HELIX NEBULA
THESCIENCECLOUD

Scientific Cloud Computing Infrastructure for Europe - Strategic Plan

Rupert Lueck
Head of IT Services
EMBL Heidelberg

5th DFN Forum
May 21th 2012, Regensburg

EMBL

Helix Nebula

- “Eye of God”
- NGC 7293
- located inside Aquarius constellation
- One of the closest planetary nebula
- ~700 light-years away

Origin of the initiative

- Conceived by ESA as a prospective for providing cloud services to space sector in Europe
- Presented to the IT working group of the EIROforum where other members (CERN, EMBL) joined
- Two workshops held during 2011
  - June: hosted by ESA in Frascati
  - October: hosted by EMBL in Heidelberg

EIROforum: CERN, EFDA-JET, EMBL, ESA, ESO, ESRF, European XFEL, ILL

Strategic Plan

for a scientific Cloud Computing Infrastructure in Europa

- Establish a sustainable multi-tenant cloud computing infrastructure in Europe
- Initially based on the needs for the European Research Area & space agencies
- based on commercial services from multiple IT industry providers
- which adhere to internationally recognised policies and quality standards to be adopted by the governance structure involving all stakeholders

Lengert, Maryline, Jones, Robert (2011) CERN-OPEN-2011-036
http://cdsweb.cern.ch/record/1374172/

Objectives of the initiative

1. Set up a cloud computing infrastructure for all European Research Area
2. Identify and adopt policies for trust, security and privacy on a European-level
3. Create a light-weight governance structure involving all stakeholders
4. Define a short and medium term funding scheme

A Collaboration Initiative

HELIX NEBULA
THESCIENCECLOUD

European Commission
& relevant projects

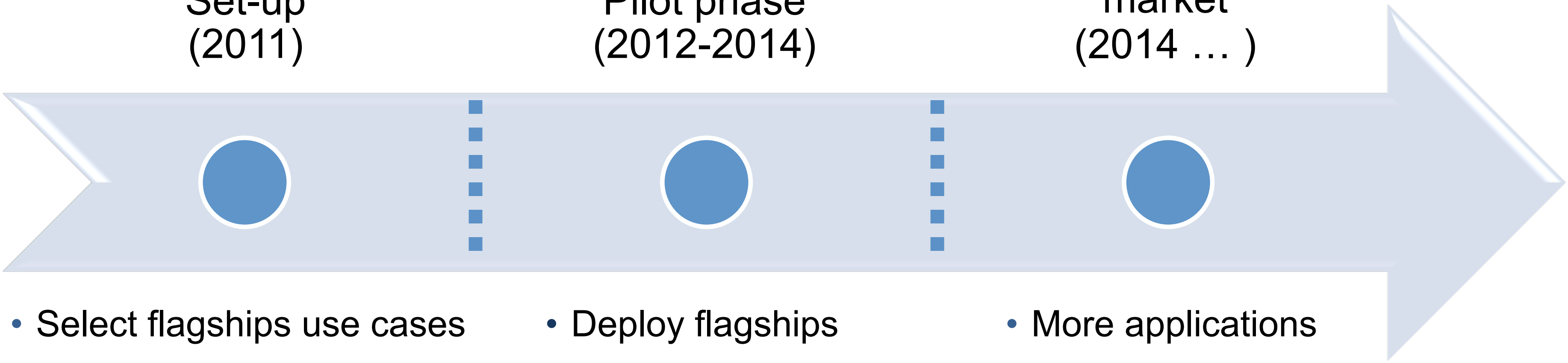
User organisations
Demand-side

**European
Cloud Computing
Strategy**

**Commercial Service
Providers**
Supply-side

Bringing together all the stakeholders to establish a public-private partnership

Timeline



Set-up
(2011)

- Select flagships use cases
- Identify service providers
- Define governance model

Pilot phase
(2012-2014)

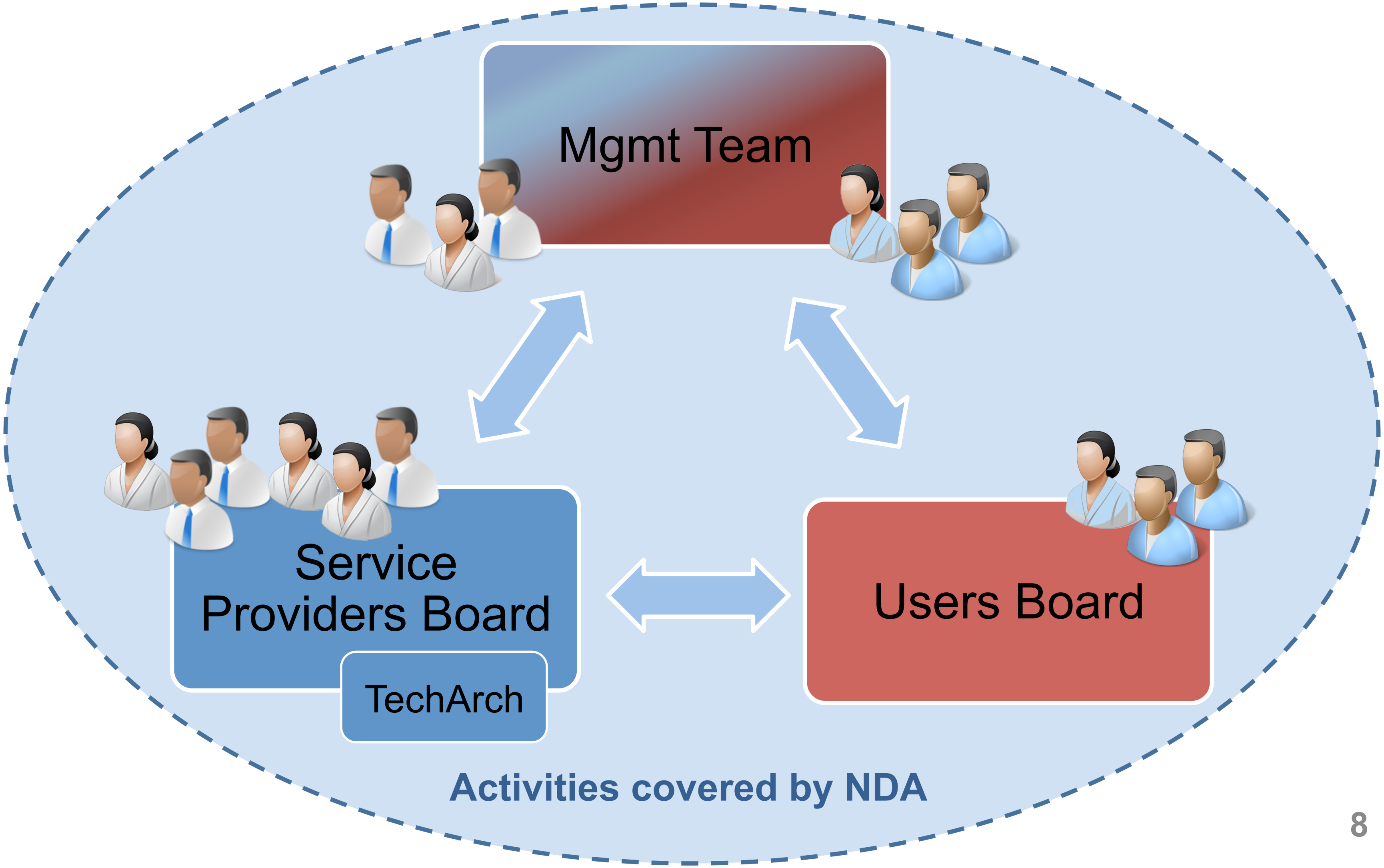
- Deploy flagships
- Analysis of functionality, performance & financial model
- Success Stories

Full-scale
cloud service
market
(2014 …)

- More applications
- More services
- More users,
- More service providers

Governance Model

Proof of Concept stage



Activities covered by NDA

Consortium

Q1 / 2012

Demand-Side

CERN

EMBL

esa

Supply-Side

AtoS

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

CloudSigma

egi

interoute
from the ground to the cloud

logica
be brilliant together

OpenNebula.org
The Open Source Toolkit for Cloud Computing

orange™ Business Services

CSA

SAP®

the SERVER LABS
the IT architects

sixsq.

