Cloud Computing For Biology

Guy Coates

Wellcome Trust Sanger Institute

gmpc@sanger.ac.uk

wellcome trust sanger institute

The Sanger Institute

Funded by Wellcome Trust.

- 2nd largest research charity in the world.
- ~700 employees.
- Based in Hinxton Genome Campus, Cambridge, UK.

Large scale genomic research.

- Sequenced 1/3 of the human genome. (largest single contributor).
- We have active cancer, malaria, pathogen and genomic variation / human health studies.

All data is made publicly available.

- Websites, ftp, direct database. access, programmatic APIs.

wellcome trust sanger institute

Use Cases

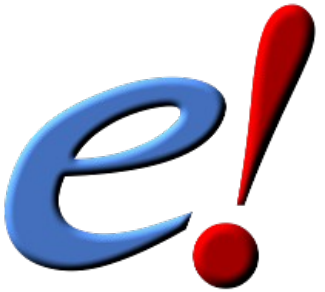
Virtual Co-location

Scientific SaaS

Smart Archives

wellcome trust sanger institute

Ensembl



Ensembl is a system for genome Annotation.

Data visualisation / Mining web services.

- www.ensembl.org
- Provides web / programmatic interfaces to genomic data.
- 10k visitors / 126k page views per day.

Compute Pipeline (HPTC Workload)

- Take a raw genome and run it through a compute pipeline to find genes and other features of interest.
- Ensembl at Sanger/EBI provides automated analysis for 51 vertebrate genomes.

- Software is Open Source (apache license).
- Data is free for download.

We have web services and HPTC workloads running on Iaas.

wellcome trust sanger institute

Virtual Co-location

Virtual Co-location for Web services

- Was hosted in a single datacentre at the Genome Campus, UK.
- 1 datacentre = Single point of failure.
- Access slow if you were not in western Europe.

Cloud Application

- Build worldwide network of mirrors on Amazon AWS (IaaS).
- Mirrors currently in US and Asia.

Deployment times:

- Very time consuming to put machines into a co-lo facility.
- Especially abroad. (death by FCC paperwork.)

Flexibility:

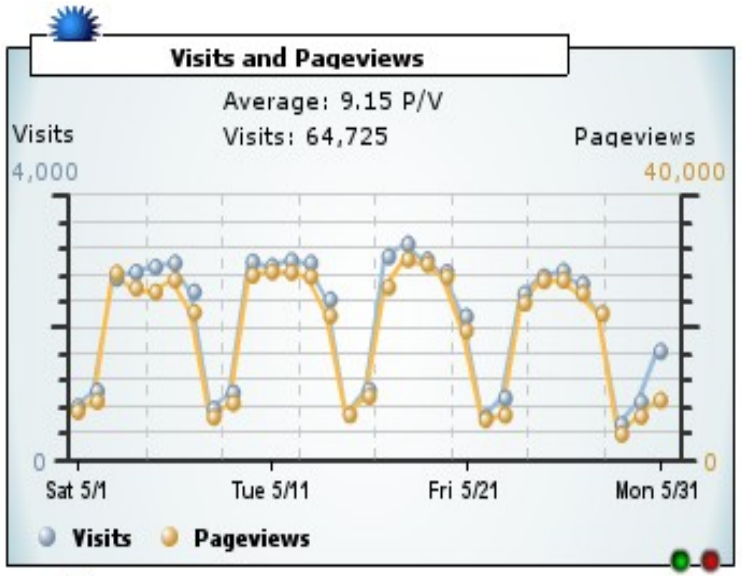
- Simple to add new machines into the virtual Colo, terminate contracts.

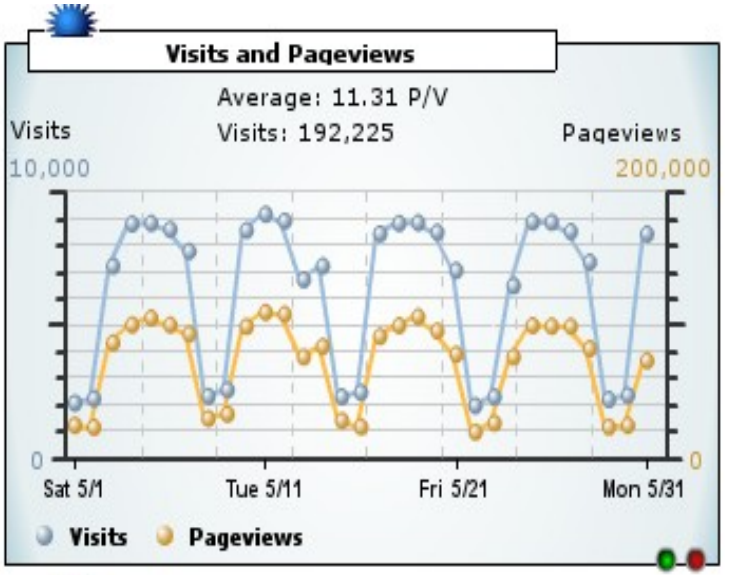
Costs:

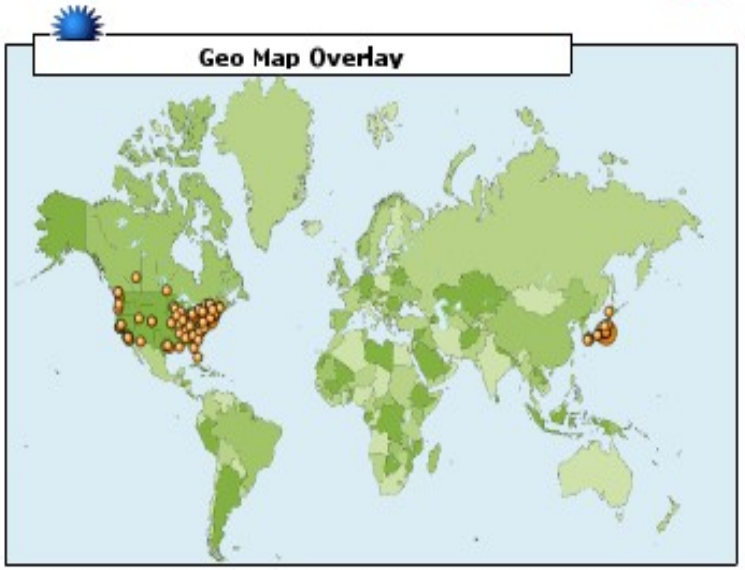
- Pay as you go model more cost effective that tiered model provided by co-lo (not the primary motive)...

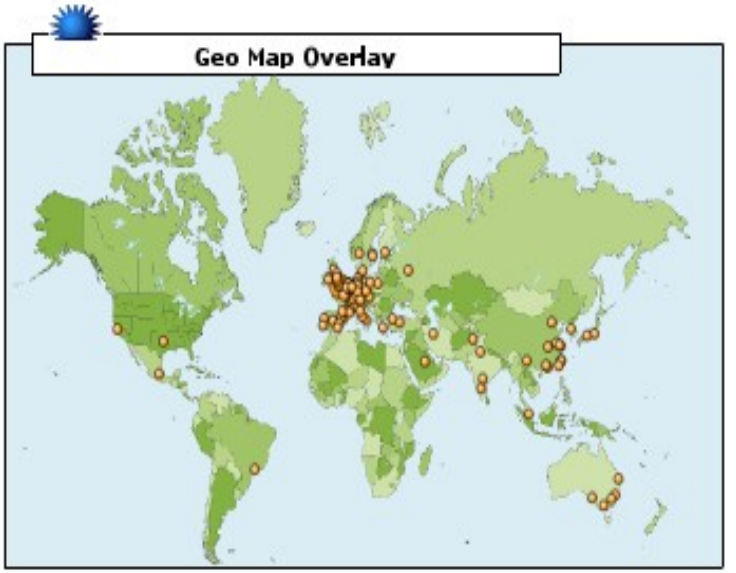
wellcome trust sanger institute

Usage









wellcome trust sanger institute

Sequencing Informatics

wellcome trust sanger institute

Economic Trends:

As cost of sequencing halves every 12 months.

- *cf* Moore's Law

The Human genome project:

- 13 years.
- 23 labs.
- $500 Million.

A Human genome today:

