vl·e virtual laboratory for e·science

e-Science and e-Bioscience

The VL-e approach

L.O. (Bob) Hertzberger

Director of Virtual Laboratory for e-Science (VL-e)

Adjunct director Netherlands Bio Informatics Centre (NBIC)

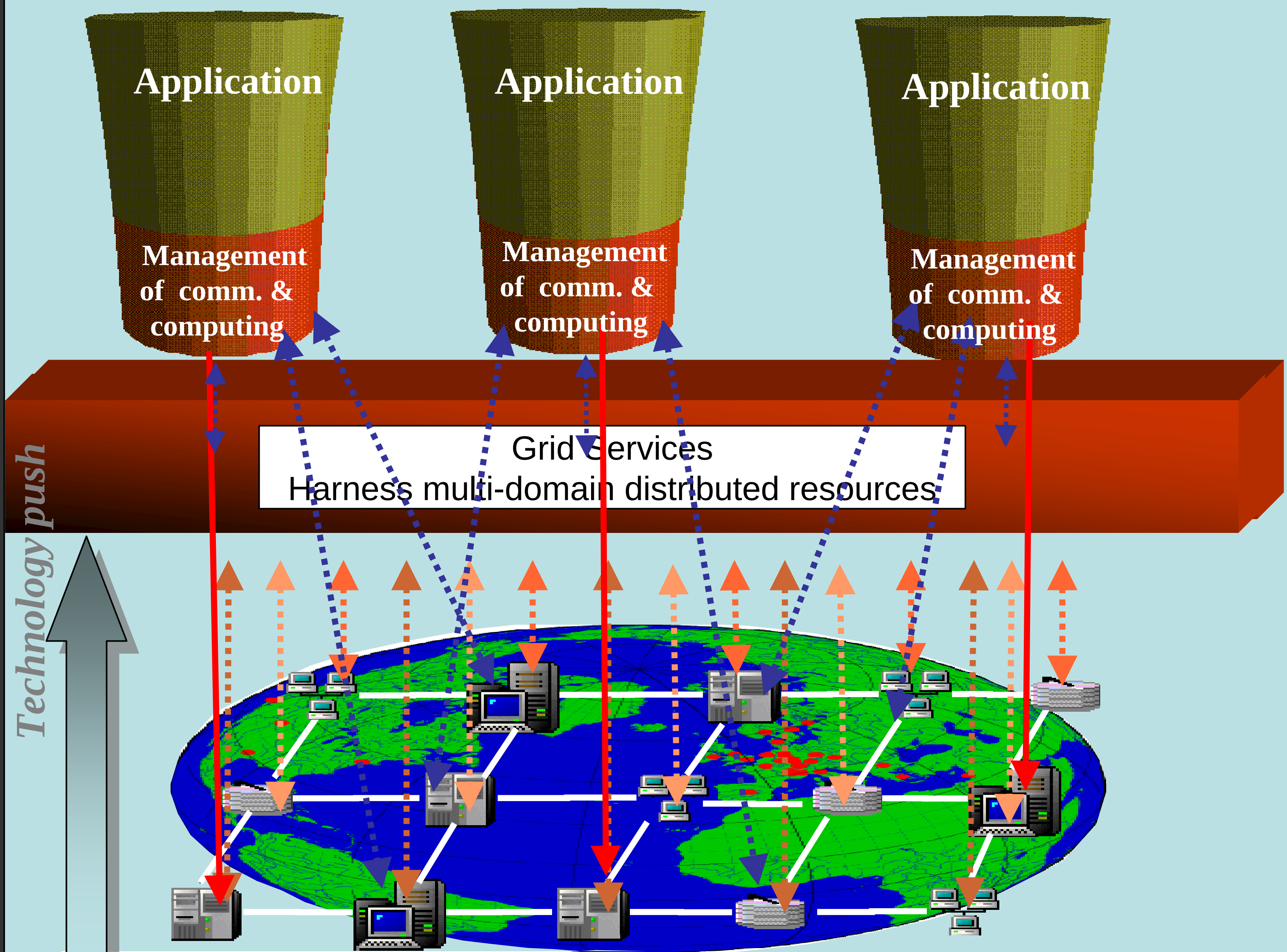
One of the directors of BIG GRID

bob@science.uva.nl

Content

- Some issues on Grid important for e-Science
- Developments towards e-Science
- The VL-e approach
- Some examples
- Conclusions

What is Grid


Application
Management of comm. & computing
Application
Management of comm. & computing
Application
Management of comm. & computing
Grid Services
Harness multi-domain distributed resources
Technology push

Grid before WSRF/OGSA

The word '*grid*' has been used in many ways

- ✓cluster computing
- ✓cycle scavenging
- ✓**cross-domain resource, data and information sharing**

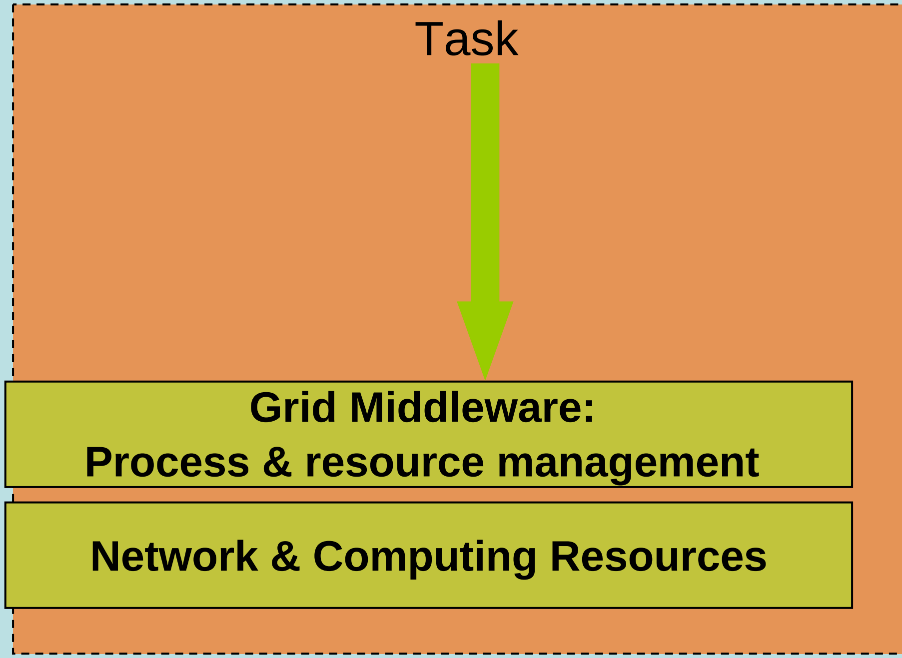
A definition for what we mean with grid

•Coordinates resources not subject to centralised control

•Using standard, open and generic protocols & interfaces

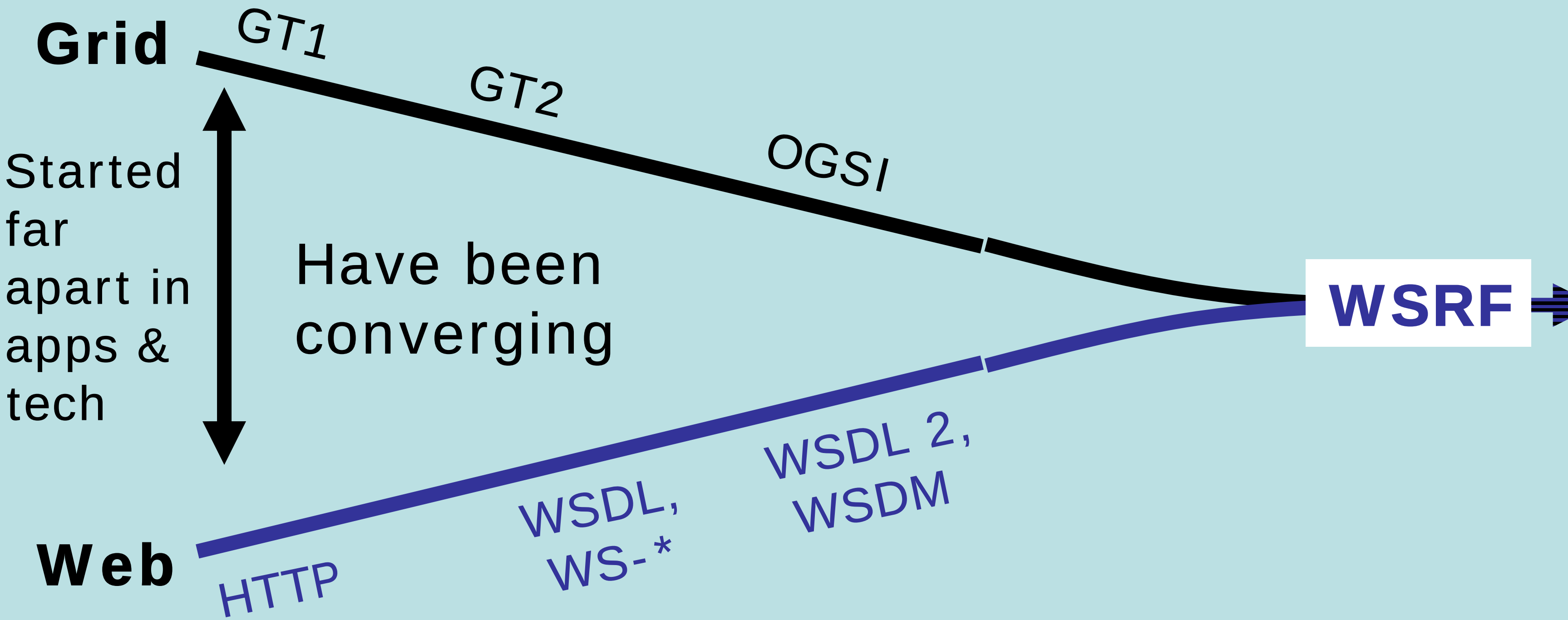
•Provides non-trivial qualities of collective service

•Virtualization of resources via among others Virtual Organizations



Definition source: Ian Foster in *Grid Today*, July 22, 2002; Vol. 1 No. 6, see http://www-fp.mcs.anl.gov/~foster/Articles/WhatIstheGrid.pdf

Grid and Web Services Convergence



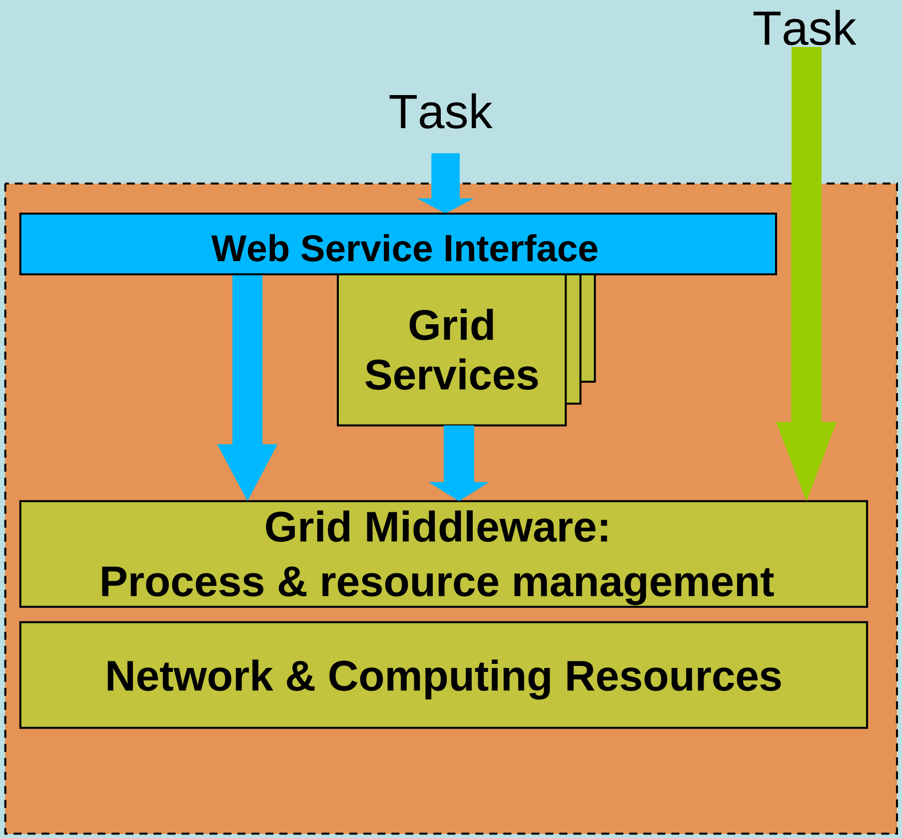
Definition of Web Service Resource Framework(WSRF) makes explicit distinction between “service” and stateful entities acting upon service i.e. the resources

Means that Grid and Web communities can move forward on a common base!!!

Ref: Foster

Grid after WSRF & OGSA

- Important aspects:
  - ✓Uniform syntax & use of WSDL or other
- ✓Problems
  - ✓Level of abstraction of service
  - ✓How does service react
  - ✓Quick standardization
    - ▪First implementation in 2008 might be late
- Advantage:
  - ✓OGSA offers a coherent set of services



What is next ???