Telefonica

terradue 2.0

THALES

T - Systems

Activities covered by NDA

Consortium membership

- Consortium includes all participating supply-side and demand-side companies / organisations
  - Member status and adopter status
  - All sign a non-disclosure agreement

- Initial membership is defined
  - More members and adopters will be added following the Proof of Concept stage within the Pilot Phase (summer 2012)

Pilot Phase

Explore / push a series of perceived barriers to Cloud adoption:

- **Security**: Unknown or low compliance and security standards
- **Reliability**: Availability of service for business critical tasks
- **Data privacy**: Moving sensitive data to the Cloud
- **Scalability / Elasticity**: Will the Cloud scale-up to our needs
- **Network performance**: Data transfer bottleneck; QoS
- **Integration**: Hybrid systems with in-house / legacy systems
- **Vendor lock-in**: Vendor dependency once data & applications are transferred to the Cloud
- **Legal concerns**: Such as who has legal liability
- **Transparency**: Clarity of conditions, terms and pricing

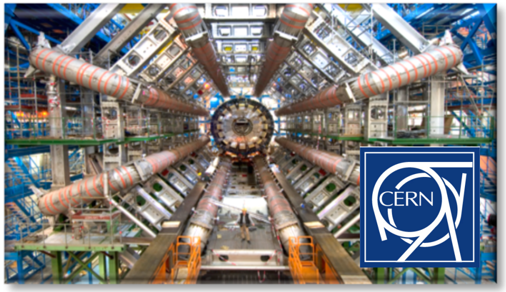
Flagship use cases

- Proposed by demand-side user organisations
- Addressing scientific challenges with societal impact
  - High-profile applications
  - Catching the public imagination & encourage others to use the services
  - Innovate in terms of functionality, performance, scope, business opportunities or impact
- Sponsored by user organisations
  - Need to contribute their own resources during the pilot phase to port application (manpower) and contribute to the cost of procuring required services from the supply-side (cash)
  - Must participate in a costing exercise where the total cost of deploying and operating the flagship application in-house can be compared to the cost of procuring the services via Helix Nebula
- Want to propose a flagship?
  - Send email to contact@helix-nebula.eu

Initial flagships use cases

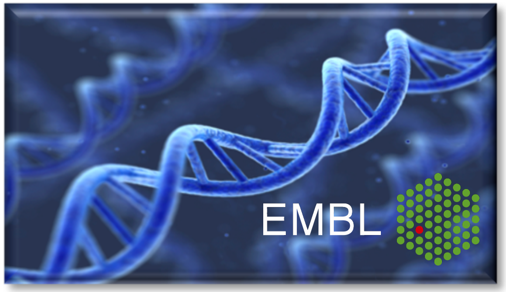
HELIX NEBULA THESCIENCECLOUD

ATLAS High Energy Physics Cloud Use



To support the computing capacity needs for the ATLAS experiment

Genomic Assembly in the Cloud



A new service to simplify large scale genome analysis; for a deeper insight into evolution and biodiversity

SuperSites Exploitation Platform

To create an Earth Observation platform, focusing on earthquake and volcano research

Call for proposals

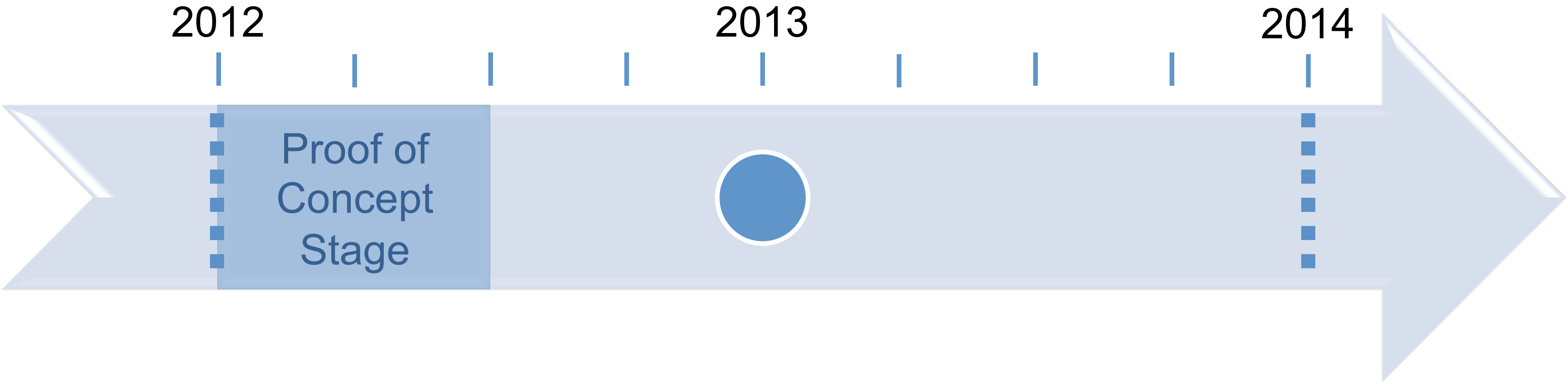
- Template agreed by demand and supply side
- Eligibility review and analysis with cloud service suppliers

Flagship use cases

	ATLAS H.E.P. Cloud Use (CERN)	Genomic Assembly in the Cloud (EMBL)	SuperSites Exploitation Platform (ESA/CNES/DLR)
Scientific goal, society impact, photogenic	✔	✔	✔
Scale of resources used	✔	✔	
Federation / aggregation of datasets		✔	✔
Long-term archiving of data			✔
On-demand processing	✔	✔	✔
Impact on community & benefits	✔	✔	✔
Potential increase of users	✔	✔	✔
Interoperability	✔	✔	✔
Data security	✔	✔	✔
Maturity	✔	✔	✔
Access to license-controlled software			✔

Flagship deployments

Pilot Phase



- Proof of Concept (PoC) stage started in January 2012
- Each flagship will be deployed with a series of providers independently
- Sequence: CERN-ATLAS, EMBL & ESA
- Initial PoC expected to be completed by summer 2012

Flagship use cases
Participating Suppliers

CloudSigma

Atos

interoute
from the ground to the cloud

logica
be brilliant together

terradue 2.0

sixsq.

