



Life Sciences and Large Data Challenges

David Fergusson

Head of Scientific Computing

The Francis Crick Institute





WHAT IS THE CRICK?

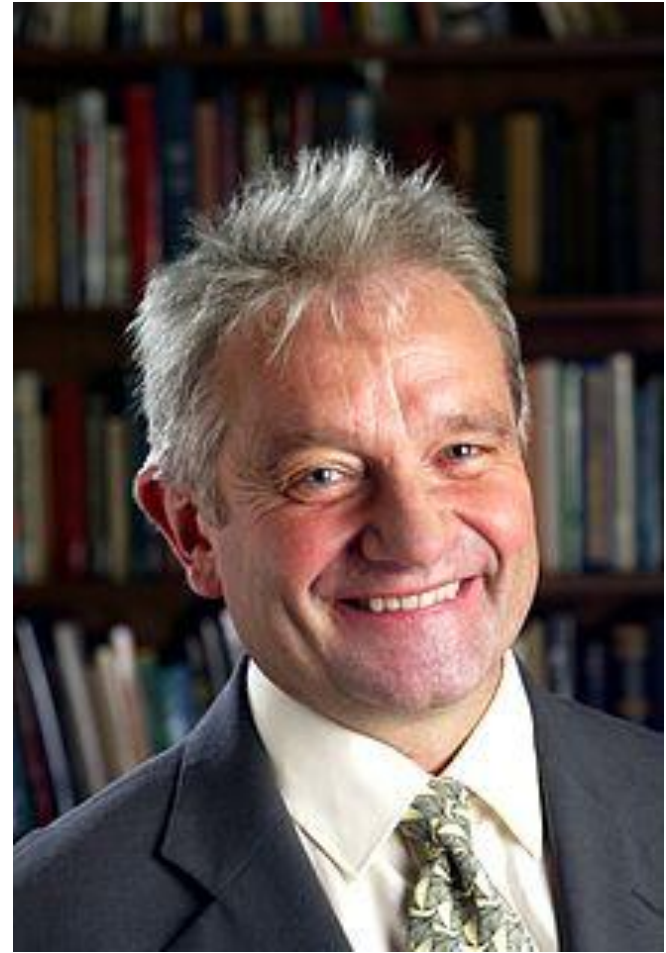
The Francis Crick Institute



Sir Paul Nurse

Nobel Prize with Hartwell and Hunt for discovery of cyclins and CDK which control the cell cycle.

President of the Royal Society and Chief Executive and Director of the Francis Crick Institute.



Synthesis of two institutes

- National Institute for Medical Research (NIMR) – MRC

- Nobel Laureates - Sir Peter Medawar, Sir Frank Macfarlane Burnett, Sir Henry Hallett Dale, Archer John Porter Martin
- Dame Margaret Thornton



- London Research Institute (LRI) - CRUK

- Nobel Laureates - Renato Delbecco, Paul Nurse, Tim Hunt





Partners

- Wellcome Trust
- Medical Research Council
- Cancer research UK

- University College London
- King's College London
- Imperial College London

Crick Vision

- 1) Pursue discovery without boundaries
- 2) Create future science leaders
- 3) Collaborate creatively to advance UK science and innovation
- 4) Accelerate translation for health and wealth
- 5) Engage and inspire the public

Scientific Computing Vision

Support Scientists

- Platforms to accelerate transformations
- Improve analysis
- Data Security
- Cooperate to develop novel methods

Collaboration

- Secure shared data
- Shared best Practise
- Platforms for national & International biomedical collaboration