Knowledge and the knowledge producing & consuming protocols & patterns are already in Grid Middleware and Grid Applications

Embedded in middleware code, in schemas, in catalogues, in applications and in practice

They are often called metadata.

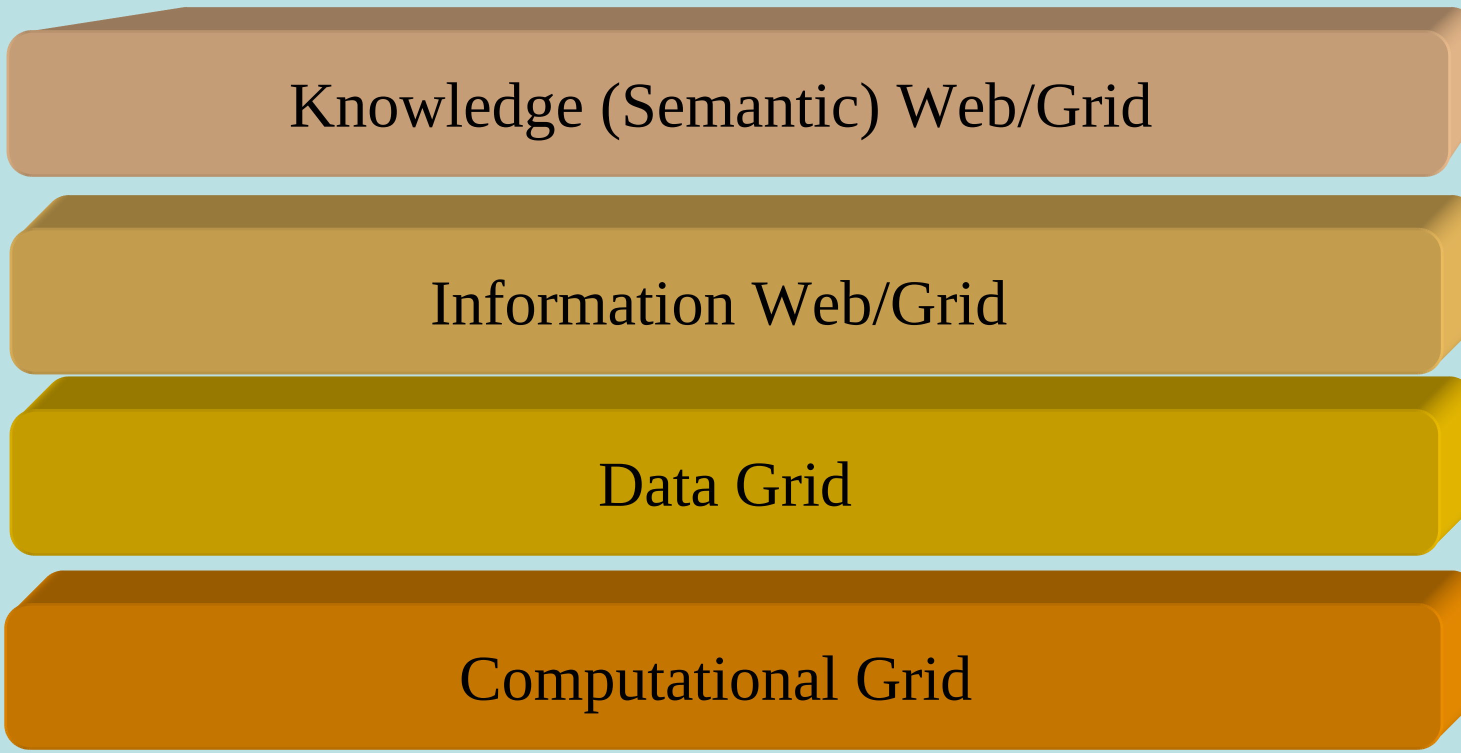
Carol Goble.

Issue is to make this knowledge explicit

Semantic WEB/GRID

vl-e

Levels of Grid abstraction



vl·e

Background information experimental sciences

- There is a tendency to look ever deeper in:
  - ✓ Matter e.g. Physics
  - ✓ Universe e.g. Astronomy
  - ✓ Life e.g. Life sciences
- Instrumental consequences are increase in detector:
  - ✓ Resolution & sensitivity
  - ✓ Automation & robotization
- Therefore experiments become increasingly more complex

vl-e

Impact in the life sciences

- Impact of high throughput methods e.g. Omics experimentation
  - ✓ genome ===> genomics

- Instrumentation being used in omics experimentation:
  - ✓ Transcriptomics via among others; micro-arrays
  - ✓ Proteomics via among others; Mass Spectroscopy (MS)
  - ✓ Metabolomics via among others; MS & Nuclear Magnetic Resonance (NMR)

vl·e

Background information experimental sciences

- There is a tendency to look ever deeper in:
  - ✓ Matter e.g. Physics
  - ✓ Universe e.g. Astronomy
  - ✓ Life e.g. Life sciences
- Instrumental consequences are increase in detector:
  - ✓ Resolution & sensitivity
  - ✓ Automation & robotization
- Therefore experiments become increasingly more complex
- Results in an increase in the amount and the complexity of data

Data explosion

- In 2005 more data is being produced than during existance of human
- In 2005 increase from 3 Mexabytes towards more than 40 Mexabytes

Source: Disaster Recovery Journal Autum 2004

  - Impact on science
  - How do we deal with
  - How to extract information

Data explosion results in the Life sciences in a Paradigm shift

- Past experiments where hypothesis driven
  - ✓Evaluate hypothesis
  - ✓Complement existing knowledge

- Present experiments are data driven
  - ✓Discover knowledge from large amounts of data
  - ✓More integration necessary

Background information experimental sciences

- There is a tendency to look ever deeper in:
- Instrumental consequences are increase in detector:
- Therefore experiments become increasingly more complex
- Results in an increase in the amount and complexity of data
- Something has to be done to harness this development
  - ✓ Virtualization of experimental resources enabling sharing & leading to e-Science

vl-e

The what of e-Science

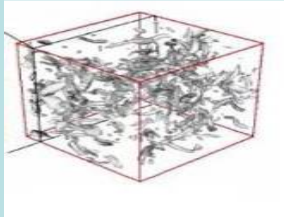
- e-Science is a new science paradigm complementing theoretical and experimental science
  - ✓ More than only coping with data explosion
  - ✓ A multi-disciplinary activity combining human expertise & knowledge

e-Science: A New Science Paradigm

- Thousand years ago:
  Experimental Science
  - description of natural phenomena
- Last few hundred years:
  Theoretical Science
  - Newton's Laws, Maxwell's Equations …

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

- Last few decades:
  Computational Science
  - simulation of complex phenomena



- Today:
  e-Science or Data-centric Science
  - unify theory, experiment, and simulation
  - using data exploration and data mining
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks
  - ✓ Scientist analyzes databases/files

(With thanks to Tony Hey and Jim Gray)

vl·e

The what of e-Science

- e-Science is a new science paradigm complementing theoretical and experimental science
  - ✓ More than only coping with data explosion
  - ✓ A multi-disciplinary activity combining human expertise & knowledge
- e-Science should
  - ✓ Apply and integrate Web/Grid methods where and whenever possible
  - ✓ Will also be a driver for new Web/Grid methods (semantics)
  - ✓ Apply an optimal ICT infrastructure
- e-Science demands a different approach to science
  - ✓ At the end-user level: application or domain system level science
  - ✓ At the e-Science technology level: integrative science

vl·e

e-Science Objectives

- It should enhance the scientific process by:
- Stimulating collaboration by sharing data & information
  - ✓ Result is re-use of data & information

The data sharing potential for Cognition

- Collaborative scientific research
  - ✓ Information sharing
  - ✓ Metadata modeling
- Allows for experiment validation
  - ✓ Independent confirmation of results
- Statistical methodologies
  - ✓ Access to large collections of data and metadata
- Training
  - ✓ Train the next generation using peer reviewed publications *and* the associated data

nature neuroscience

Sharing neuroimaging studies of human cognition

John Darrell Van Horn, Scott T Grafton, Daniel Rockmore & Michael S Gazzaniga

After more than a decade of collecting large neuroimaging datasets, neuroscientists are now working to archive these studies in publicly accessible databases. In particular, the fMRI Data Center (fMRIDC), a high-performance computing center managed by computer and brain scientists, seeks to catalogue and openly disseminate the data from published fMRI studies to the community. This repository enables experimental validation and allows researchers to combine and examine patterns of brain activity beyond that of any single study. As with some biological databases, early scientific, technical and sociological concerns hindered initial acceptance of the fMRIDC. However, with the continued growth of this and other neuroscience archives, researchers are recognizing the potential of such resources for identifying new knowledge about cognitive and neural activity. Thus, the field of neuroimaging is following the lead of biology and chemistry, mining its accumulating body of knowledge and moving toward a 'discovery science' of brain function.

The observation that changes in regional cerebral blood flow accompany neural activity during cognition[1–3] has been a boon to the cognitive and brain sciences, most notably through the use of brain mapping technologies such as functional magnetic resonance imaging (fMRI). Current research efforts for imaging the brain 'in action' are underway to rigorously explore the structure and function of cognitive brain processes, thereby characterizing the mental properties that make us uniquely human[4]. The fMRI studies range from the examination of familiar cognitive processes such as human memory and language processing to novel studies of racial threat[5] and the neurofunctional components of humor[6].

This increasing dependence on brain mapping for exploring cognition has led to an unprecedented data explosion that is pressing neuroscientists to manage and analyze data on scales never before imagined. Complete fMRI study data sets now routinely reach several gigabytes in size, with the amount of brain image data collected in some articles[7,8] beginning to rival the current size of many biological and physical science databases[9,10]. What is more, the size of fMRI studies has grown over time, and what is now considered a large fMRI study will seem relatively small within only a few years, as new technological developments occur in scanner physics, engineering and protocol design.

Unfortunately, despite this progress, much of these fMRI data are not readily available to anyone beyond the original research team that collected them. There are several reasons behind the fact that other investigators do not typically get to work with the actual data that went into the heavily processed images appearing in a published article: (i) limitations of publication space on the complete representation of fMRI methods and findings, (ii) the proprietary feelings of investigators against letting others view their data, (iii) the immensity of data set size and (iv) the convention of only reporting tabular representations of activity in individual image voxels. However, given recent success stories from genomics[11] and proteomics[12] for organizing, archiving and mining large amounts of data from their communities, it may come as no surprise that cognitive neuroscientists are now looking to unfettered data sharing and study archiving to better understand these rich collections of dynamic brain data.

Data sharing sociology in neuroimaging

In 2000, with precisely such a goal, we founded the fMRIDC (www.fmridc.org) at Dartmouth College. We sought to facilitate progress in understanding cognitive processes through the collection, archiving and open distribution of neuroimaging data sets in the peer-reviewed literature[13]. We reasoned that there could be several positive outcomes to making the complete study data sets available to others. First, the study findings could be independently confirmed, helping to strengthen the findings drawn by the original authors. Second, new statistical methodologies could be applied to the data, providing novel insights into cognitive processes. Different studies could be compared, possibly identifying unanticipated functional homologies between seemingly different cognitive tasks. Moreover, these studies could be used to train the next generation of neuroscientists by using fMRI data that had already undergone interpretation by those who collected it and had published it in leading journals. We decided to focus on fMRI data from published articles and not to be concerned with unpublished data. This allowed us to focus the enormous chore of collecting and managing the data, as well as to construct an archive that was representative of the field's body of work.

