



HealthGrid, the EU Share project and the Embrace project

V. Breton, CNRS-IN2P3, France

<http://www.embracegrid.org>

- **Where to find good information about diseases and integrate this with information available in biomedical data resources ?**
- **Goals of this talk**
 - Inform you on present attempts to achieve integration of distributed and heterogeneous molecular biology data
 - Embrace
 - Inform you on present attempts to achieve collection and integration of disease related data using grid technology
 - Inform you about an international initiative to foster the adoption of grid technology for medical research and healthcare
 - The HealthGrid initiative
 - Share, a roadmap to the adoption of grids for health

- **Ontologies are about defining common vocabulary and semantics to handle data**
 - Ontology defines data entities, data attributes, relations and possible functions and operations
- **Agreed ontologies are needed but are not sufficient**
 - Data have to be exposed according to the ontology
 - How to foster adoption of ontologies ?
 - What about the existing data sources which are not compliant with the ontology ?
 - More generally, data must be accessible
 - Data exposed according to the ontology must be accessible
 - private medical data vs public molecular biology databases
- **Need for interoperability to enable the deployment and adoption of ontologies**

Embrace approach to information integration



- **Embrace aims at building a distributed infrastructure allowing integrated exploitation of biomolecular data**
 - collection, curation and provision of biomolecular information
 - Availability of most of the popular databases and software products
 - tools and programming interfaces to exploit that information
 - taking away the need for maintaining local copies of databases and software

What are the foundations of the Embrace Grid ?



- **Technology: Web Services**
 - Interoperability
 - Strong support from computer science industry
 - Wide adoption in the Grid community
- **Implementation standards: WSDL and WS-I**
 - Interoperability
 - Strong support from computer science industry
 - Wide adoption in the grid community
- **Communication protocol: SOAP**
 - Interoperability
 - Strong support from computer science industry
 - Wide adoption in the grid community

The Embrace objective

Web service interfaces to the tools

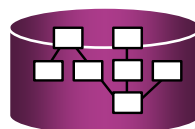


Client: TAVERNA, TRIANA,...

