Big Data in Health asks for new approaches across disciplines

Dr David Fergusson, Head of Scientific Computing.

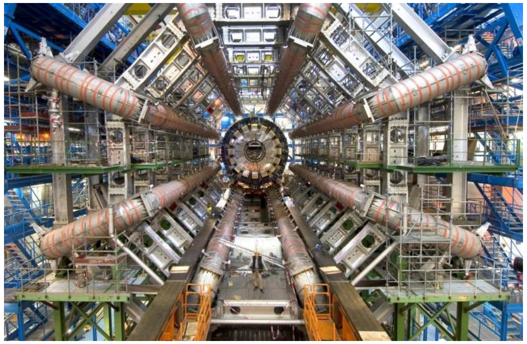
The Francis Crick Institute

The challenges…

THE FRANCIS CRICK INSTITUTE

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

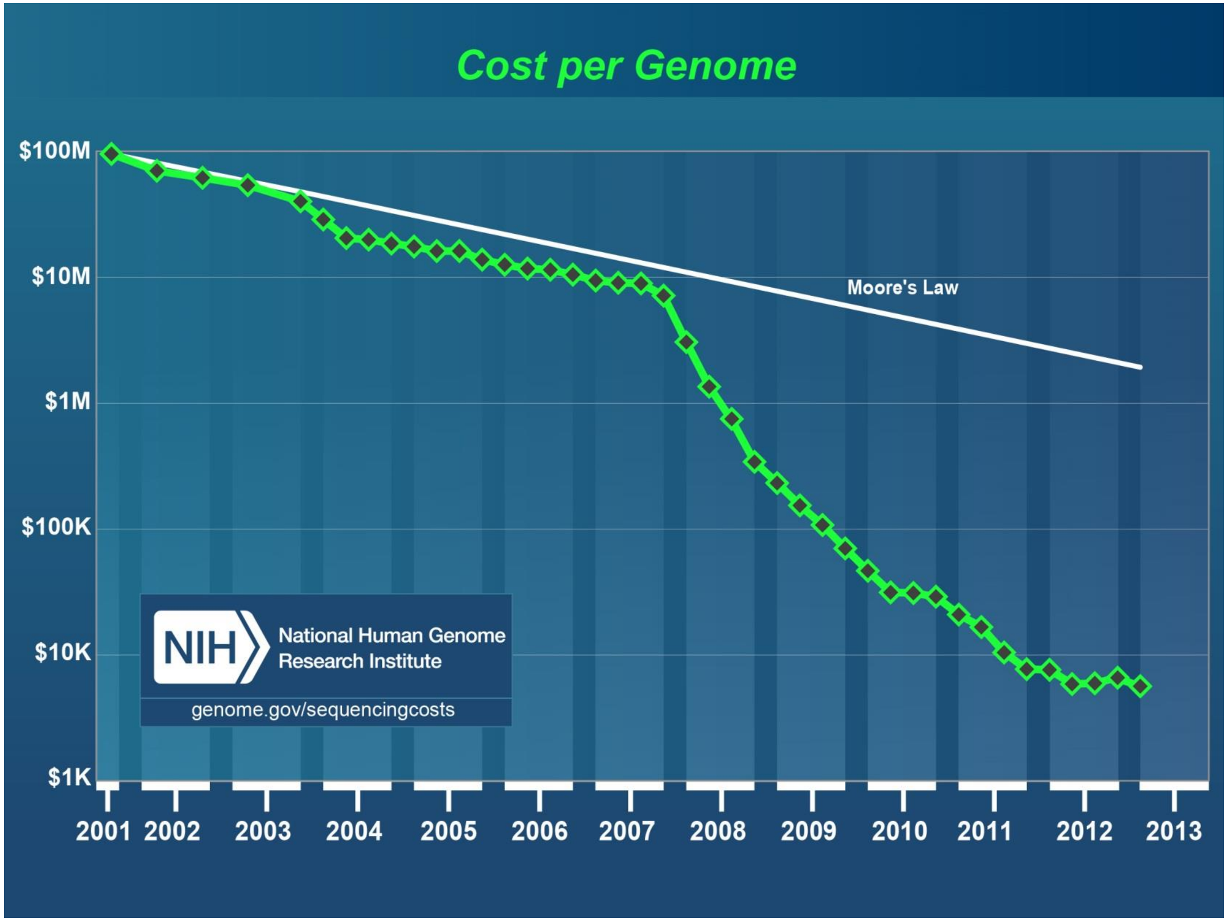
- Life/Biomedical Sciences are generating a lot of data

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

SGI DMF: Addressing data explosion for over 20 years

IVEC, Square Kilometer Array	- 100.0 PB
NASA Ames (40 GB/sec) (21 years online)	- 60.0 PB
GFDL/NOAA (300 TB/day I/O – 105GB/s NFS throughput) (Weather)	- 50.0 PB
CSC, Finland	- 30.0 PB
Double Negative – (Movie visual effects)	- 30.0 PB
WETA Digital Ltd. (Movie visual effects – 1.8 Billion files)	- 24.0 PB
NASA Goddard (21 years online)	- 20.0 PB
Australian National University	- 20.0 PB
NBA Digital Media Management (~40TB/day ingest)	- 18.0 PB
CESNET (Czech Republic)	- 15.0 PB
Météo France (13TB/day) – With Lustre	- 10.0 PB
CSIRO Australia (21 yrs in prod, always online)	- 8.0 PB
National Geographic Film Library	- 7.0 PB
DERM, Queensland, Australia	- 7.0 PB
TOTAL - French Oil and Gas	- 5.0 PB
Monash University, Australia	- 5.0 PB
INA (French National Institute for Audio & Video)	- 4.5 PB
LHC Tier-1 Site, SARA (Netherlands)	- 4.0 PB
IDRIS (French National Research Agency)	- 4.0 PB
CINES (GENCI)	- 4.0 PB
British Petroleum	- 2.7 PB
Boeing	- 2.0 PB
Earth Data	- 1.7 PB
Pittsburgh Super Computing	- 1.6 PB
SARA Computing and Network Services (Netherlands)	- 1.5 PB
ICR, UK	- 1.1 PB