T-Systems

the SERVER LABS
the IT architects

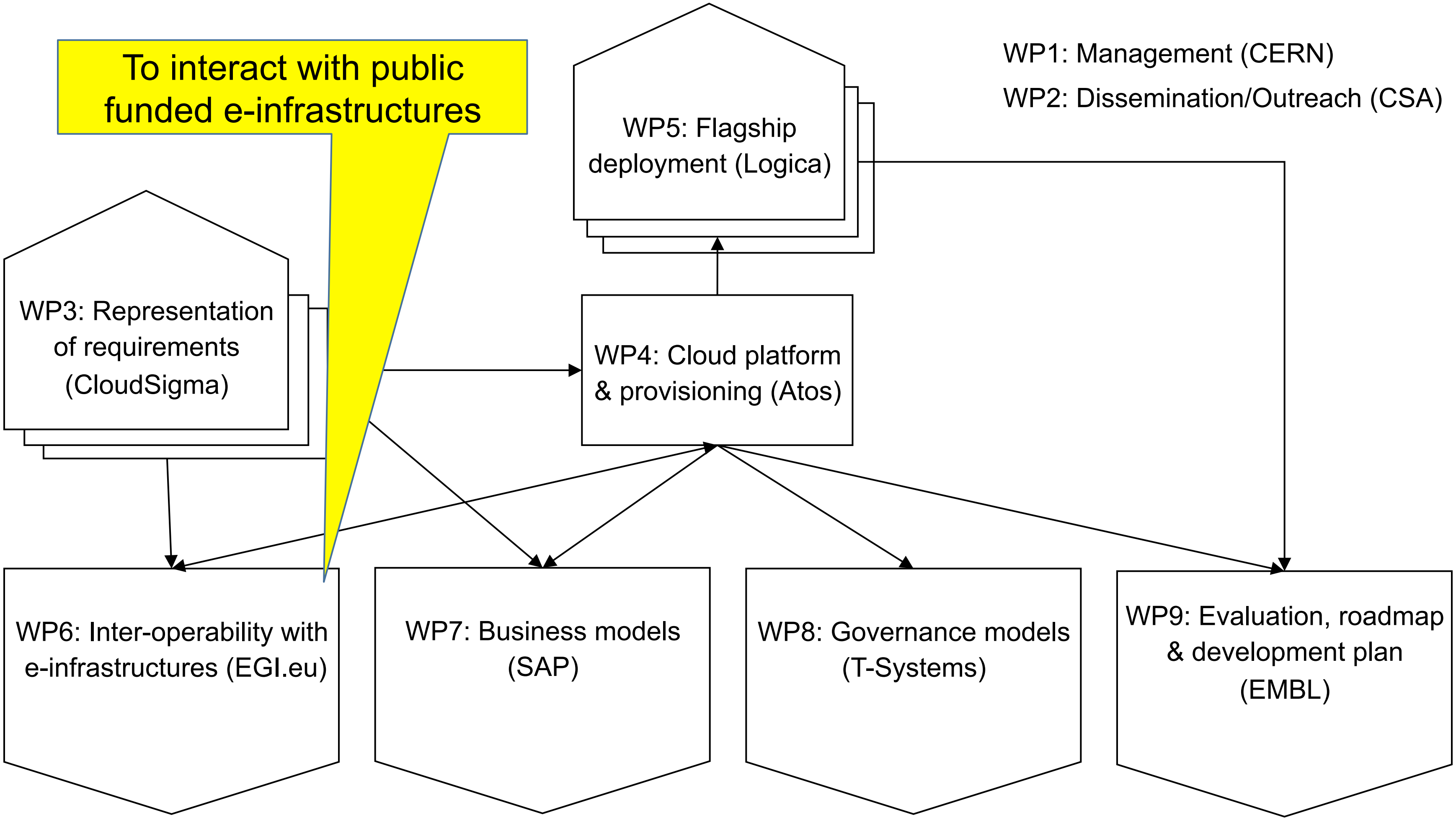
Helix Nebula EC project

Coordination action under call INFRA-2012-3.3

- Start-date 1st June 2012, duration 24 months
- Total budget ~3M€ (1.8M€ EC funding)

	Short name	Organisation	Country
1	CERN	European Organization for Nuclear Research (coord.)	CH
2	EGI.eu	STICHTING EUROPEAN GRID INITIATIVE	NE
3	EMBL	European Molecular Biology Laboratory	DE
4	Atos	ATOS	NE
5	T-Systems	T-Systems International GMBH	DE
6	CloudSigma	CloudSigma AG	CH
7	SAP	SAP AG	DE
8	Logica	Logica Deutschland GmbH & Co KG	DE
9	CNR	CONSIGLIO NAZIONALE DELLE RICERCHE	IT
10	CSA	Cloud Security Alliance Europe	UK

Helix Nebula proposal


To interact with public funded e-infrastructures
WP5: Flagship deployment (Logica)
WP1: Management (CERN)
WP2: Dissemination/Outreach (CSA)
WP3: Representation of requirements (CloudSigma)
WP4: Cloud platform & provisioning (Atos)
WP6: Inter-operability with e-infrastructures (EGI.eu)
WP7: Business models (SAP)
WP8: Governance models (T-Systems)
WP9: Evaluation, roadmap & development plan (EMBL)

Relevance of Helix Nebula for Network Community

- Helix Nebula to build hybrid cloud (public & private data centres)

- Research community moving towards using commercial cloud services (Helix Nebula not the only initiative in this domain)

- E-IRG response to GEANT 2020 vision paper:
  Importance of participation of private research in the use of research networks
  (http://www.e-irg.eu/images/stories/e-irgs_reaction_geg_a5.pdf)

e-IRG: e-Infrastructure Reflection Group

Relevance of Helix Nebula for Network Community (2)

Helix Nebula pilot phase provides

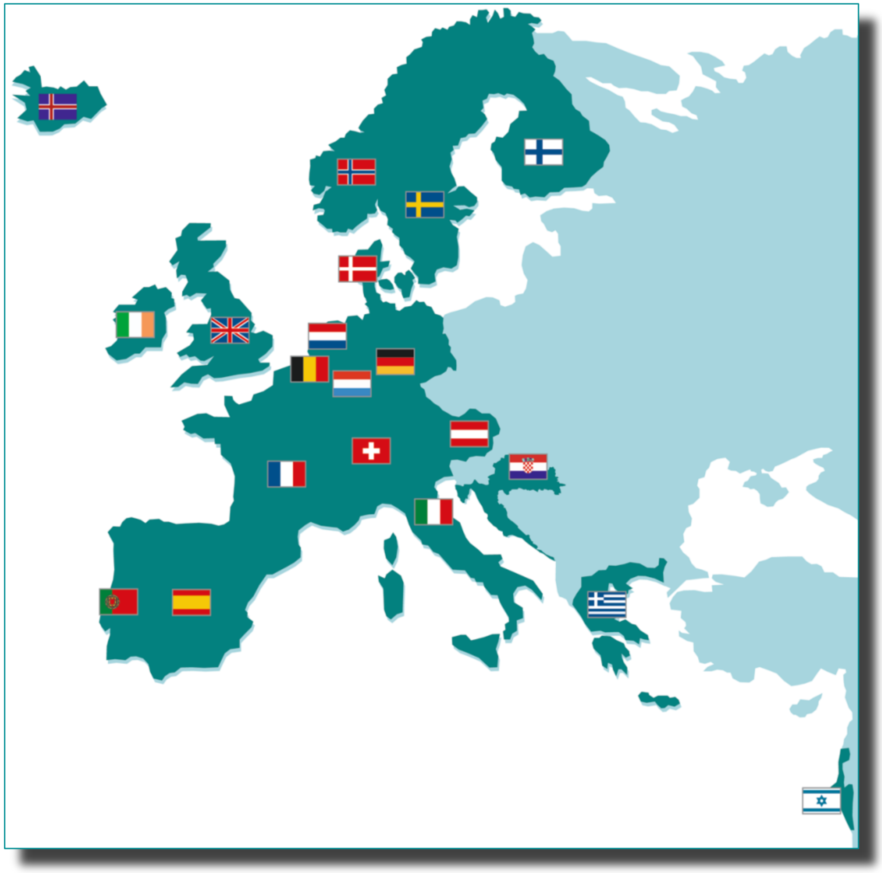
- Opportunity for NRENs & network community
- to work with the research communities and commercial cloud service providers to deploy flagship applications
- to evaluate ability of NRENS to offer access to commercial data centres
- to investigate how a public-private cloud serving the research community could exist

e-IRG: e-Infrastructure Reflection Group

Genomic Assembly in the Cloud

EMBL Flagship Use Case

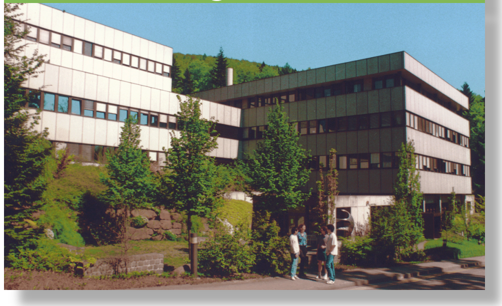
EMBL: European Molecular Biology Laboratory



- Intergovernmental Research Organization
- Supported by 20 Member States (+1 associated: 🇦🇺)
- One of the world‘s foremost life science institutions
- EIROforum member
- 1500 staff
  >70 nationalities

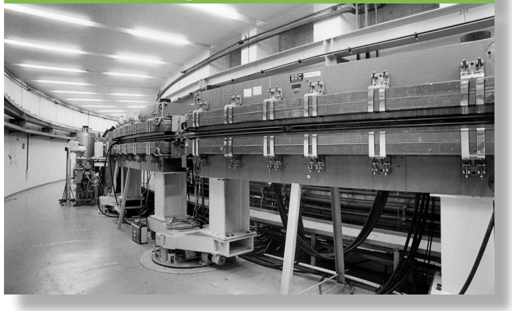
The Five Branches of EMBL

Heidelberg



Basic Molecular Biology Research
Main Lab / Headquarters

Hamburg



Structural Biology
DESY

Hinxton

European Bioinformatics Institute (EBI)
Sanger Centre

Grenoble

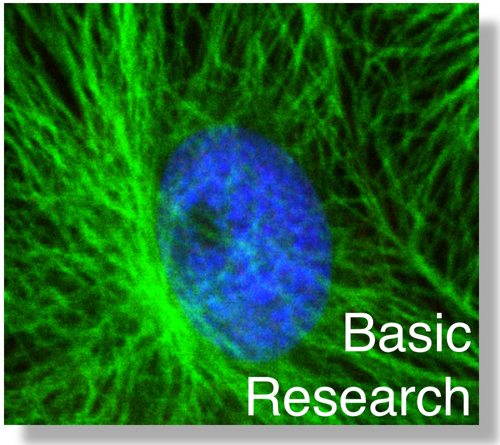
Structural Biology
ILL, ESRF, IBS, UVHCI