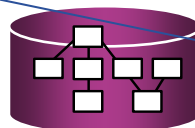
Input data

Web service interfaces to the databases

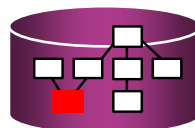
Database 1



Database 2



Database 3



algorithm 1

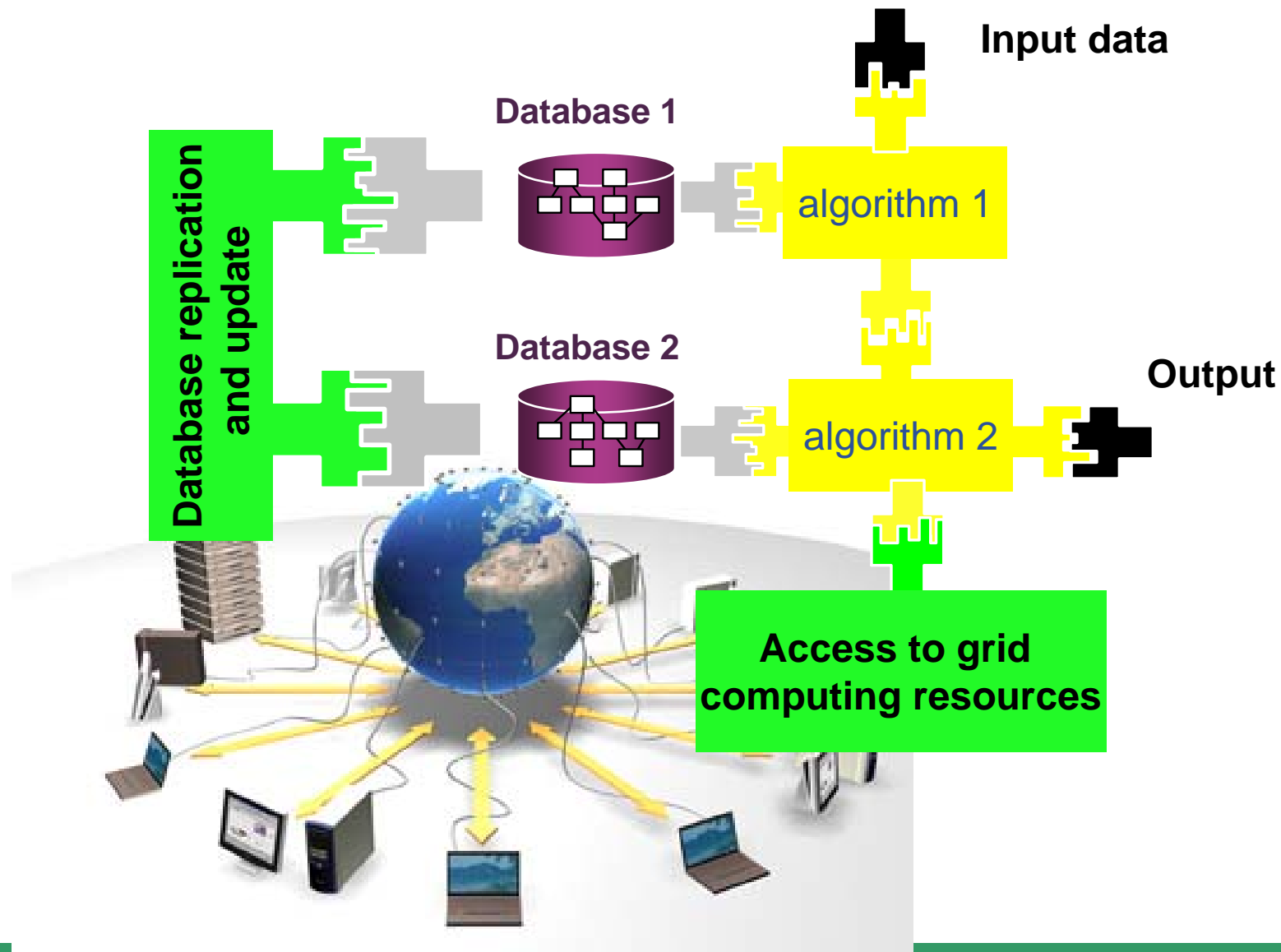
algorithm 2

algorithm 3

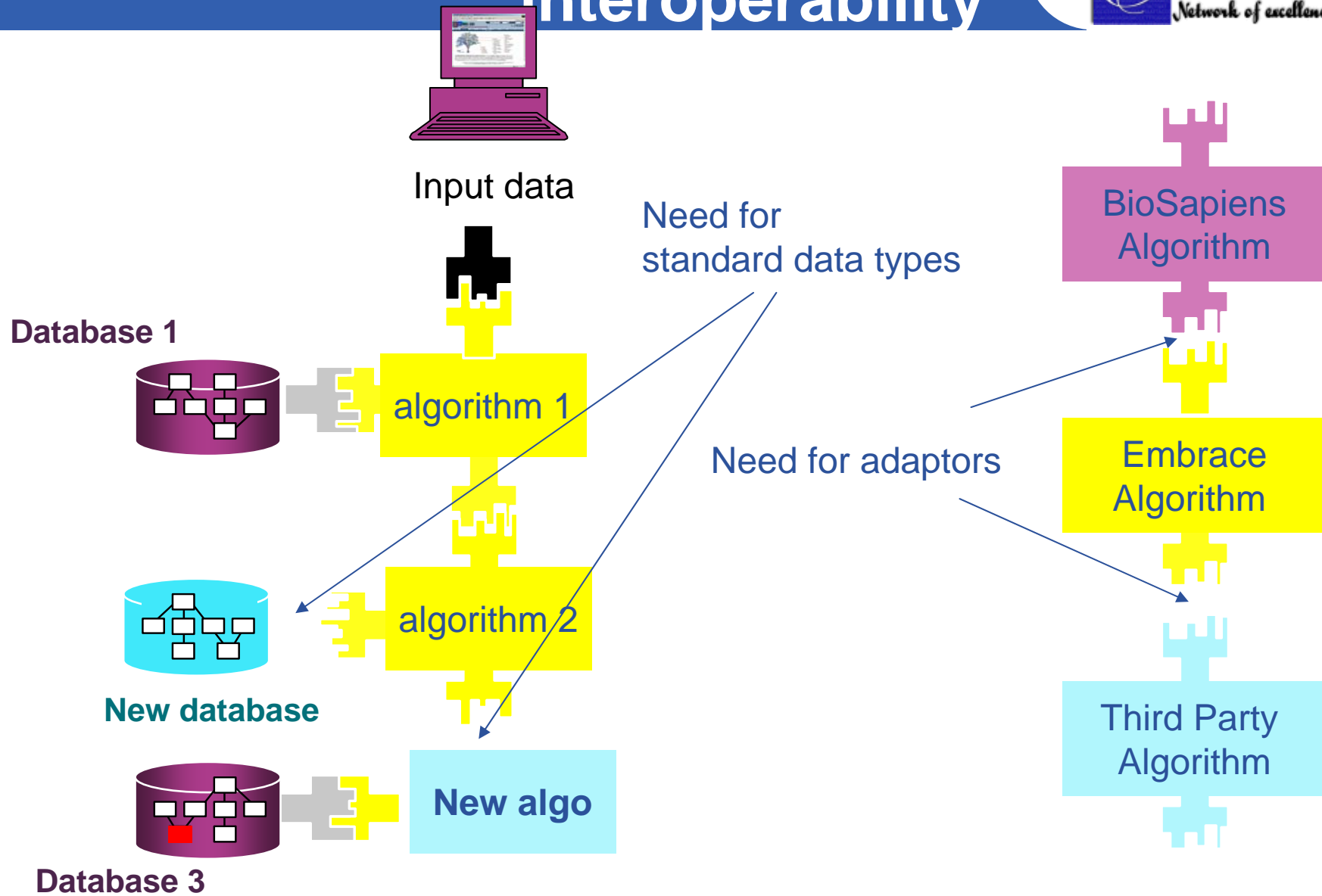
biological test case

Output

The grid added value today

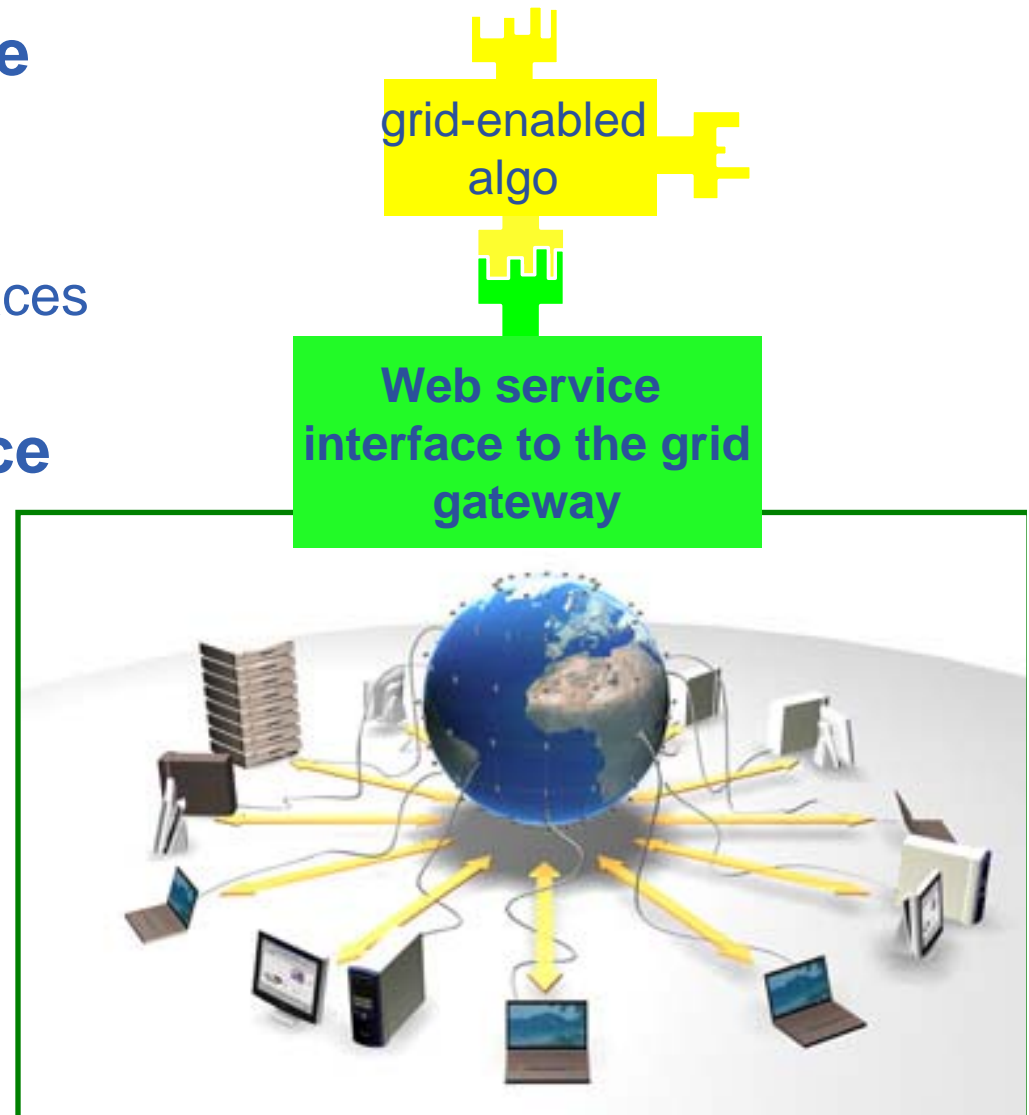


Challenge : achieving interoperability



Challenge: web service interfaces to grid services

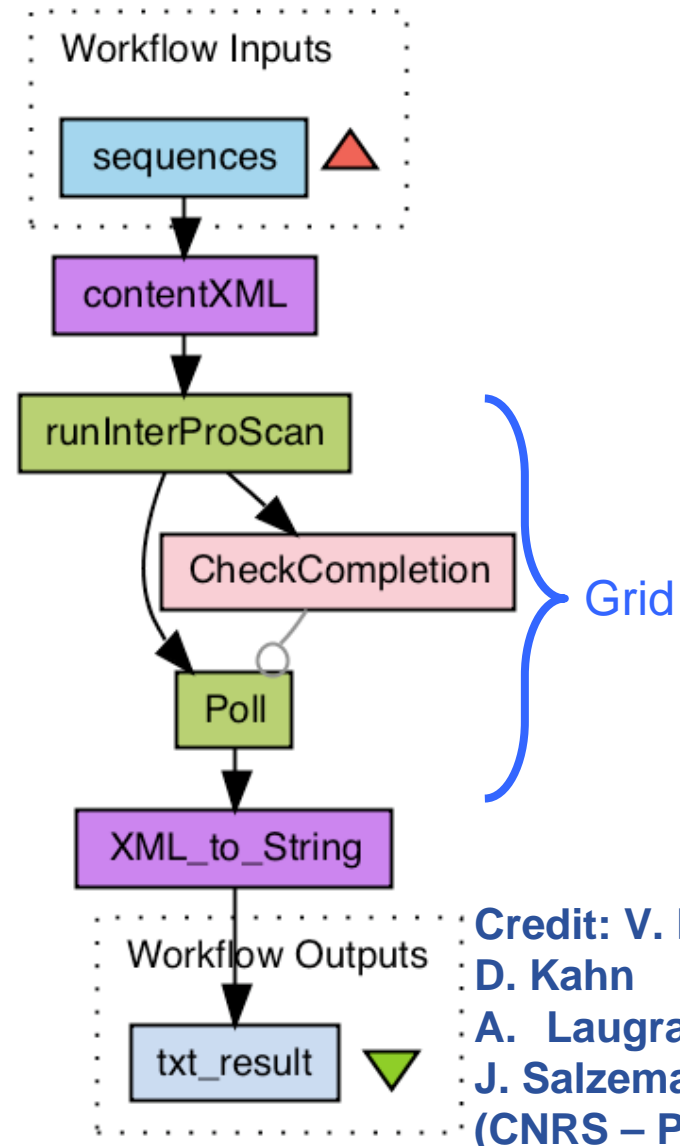
- Existing grid infrastructures in Europe are not
 - WSDL compliant
 - Using web services interfaces for calling grid services
- Need to build web service interface to the grid services through its gateway



Application: protein domain annotation



- **InterProScan is a service that combines different protein signature recognition methods specific to InterPro databases.**
- **InterProScan is very CPU demanding**
 - users cannot submit many sequences at once
- **InterProScan on the grid**
 - Fast response to punctual user request
 - Allow users to submit many sequences at once
 - Workflow enabled service

