# BIOBANKCLOUD
The PaaS for Biobanking

# A Platform-as-a-Service for Biobanking

**www.biobankcloud.com**


- **Presented by: Paulo Esteves Veríssimo**
- ***Univ. de Lisboa, Faculdade de Ciências (FCUL), LaSIGE, Portugal,***
- ***pjv@di.fc.ul.pt*** http://www.di.fc.ul.pt/~pjv


- **64th IFIP WG 10.4 Meeting, Visegrad - Hungary**
- **June 27-30, 2013**


Financed by the European Commission 7th Framework Programme.


# BiobankCloud – a PaaS for Biobanking

**Duration:** Dez 2012-2015

**Budget:** ~3Mio Euros

A Biobank is :
*"biological material from one or several human beings collected and stored indefinitely or for a specified time and whose origin can be traced to the human or humans from whom it originates"*

## Goal

Provide a viable platform-as-a-service (PAAS) for Biobanking to support the secure storage, analysis, and sharing of genomic data.

## Research Challenges

Design and build services for cloud platforms that securely store and analyze sensitive personal data, while conforming to a regulatory framework.

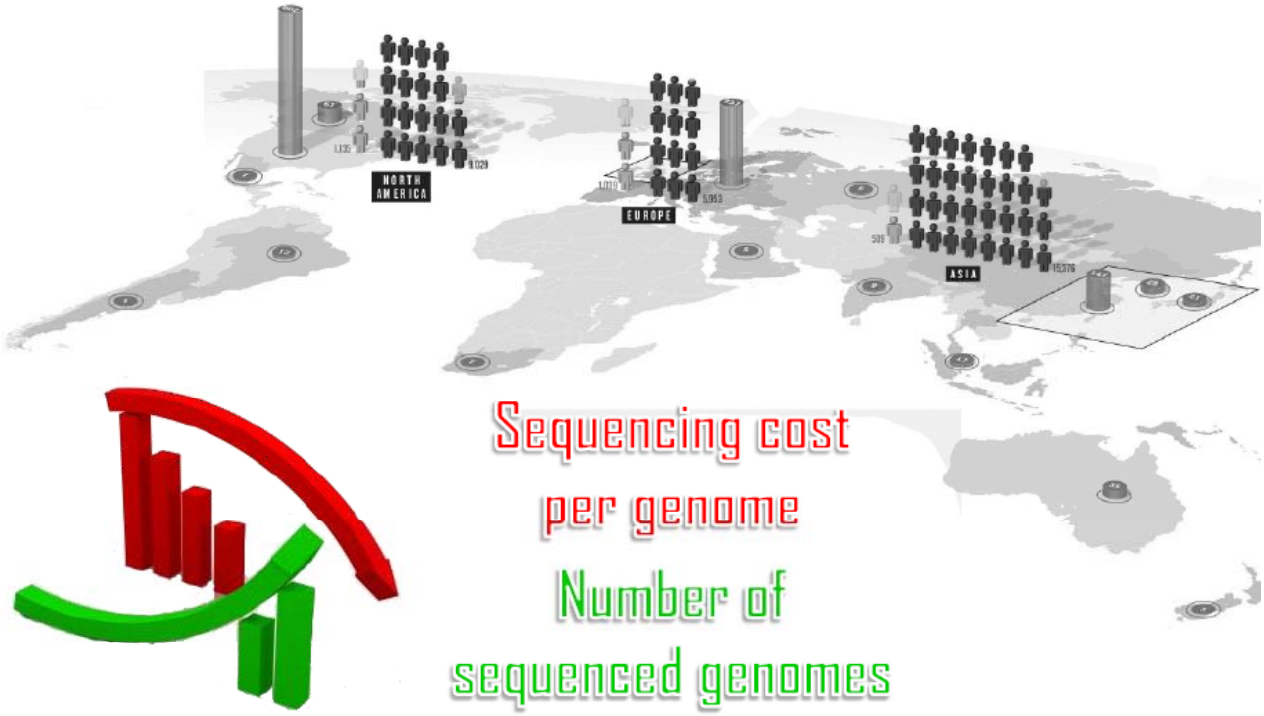
# Genomics and Big Data

- We will soon be generating more genomic data than we can currently securely store and process
- Handling terabytes of information has b… e the norm for genomics, b… up.
- Th… tems res… icient sto…