Engage & Inspire Public

- Focussed examples
- Support science curricula
- Support Crick comms activity

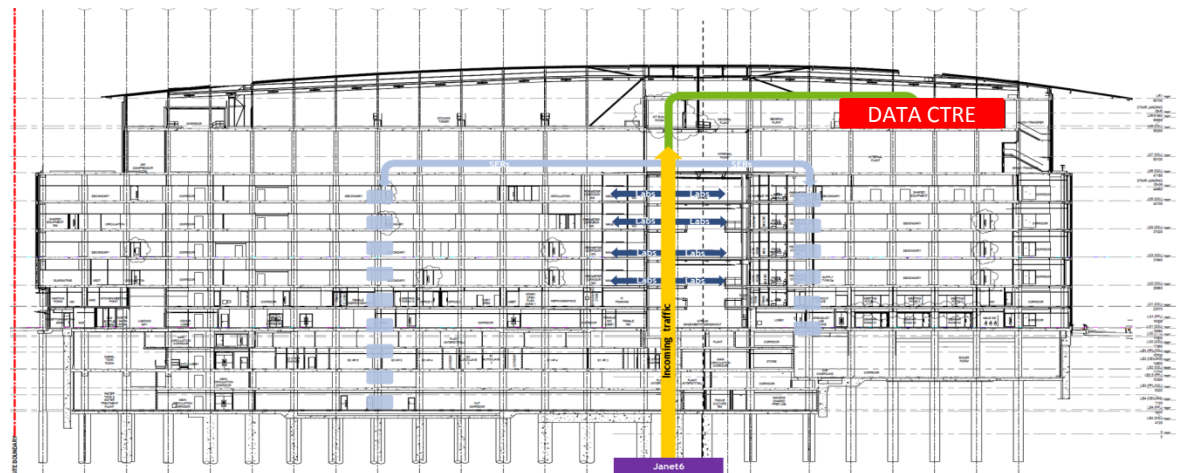
Promote Training

- Safe and sophisticated teaching environments
- Best Practise exemplars
- Expand experimental horizons
- Streamline workflows

Data Centre

On-site/Off-site Strategy

- **On-site/Off-site strategy** to be implemented from day one
- Purposely designed so that the physical building does not limit the computational needs of the science
- On-site Data Centre - modest size, designed to hold just 40 high density racks
- Location: rooftop to take free-air cooling, 750kW of power. Average 18kW per rack.
- Immediate data collection and processing, data staging before transfer or replication to offsite data centre
- It will also host key services for our users and for the building itself.