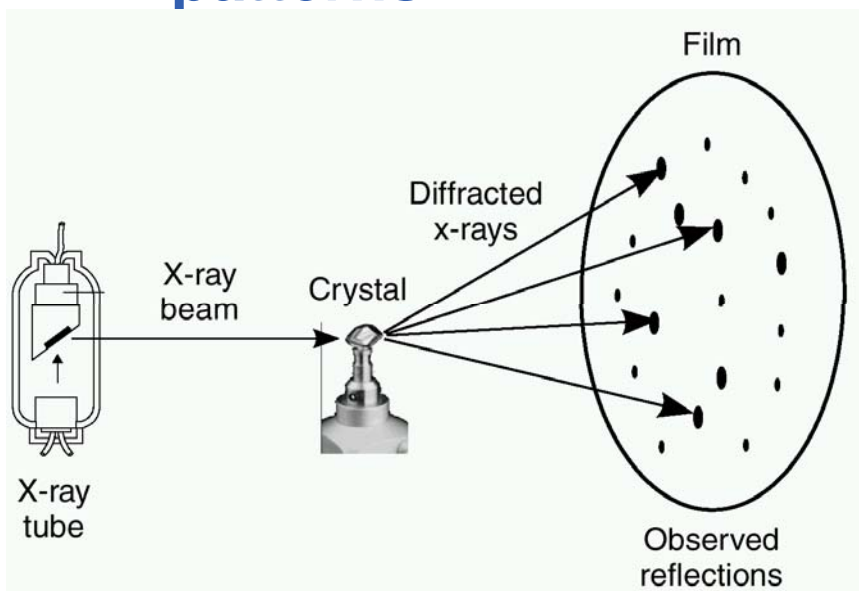
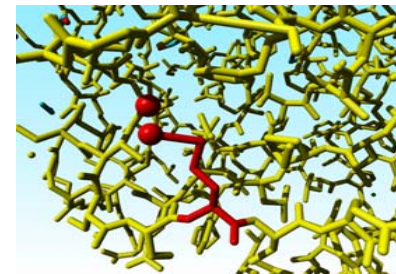
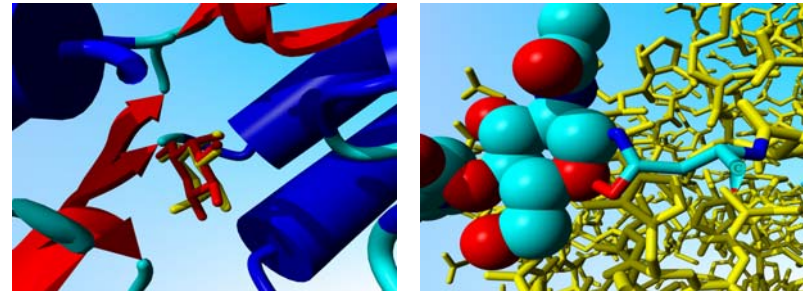


Credit: V. Bloch,
D. Kahn
A. Laugraud,
J. Salzemann
(CNRS – PRABI)

Application: recalculating protein 3D structures in PDB



- The PDB data base gathers publicly available 3D protein structures
 - Full of bugs
- Goal: redo the structures by recalculating the diffraction patterns



PDB-files	42.752
X-ray structures	36.124
Successfully recalculated	~36.000
Improved R-free	12.500/17000
CPU time estimate	21.7 CPU years
Real time estimate	1 month on Embrace Virtual Organization on EGEE

Credit: G. Vriend, R. Joosten CMBI

FlexX,
Gold,
Autodock

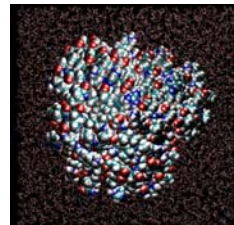


Millions

Molecular docking

- Large scale virtual screening of drug-like compounds

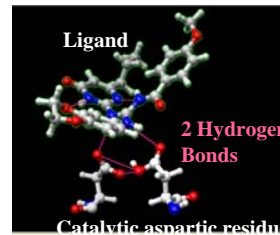
- Very high throughput: up to 100.000 docked compounds per hour on the grid