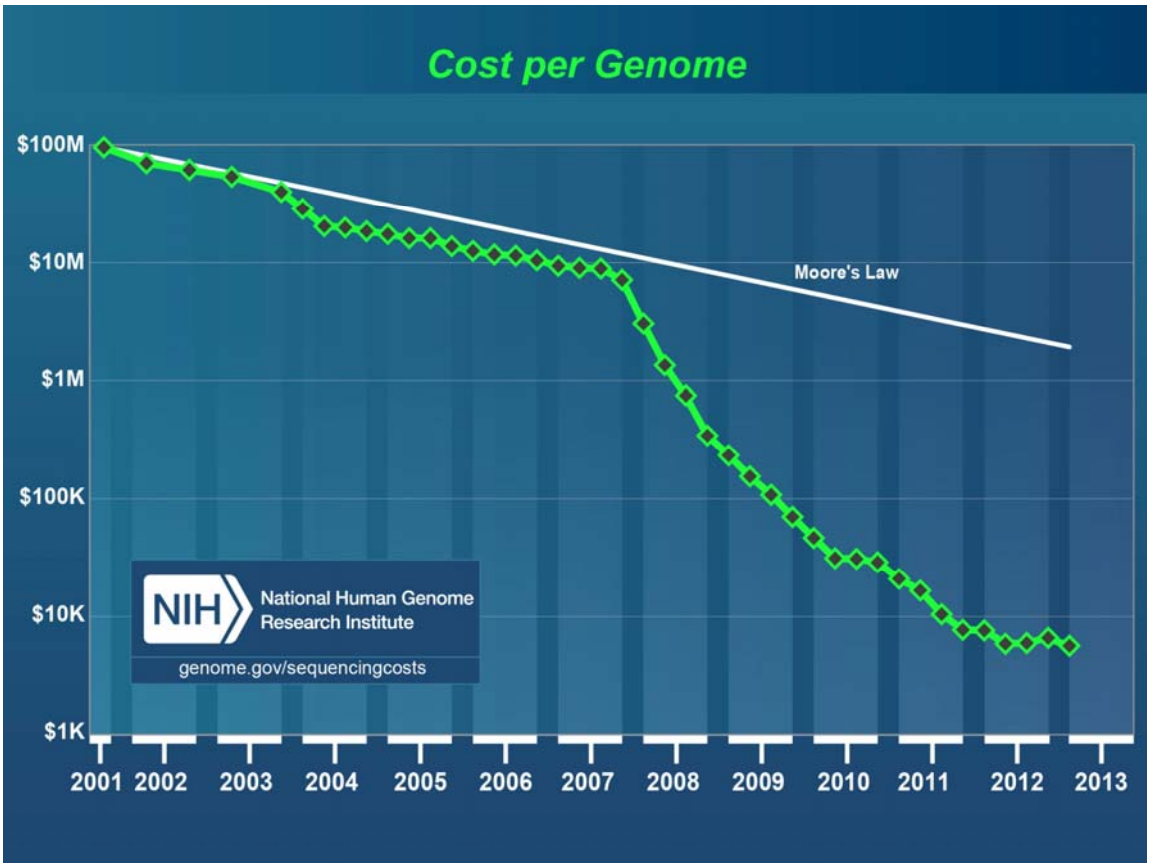
**Urgent need for better systems to store and process genomic data**