Monterotondo

Mousebiology
CNR, EMMA

EMBL’s Missions

Services


Basic Research

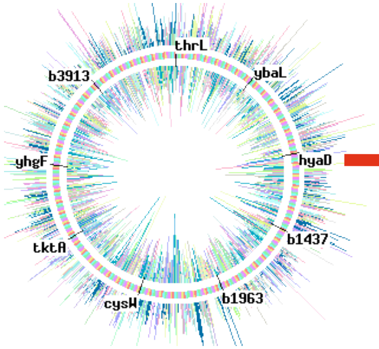
Technology Transfer

Advanced Training

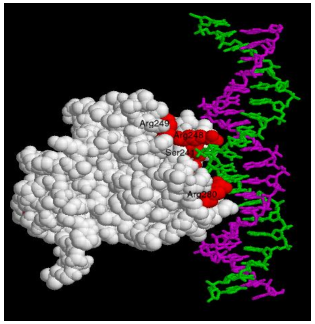
European Integration

Instrument and Technology Development

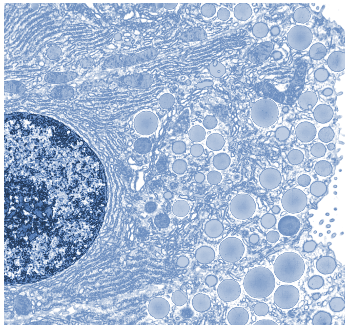
Systems Biology: From Molecules to Organisms



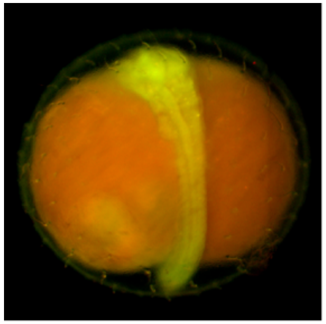
Genome



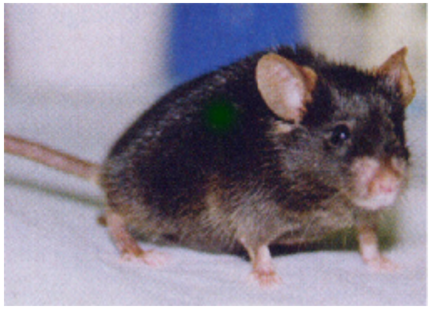
Protein/DNA



Cell

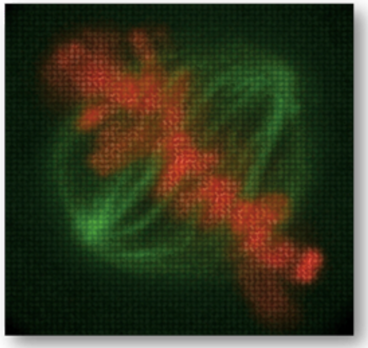


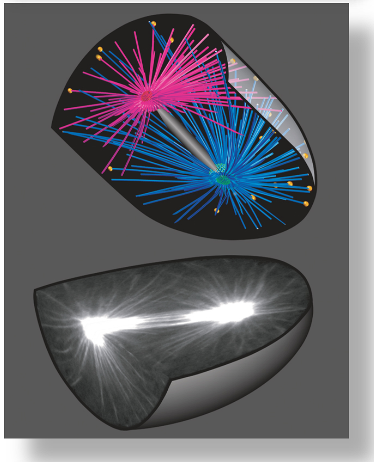
Embryo

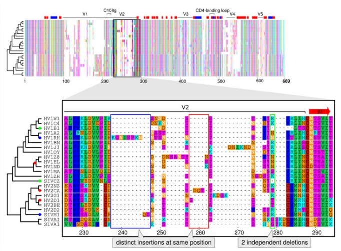


Development Organisms Complexity Aging Disease

Research Directions & Key Technologies







Imaging

- Bridging scales of biological organisation: combine low- and high-resolution techniques
- Biology in four dimensions: live imaging to study dynamic processes in space and time
- Generate quantitative data

Computational Biology

- Analysing, integrating and exploiting quantitative data
- Build predictive networks and models of biological processes

Next Generation Sequencing

- Inter-species variation: comparative sequence analysis to study evolution
- Intra-species variation: link genetic variation to phenotype

Disease Models and Mechanisms

- Decipher the molecular basis of genetic and infectious diseases with the help of animal and cellular models

Exemplary Big Data Challenge

NEXT GENERATION SEQUENCING (NGS)

DNA and Life on Earth

The Sequence Holds the Code for the Organism

Next Generation Sequencing (NGS) Revolution

1990 1 thousand DNA bases / day

2000 1 million DNA bases / day

2010 1 billion DNA bases / day

NGS Impact on Human Genome Sequencing

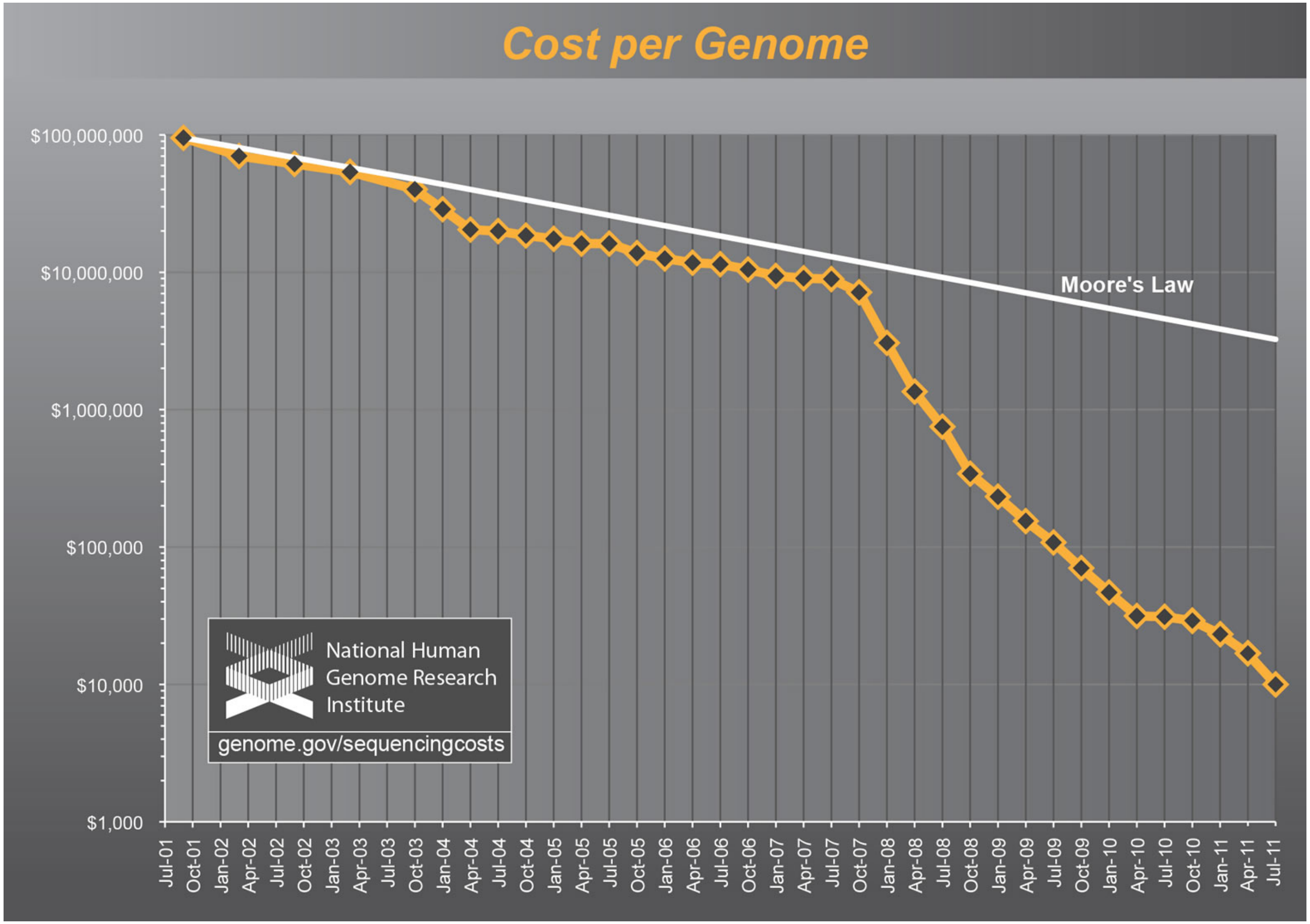
- Human genome project
  - 10 years
  - Large International Consortium
  - Thousands of Sequencers
  - $3,000,000,000