$1,000 Genome?? Not yet…but…



Data from NHGRI Sequencing Program – April 11th 2013

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

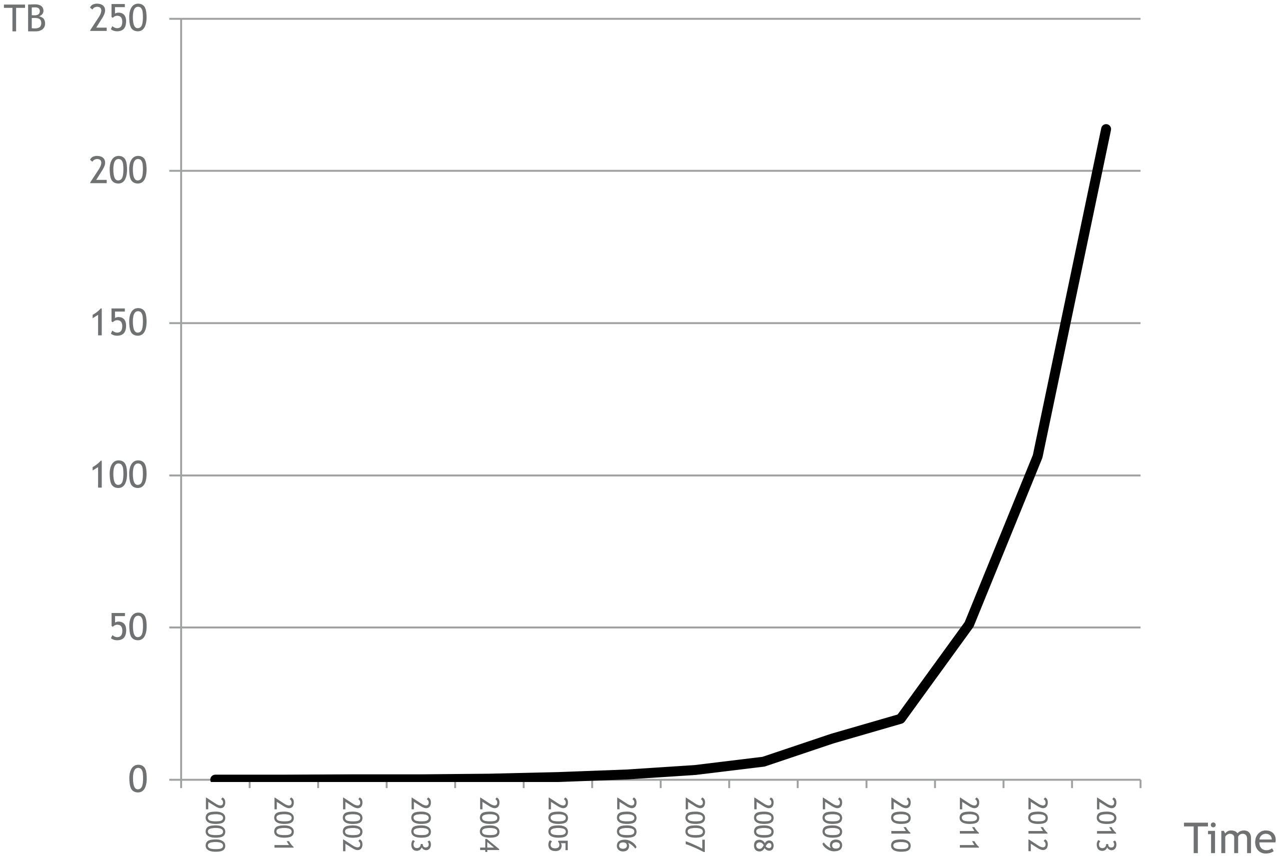
Cost of sequencing is falling

2003

2008

Acknowledgement: Dame Janet Thornton, EBI

Institute Storage Growth Rate



Sequencing…and more…

Merged Three-Channel Image

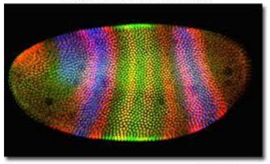
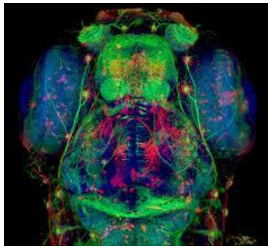