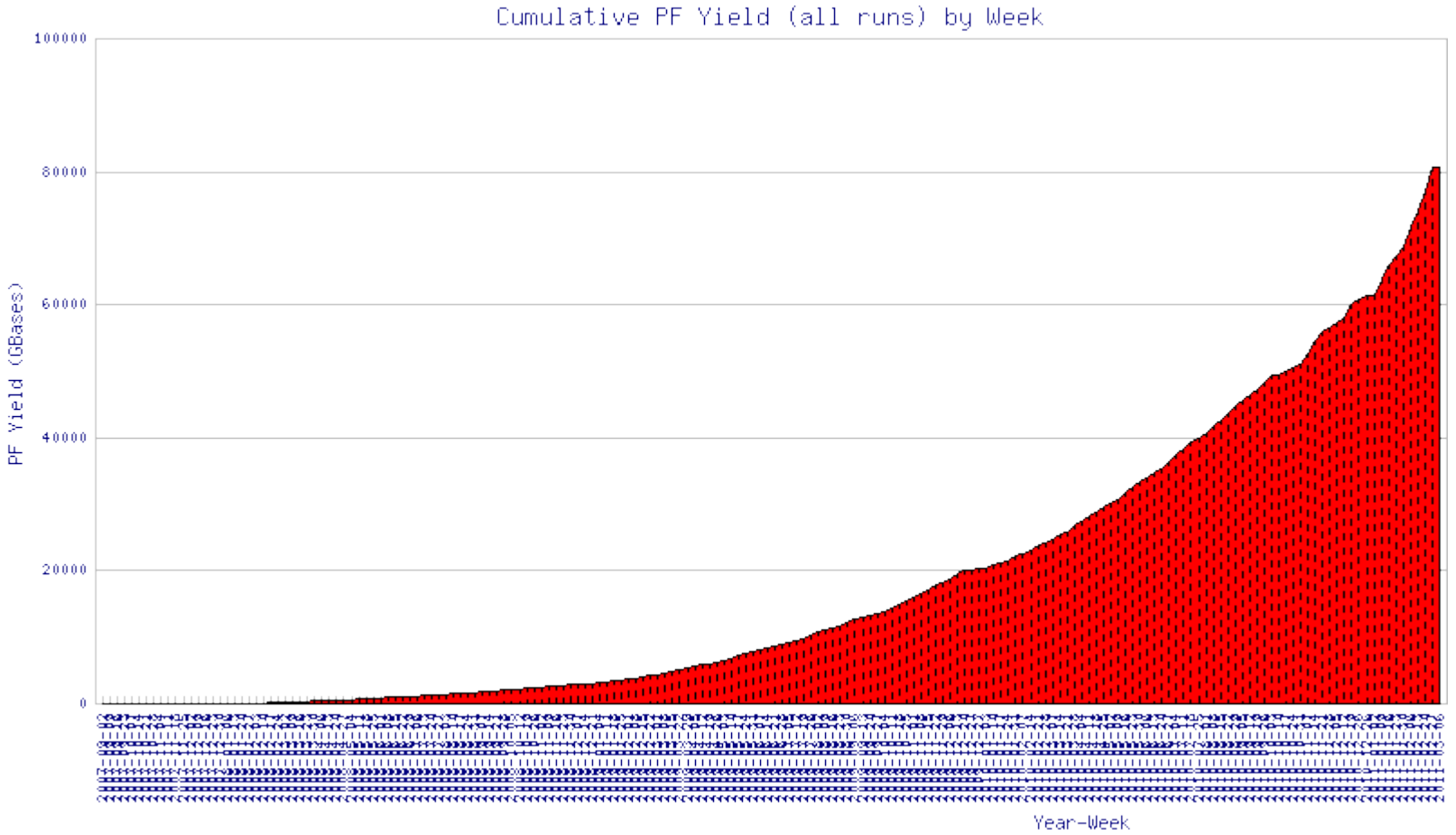
- 3 days.
- 1 machine.
- $10,000.
- Large centres are now doing studies with 10,000s of genomes.

Trend will continue:

- Generation 3 sequencers are on their way.
- $500 genome is probable within 5 years.

wellcome trust sanger institute

The Scary Graph



wellcome trust sanger institute

Managing Growth & Change

We have exponential growth in storage and compute.

- Storage /compute doubles every 12 months.
  - 2010 ~12 PB raw.

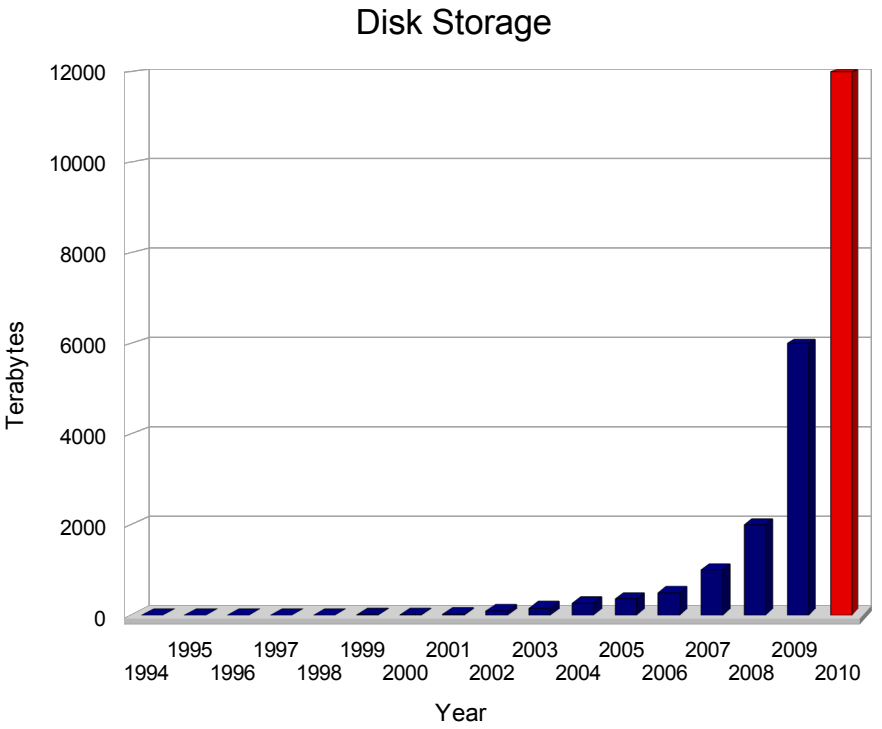
Processes change constantly.