5000

Molecular dynamics

AMBER



Re-ranking
MMPBSA-GBSA

180

Complex
visualization

- Compounds patented against malaria and avian flu

WET LABORATORY



30

In vitro
tests

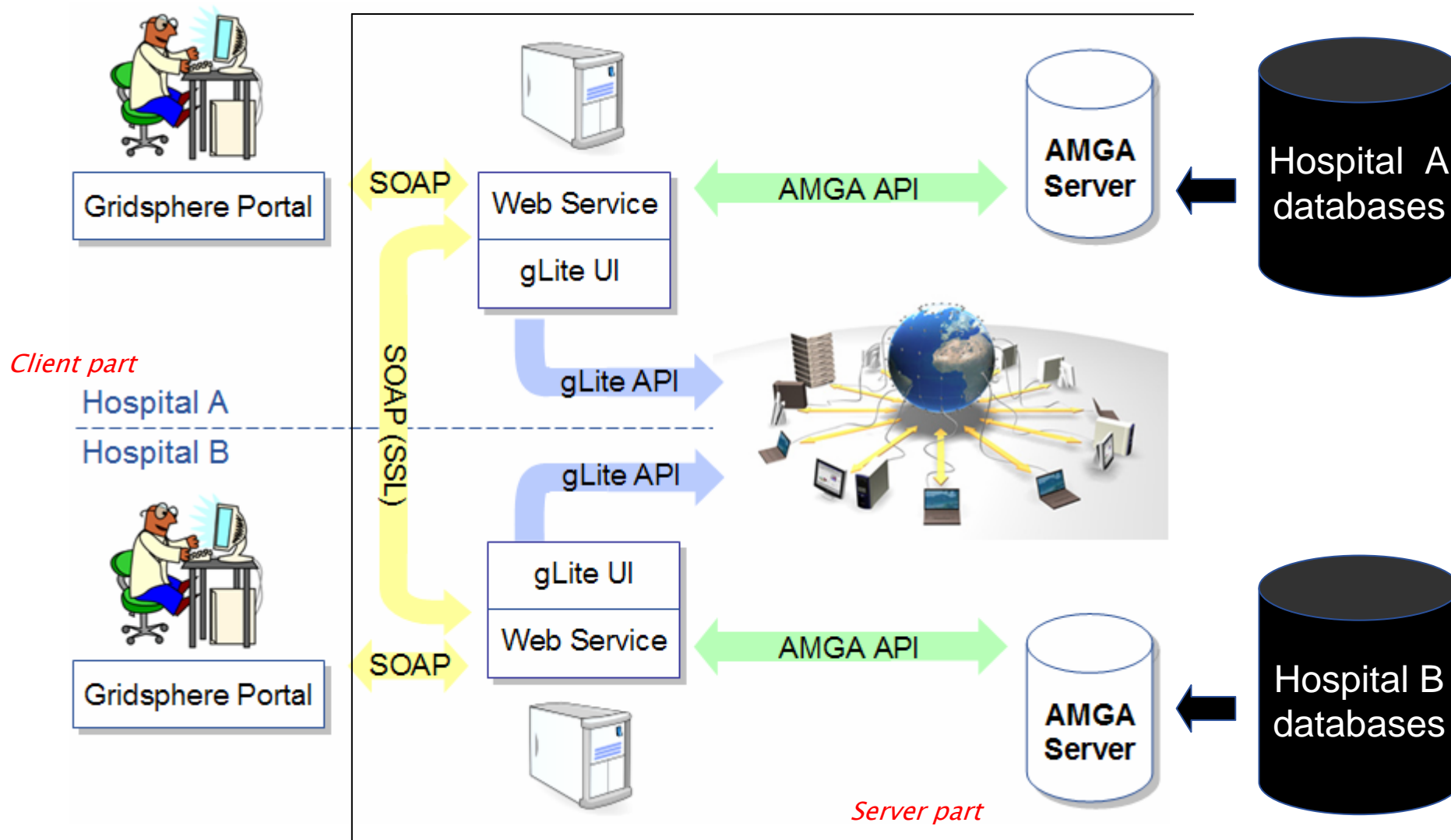
CHIMERA

Integration of medical data: grid added value



- **Grid technology offer today secured data management services**
 - AMGA and MDM in the case of gLite middleware, Medicus in the case of Globus Toolkit 4
- **These services open up a whole family of applications handling distributed biomedical data**
 - Offer a set of different components for handling distributed medical data
 - Not a ready-to-use application
 - Ensure data security and keep privacy.
 - Crucial in medical world
 - Leave the data where they are produced
 - Typically in hospitals or medical structures
- **Take advantage of grid services: storage, computation and communication**