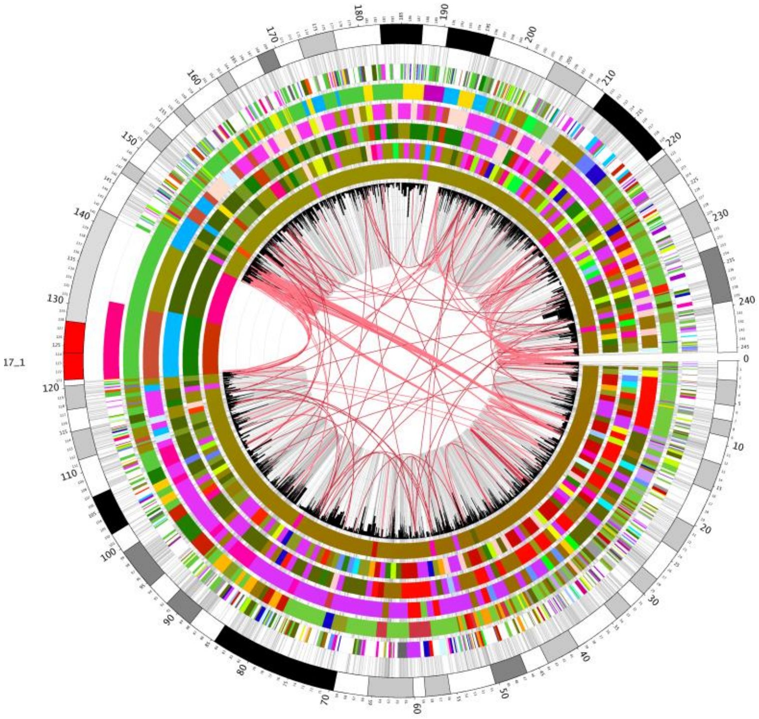
Figure 1



Developing techniques…

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



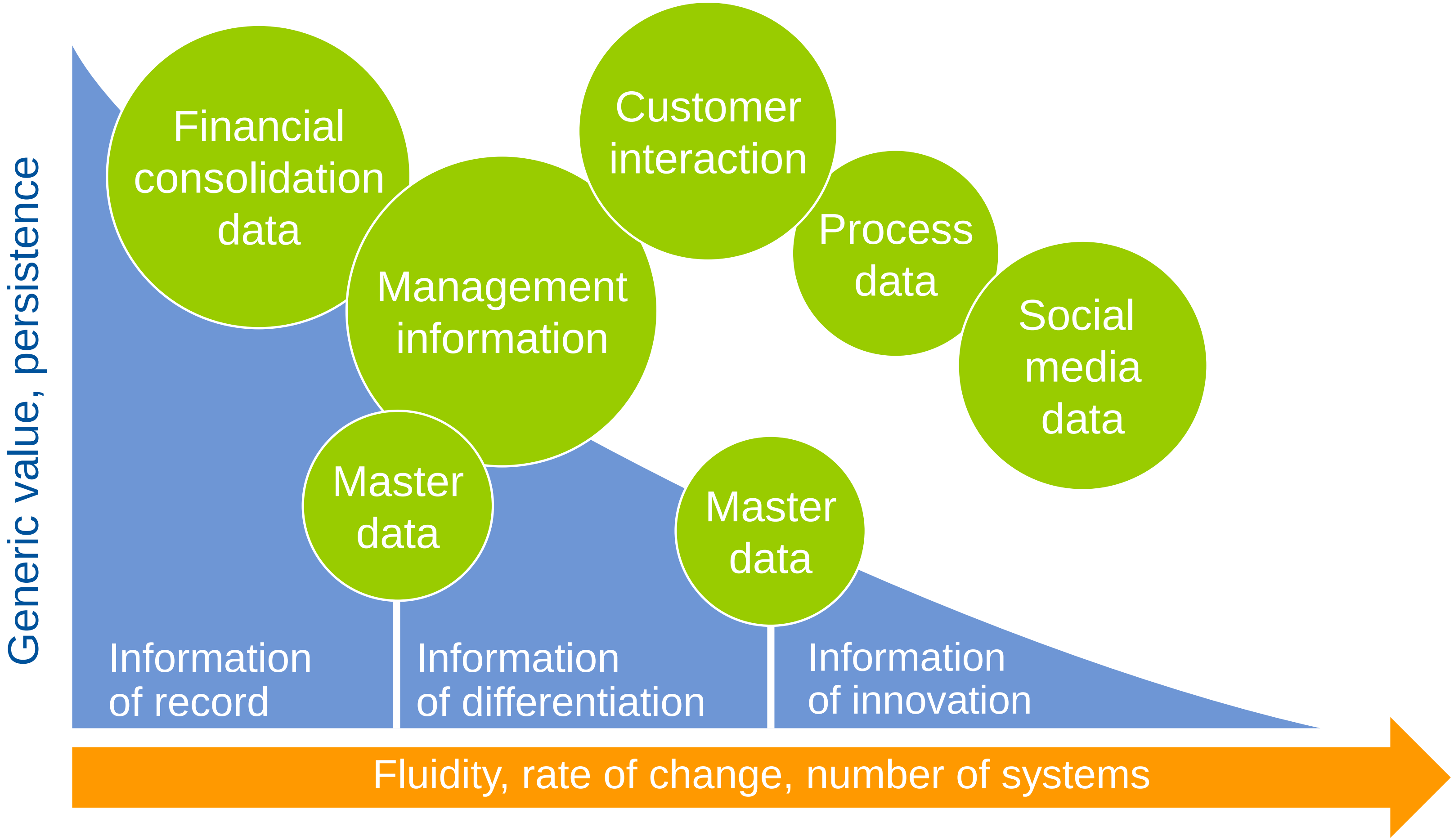
Challenge: To develop the tools/infrastructure/middleware in a common way as opposed to the many groups developing strategies independently and across the globe.