- Driven by changes in machine chemistry.
- LIMs, compute pipelines and algorithms in constant flux.

Moore's law will not save us.

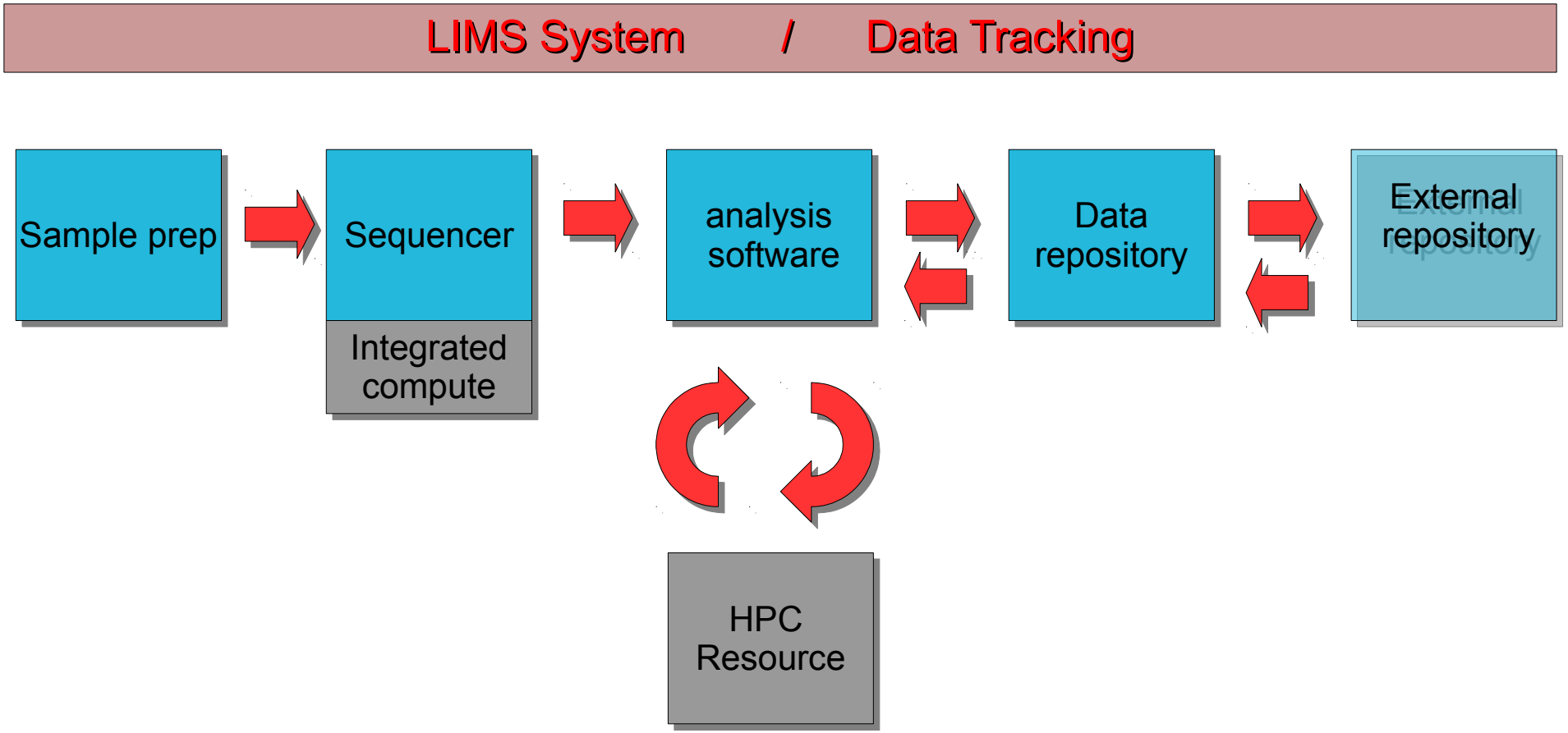
- Transistor/disk density: T_d=18 months
- Sequencing cost: T_d=12 months
- Sequencing output: Td=3-6 months

This is hard!