# A scent of the future trends ….

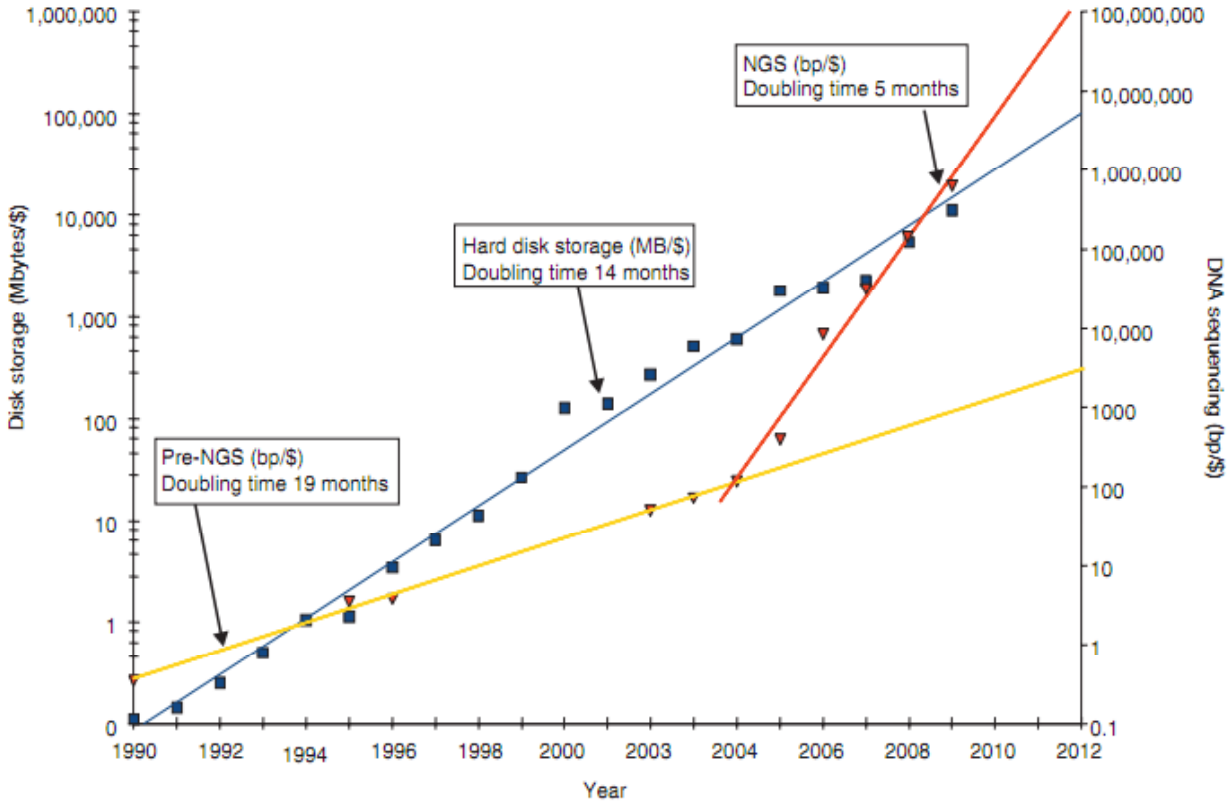
# The cost per ger genome is bound to decrease




# Cost per genome vs. Moore's Law



# Growth in NGS capability vs. Kryder's law




- Growth in disk storage capacity vs. growth in NGS data generation capability

# Current reality already significant ....

# Increasing number of research databases

1000 Genomes
A Deep Catalog of Human Genetic Variation

Complete genomics
A BGI Company

NCBI GenBank

UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

GENOME 10K

wellcome trust sanger institute

Personal Genome Project

# LifeGene: ~500,000 people

LifeGene

## Sweden has 73 Opt-Out "Quality Registries"

Nationella Kvalitetsregister

http://www.kvalitetsregister.se/om_kvalitetsregister/quality_registries

## A Coalition of the Willing

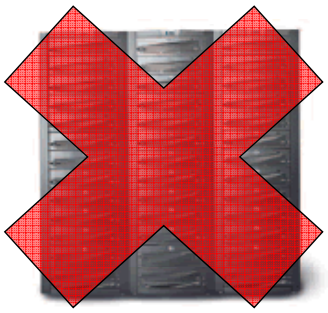
- In Europe, BBMRI is a trans-national network of Biobanks with **the goal of sharing study data** for research