- Sequencing today
  - < $10,000
  - A few hours
  - One machine

2000

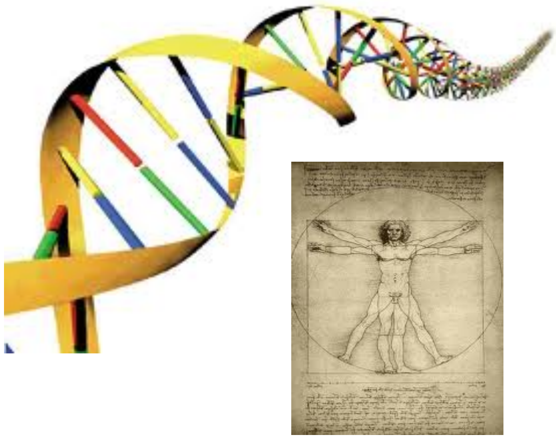
2010

Cost of Sequencing Decreasing Rapidly

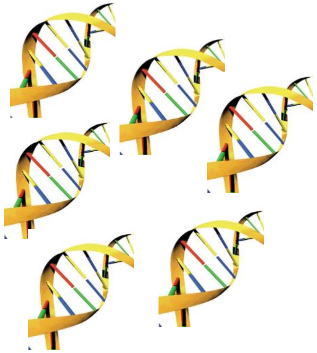


Read the Sequence to Study the Organism

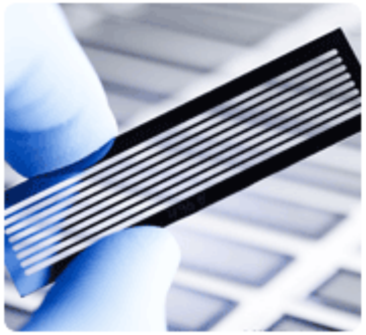
Extract DNA



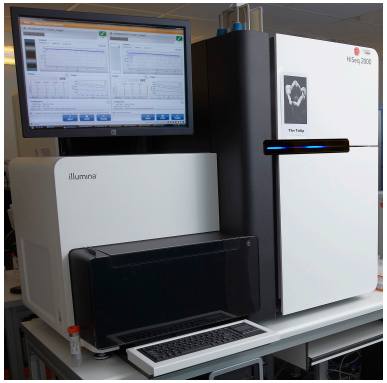
Fragment



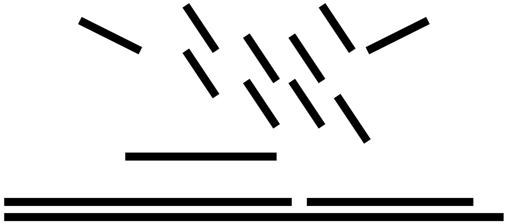
Prepare



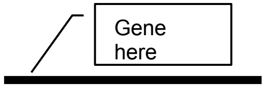
Sequence



Assemble



Annotate

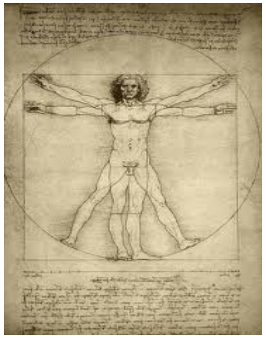


Requires Computing Infrastructure & Expertise

Genomic Sequencing is Now an Affordable Solution

Academic Research Groups

Medical Research



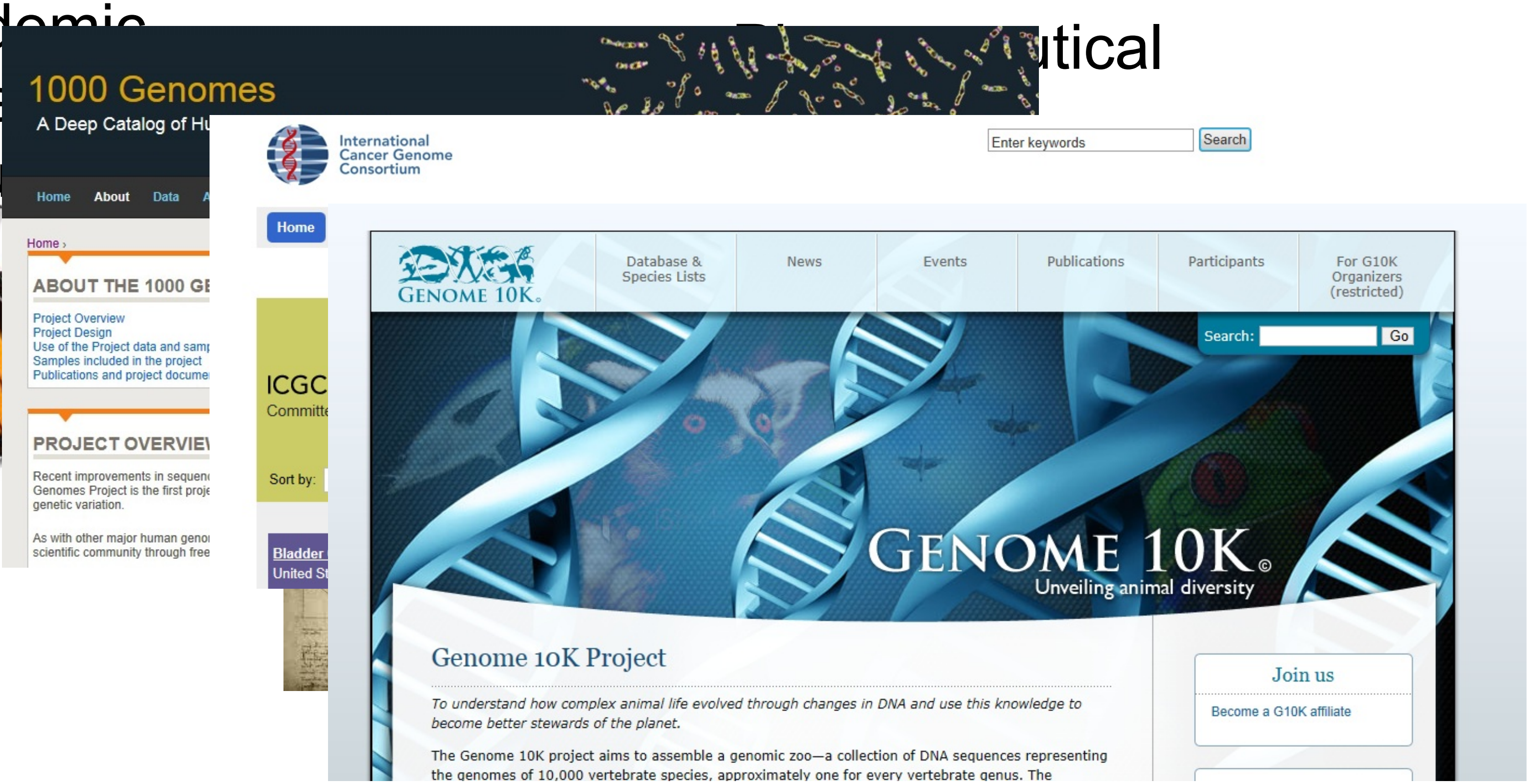
Pharmaceutical Companies

Agricultural Research

Genomic Sequencing is Now an Affordable Solution

Academic
Rese
Grou

utical



1000 Genomes
A Deep Catalog of Hu

Home ›

ABOUT THE 1000 GE

Project Overview
Project Design
Use of the Project data and samp
Samples included in the project
Publications and project docume

PROJECT OVERVIE

Recent improvements in sequen
Genomes Project is the first proje
genetic variation.

As with other major human geno
scientific community through free

International Cancer Genome Consortium

ICGC
Committe

Sort by:

Bladder
United St

Genome 10K Project

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The

Join us

Become a G10K affiliate

Genomic sequencing is now an affordable solution

but ...

