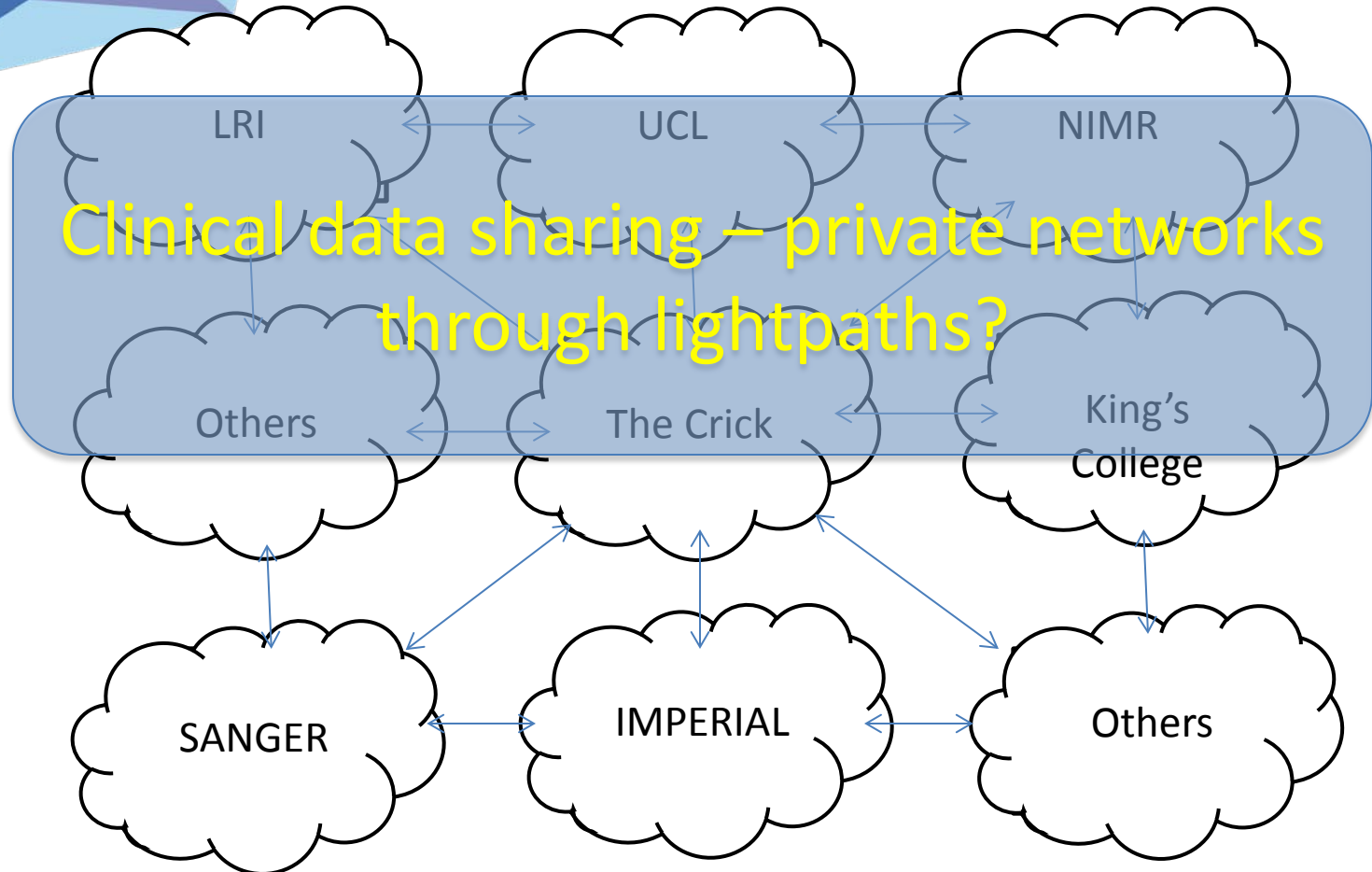
Trust networks

- Trust networks to support “big computation” have been created and shown to work.
- Big Data is a new opportunity to base these around shared data resources.
- Just as “big computation” was (and is) out of reach for many organisations – so is big data for many.



Into the Cloud

Community Cloud Model Offsite Data Centre



UK JANet pilot projects expected this year.

ELIXIR/CSC (Finland) have come to the same technical solutions independently. Hope to collaborate between UK and Finland to extend the connections.