We approached the editors of several leading journals and were pleased by their initial support. To form the first corpus of data sets and accompanying study material, a special issue of the *Journal of Cognitive Neuroscience* (*JOCN*) was published containing a collection of articles from leading laboratories (Vol. 12, Suppl. 2, 2000). The authors of these articles generously provided the raw, processed and structural images and study meta-data

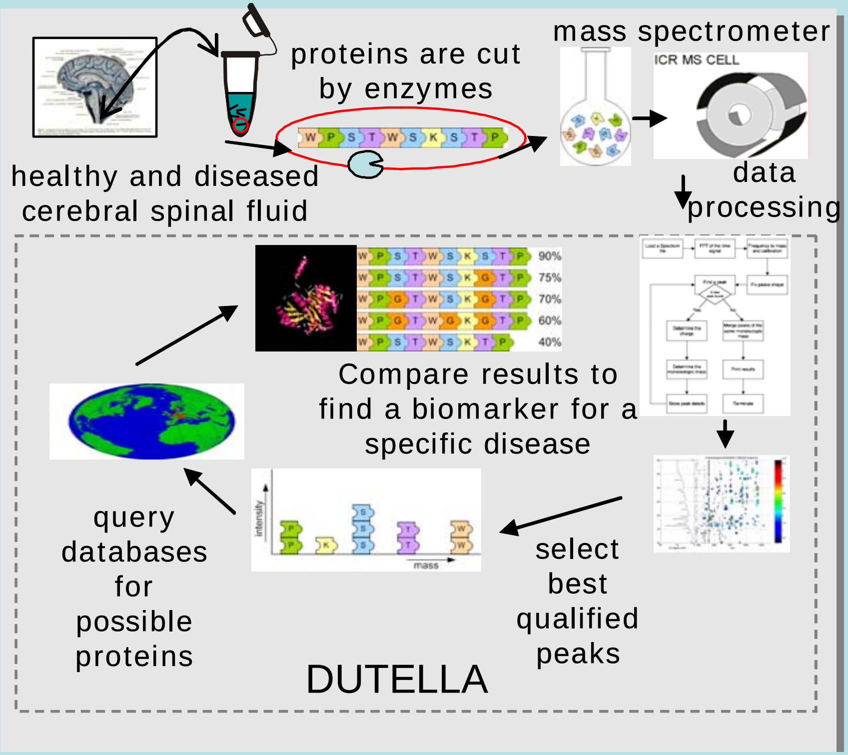
All authors are at Dartmouth College, Hanover, New Hampshire 03755, USA. John Darrell Van Horn and Scott T. Grafton are at the Center for Cognitive Neuroscience, the Dartmouth Brain Imaging Center and the fMRI Data Center, Daniel Rockmore is in the Department of Mathematics and the fMRI Data Center, and Michael S. Gazzaniga is at the Center for Cognitive Neuroscience and the fMRI Data Center.
e-mail: John.D.Van.Horn@dartmouth.edu

e-Science Objectives

- It should enhance the scientific process by:
- Stimulating collaboration by sharing data & information
  - ✓ Improve re-use of data & information
- Combing data and information from different modalities
  - ✓ Sensor data & information fusion

vl-e

In Biomarker Research Multiple Resources have to be integrated



- Samples
  - ✓ Patients
  - ✓ Hospitals
  - ✓ Diseases
- Analytical instrumentation
- Researchers and expertise
- Processing tools
- Information
- Data storage facilities
- Computational resources

vl·e

e-Science Objectives

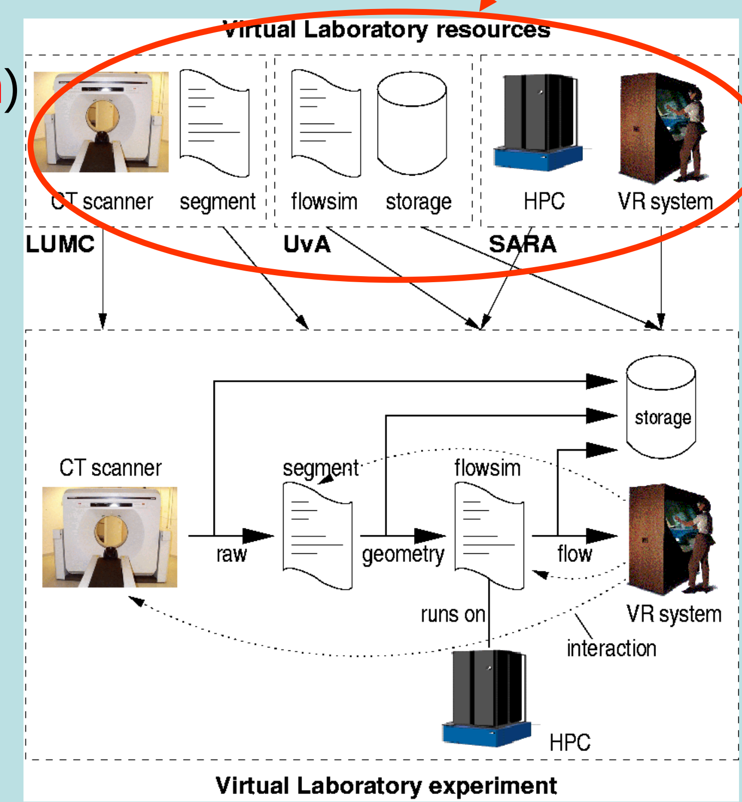
- It should enhance the scientific process by:
- Stimulating collaboration by sharing data & information
  - ✓ Improve re-use of data & information
- Combing data and information from different modalities
  - ✓ Sensor data & information fusion
- Realize the combination of real life & (model based) simulation experiments

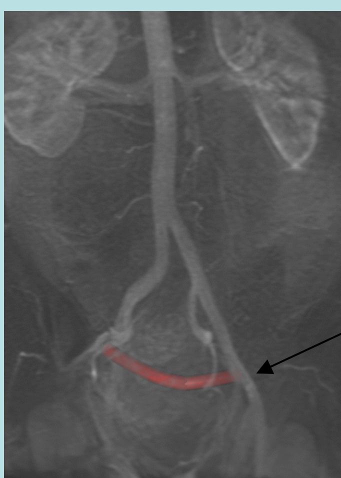
vl-e

Simulated Vascular Reconstruction in a Virtual Operating Theatre

- *An example of new models and interaction*
  - patient specific vascular geometry (from CTA)
  - segmentation
  - blood flow simulation (Latice Bolzmann)
- Pre-operative planning (interaction)
- Suitable for parallelization through functional decomposition

Grid resources





Patient's vascular geometry (CTA)

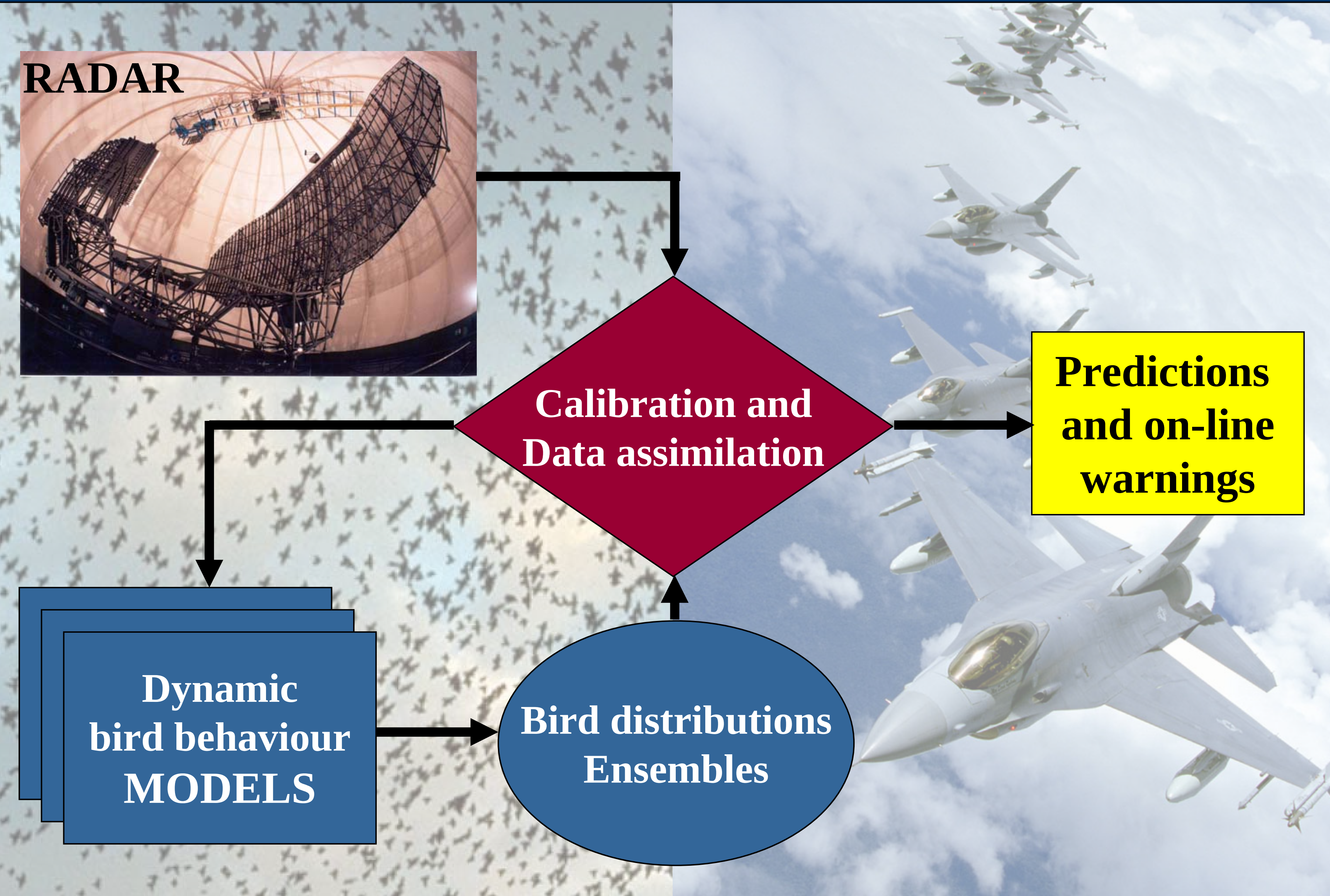
Simulated "Fem-Fem" bypass

e-Science Objectives

- It should enhance the scientific process by:
- Stimulating collaboration by sharing data & information
  - ✓ Improve re-use of data & information
  - ✓ Combing data and information from different modalities
    - Sensor data & information fusion
- Realize the combination of real life & (model based) simulation experiments
- Allow for modeling of dynamic systems

vl-e

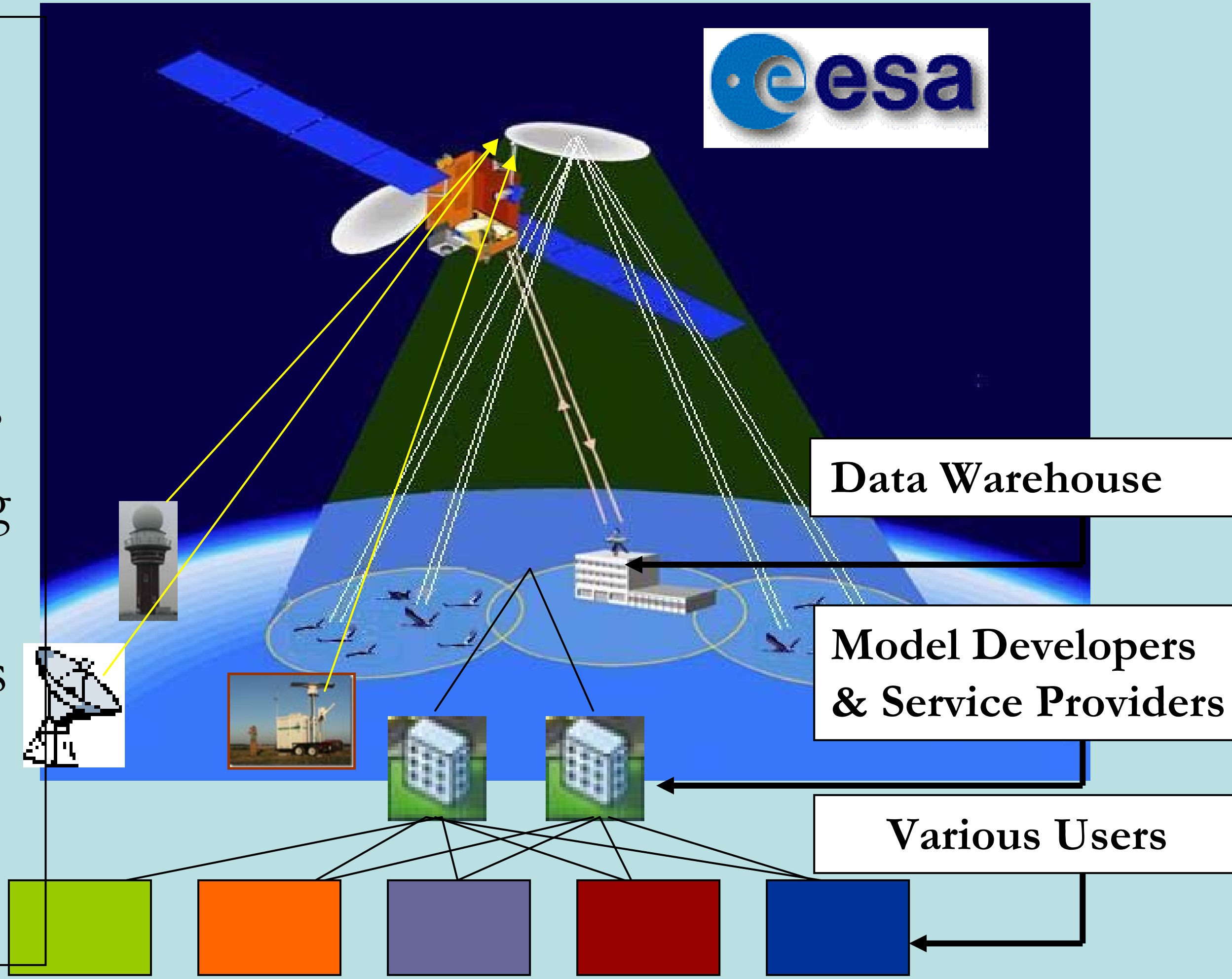
Bird behaviour in relation to weather and landscape



System of Systems: Avian Allert (2006-2013)
Monitoring and Modelling Bird Behaviour and Migration