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

## Storing and computing with this data …

## Scalable storage on commodity hardware

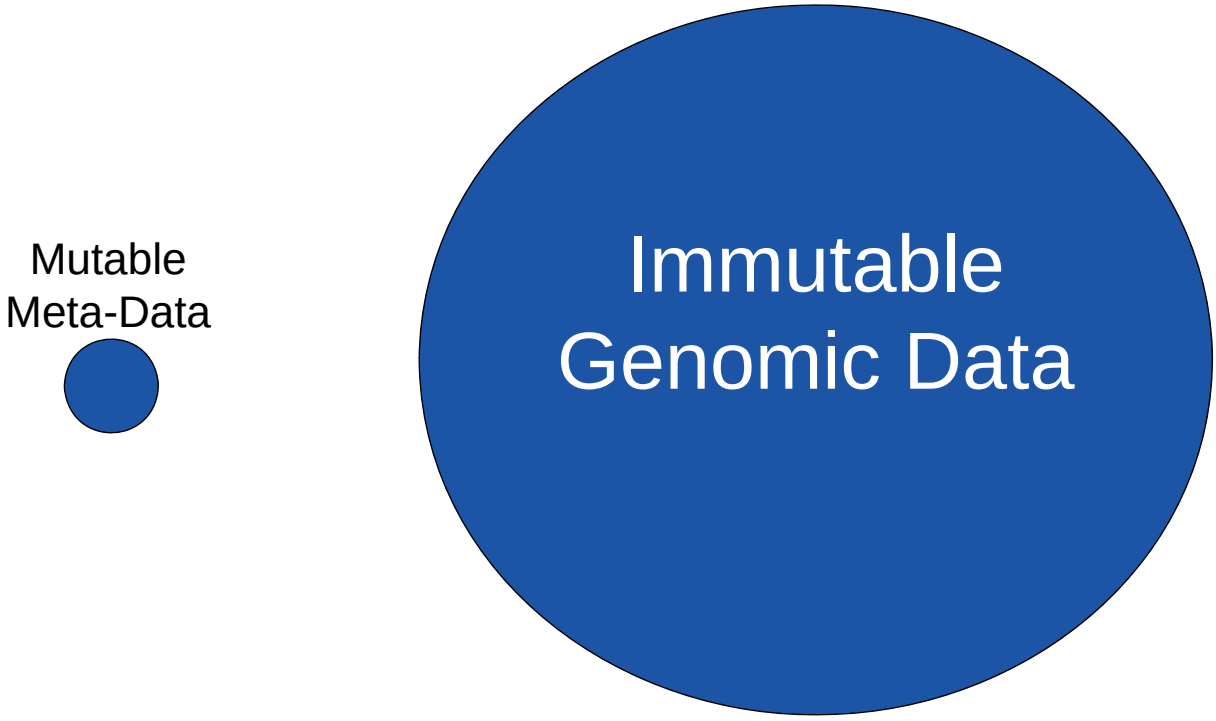
- Many genomics research groups still run expensive SANs (storage area networks) or file systems that require expensive interconnects.