The HOPE platform

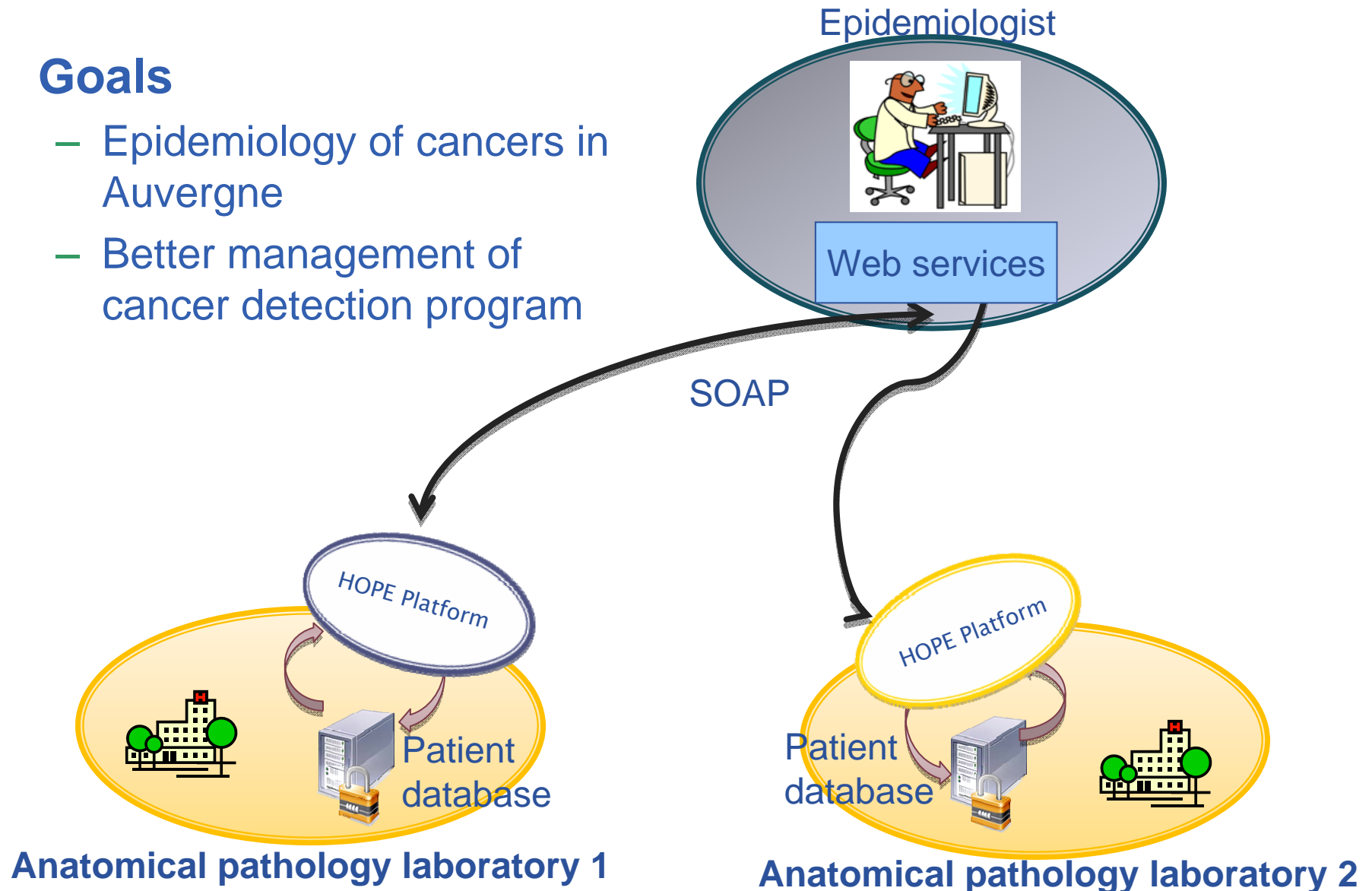


Application: cancer surveillance network



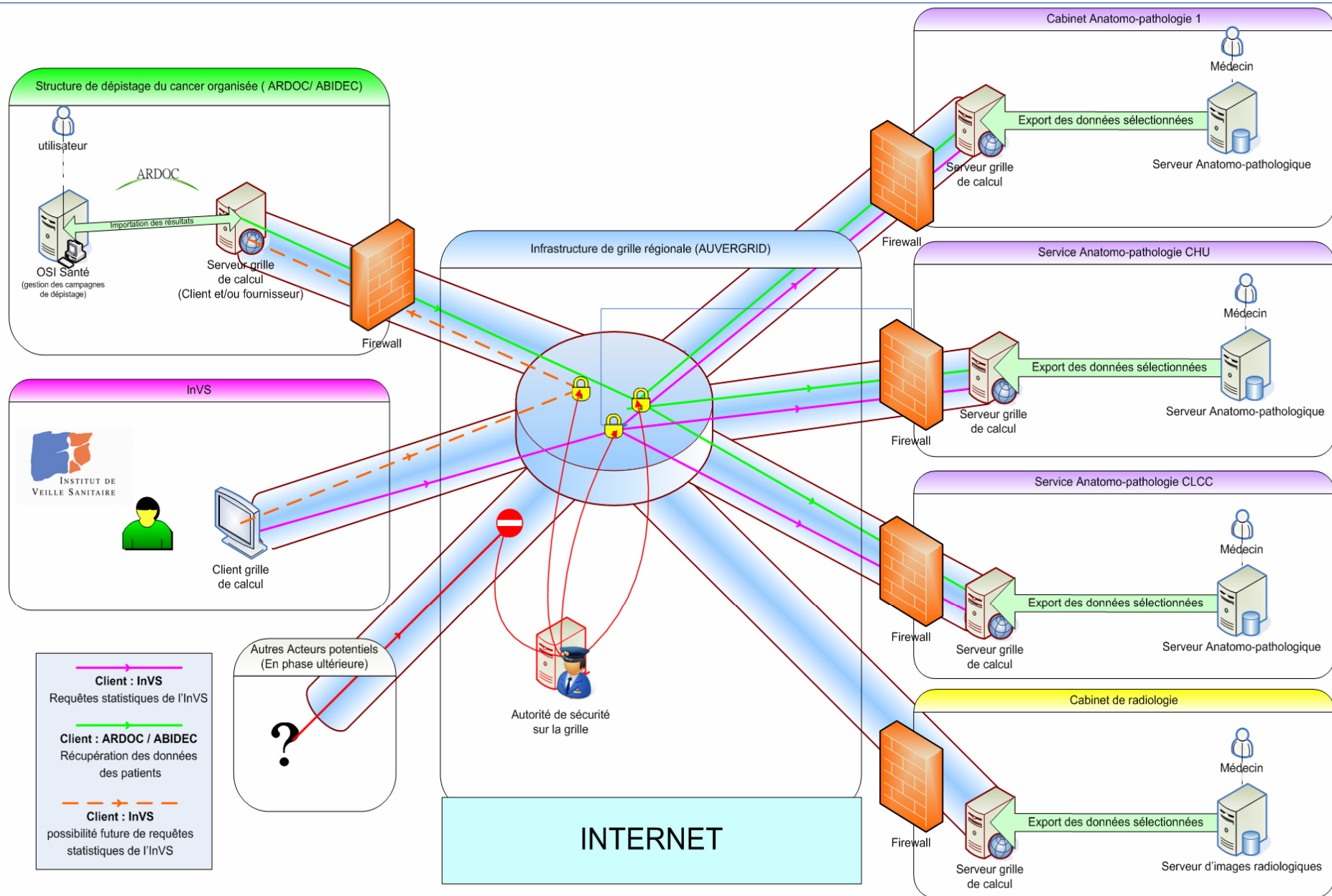
- **Goals**

- Epidemiology of cancers in Auvergne
- Better management of cancer detection program