<u>Systems</u>

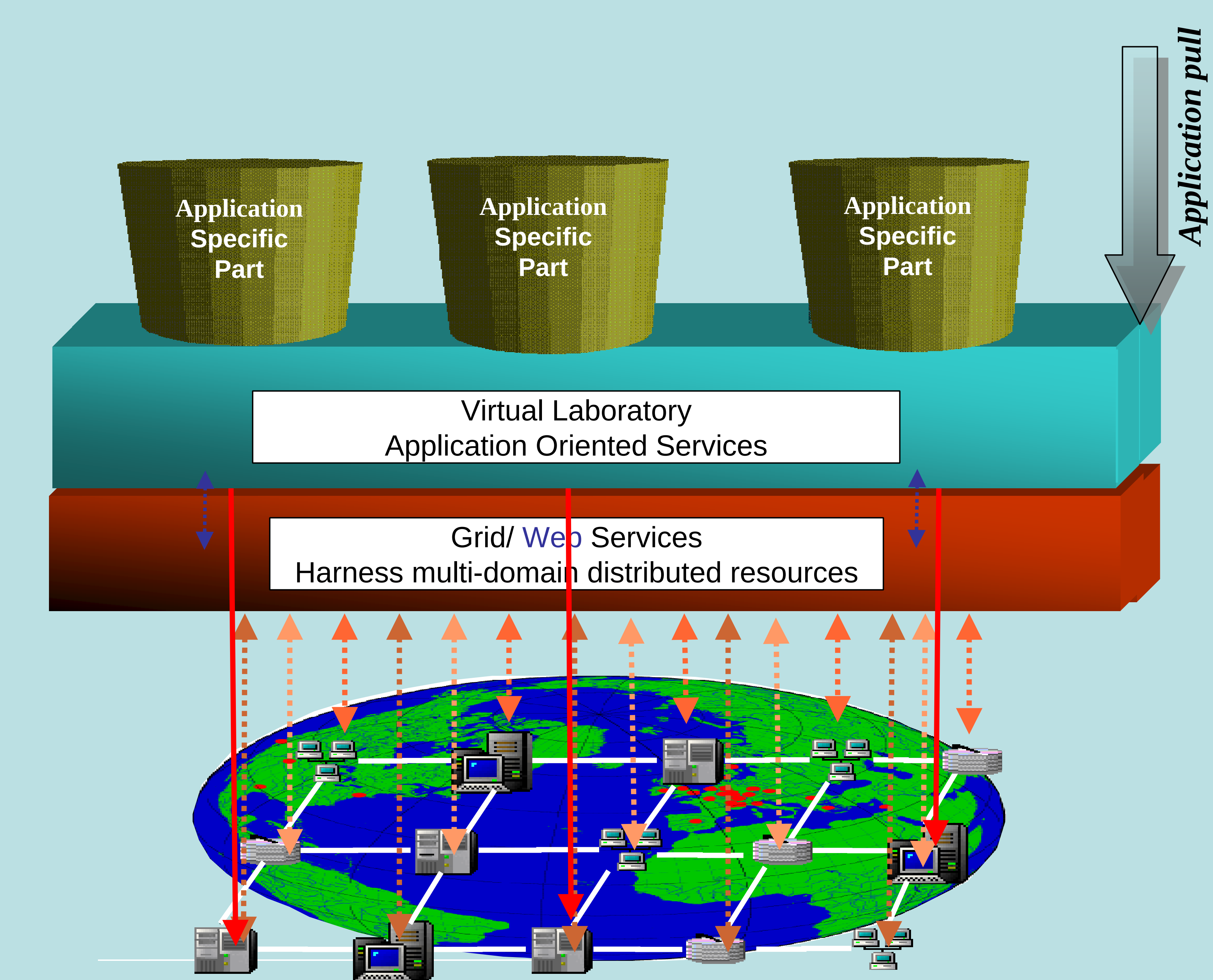
1. Military Surveillance Radars
2. Weather Radars
3. Remote Sensing
4. GPS on Individual Birds
5. Virtual Lab. for Model Development



Virtual Lab for e-Science research Philosophy

- Generic application support
  - ✓ Application cases are drivers for computer & computational science and engineering research
  - ✓ Problem solving partly generic and partly specific
  - ✓ Re-use of components via generic solutions whenever possible

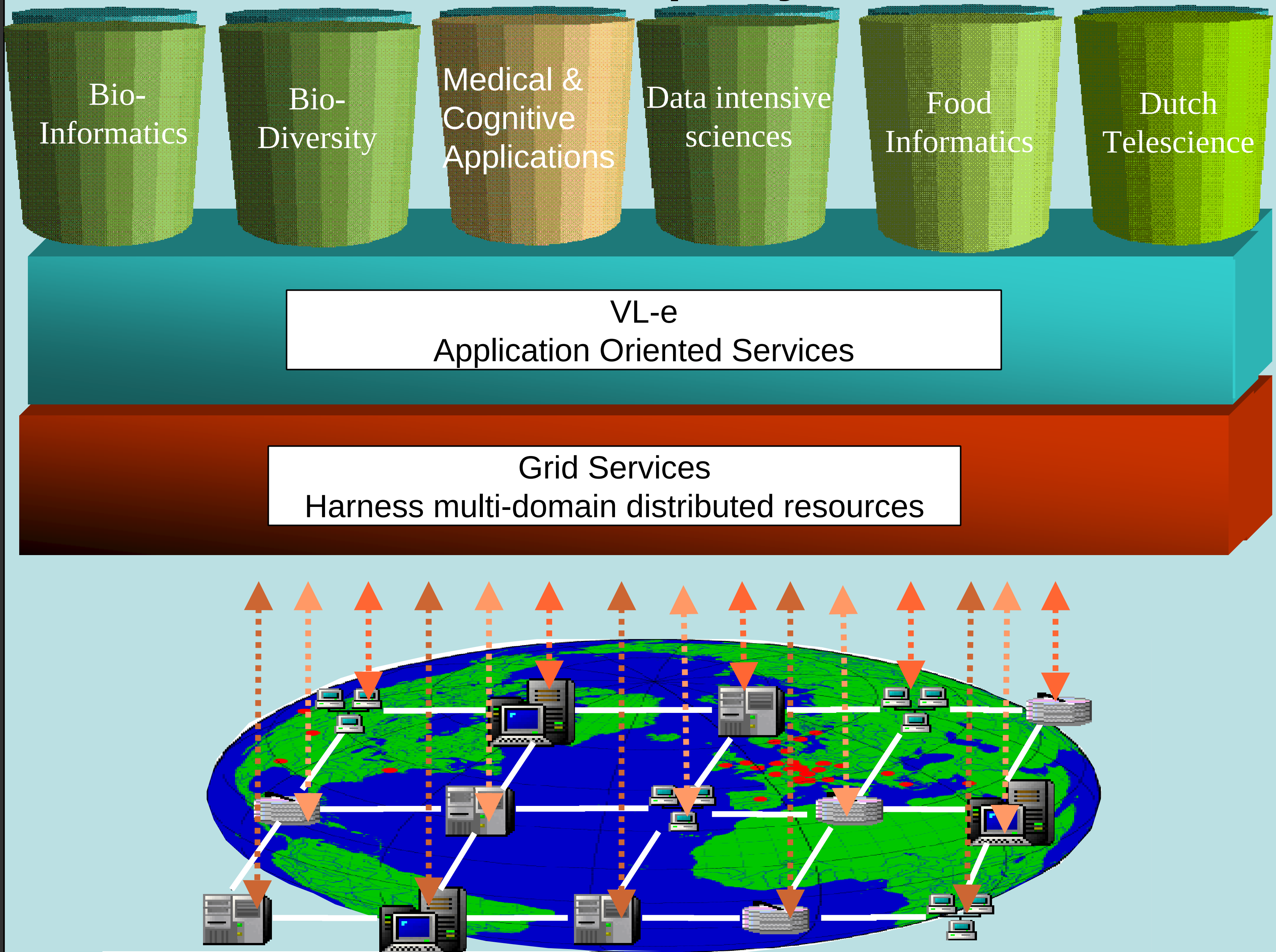
vl·e


Application Specific Part
Application Specific Part
Application Specific Part
Application pull
Virtual Laboratory
Application Oriented Services
Grid/ Web Services
Harness multi-domain distributed resources
vl·e

Potential for generic e-Science services

- Virtual Reality Visualization & user interfaces
- Modeling & Simulation
  - ✓ Interactive Problem Solving
- Data & information management
  - ✓ Data modeling
  - ✓ dynamic work flow management
- Content (knowledge) management
  - ✓ Semantic aspects
  - ✓ Meta data modeling
    - Ontologies
- Wrapper technology
- Design for Experimentation

VL-e project



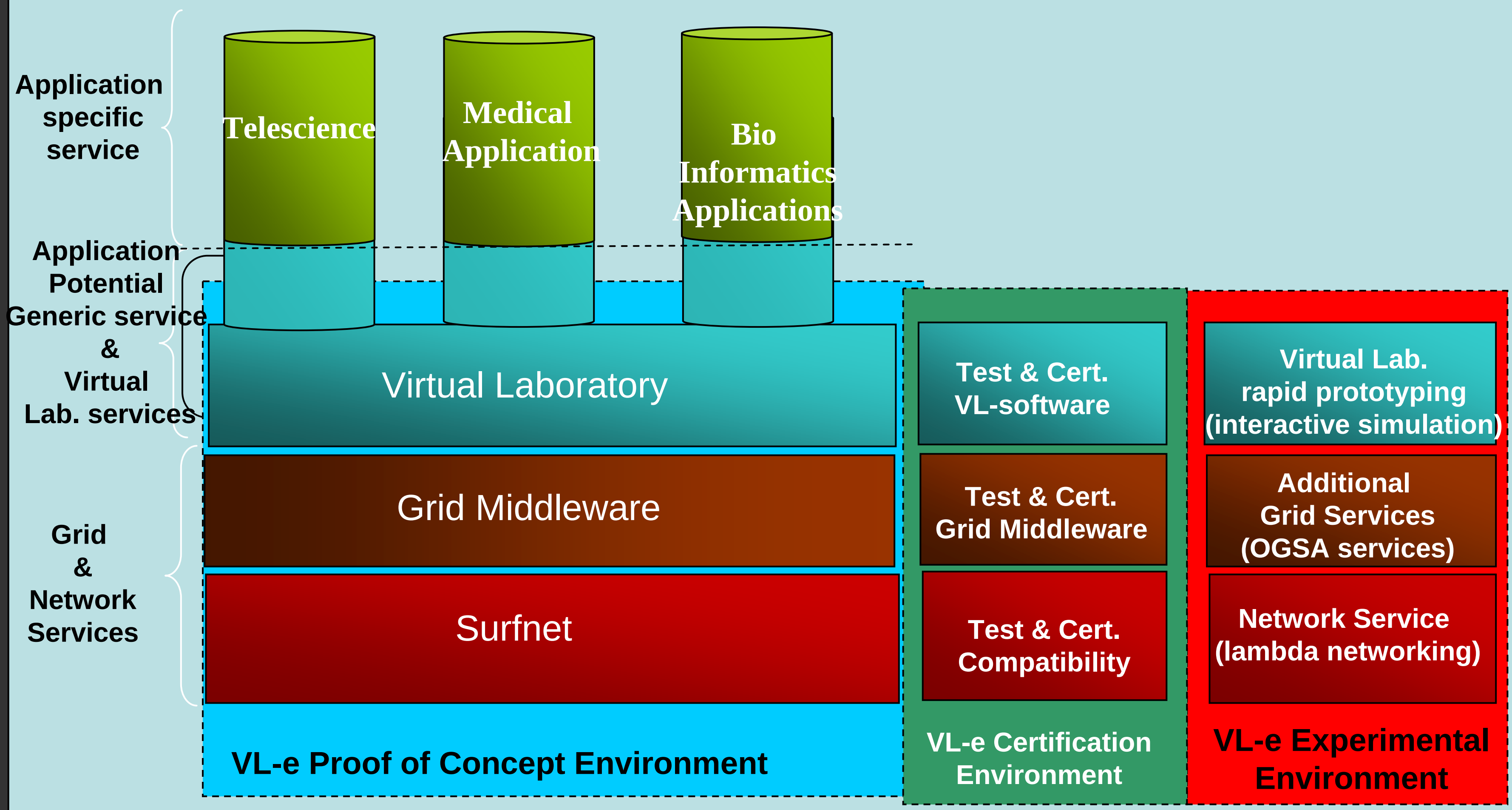
Bio-Informatics
Bio-Diversity
Medical & Cognitive Applications
Data intensive sciences
Food Informatics
Dutch Telescience
VL-e
Application Oriented Services
Grid Services
Harness multi-domain distributed resources