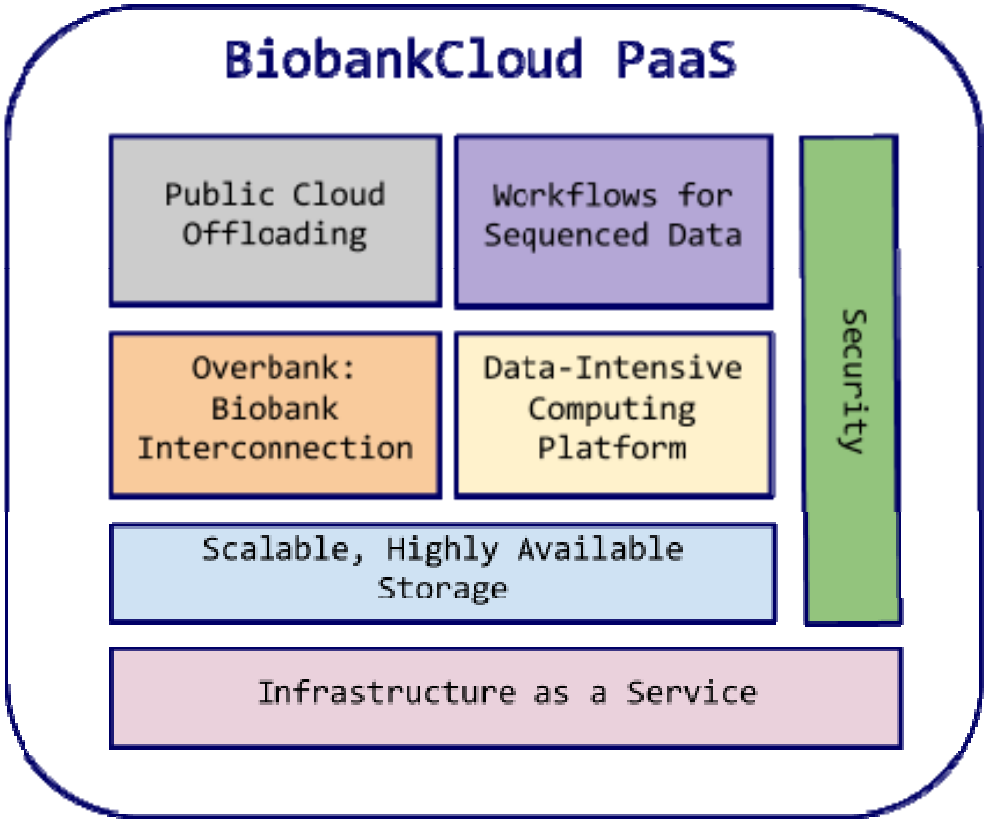

Reliable Storage Service

Unreliable Commodity Servers

# Immutable Genomic Data




# BiobankCloud PaaS



# But we can't let the data leave our site....

- You buy the NGS machines
- You might want to use affordable compute and storage
- .....but you have to keep your genomic data on-site.


[Taken from Illumia Basespace]

# What is the real privacy?

| Private | Public |
| --- | --- |
| • Donor name<br>• Identity number<br>• Social security number<br>• Home address | • DNA sequence<br>• Age<br>• State |

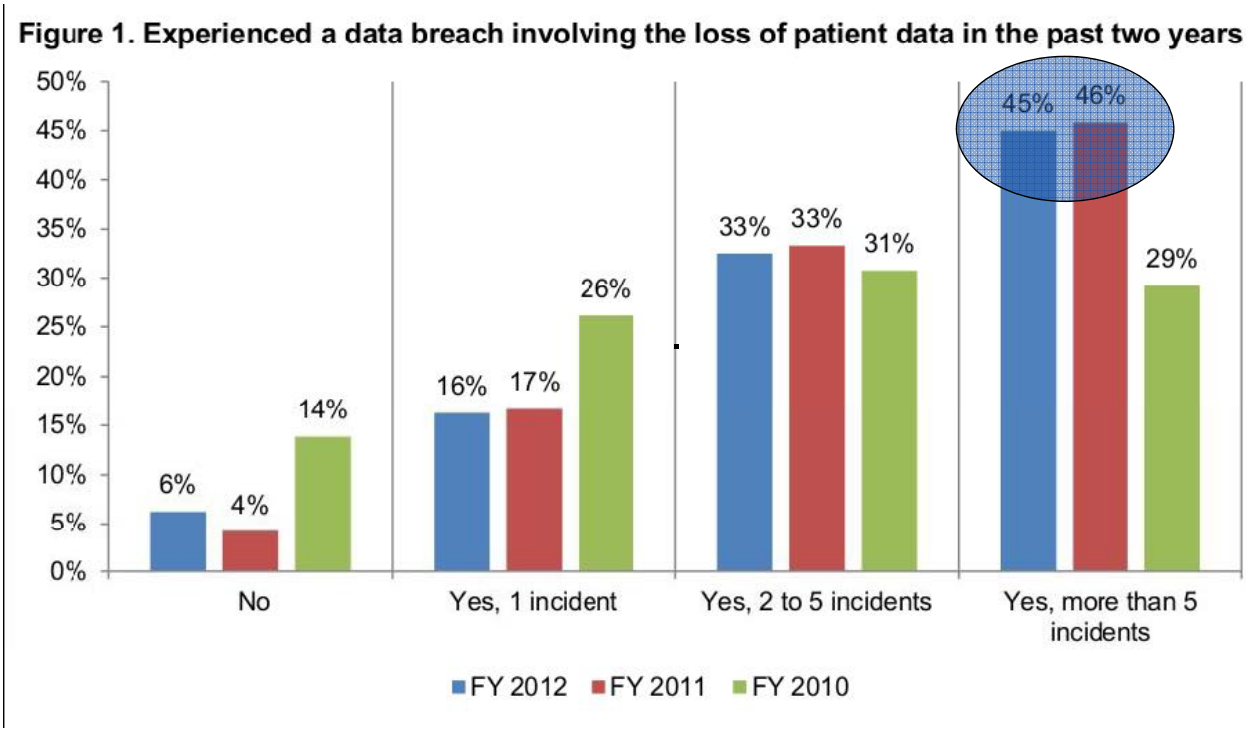
1000 Genomes
A Deep Catalog of Human Genetic Variation