Modélisation du projet Réseau Sentinelle

mercredi 18 juin 2008



The concept of HealthGrid



- **Environment where data of medical interest can be stored, processed and made easily available,**
- **To different actors in healthcare**
 - citizens,
 - physicians,
 - healthcare centres & administrations,
 - medical & biological research centres,
- **With all necessary guarantees in terms of**
 - security,
 - respect for ethics,
 - observance of regulations

The HealthGrid initiative



- **Promote the application of advanced information technology to solve cutting-edge problems in Biomedical Science and Healthcare.**
 - focus on Grid Technology
 - Commitment to open, interoperable systems and standards
 - technology neutral
- **Means**
 - HealthGrid associations: HealthGrid Europe and HealthGrid US
 - Yearly HealthGrid conferences
- **Key documents**
 - HealthGrid White Paper
 - **SHARE Roadmap for Healthgrid adoption**



<http://www.healthgrid.org>



From data collection to knowledge management: requirements



Innovative Medicines Initiative (IMI) Strategic Research Agenda

- **The capture, analysis and interpretation of knowledge generated regarding the physiology and pathophysiology related to disease stage or toxicological targets**
- **The capture, analysis and interpretation of knowledge generated for one potential drug candidate from discovery, non-clinical and clinical development all the way to lifecycle management.**

What are the requirements ?



- Capacity to search, query, extract, integrate and share data in a scientifically and semantically consistent manner across heterogeneous sources (public and proprietary) ranging from chemical structures and “omics” to clinical trial data
- Capacity to integrate and share scientific tools (e.g., modelling, simulation) as modules in a generic framework and apply them to relevant dynamic data sets,
- Expressive data representation and exchange standards,
- Dynamic and customizable configuration of applications,
- Encapsulation of validated physiological models, when applicable,
- Flexible, secure (covering all aspects of data protection encountered in a biomedical context), and scalable IT infrastructure.

Innovative Medicines Initiative (IMI) Strategic Research Agenda

Grids are the answer provided technical challenges are overcome



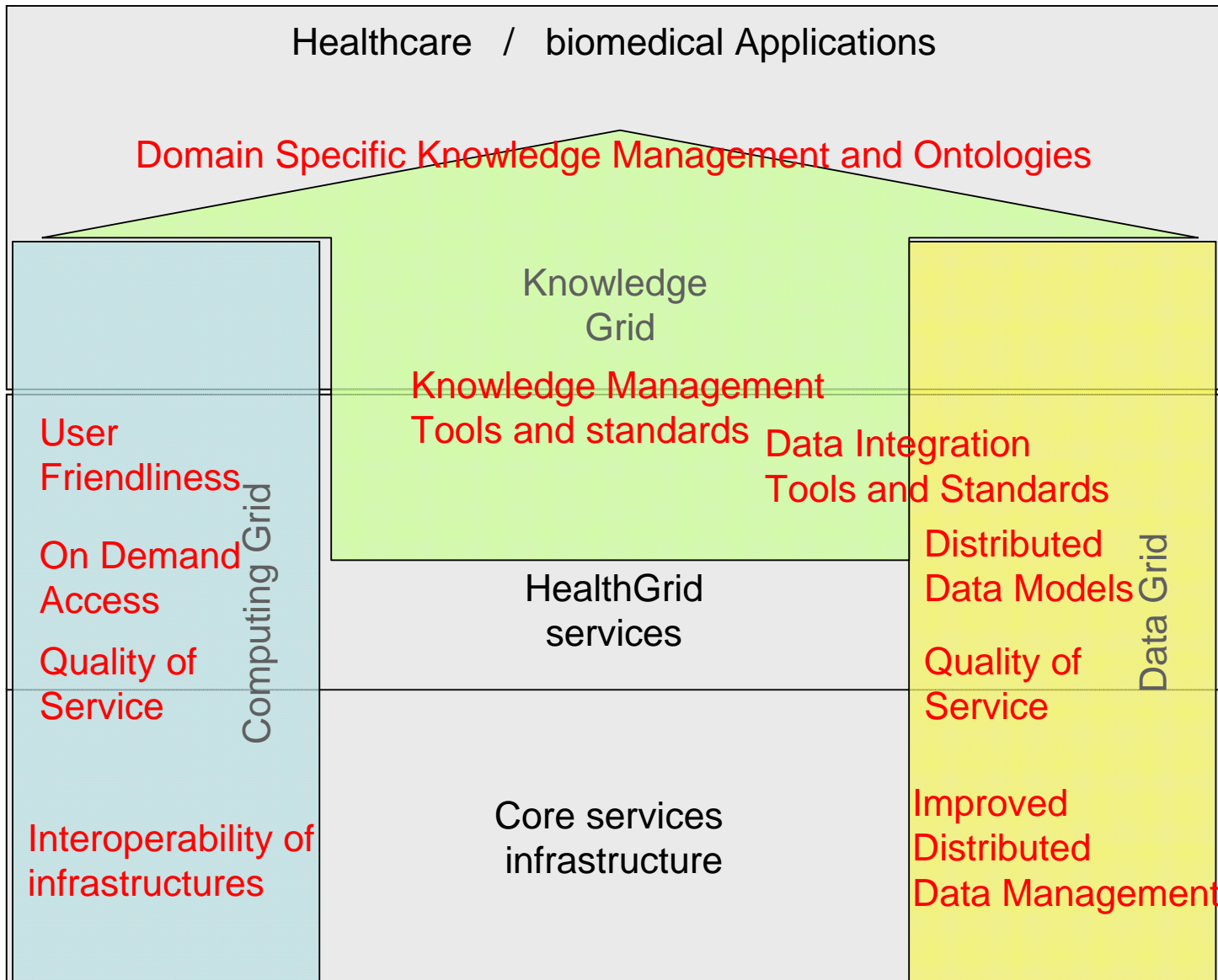
- **Distributed data integration and computing**
 - Security
 - Performance
 - Usability
- **Standards**
 - Need for **reference** implementations of **standard grid services**
 - Lack of connection between medical informatics standards and grid **standards** (e.g. grid-enabled DICOM)
 - Lack of **standard open source ontologies** in medical informatics
- **Grid deployment in medical research centres**
 - Easy installation of **secure grid nodes**
 - Friendly **user interface**