R&D big data is different…sometimes…

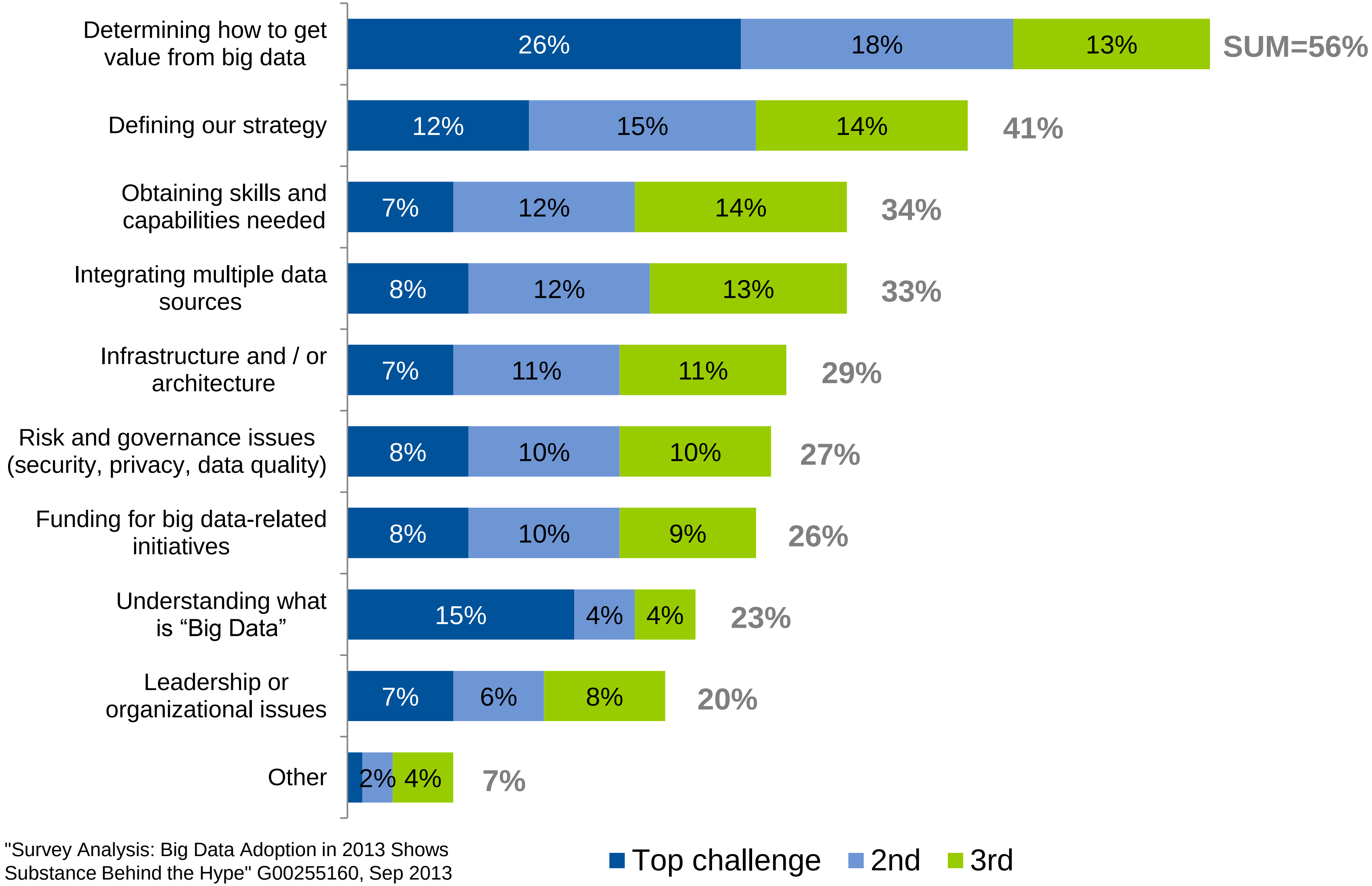
R&D data versus commercial data

R&D Data	Commercial Data
Huge volume	High Volume
High velocity – but inconsistent	Consistent velocity
High variety	Lower variety
Veracity tested by analysis	High veracity is desirable
Analytics add to the data volume	Analytics simplify the data volume

R&D data is not always on the radar…


Financial consolidation data
Management information
Customer interaction
Process data
Social media data
Master data
Master data
Information of record
Information of differentiation
Information of innovation
Generic value, persistence
Fluidity, rate of change, number of systems