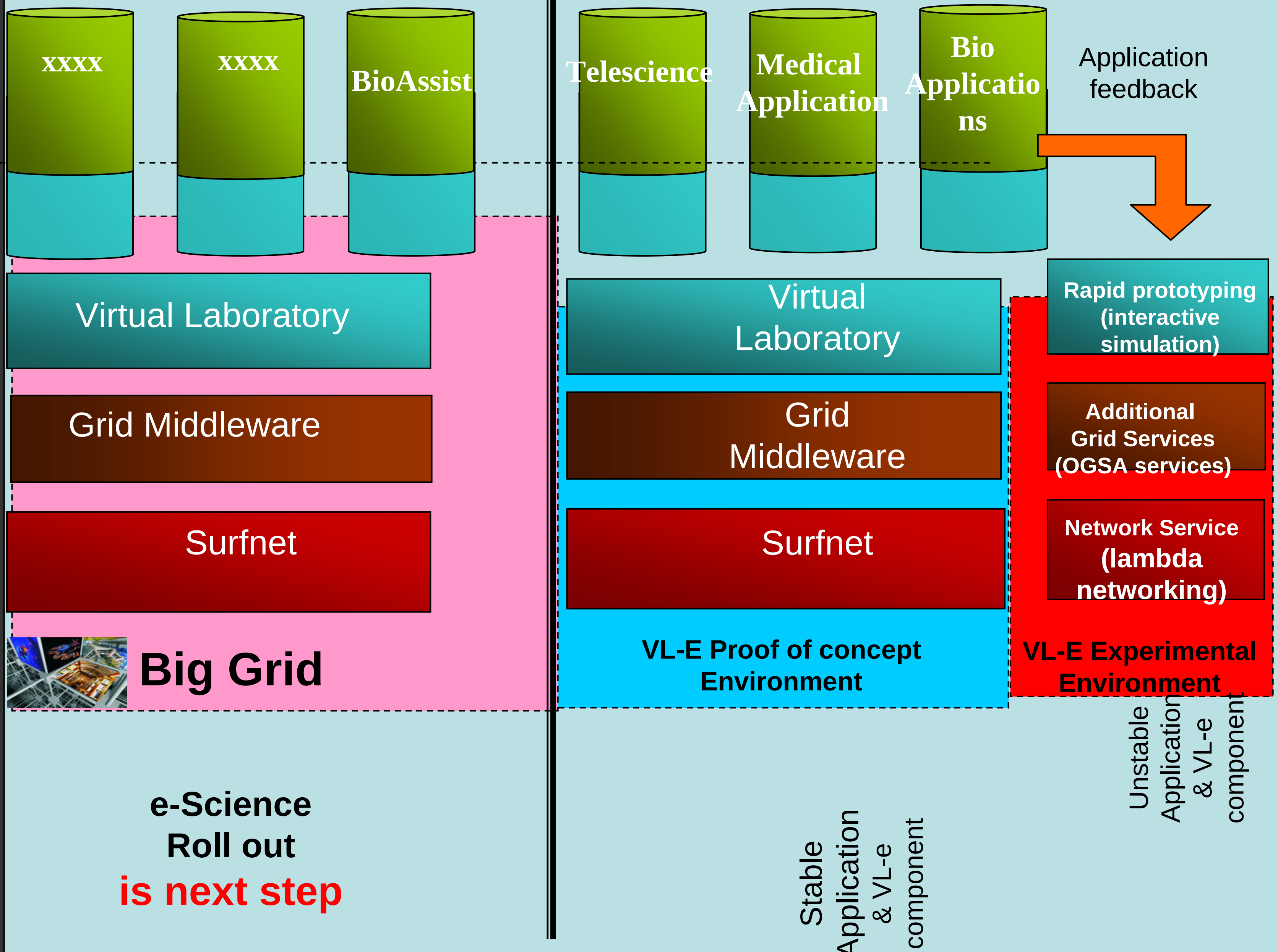
Virtual Lab for e-Science research Philosophy

- Multidisciplinary research and development of related ICT infrastructure
- Generic application support
  - ✓ Application cases are drivers for computer & computational science and engineering research
  - ✓ Problem solving partly generic and partly specific
  - ✓ Re-use of components via generic solutions whenever possible
- Rationalization of experimental process
  - ✓ Reproducible & comparable
- Two research experimentation environments
  - ✓ Proof of concept for application experimentation
  - ✓ Rapid prototyping for computer & computational science experimentation

vl·e

The VL-e infrastructure




xxxx
xxxx
BioAssist
Telescience
Medical Application
Bio Applications
Application feedback
Virtual Laboratory
Grid Middleware
Surfnet
Big Grid
Virtual Laboratory
Grid Middleware
Surfnet
VL-E Proof of concept Environment
Rapid prototyping (interactive simulation)
Additional Grid Services (OGSA services)
Network Service (lambda networking)
VL-E Experimental Environment
Unstable Application & VL-e component
e-Science Roll out
is next step
Stable Application & VL-e component

Conclusion

- e-Science is not about porting current applications towards e-Science infrastructures
- To fully exploit the potential of e-Science and its ICT infrastructures one has to do integrative experiments
- Because of its potential to do system level end user science e-Science might well lead to a radical change in science methodology
  - ✓ The VL-e generic approach will help

vl-e

Wetenschap & Technologie Centrum Watergraafsmeer
wtcw

Universiteit van Amsterdam

vrije Universiteit amsterdam

nbic

amc

VU medisch centrum

NIKHEF

sara

CWI

AMOLF

FOM

TNO

vrije Universiteit amsterdam

Unilever

frieslandfoods

AGROTECHNOLOGY & FOOD SCIENCES GROUP
WAGENINGEN UR

PHILIPS

TU Delft

FEI COMPANY™
TOOLS FOR NANOTECH

logicaCMG

IBM

DSM

vl-e

Netherlands Bioinformatics Centre

E-bioscience: a new way of life (science)

May 23, 2007

Antoine van Kampen
Scientific Director
Netherlands Bioinformatics Centre
Antoine.van.kampen@nbic.nl
www.NBIC.nl

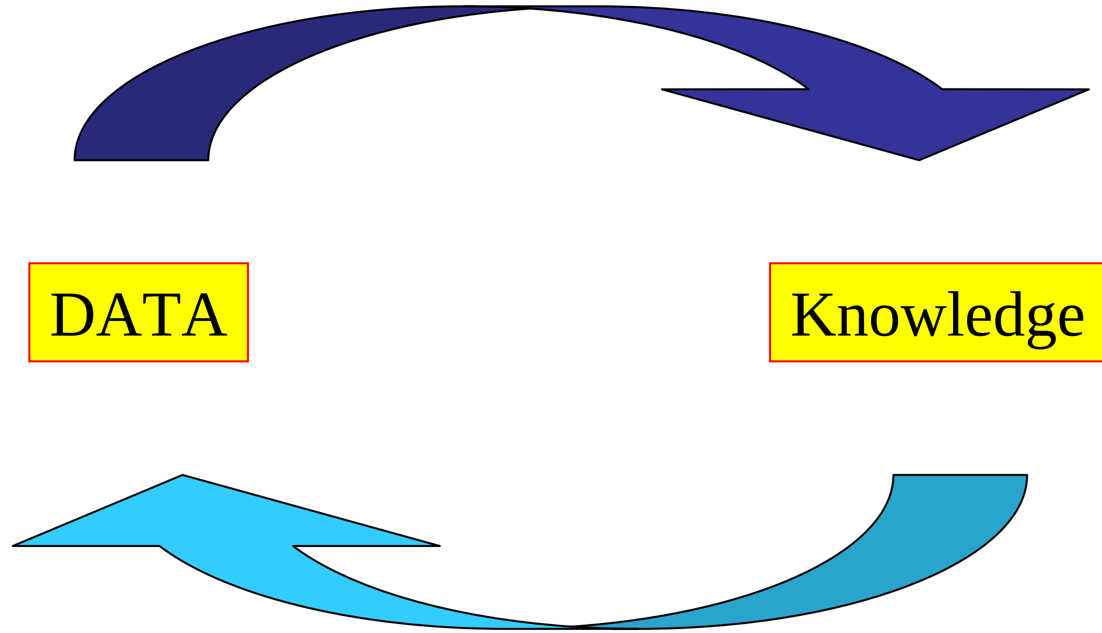
ICTDELTA
het ICT Innovatiecongres

What is Bioinformatics?

The development and application of informatics, mathematical, and statistical methods in life sciences.

Enabling science for genomics

*convert data to knowledge
*generate new hypotheses

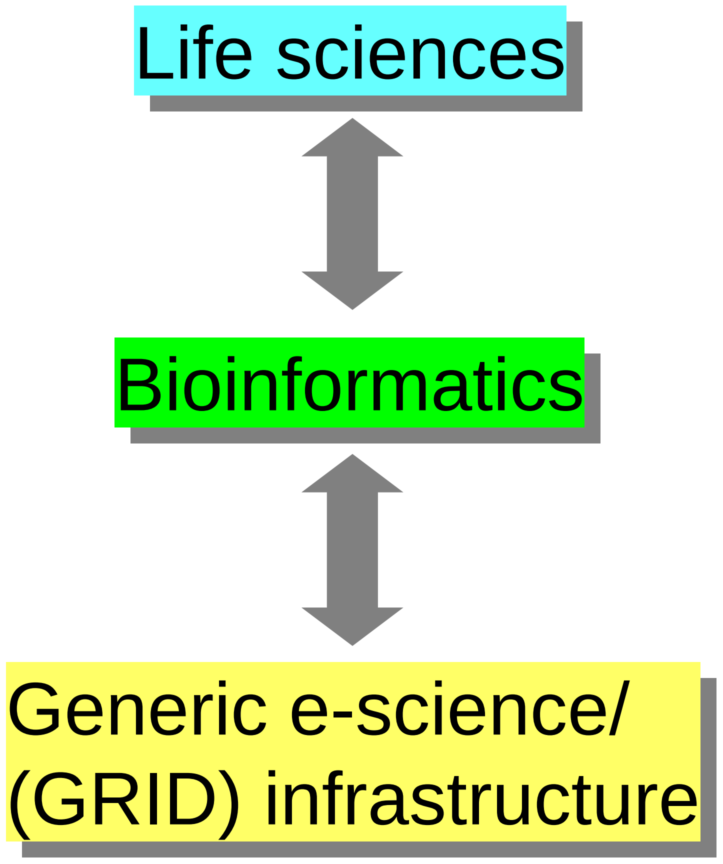


*Design new experiments

nbic Netherlands Bioinformatics Centre

Bioinformatics as interface

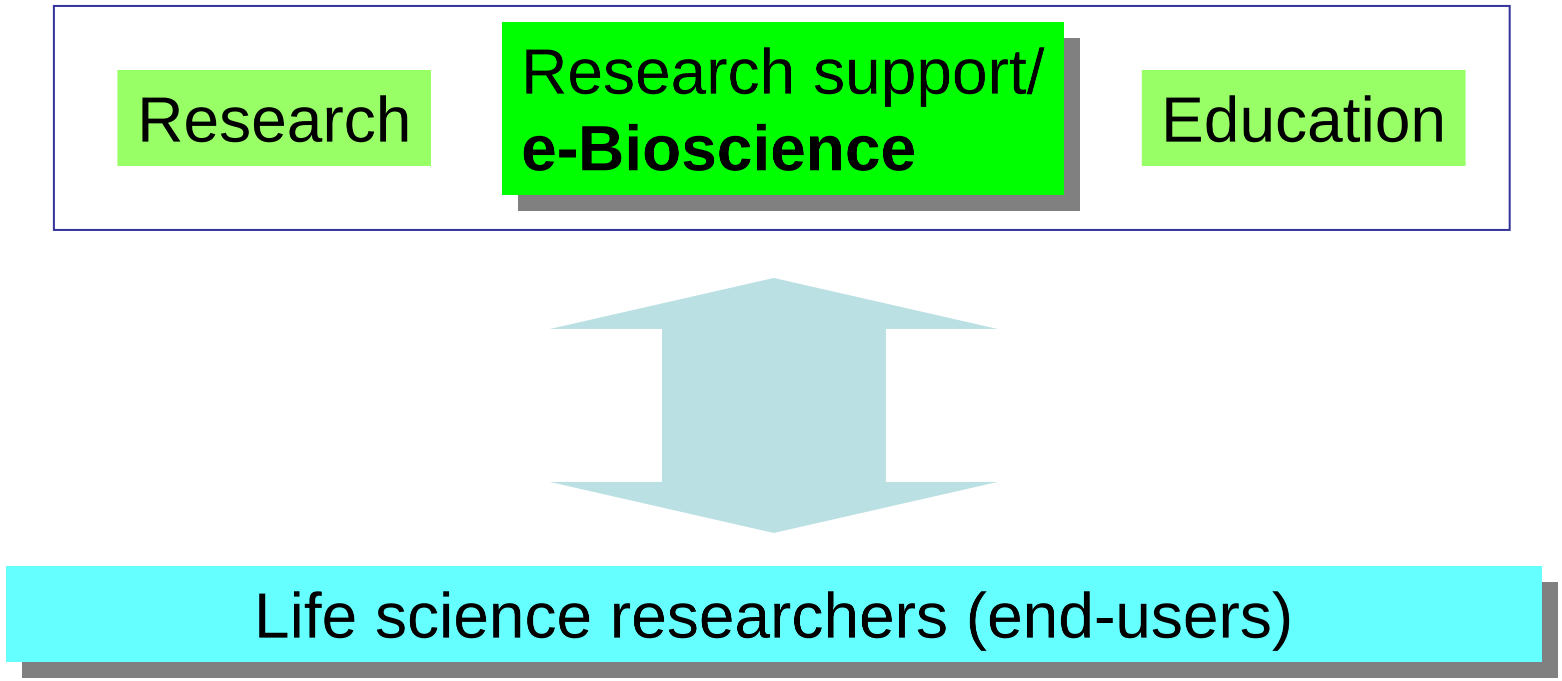
- e-Bioscience
- How can we make generic e-science methodologies and (GRID) ICT infrastructure of benefit to life sciences?