Problem – 1: Assembly

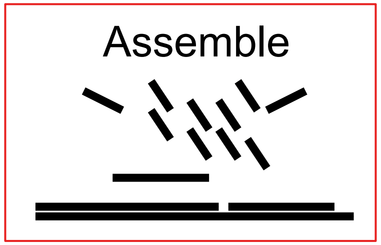
ATGCATT... 200,000,000 ...TGCGGATC

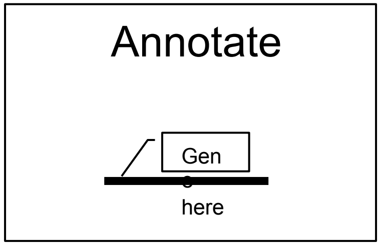
Genomes contain long strings of bases

ATGCATT... 105 ... GTATTCC

NGS output: millions of very short sequence reads

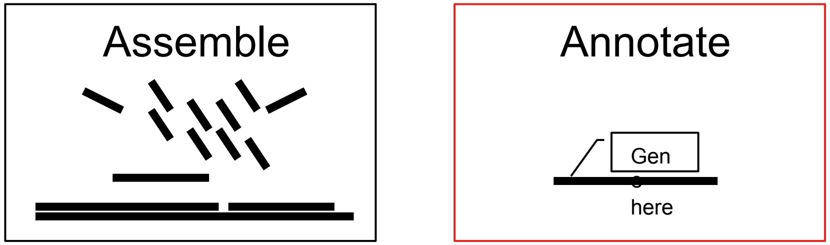
- The short reads have to be assembled into genomes
- Up to 1TB RAM required to solve puzzle



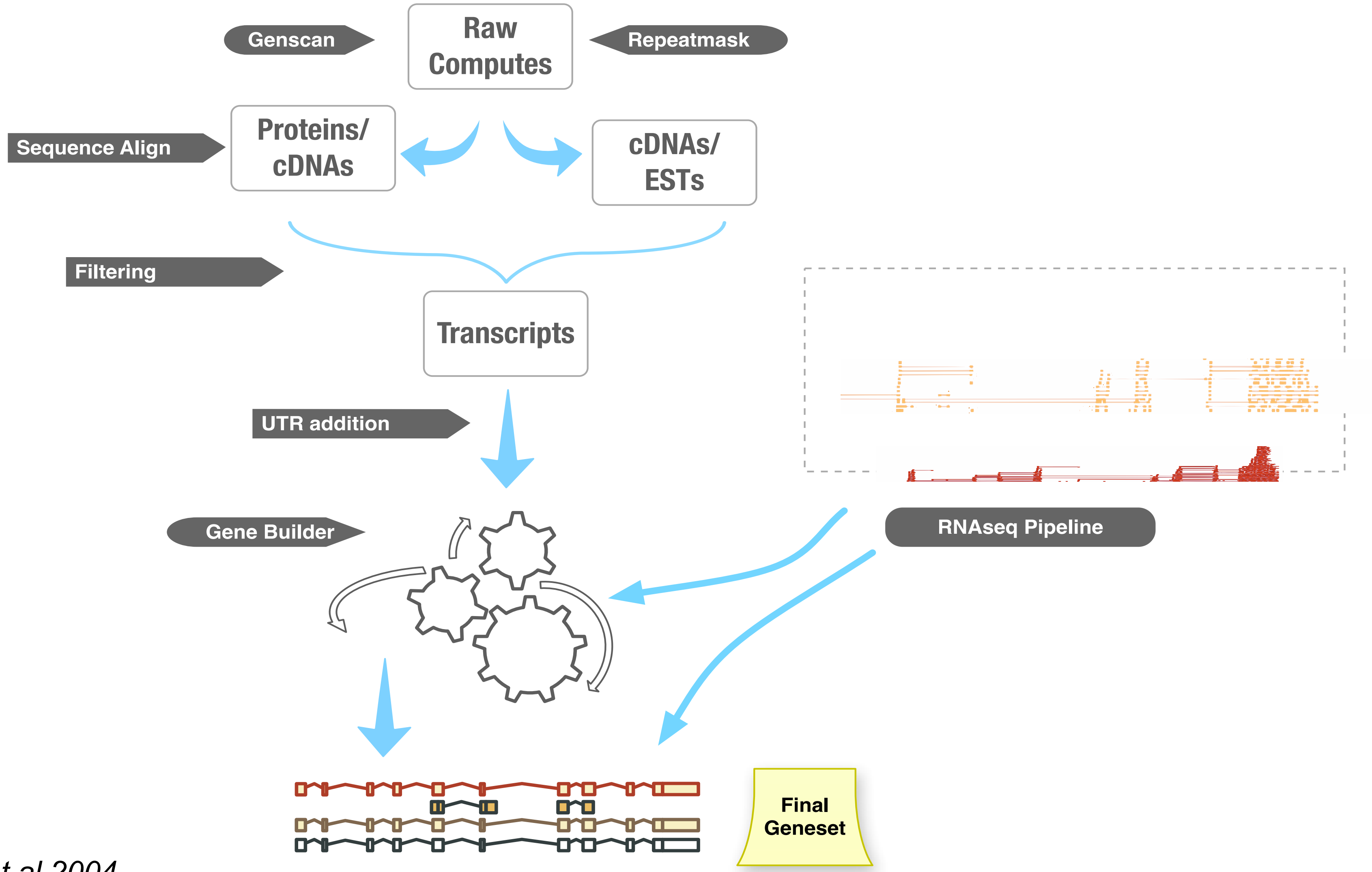


Problem – 2: Annotation

- Strings of assembled bases need to be annotated
- Document features inside the code
- 3 billion bases ~25k genes
- Looking for genes, gene and promoter sequences
- Requires multiple pipelines and databases