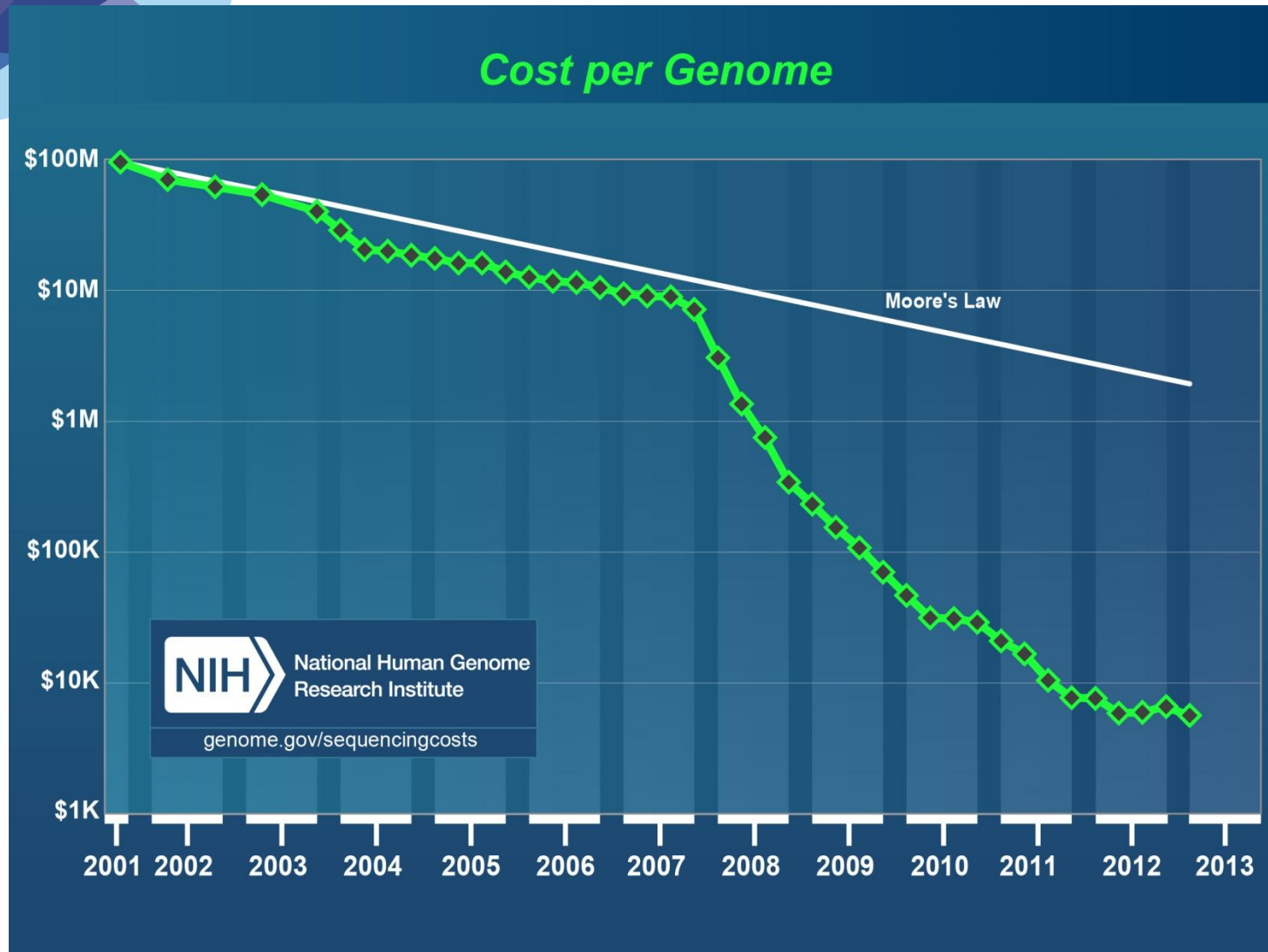


**CREATING A NEW BIOMEDICAL
INSTITUTE:
THE CONTEXT**



CHALLENGES

\$1,000 Genome?? – Not yet



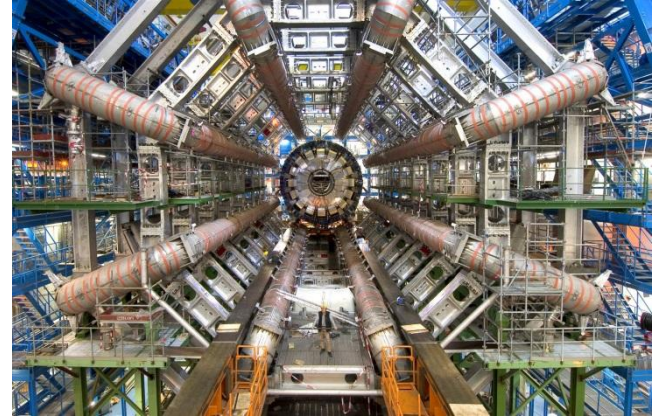
Data from NHGRI Sequencing Program – April 11th 2013

<http://www.genome.gov/sequencingcosts/>

Big Data

- High Energy Physics - CERN Hadron Collider generates big data, > 1Pb per month
- Astronomy – will generate extremely big data (SKA) potentially many Petabytes per day.....Exascale computing
- Life/Biomedical Sciences are generating **a lot of data**

But the potential to generate ever growing volumes of data exists and is set to increase rapidly.



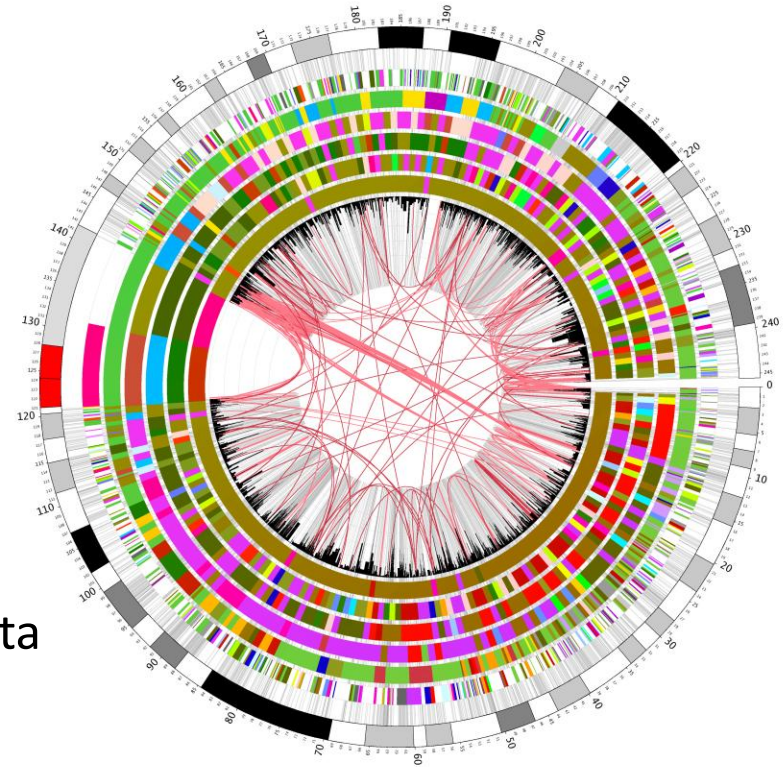


Trust networks

- 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.