nbic Netherlands Bioinformatics Centre

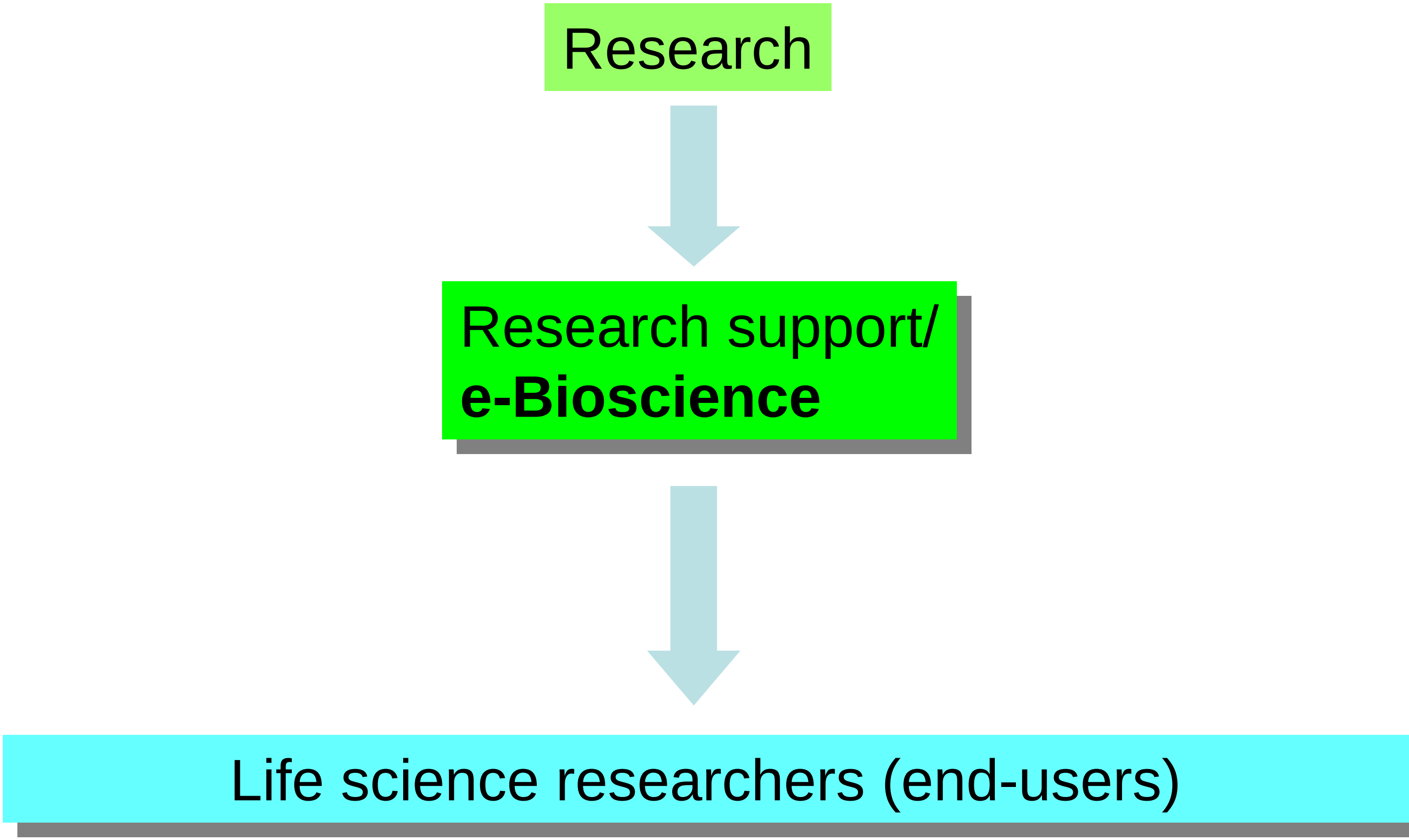
Netherlands Bioinformatics Centre (NBIC)

Programmes



nbic Netherlands Bioinformatics Centre

Deliver tools and databases to end-users

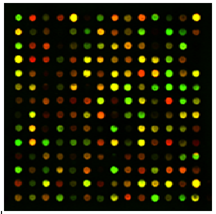


High-throughput experimental technologies in life sciences

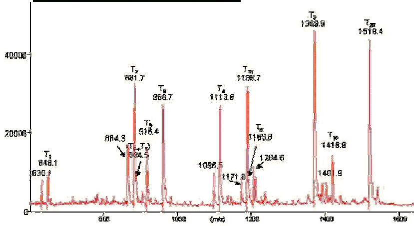
Cell/Tissue

Determine complete DNA sequence of organism (e.g., mutations)

Measure expression of all genes.

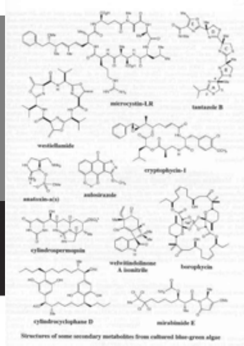


Identify many proteins or their expression level.

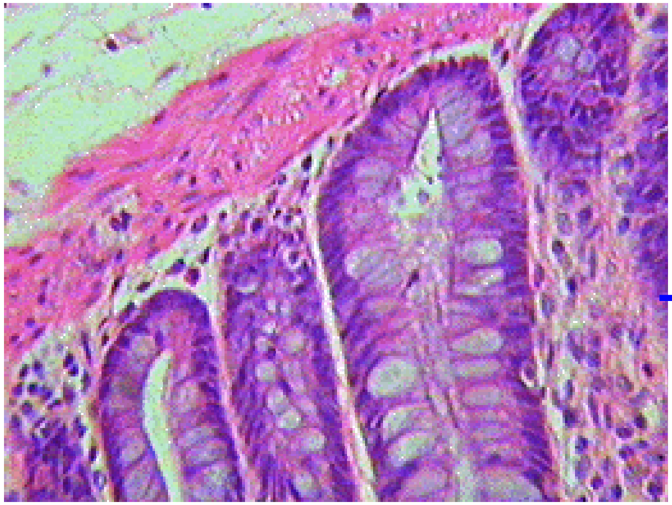


Maldi Tof mass spectrum of the tryptic digest of a high molecular weight (85,000) glutenin (DY10)

Identify many metabolites or their concentration.