Big Data Challenges



"Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype" G00255160, Sep 2013

The responses…

Changing the dynamic

- **Data centric not compute centric.**

- **Data problems are harder to deal with than compute problems.**

- **Data is hard (expensive) to move.**

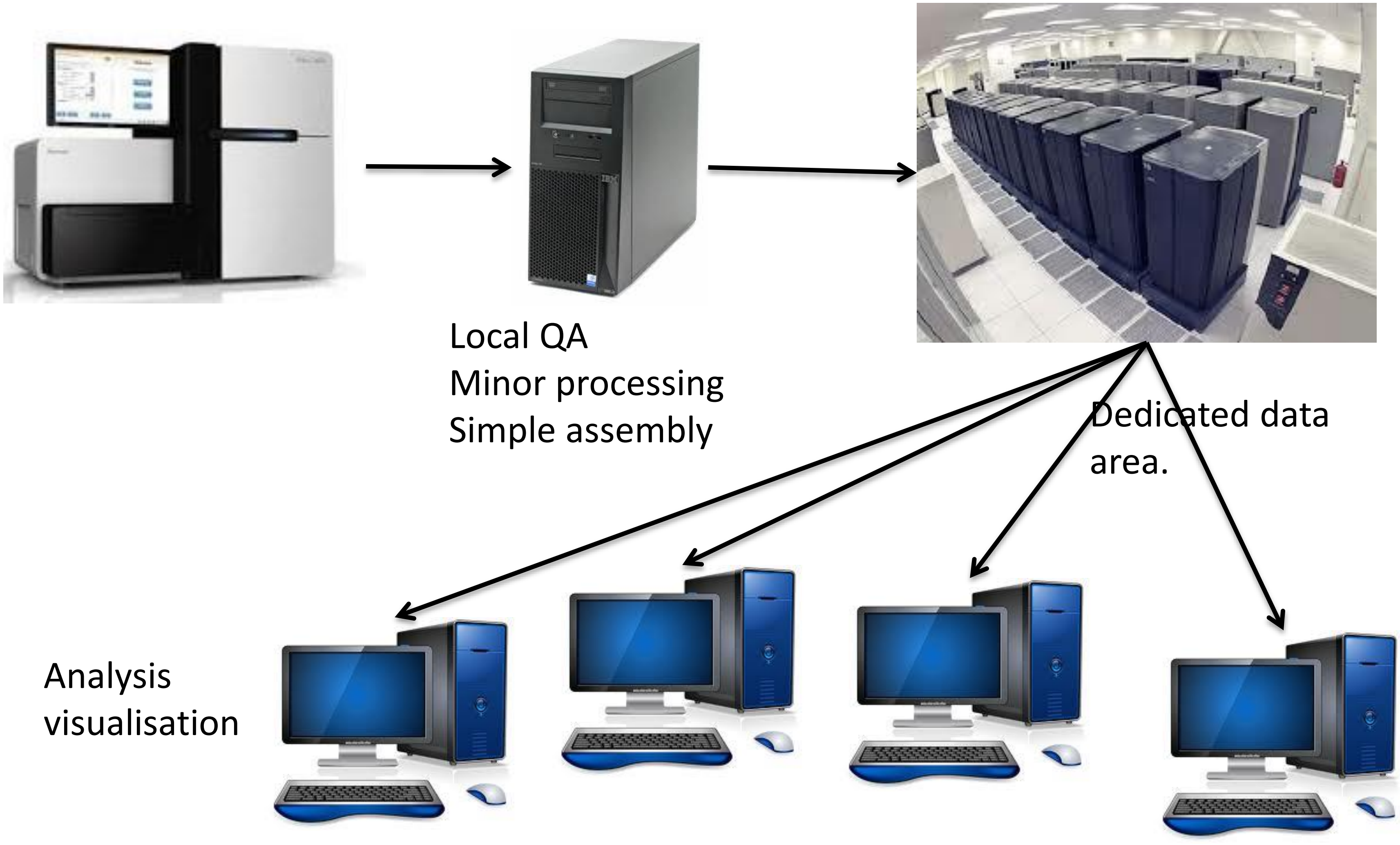
- **Data requires curation (provenance).**

- **Big data silos – trusted data suppliers**

- **Move the compute to the data**

- **Provide services around data (SaaS)**
  - Improve speed
  - Streamline worksflows
  - Support better data practice

Sequencing pipeline



Complex analysis, de novo assembly

Local QA
Minor processing
Simple assembly

Dedicated data area.