Shared Co-location

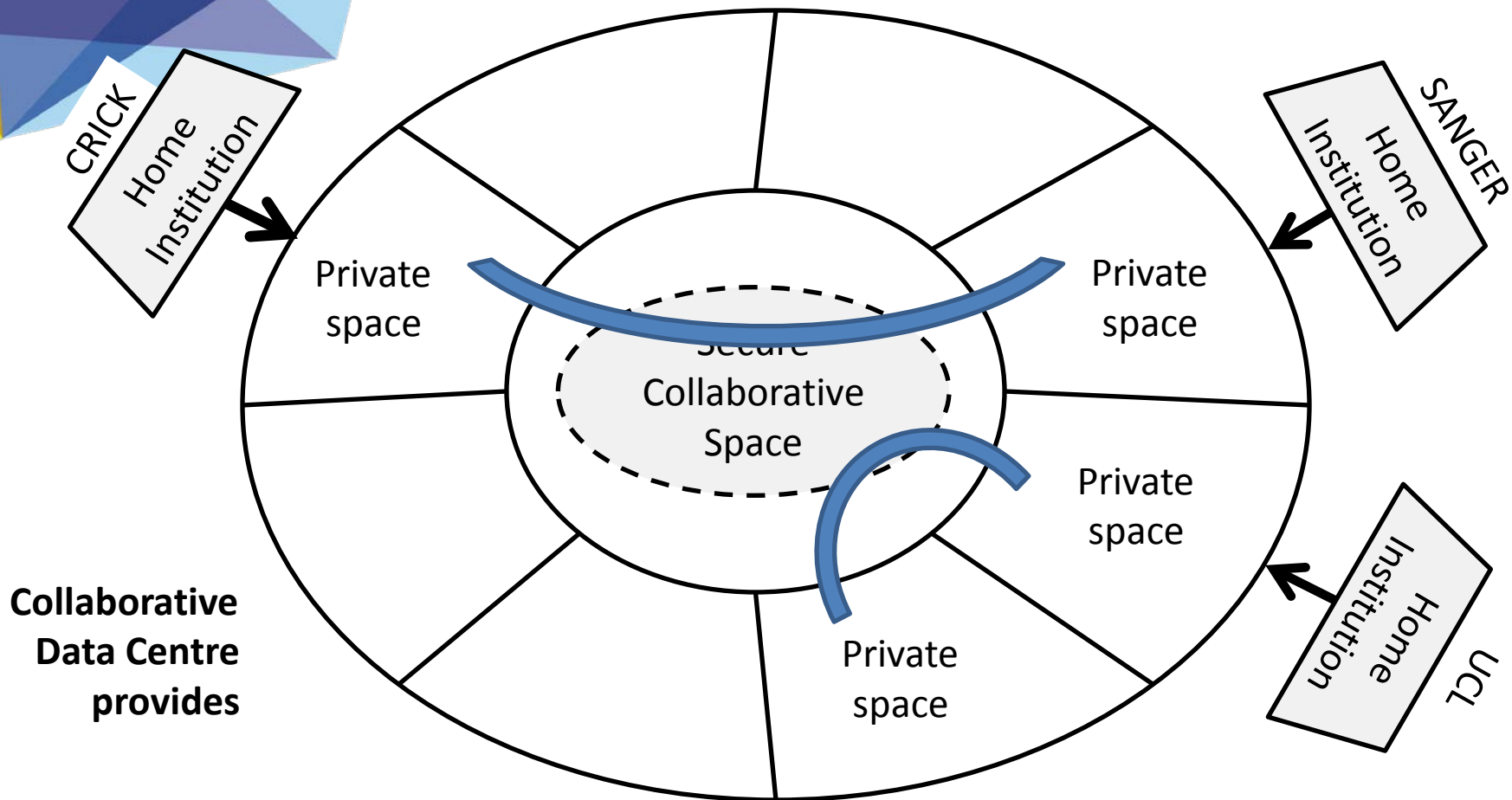
- JANet framework – any research organization can contract with supplier without full OJEU process.
- Anchor tenants:
 - UCL, Kings, LSE, QMUL, Crick, Sanger
- Interested:
 - Bristol, Cancer Research Institute, Imperial
 - Genomes England?
- Physically co-locating large data sets to allow secure shared computation across them.