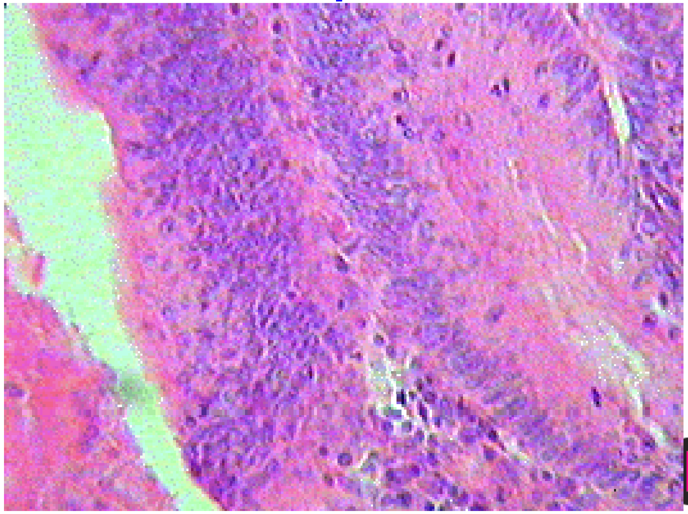
Compare normal versus cancer



High magnification of a normal human colon cell

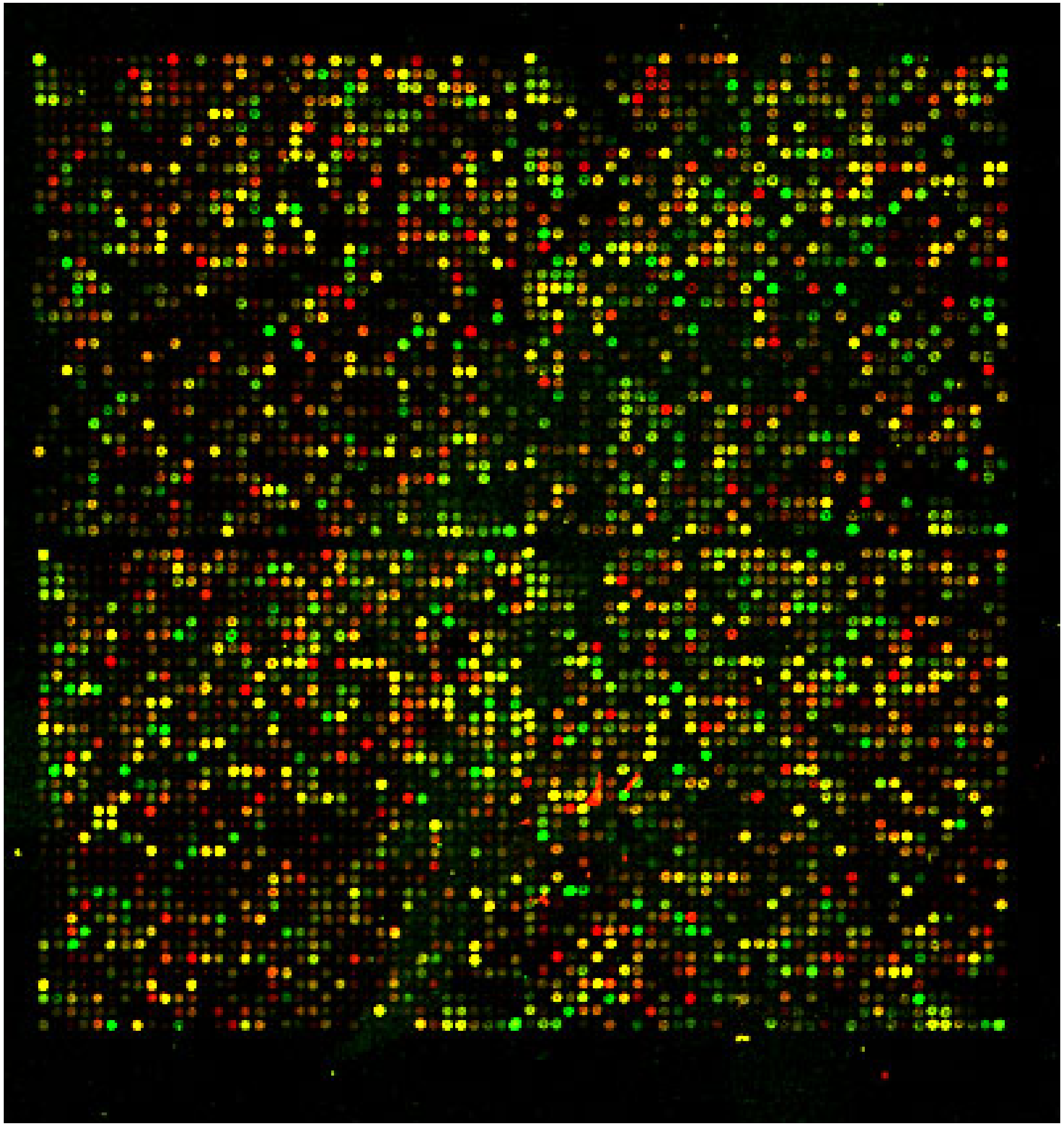
Data

Data



High magnification of a human colon cell with carcinoma

DNA microarrays

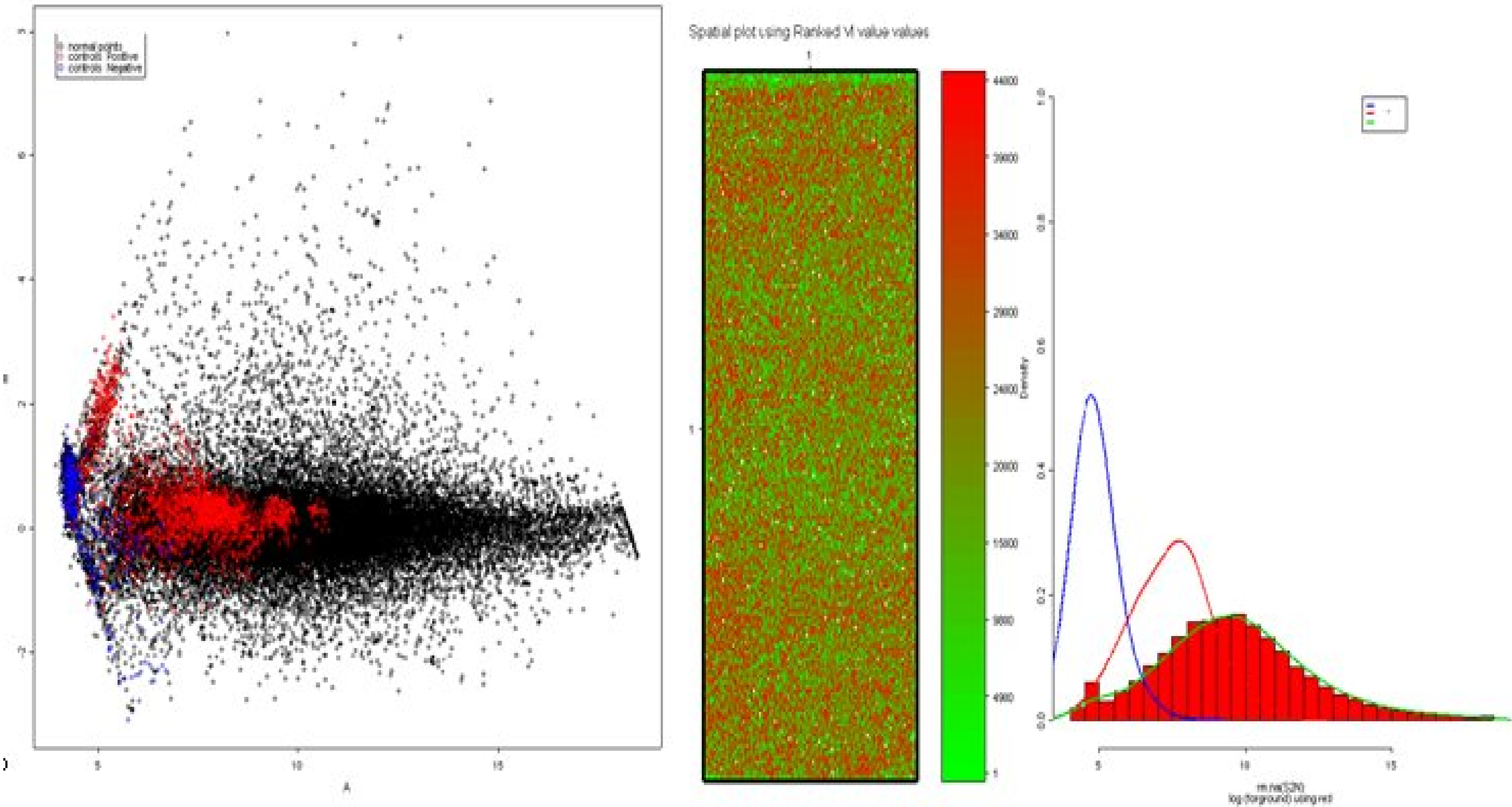


Single experiment:
30.000 – 40.000 genes

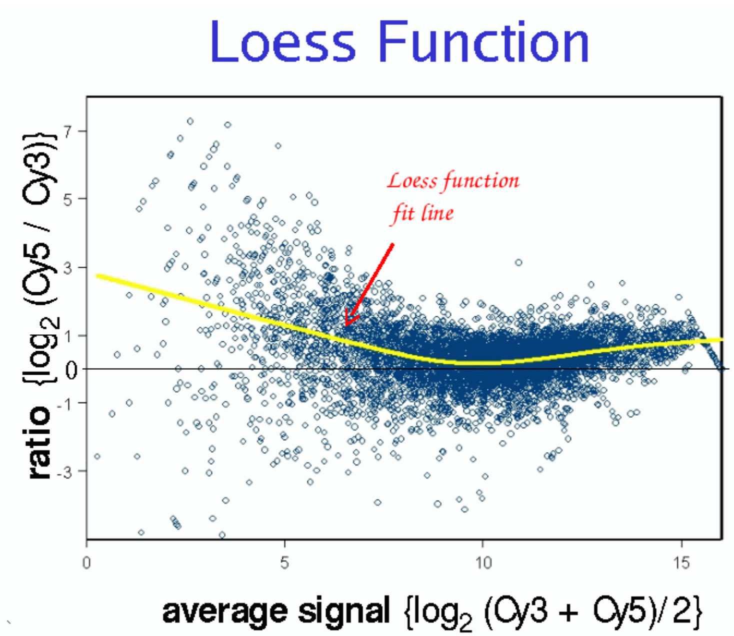
Requires dedicated approaches for analysis