Ensembl Annotation Pipeline



Genscan

Raw Computes

Repeatmask

Sequence Align

Proteins/ cDNAs

cDNAs/ ESTs

Filtering

Transcripts

UTR addition

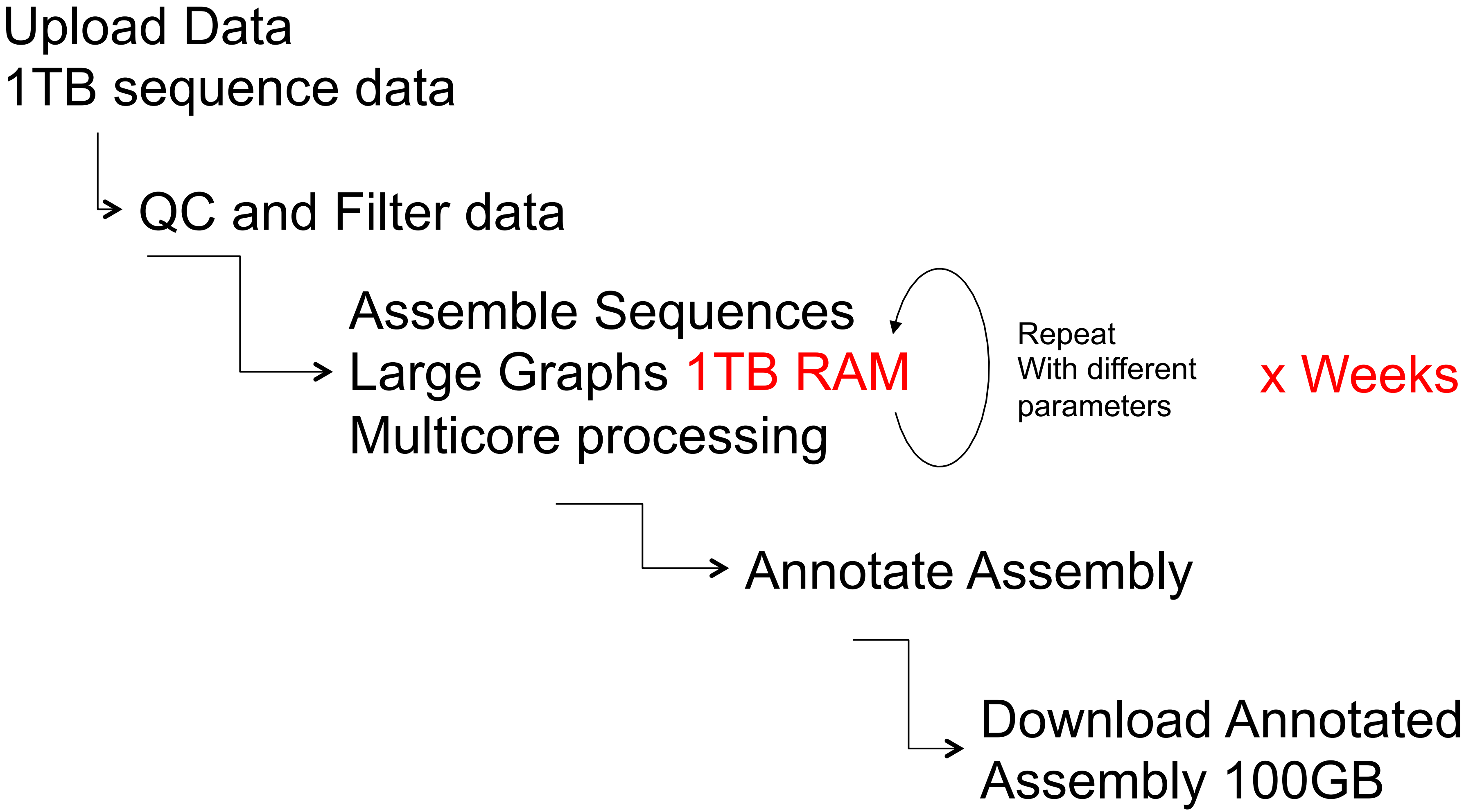
Gene Builder

RNAseq Pipeline

Final Geneset

Potter et al 2004

Computational Steps Involved



Upload Data
1TB sequence data

QC and Filter data

Assemble Sequences
Large Graphs 1TB RAM
Multicore processing

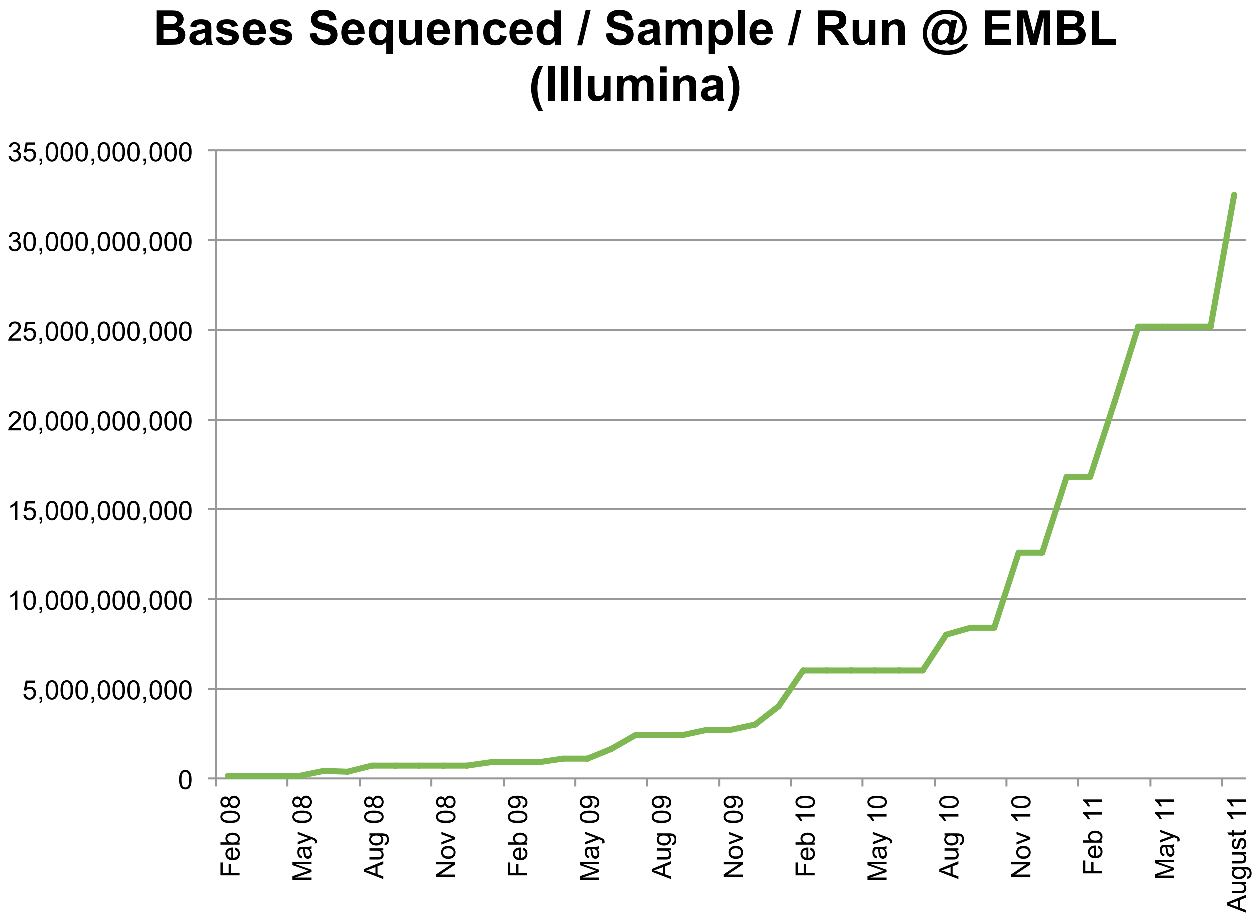
Repeat
With different
parameters

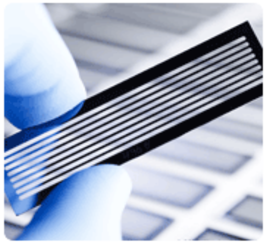
x Weeks

Annotate Assembly

Download Annotated
Assembly 100GB

Problem - Technology Explosion with NGS



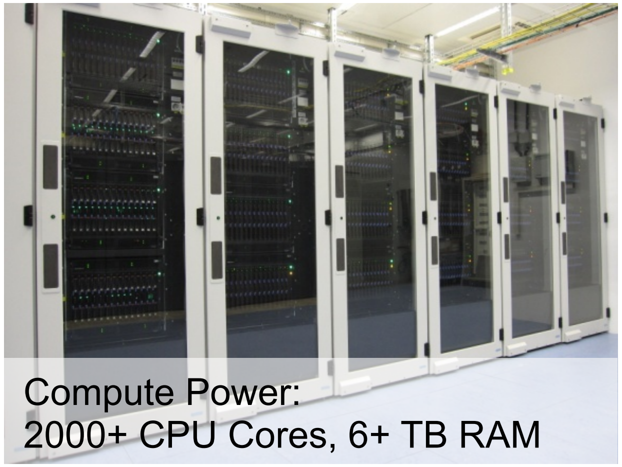


Sequence Production & IT Infrastructure at EMBL

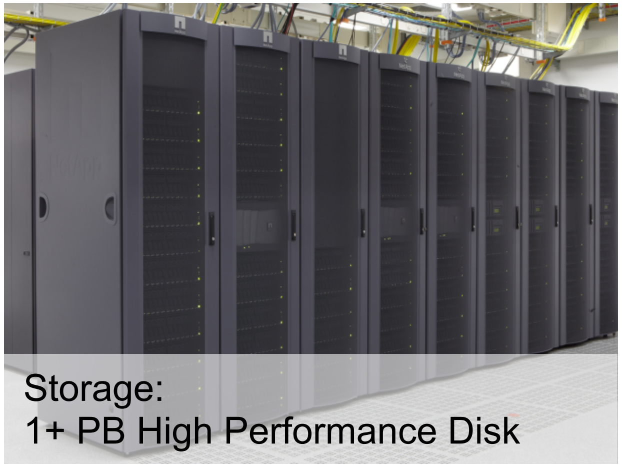
4 x Ilumina HiSeq2000

2 x Ilumina GAIIx

25 TB data each week



Compute Power:
2000+ CPU Cores, 6+ TB RAM

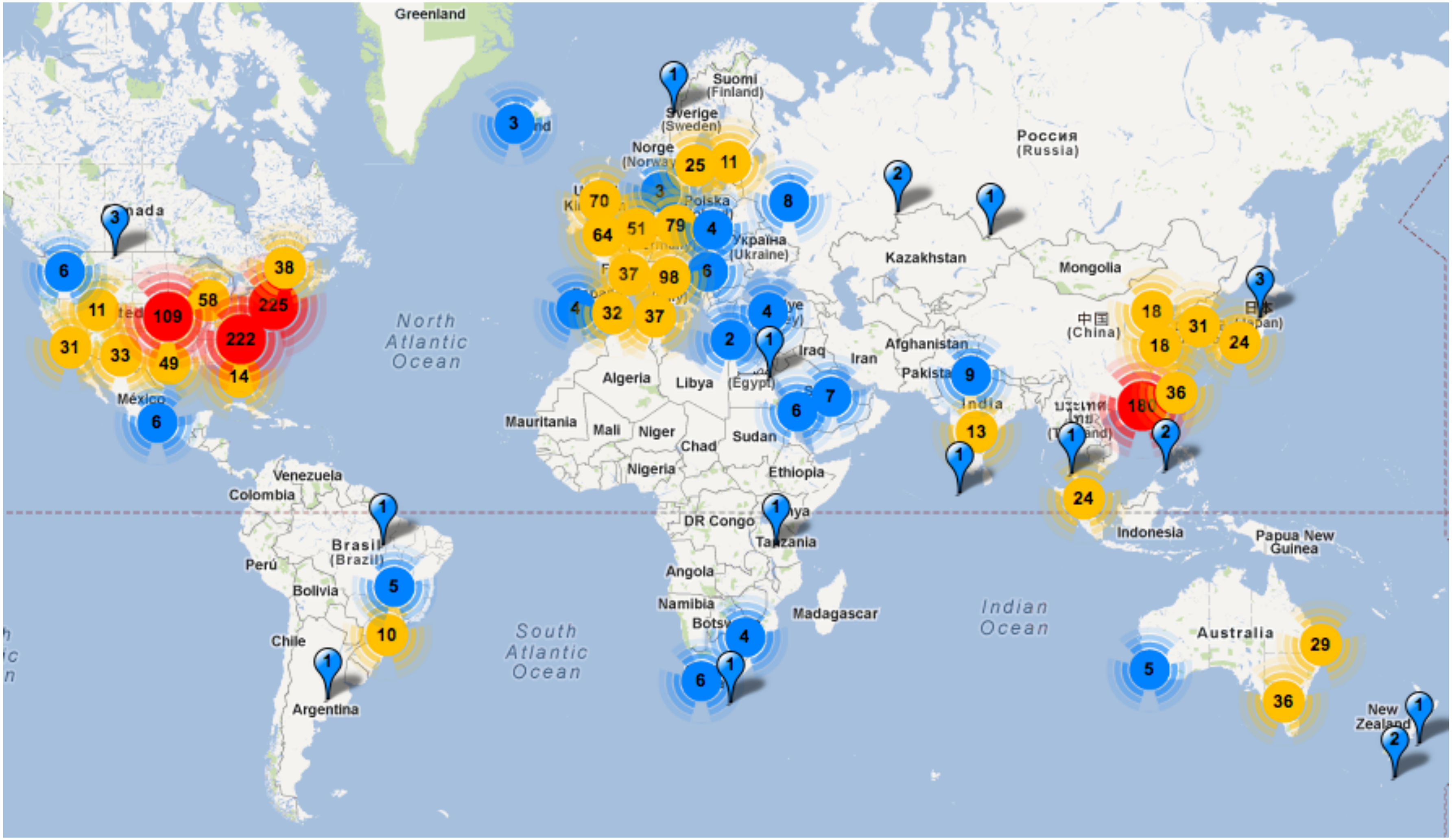


Storage:
1+ PB High Performance Disk

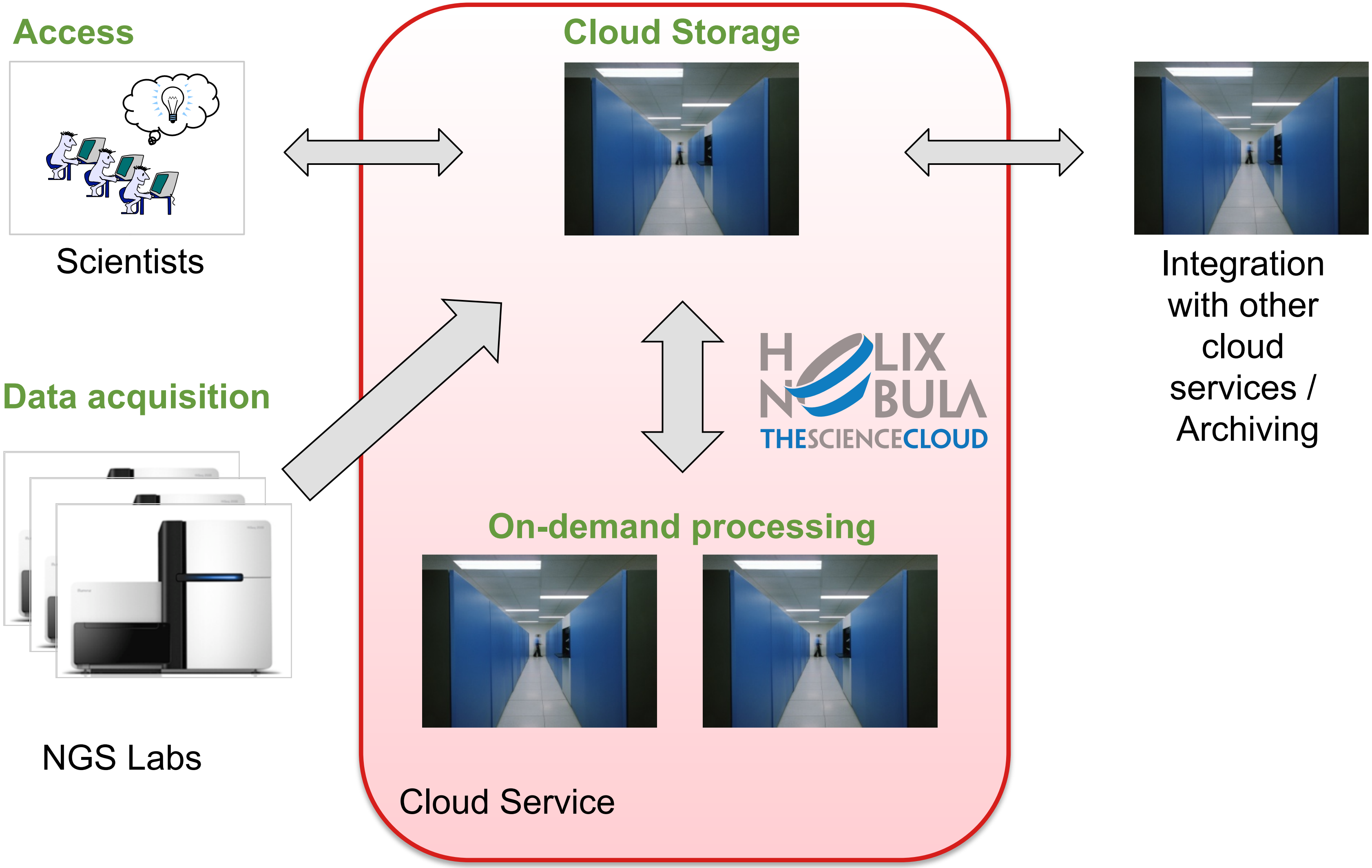
NGS - The Big Picture

- ~ 8.7 million species in the world (estimate)
- ~ 7 billion people
- Sequencers exist in both large centres & small research groups
- 200+ Ilumina HiSeq sequencers in Europe alone
    - capacity to sequence 1600 human genomes / month
- Largest centre: Beijing Genomics Institute (BGI)
    - 167 sequencers, 130 HiSeq
    - 2,000 human genomes / day