Offsite Data Centre/ Collaborative Data Centre

- We will also have the ability to offer collaborative space for stakeholders and others
- In the future we will want to analyse distributed data sets but this needs work and is a way off
- A joint data centre model provides a platform to not only share data but it acts as a catalyst for collaboration particularly at the infrastructure level
- I believe this is the biggest win initially and that the science will inevitably benefit from this collaborative model
- Examples of this happening in the U.S include:-
 - **CGHub** – David Haussler - Santa Cruz – have installed a cluster local to the hub to provide an analysis engine close to the data
 - **New York Genome Centre** - Identical IT strategy – onsite/offsite and providing central computation for 10+ stakeholders

Collaborative Data Centre - eMedLab



Private colocation (traditional) – Logical Extension to local LAN
Collaborative/Shared space, **Secure space** for sensitive data (patient data)

Unique, powerful centre to build, test, deploy new infrastructure tools between Organisations. **HPC where the data resides!!!!**



Technologies

- Cloud Technologies

Cloud – usually means public cloud..... Amazon, Google, MS-Azure etc....

OpenStack – current leading contender, also vCloud

but OpenHelix/Nebula out there too

- All Clouds are big data centres with excellent engineering, good networks and middleware for job distribution, compute virtualisation, data management and accounting.
- In the future it is highly likely that the ‘cloud’ will become easier to use and we will develop tools for large scale sensitive data to be contained and analysed securely.
- Cloud is not the panacea for all our needs and in fact for many ‘sensitive data sets’ Public Cloud may **not** be the answer at all.....



Collaborative Data Centre

A central place where collaborating institutes can:

Expand their own user space Choose to work with others in a shared environment

While distributed datasets and distributed analysis tools are the eventual aim we need stepping stones to get there. CDC provides a real world environment.

Improving algorithms often requires large memory machines – once understood the problem can be split into smaller parts for high throughput analysis.

Secure data management needs development, agreement on common standards and testing between organisations.

Secure data environments can be purpose built from the ground up

Life sciences institutes and NHS must work together to ensure a truly joined up approach otherwise we will all be chasing an elusive goal.




SAFESHARE

- Drivers
- Requirement for connectivity to move and access electronic sensitive data securely
- Challenge to give public confidence that data is appropriately protected
- Provide economies of scale in secure connectivity
The safe share project
- Jisc management and funding of £865k to pilot potential solutions with the aim of developing a service in 2016/17




Partners

- The Farr Institute
The MRC Medical Bioinformatics Initiative
The Administrative Data Research Network
- Pilot institutions
- University of Bristol University of Leeds University of Oxford Swansea University
- University of Dundee UCL
University of Sheffield
- Francis Crick Institute University of Manchester
University of Southampton



1. Secure connectivity with a higher assurance network (HAN)

- Use Cases:
- Inter-Farr – initial trial between Farr centres at Manchester and Leeds, then extending to other Farr centres (London, Scotland and Swansea)
- Intra-Farr – to support the ALSPAC project between Swansea and Bristol
- ADRC / Farr Pod to Data Centre – connectivity between accredited secure rooms that can be connected to ADRC data centres for remote working, University of Southampton




2. Authentication, Authorisation and Accounting Infrastructure (AAAI)

- Use Cases:
- Dementia Study – proof of concept– objective to enable researchers to use home institution credentials to authenticate to request access to datasets, University of Oxford
- HeRC,N8, HPC, DiRAC – access between facilities using home institution credentials
- eMedLab – partners will be able to use a common AAAI to access this new system (for analysis of for instance human genome data, medical images, clinical, psychological and social data)
- Swansea University Health Informatics Group – investigating Assent (Moonshot technology) as an authentication mechanism to allow use of home institution credentials




IN CONCLUSION



Collaborative Space – Life Science Hub – eMedLab & beyond (?)

- Promote Skills Development (Systems, Informatics)
- Prototyping and deploying standards across multiple entities (Global Alliance)
- Promotes collaboration (both at IT and Informatics levels – faster development, less duplication of effort – de-facto standards)
- Produce real world infrastructure tools (production use across collaborating partners)
- Provide Sandboxes (testing development)
- Attractive to Industry partners (hardware evaluations, new technology deployment)
- Prototype public cloud techniques in private setting (safe environment)
- Safe Haven for sensitive data that should not move to public cloud
- Provide easier access to larger data sets.
- Pooled resources maximise Capital investment – benefits for small and large user



AAI, Networks and sharing

- Trust = robust AAI
- Robust AAI required for sharing
- Secure information cannot be moved
 - *Results as a Service*
- Sharing data within a “separate” secure network layer
- Sharing and synthesising data is crucial to scientific progress.
- Infrastructure’s job is to enable this.

Thank You