Complete genomics

wellcome trust sanger institute

Personal Genome Project

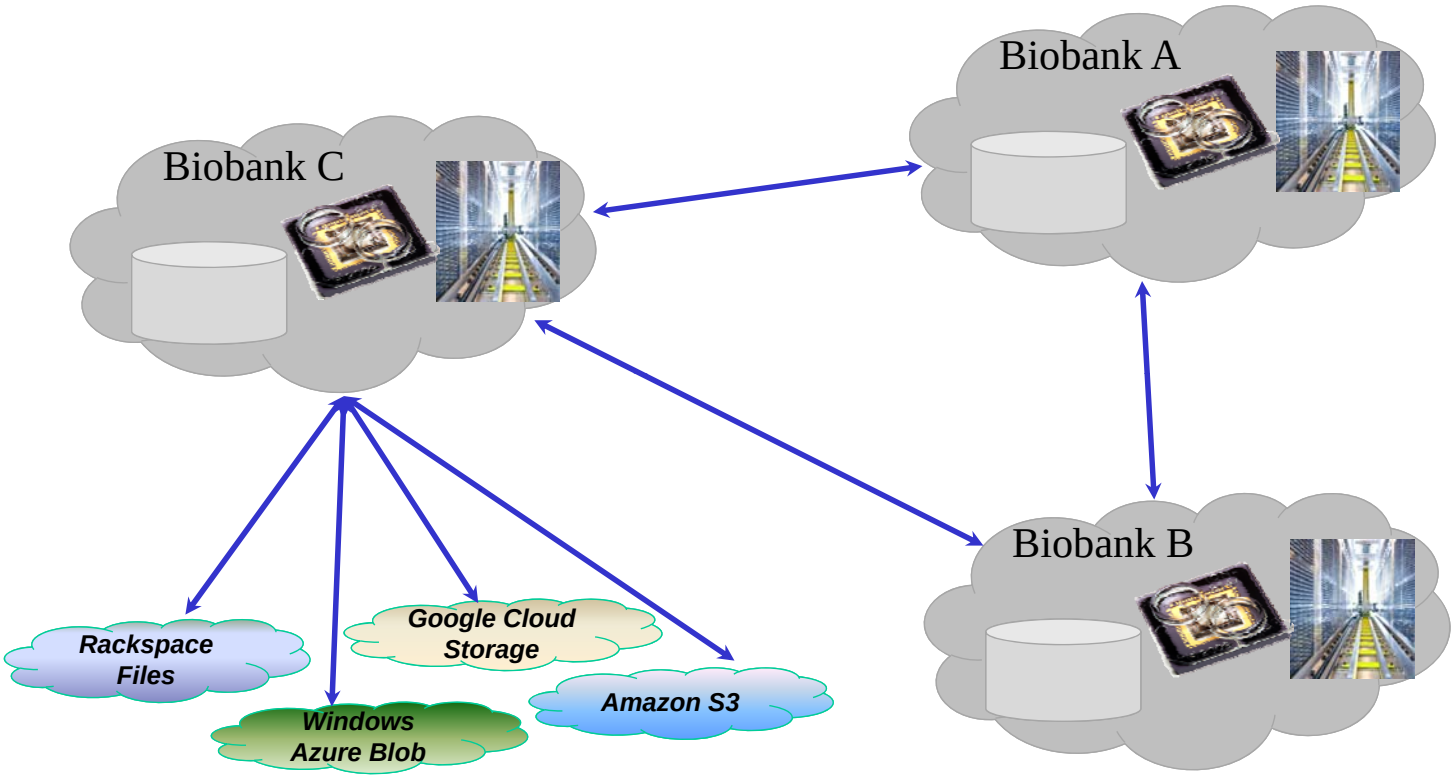
# Data Leakage




Ponemon Institute 3rd Annual benchmark study on patient privacy and data security - 80 health care organisations