Analysis visualisation

Imaging pipeline

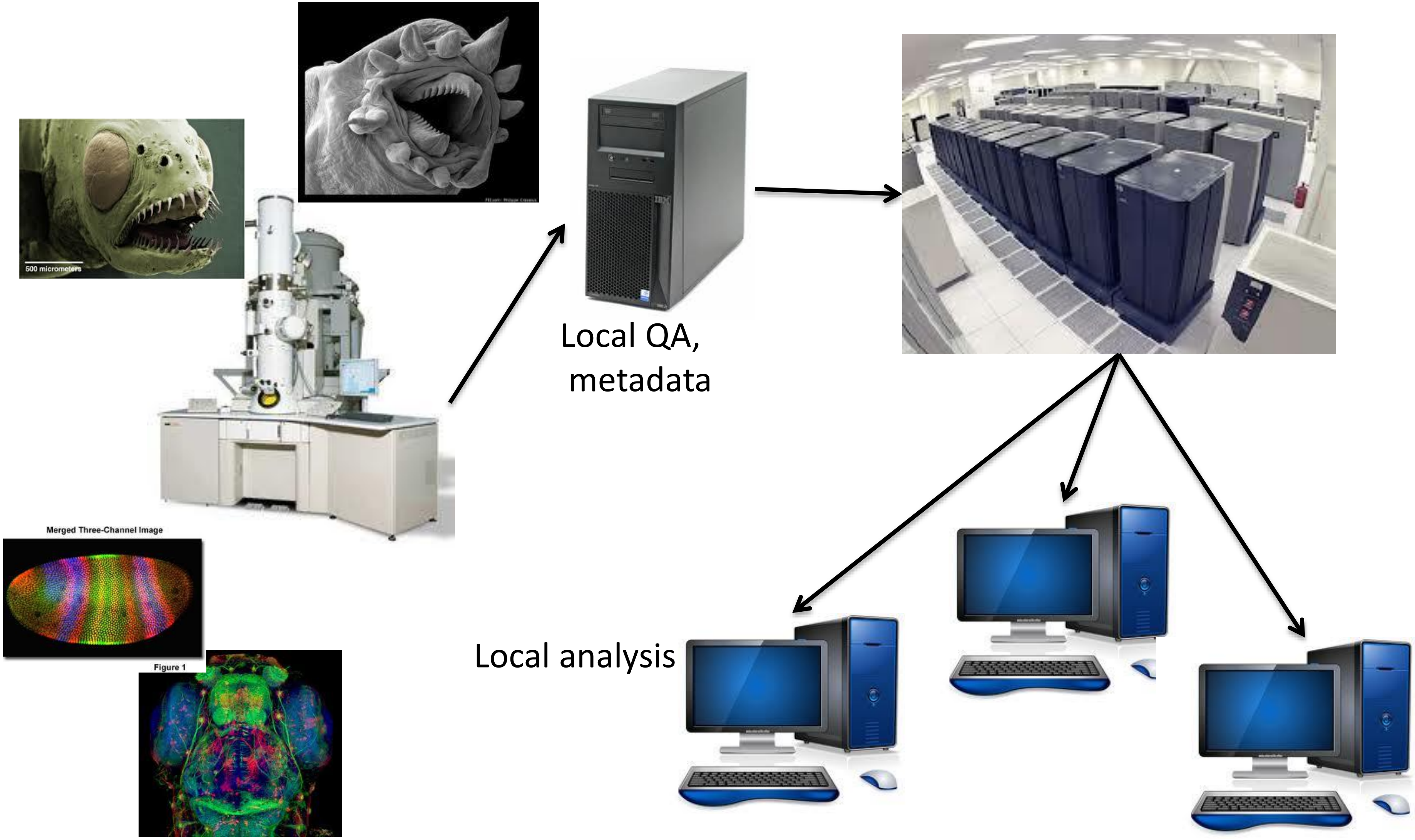


Image DB, re-indexation

Local QA, metadata

Local analysis

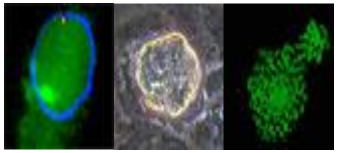
Science and IT

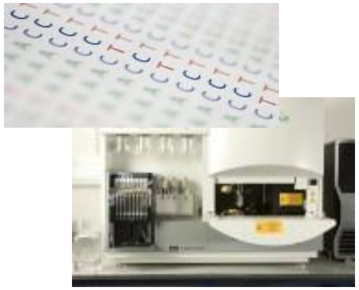
Mouse model Organisms

Electron Microscopy (Image Data)

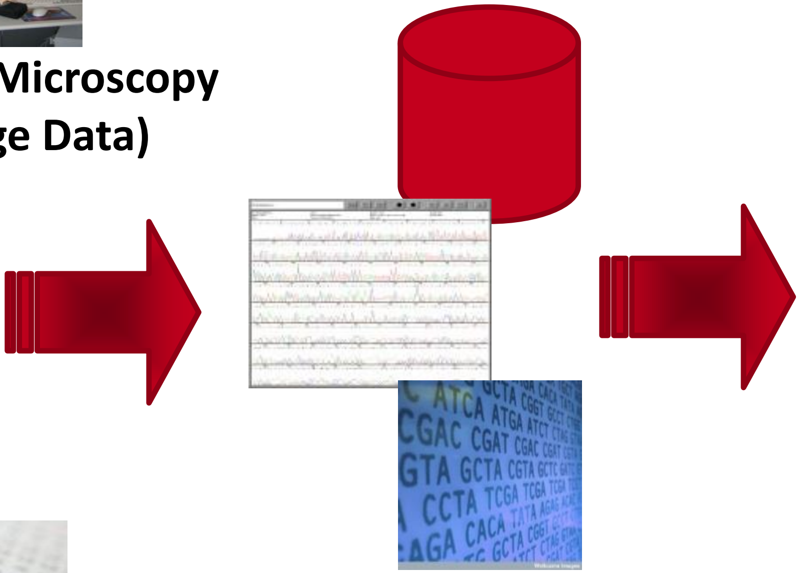
Pathogens

Cellular Science





Sequencing (Genomic Data)



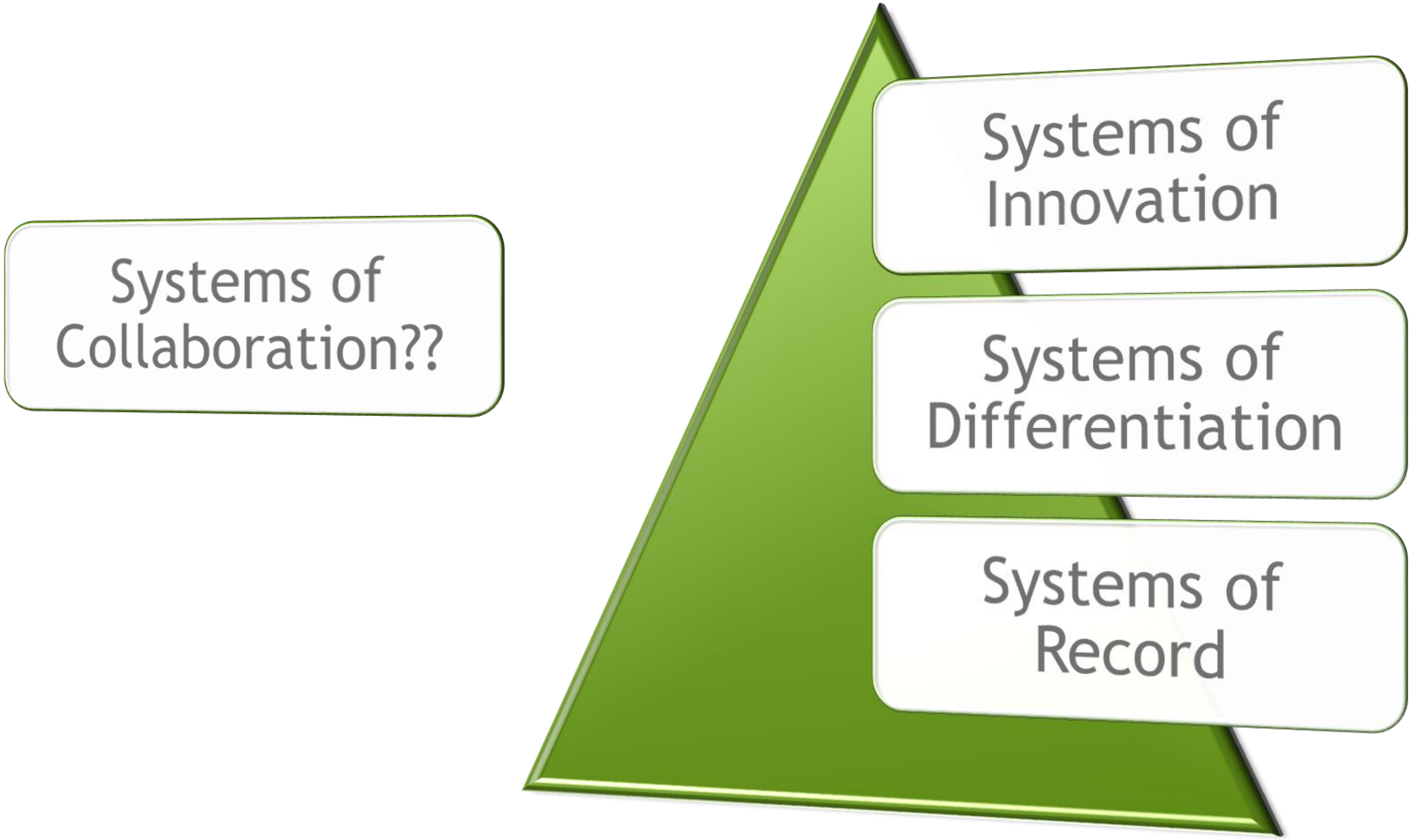
Bioinformatics, IT, Databases, Systems, HPC