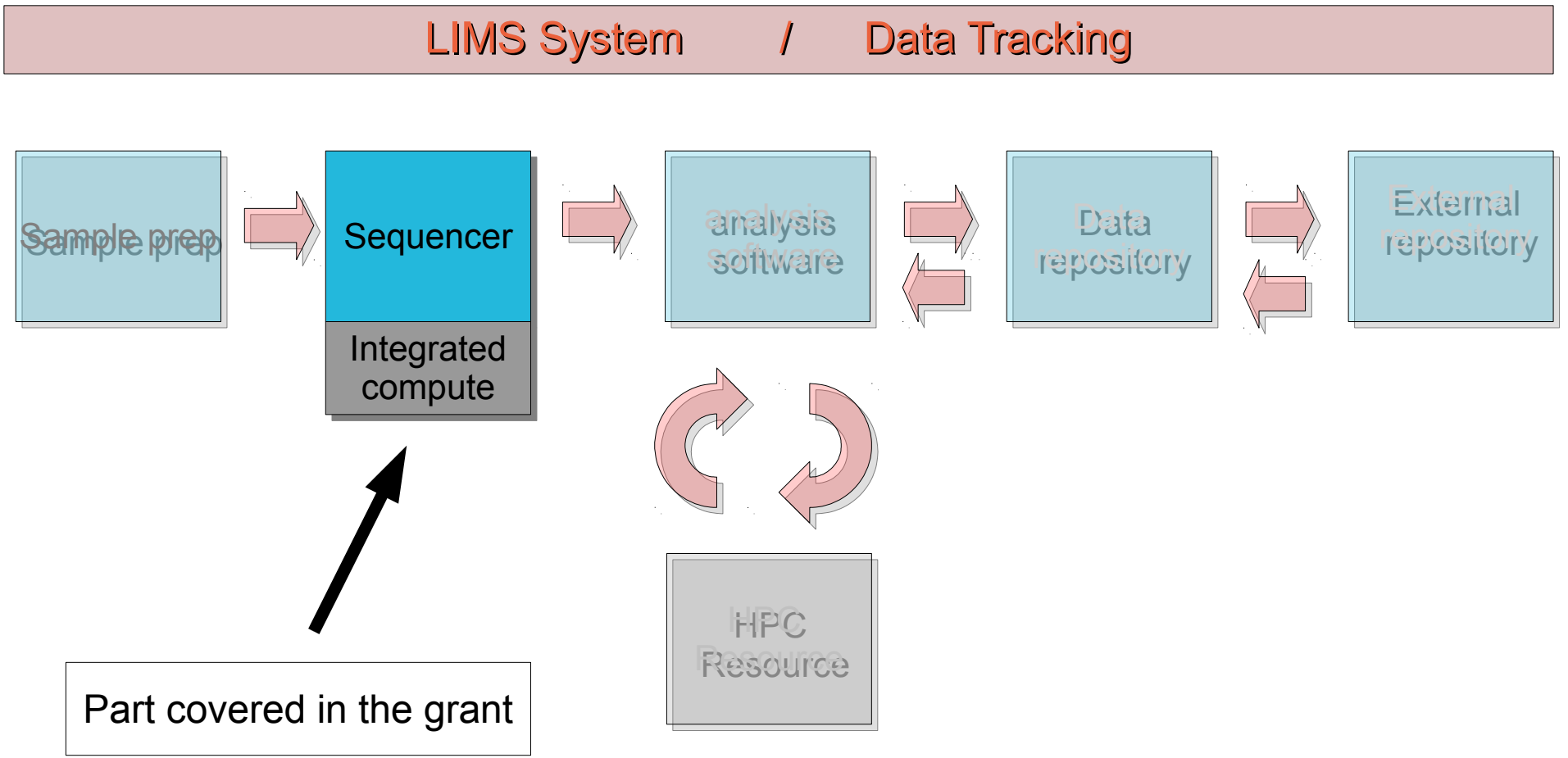
wellcome trust sanger institute

What do you need to do sequencing?