SHARE roadmap
<http://share.healthgrid.org>

Knowledge grids are built on two pillars: data and computing grids



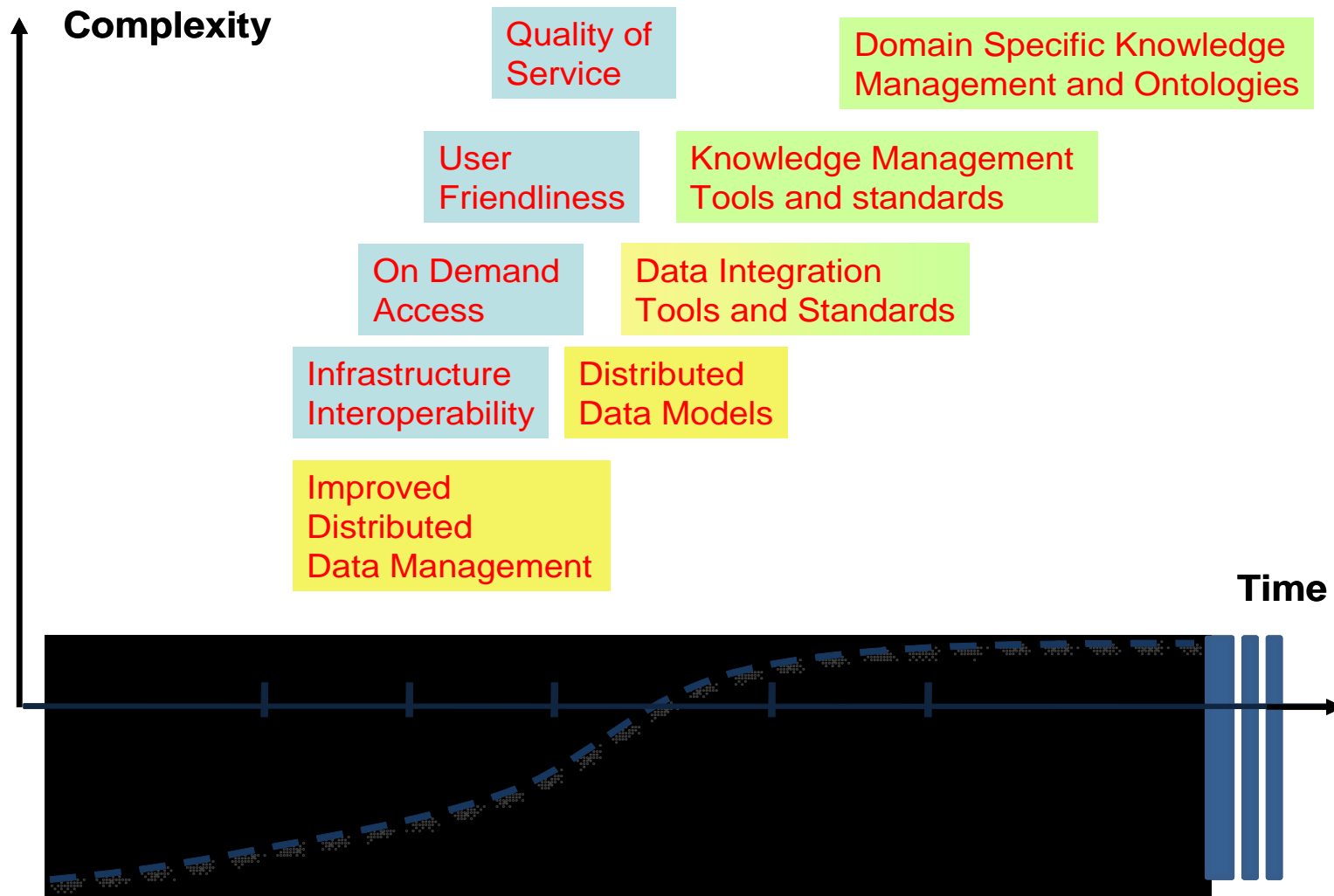
SHARE roadmap
<http://share.healthgrid.org>



The technology to build knowledge grids is not yet mature



SHARE roadmap
<http://share.healthgrid.org>



Conclusion



- **Grid technology open perspectives for the integration of disease related data and biomedical informatics data sources**
- **Embrace is currently integrating biomedical data resources using web services and grid technology**
 - Definition of standard data types -> Ontologies
 - All web services available on <http://www.embracegrid.info>
- **In a longer term, grids open the right technology and environment to enable pharmaceutical R&D**
 - The SHARE roadmap has identified the research challenges for this purpose
- **Open source disease ontologies are absolutely necessary to enable the vision**

SHARE definition of grid



- a fully distributed, dynamically reconfigurable, scalable and autonomous infrastructure to provide location independent, pervasive, reliable, secure and efficient access to a coordinated set of services encapsulating and virtualizing resources (computing power, storage, instruments, data, etc.) in order to support problem solving and knowledge generation across multiple administrative domains.

(Amalgam of definitions by Peter Coveney, The Coregrid Network of Excellence, Ian Foster and co-authors)