Share data with Scientists (Local, National & International)

Collaborate with other Scientific Institutes

Build translational relationships with Clinical Partners

Organising for big data



Systems of Collaboration??

Systems of Innovation

Systems of Differentiation

Systems of Record

Gamification of big data

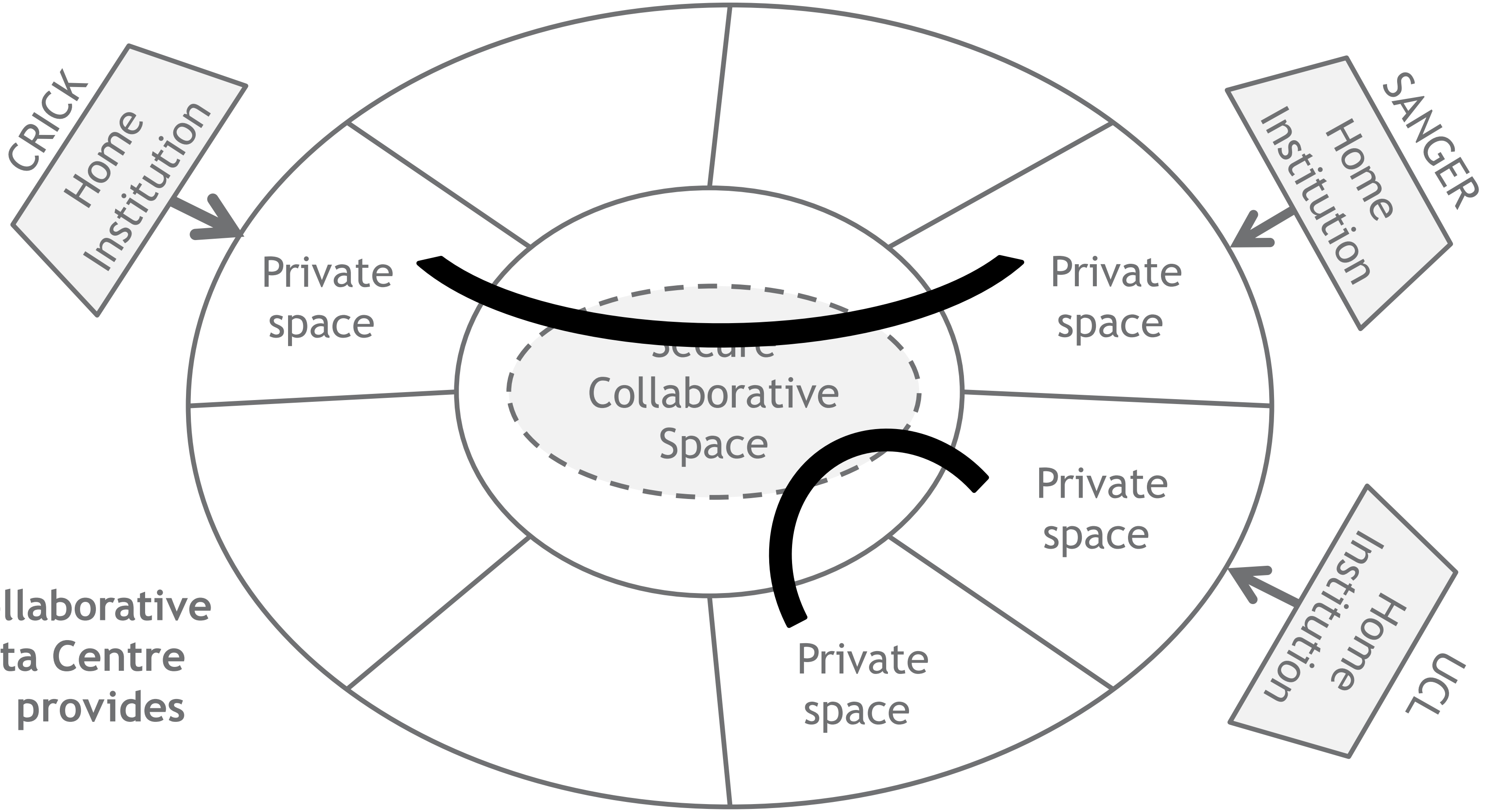
Trust networks

- 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.