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)



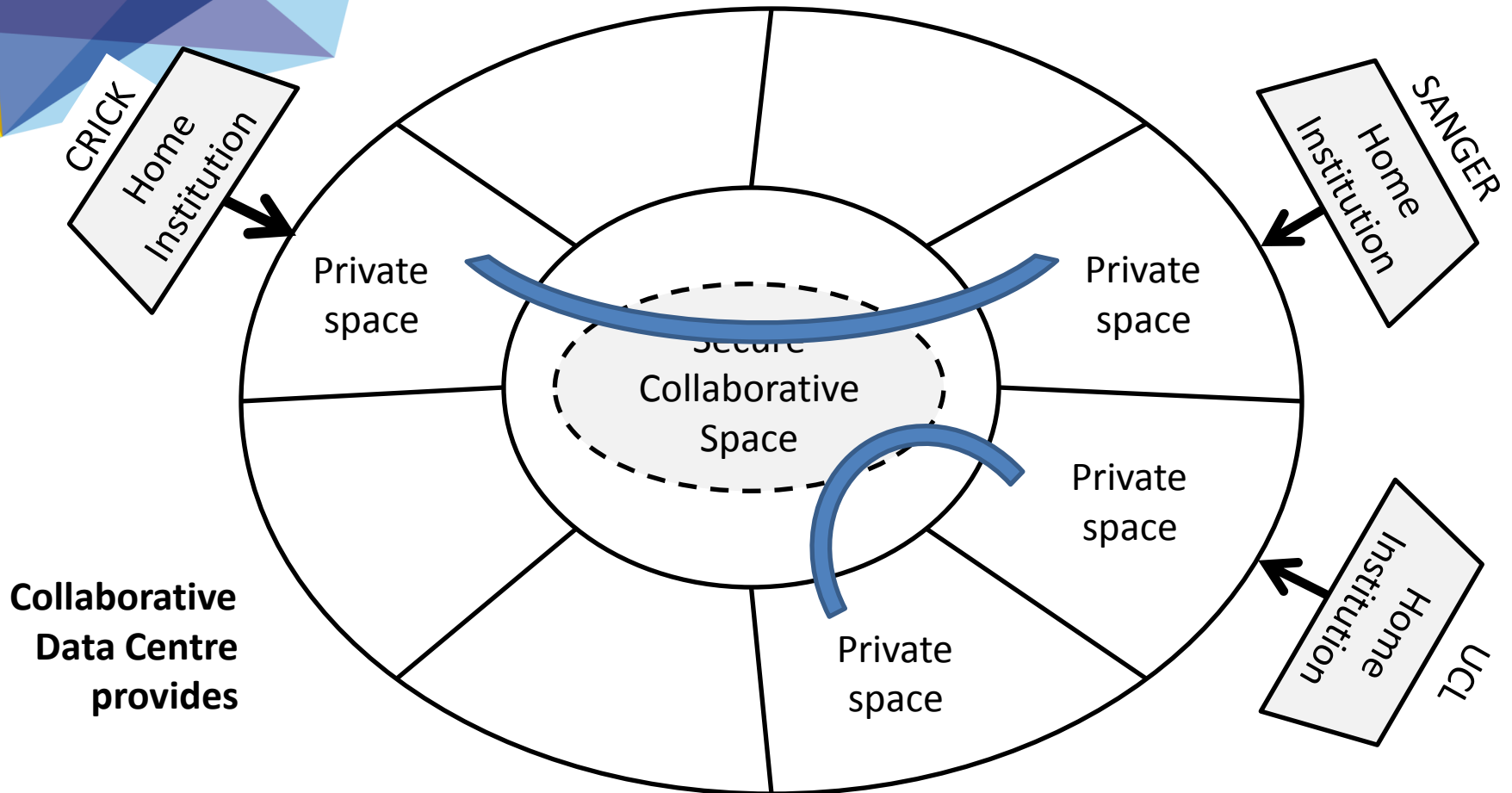
CREATING A DATA CENTRE FOR THE CRICK



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



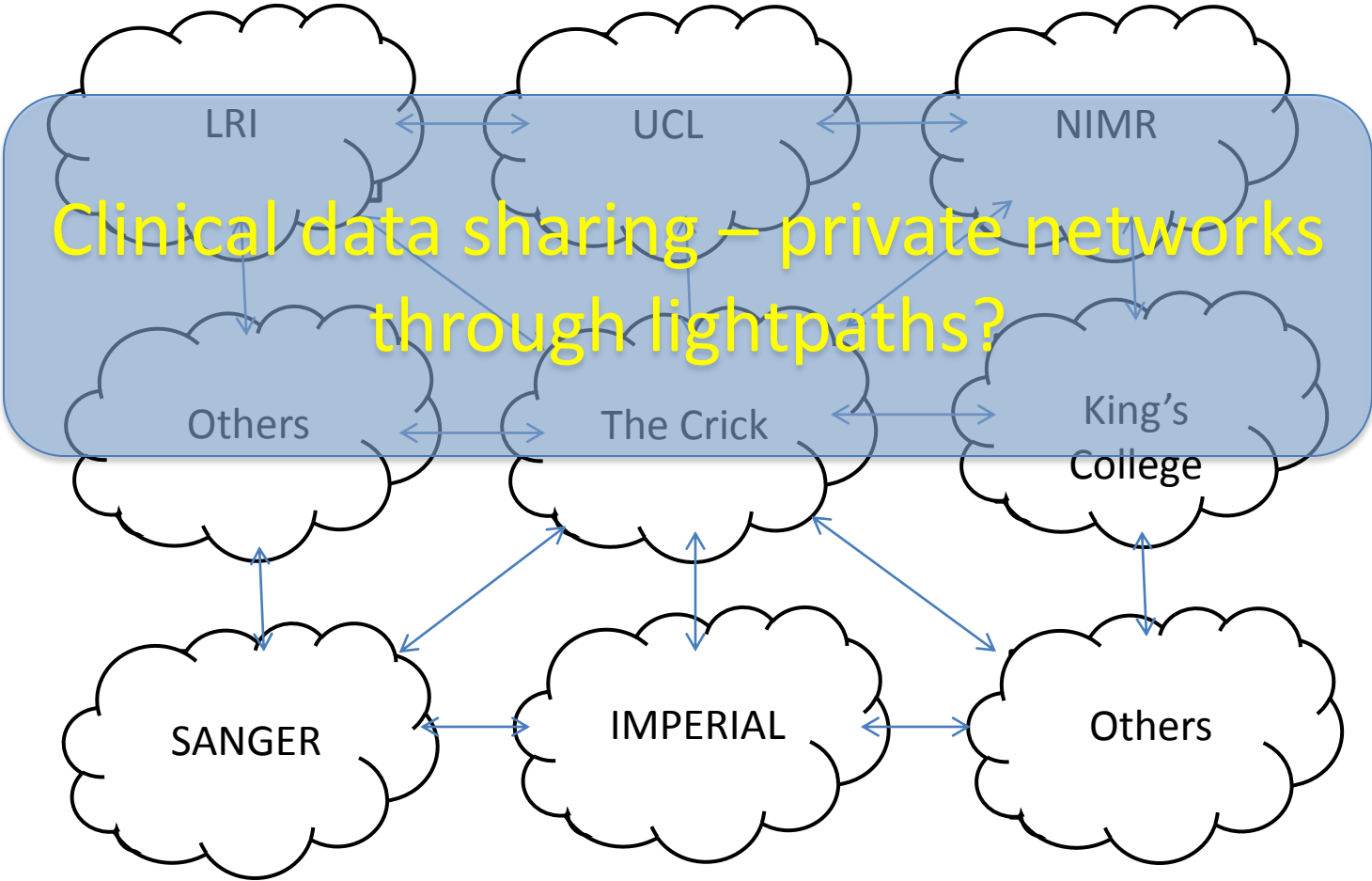
Private colocation (traditional) – Logical Extension to local LAN
Collaborative/Shared space, **Secure space** for sensitive data (patient data)

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




In the Cloud

Community Cloud Model



A decorative graphic in the top-left corner consisting of several overlapping, semi-transparent geometric shapes in shades of blue, purple, yellow, and green.

IN CONCLUSION



Collaborative Space – Life Science Hub – eMedLab (?)

- 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

Thank You