# Proposed solutions

# Overbank






# Six Research Ideas

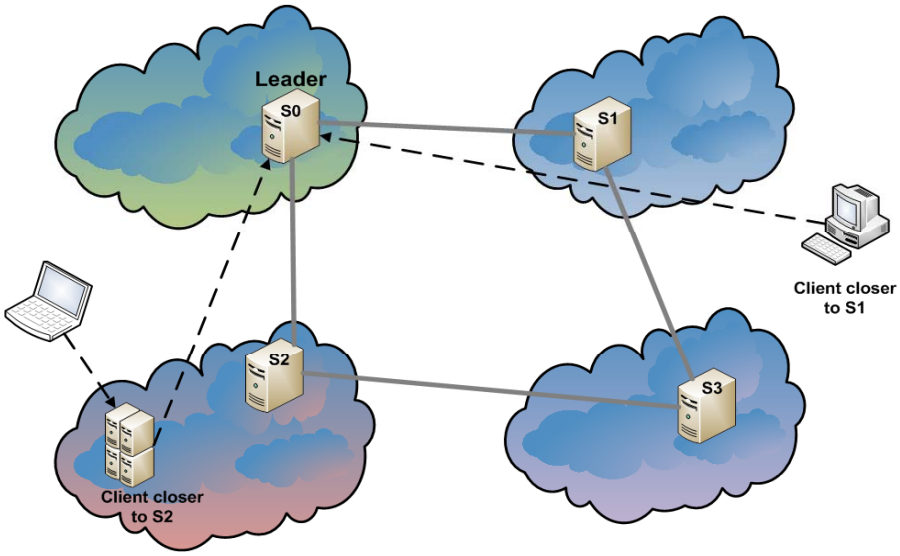
- Wide-area state machine rep.
- Cloud-of-clouds storage
- Multi-Cloud MapReduce
- Network coding
- Software-Defined Networking
- Secure two-party computation