- 500-1000 Hiseq devices worldwide today
    - 3-6 PB /day
    - 1.1 – 2.2 ExaBbytes / year

World Map of High-throughput Sequencers



EMBL Flagship project: Whole-Genome Assembly



Access

Scientists

Data acquisition

NGS Labs

Cloud Storage

HELIX NEBULA THESCIENCECLOUD

On-demand processing

Cloud Service

Integration with other cloud services / Archiving

A European cloud computing partnership: big science teams up with big business

HELIX NEBULA THESCIENCECLOUD

HELIX NEBULA THESCIENCECLOUD

Strategic Plan

- Establish multi-tenant, multi-provider cloud infrastructure
- Identify and adopt policies for trust, security and privacy
- Create governance structure
- Define funding schemes

CERN

To support the computing capacity needs for the ATLAS experiment

EMBL

Setting up a new service to simplify analysis of large genomes, for a deeper insight into evolution and biodiversity

esa

To create an Earth Observation platform, focusing on earthquake and volcano research

Atos | Capgemini CONSULTING.TECHNOLOGY.OUTSOURCING | CloudSigma | egi | interoute from the ground to the cloud

logica be brilliant together | OpenNebula.org The Open Source Toolkit for Cloud Computing | orange™ Business Services | SAP® | the SERVER LABS the IT architects

sixsq | Telefónica | terradue 2.0 | THALES | T-Systems

contact@helix-nebula.eu @HelixNebulaSC HelixNebula.TheScienceCloud