wellcome trust sanger institute

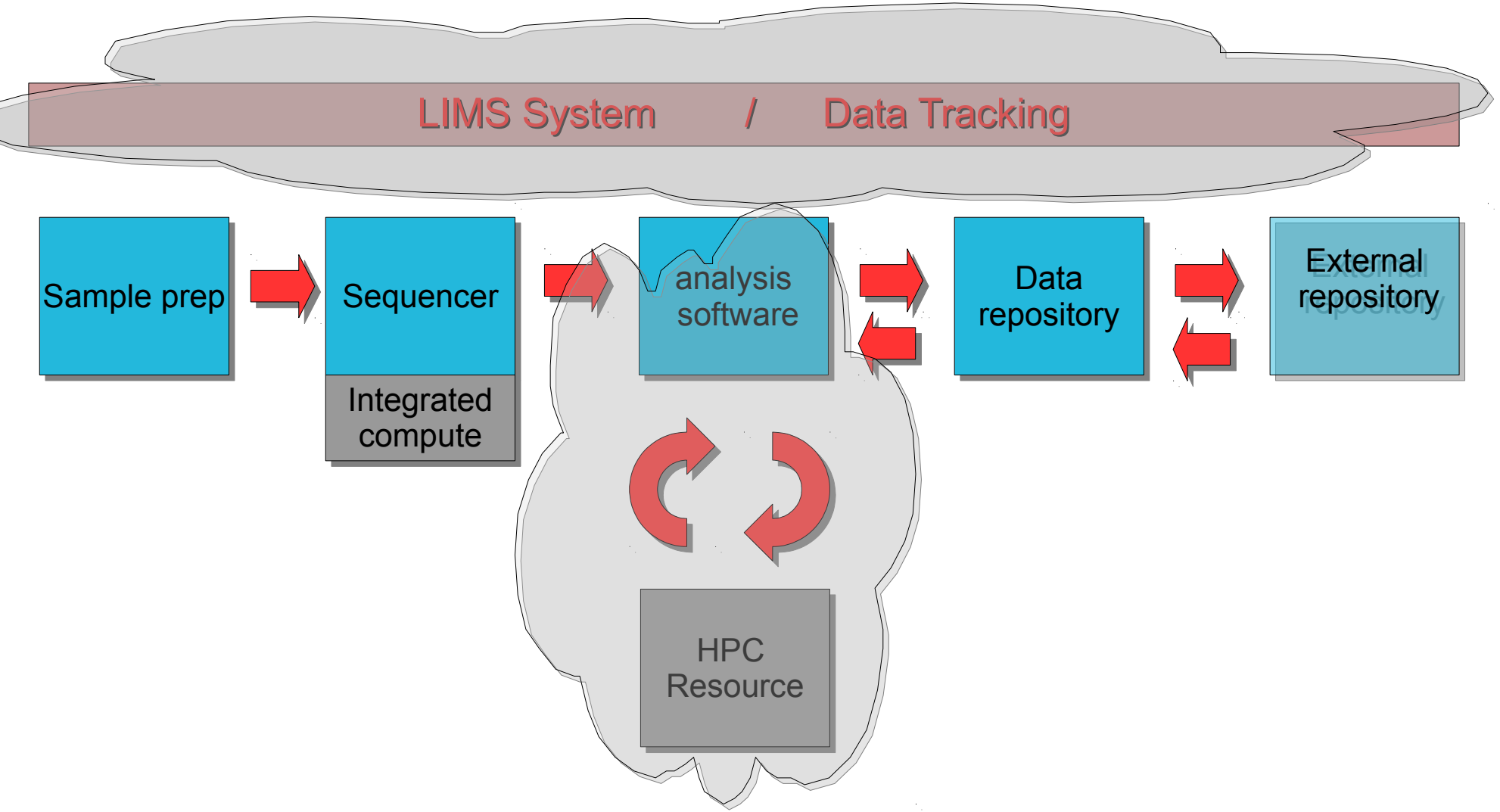
What do you need to do sequencing?



How can cloud help?

wellcome trust sanger institute

Cloud SaaS



wellcome trust sanger institute

Commercial or Academic?

Commercial sequencing SaaS exists today.

- Upload data straight from sequencer into cloud.
- Lims, data analysis, archiving.

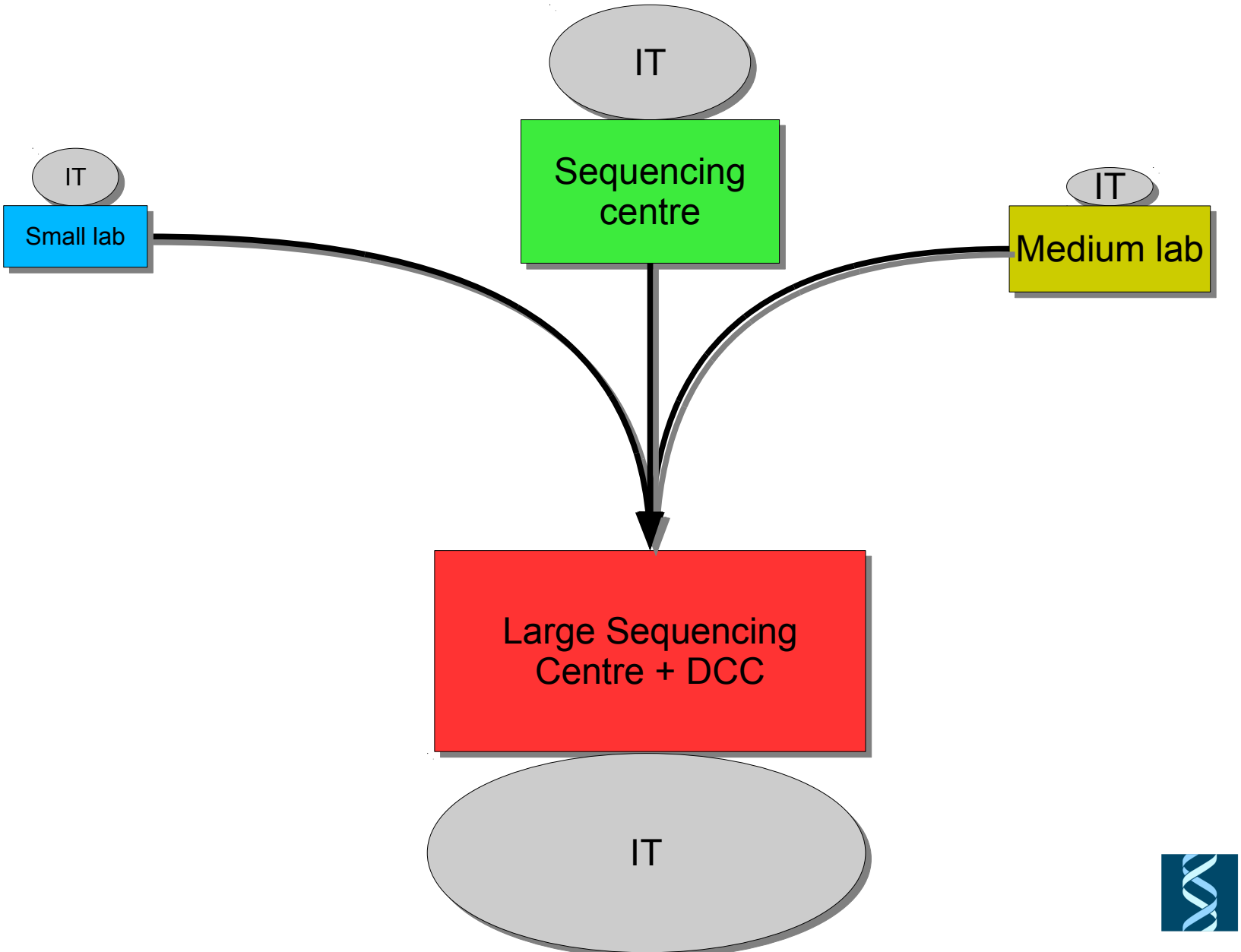
Problems:

- Data transfers across the public internet make it difficult to get data off the sequencers fast enough.
- Potential data protection issues.

There is a role for academic providers.

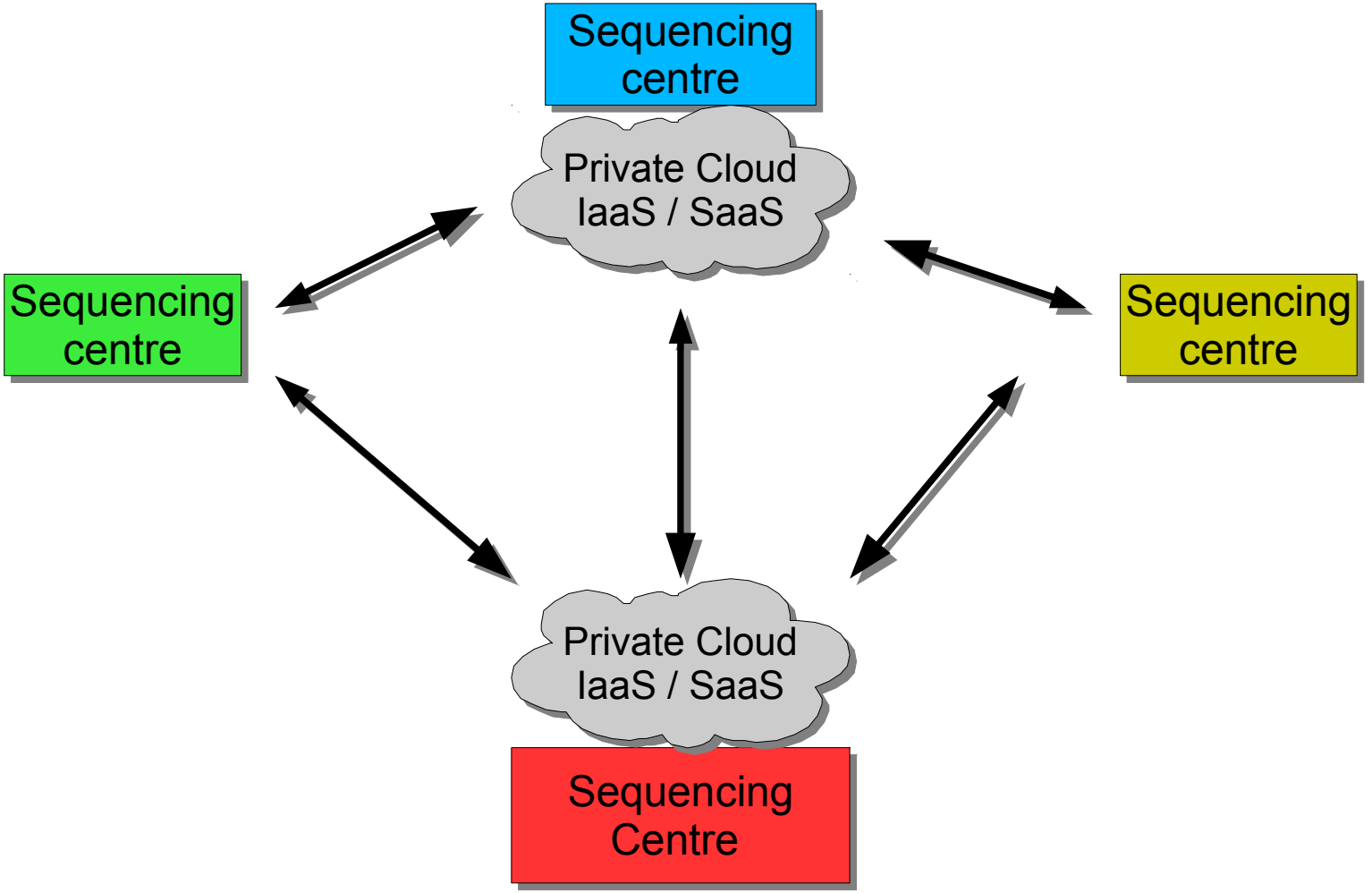
wellcome trust sanger institute

Traditional Collaboration



wellcome trust sanger institute

Cloud Collaborations



wellcome trust sanger institute

Consortium Cloud

SaaS Advantages:

- Small organisations leverage expertise of big organisations.
  - Frees up staff to concentrate on science, not IT.
- Software and IT infrastructure easily shared with consortium.
  - Ensures consistency across a large project.
- Academia tends to be linked by fast research networks.
  - Moving data is easier.
- Consortium will be signed up to data-access agreements.
  - Simplifies data governance.
- Opens up new opportunities:

wellcome trust sanger institute

Data Archives

Sequencing data is held in public archives.

- Strong community drive for open science.
- Raw data available for researchers and industry.

Managing archives is difficult.

- All the data is in one place.
- Large burden on the people looking after it.

Data in current archives is "dark".

- You cannot compute across it.

Getting data out is hard:

- Moving TBs across the internet is difficult.

wellcome trust sanger institute

Example problem:

“We want to run out pipeline across 100TB of data currently in European Genotyping Archive.”

We will need to de-stage the data to Sanger, and then run the compute.

- Extra 0.5 PB of storage, 1000 cores of compute.
- 3 month lead time.
- ~$1.5M capex.

(and the data is actually hosted on our datacentre!)

wellcome trust sanger institute

Cloud archives

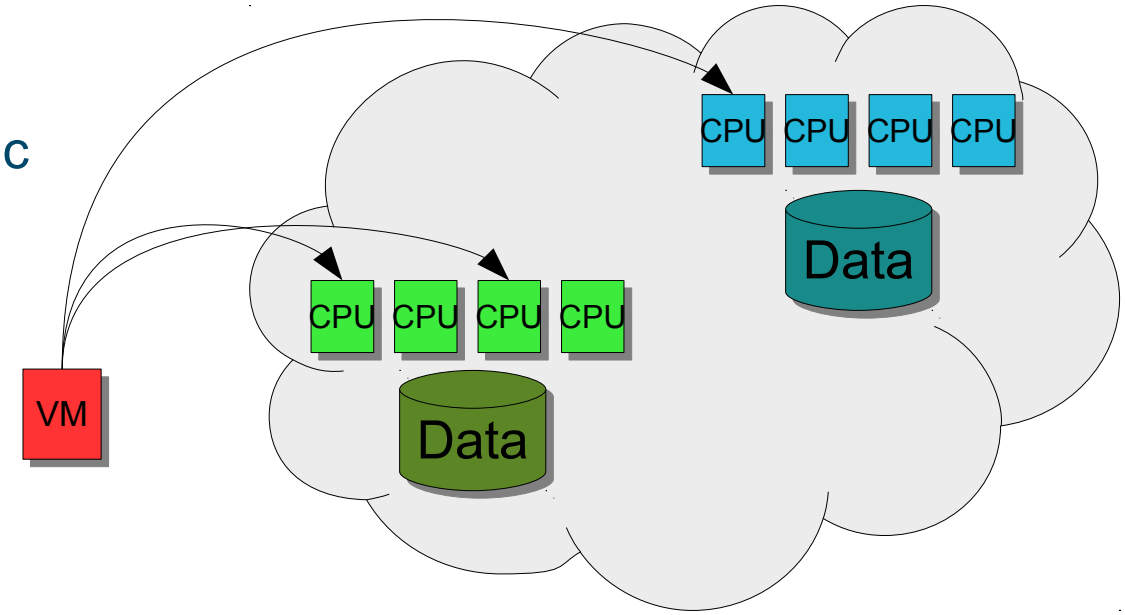
Host data on consortium cloud infrastructures.

Move the compute to the data.

- Upload workload onto VMs.
- Upload VMs into the academic cloud infrastructure.

Federated between centres

- Grid software build on top of cloud components.
- Avoids scaling problems inherent in putting large datasets all in on one place.



wellcome trust sanger institute

Challenges

None of this infrastructure exists yet!

- Academic service providers have to invest in infrastructure & people.

Funding models:

- How are cost recouped from end users?
- Selling services / charge-back can be tricky in some environments.
  - eg research charities.

Can we do it as well as the big internet companies?

- We have to provide better value than commercial operators.
- Do we build scientific services on top of public IaaS?

wellcome trust sanger institute

Acknowledgements

Sanger

- **Phil Butcher**
- James Beal
- Pete Clapham
- Simon Kelley
- Gen-Tao Chiang

- Steve Searle
- Jan-Hinnerk Vogel
- Bronwen Aken

EBI

Glenn Proctor
Steve Keenan

wellcome trust sanger institute