*Consolidated*

*Pure Research*

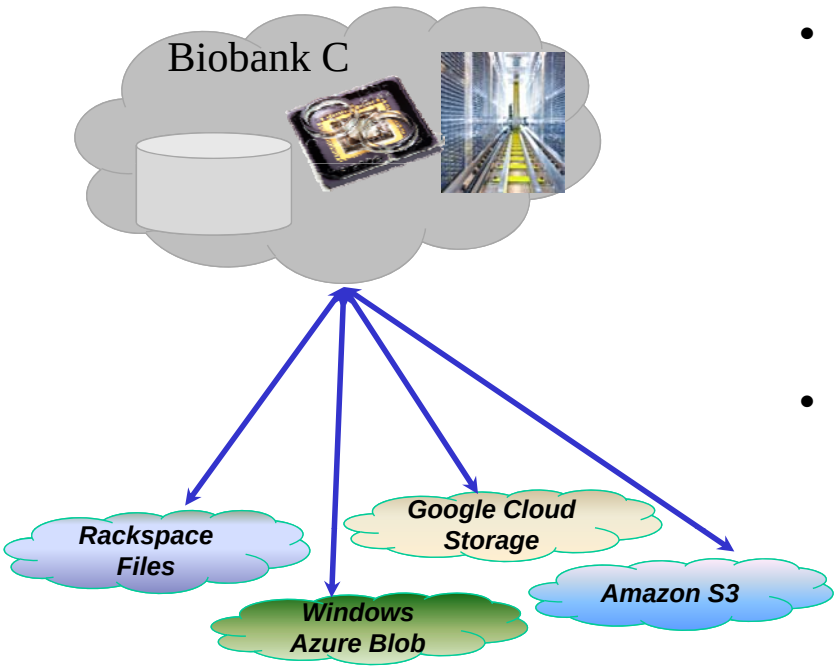
# Wide-Area Replication

- Consistent and secure replication across datacenters





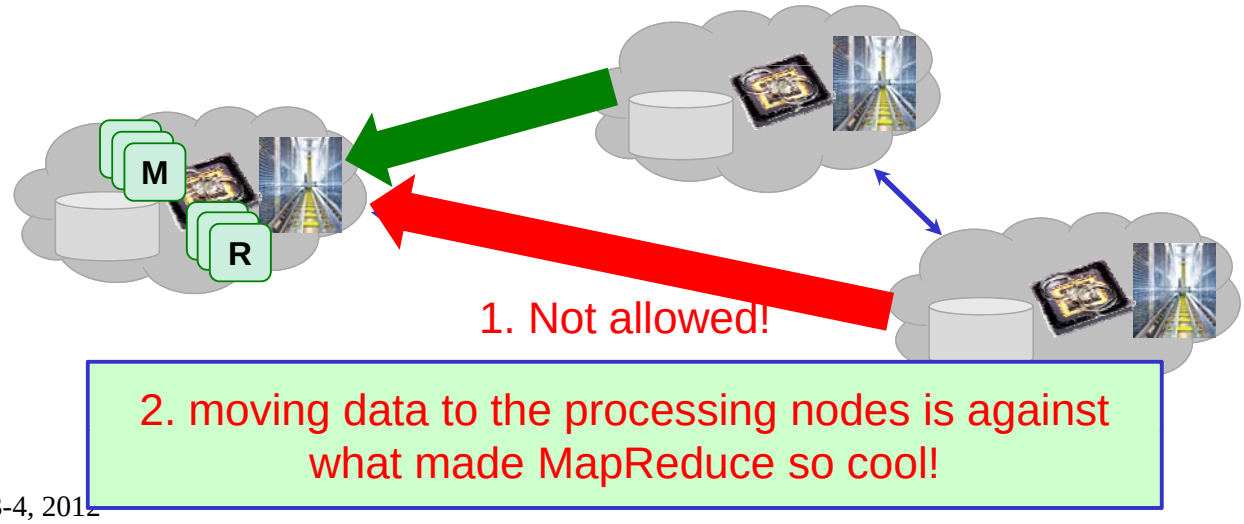

# Cloud-of-Clouds Storage




- Not everything can go to public clouds
  - Even if we encrypt and disperse the data, there are legal limitations
  - But a lot of data can!
- TClouds work: DepSky, **C2FS (file system)**
  - Can be improved for our workloads
  - Database-like system

# Multi-Cloud MapReduce

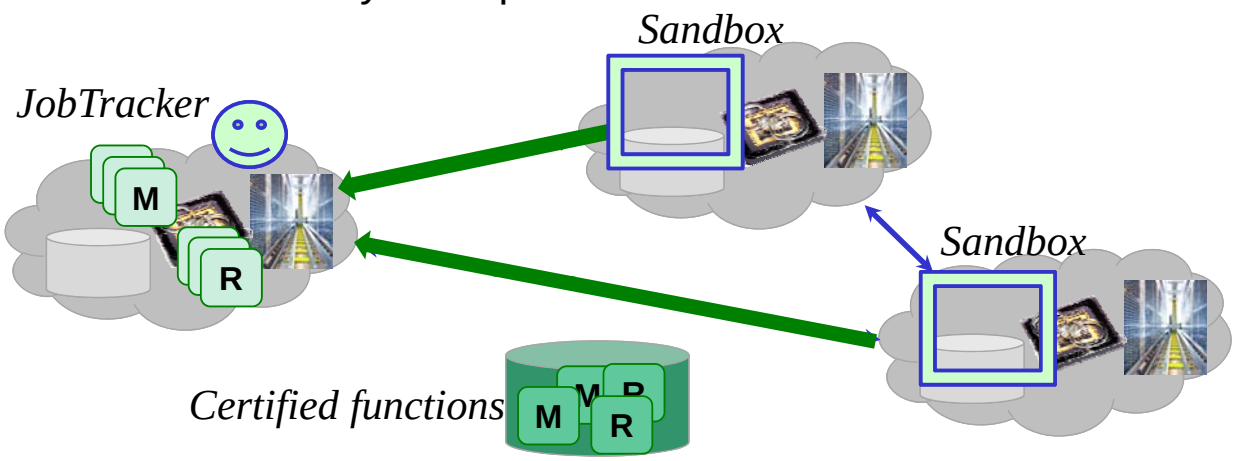
- What if you want to do some analysis but you need data from other Biobanks?






# Multi-Cloud MapReduce

- Move computation, not data, without privacy violations
  - sandboxing
  - wide-area job tracker
  - certified library of map and reduce functions



# Thank you! Tack! Danke! Obrigado!