Collaborative data approaches

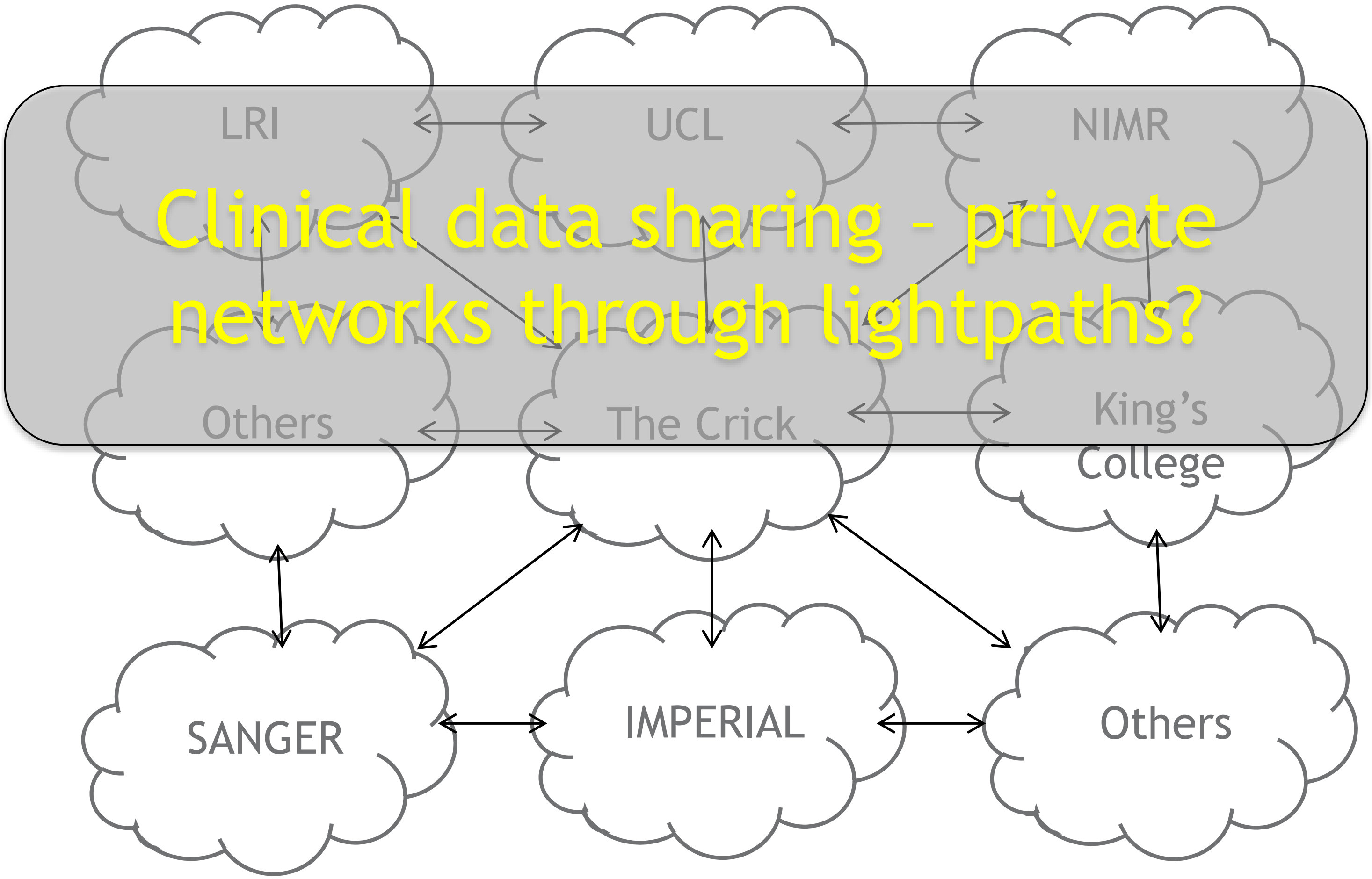
- **In the future we will want to analyse distributed data sets but this needs work**
- **A joint data centre model provides a platform to not only share data but it acts as a catalyst for collaboration particularly at the infrastructure level**
- **Believe that the science will inevitably benefit from 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 providing central computation for 10+ stakeholders

Collaborative Data Centre – eMedLab



Collaborative Data Centre provides

Community Cloud Model



Clinical data sharing - private networks through lightpaths?

The
Francis Crick Institute

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
Chief Executive and Director of the Francis Crick Institute.

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

- EBI Director: Dame Janet Thornton

London Research Institute (LRI) - CRUK

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

Nos. 44, 45, 46, Lincoln's Inn Fields in 1953

"To discover the biology underlying human health, improving the treatment, diagnosis and prevention of human disease and generating economic opportunities for the UK."

THE FRANCIS CRICK INSTITUTE

MRC Medical Research Council

UCL

Imperial College London

KING'S College LONDON

wellcome trust

CANCER RESEARCH UK

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

THE
FRANCIS
CRICK
INSTITUTE

David.fergusson@crick.ac.uk

crick.ac.uk