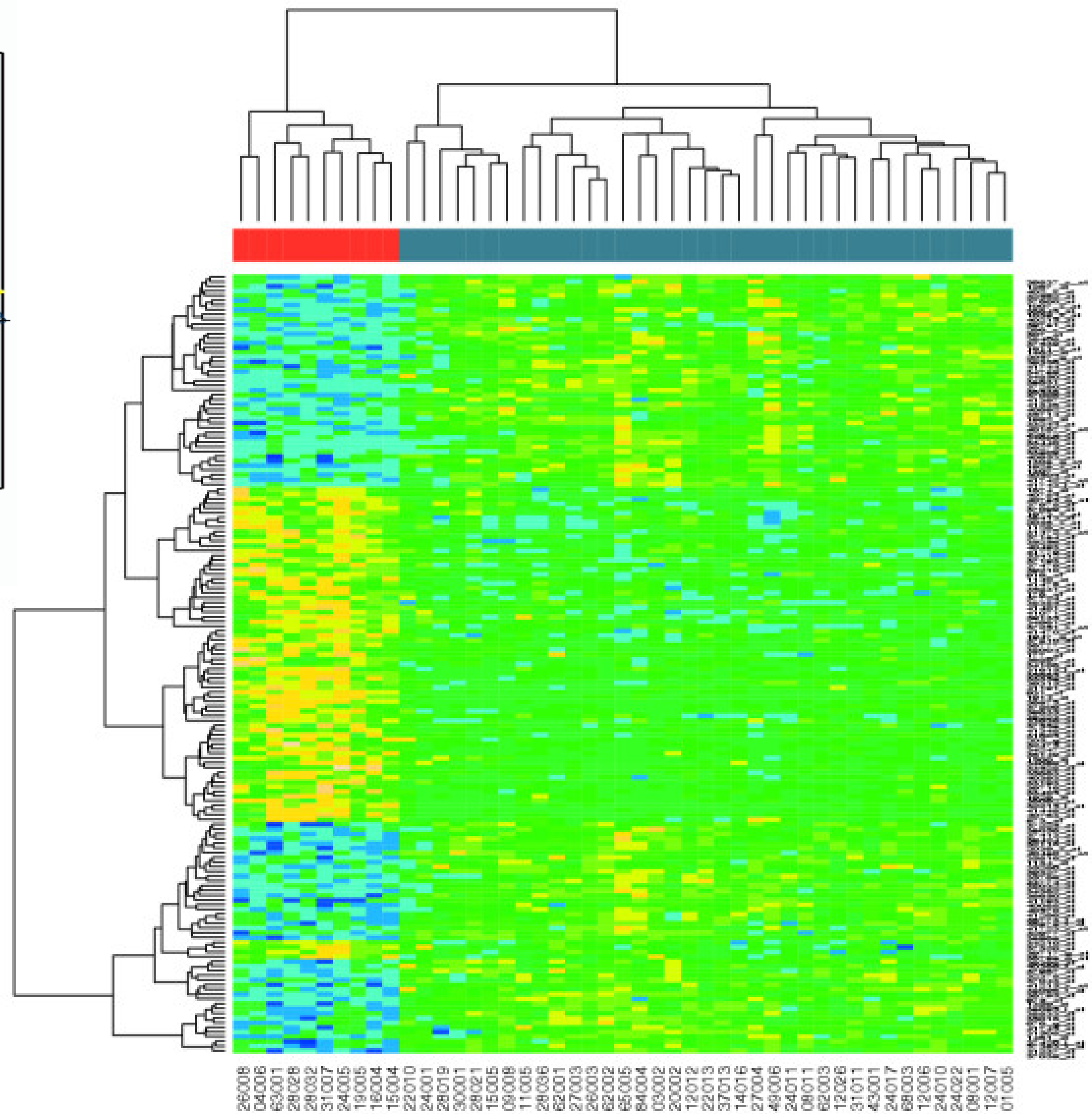
Netherlands Bioinformatics Centre

Quality control

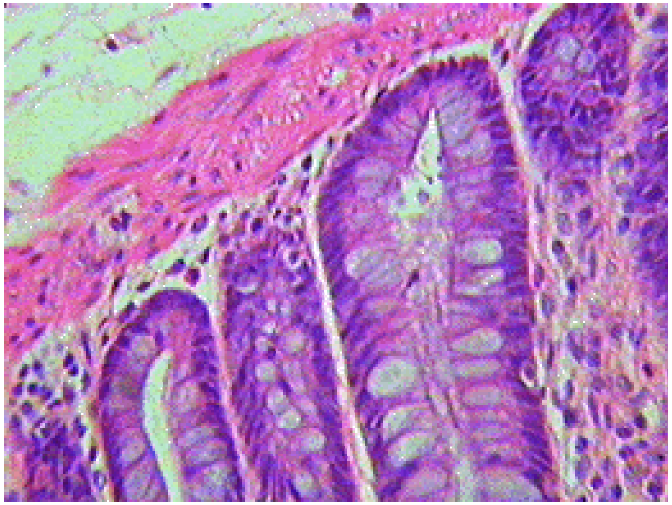


Normalization and statistical analysis

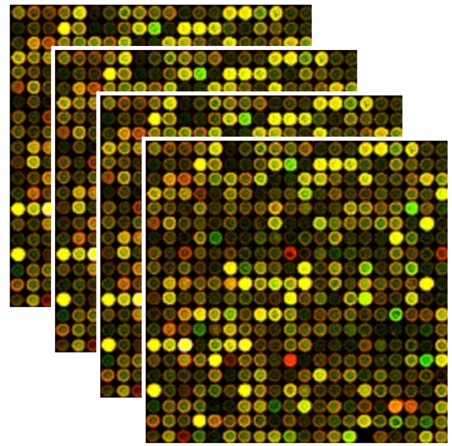




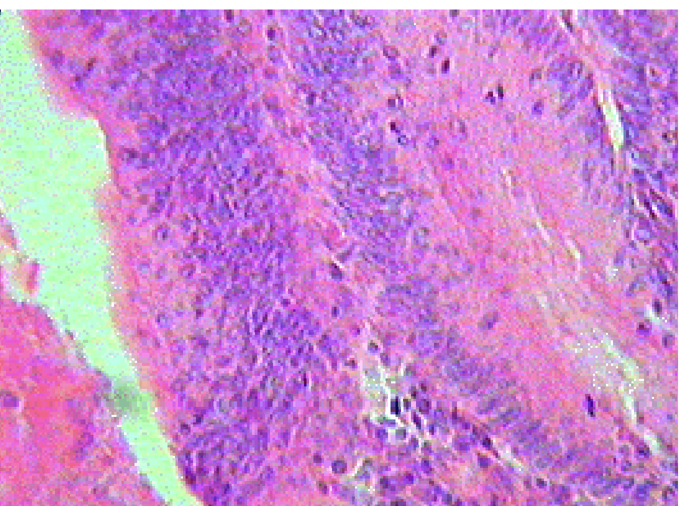
Molecular diagnostics

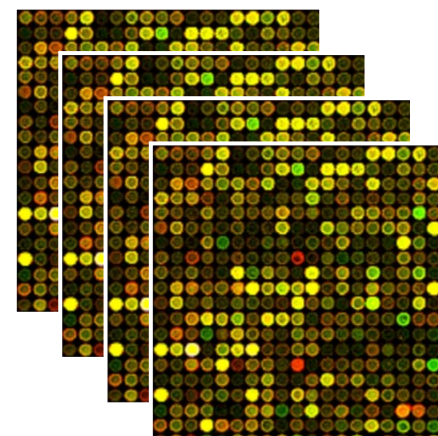


High magnification of a normal human colon cell



Which genes discriminate between normal & patient



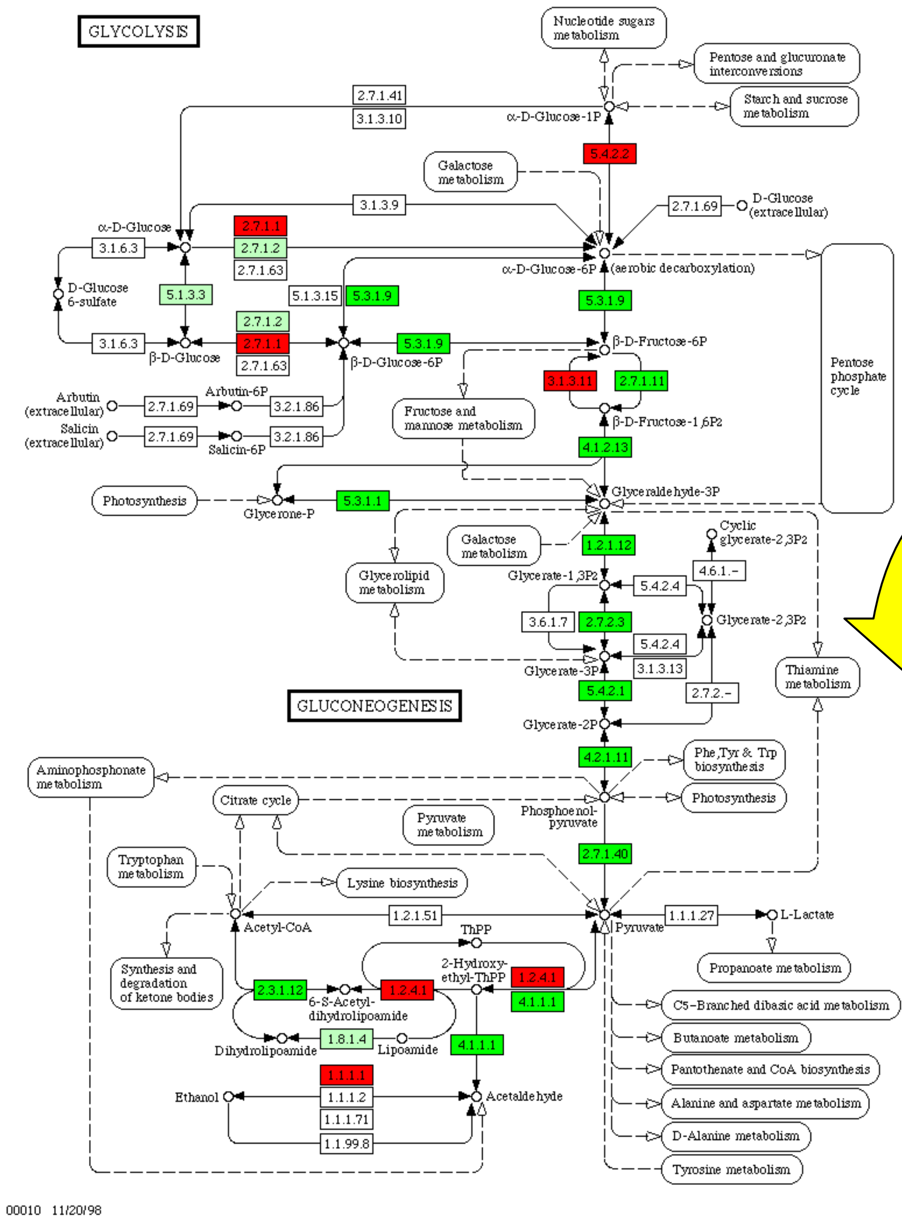


High magnification of a human colon cell with carcinoma

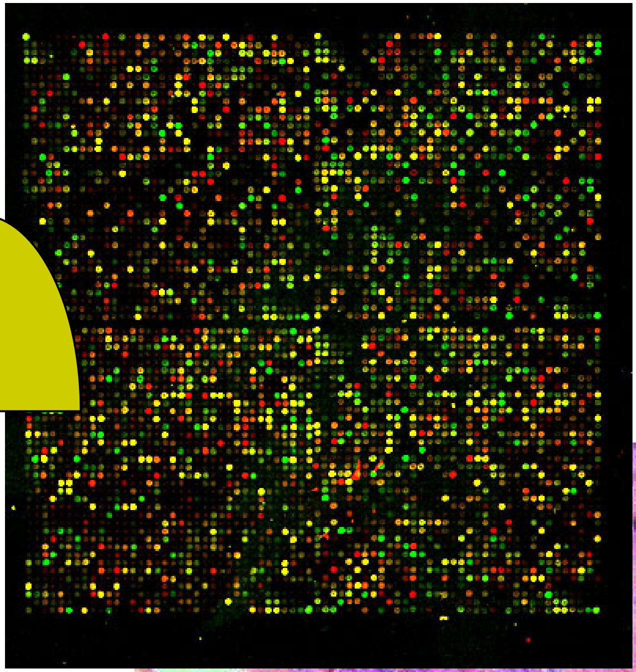
Netherlands Bioinformatics Centre

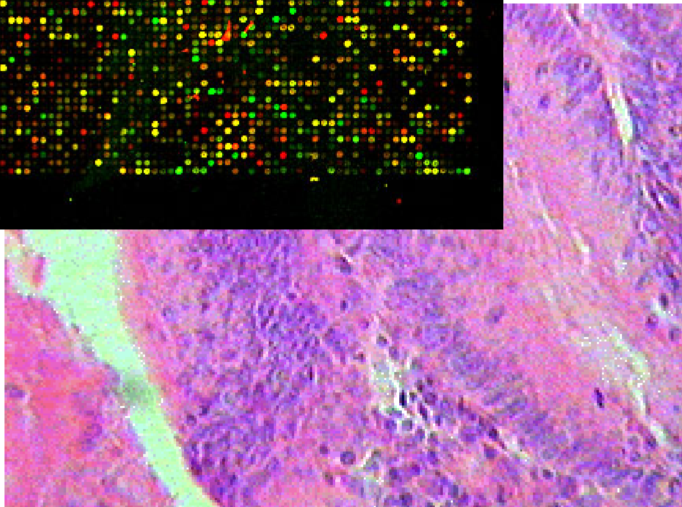
This requires statistical analysis of the data. Complex!

Understanding molecular processes

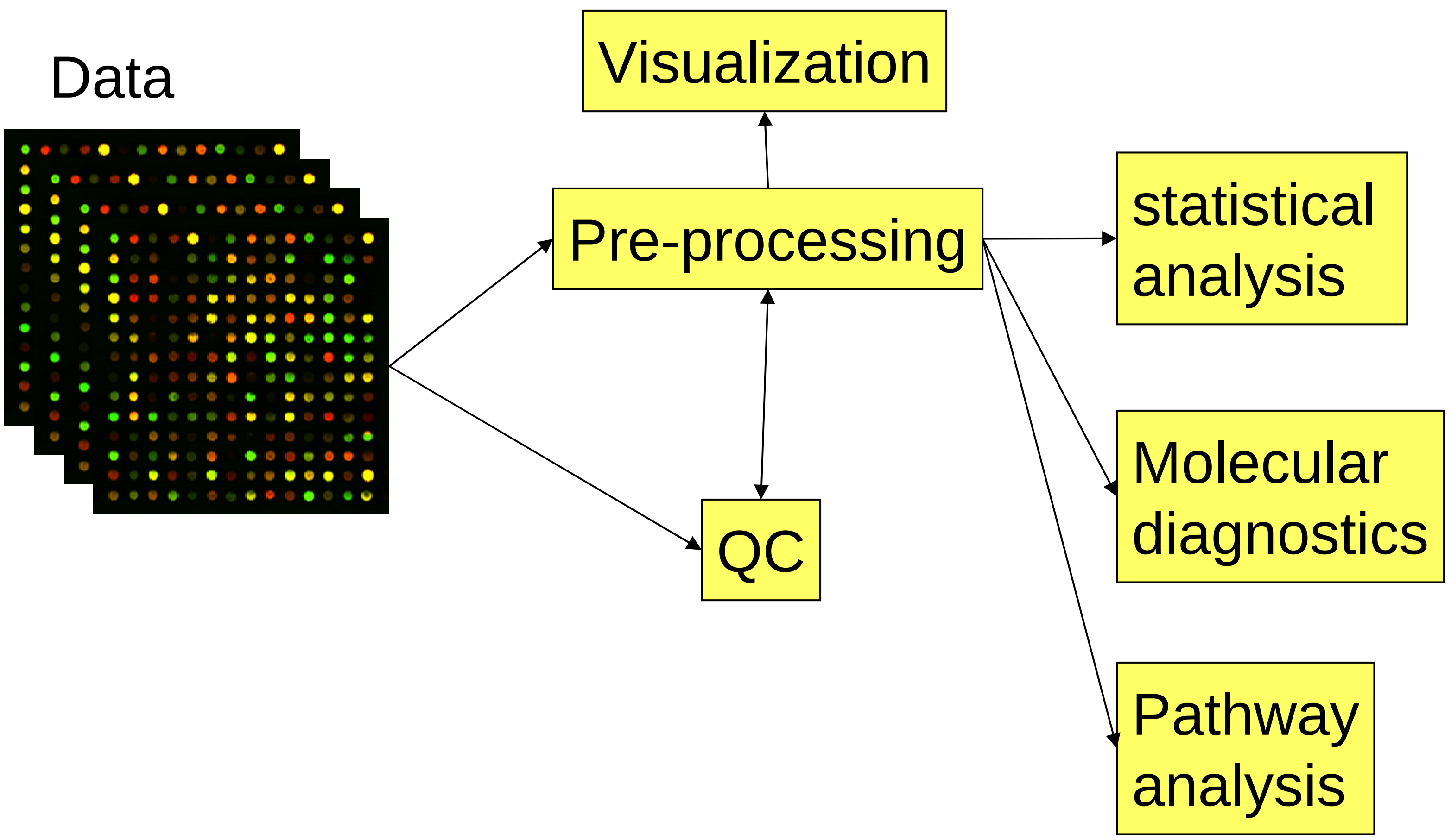


Integration with pathway databases (eg KEGG)

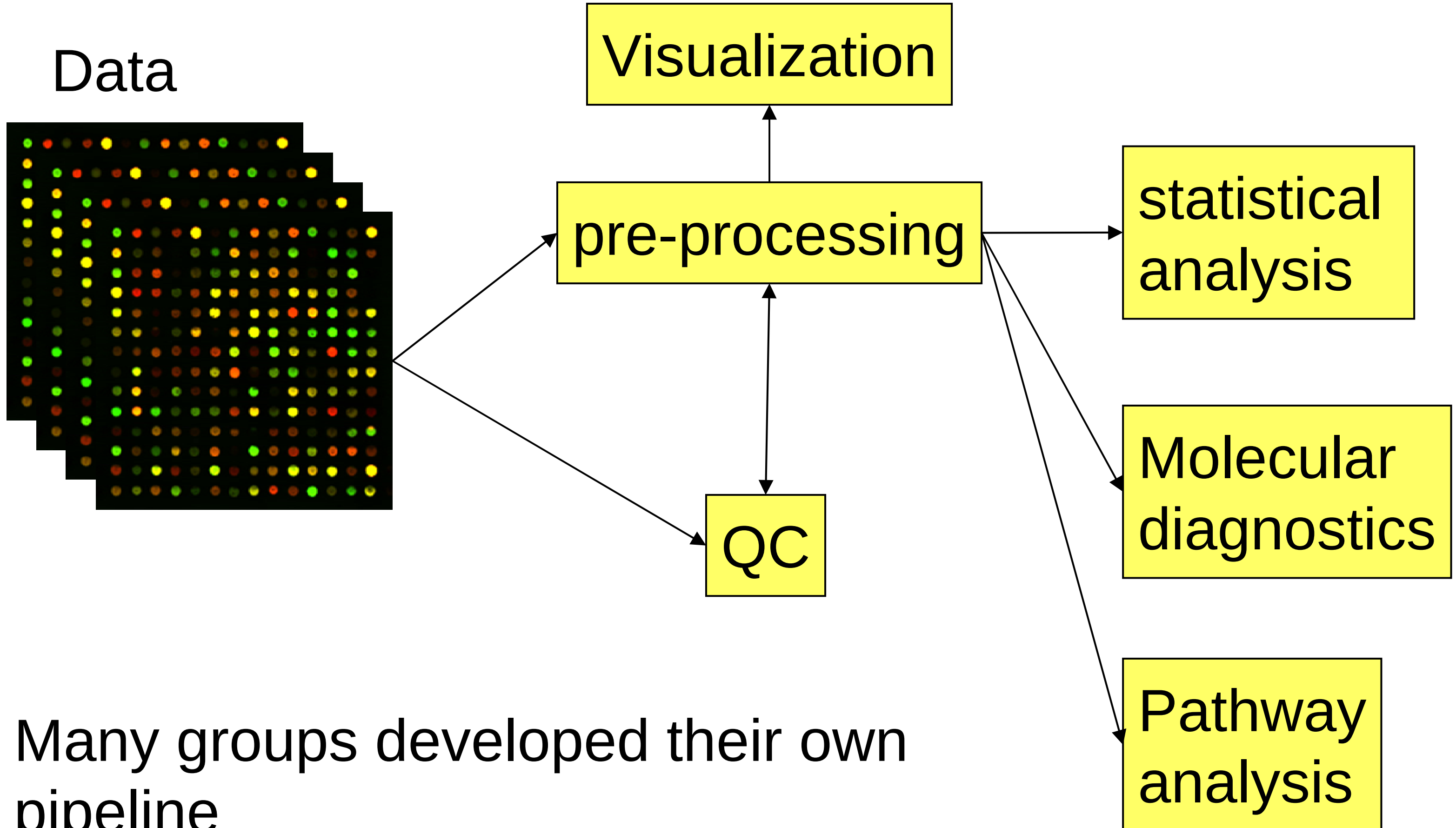




Microarray in-silico experimentation pipeline



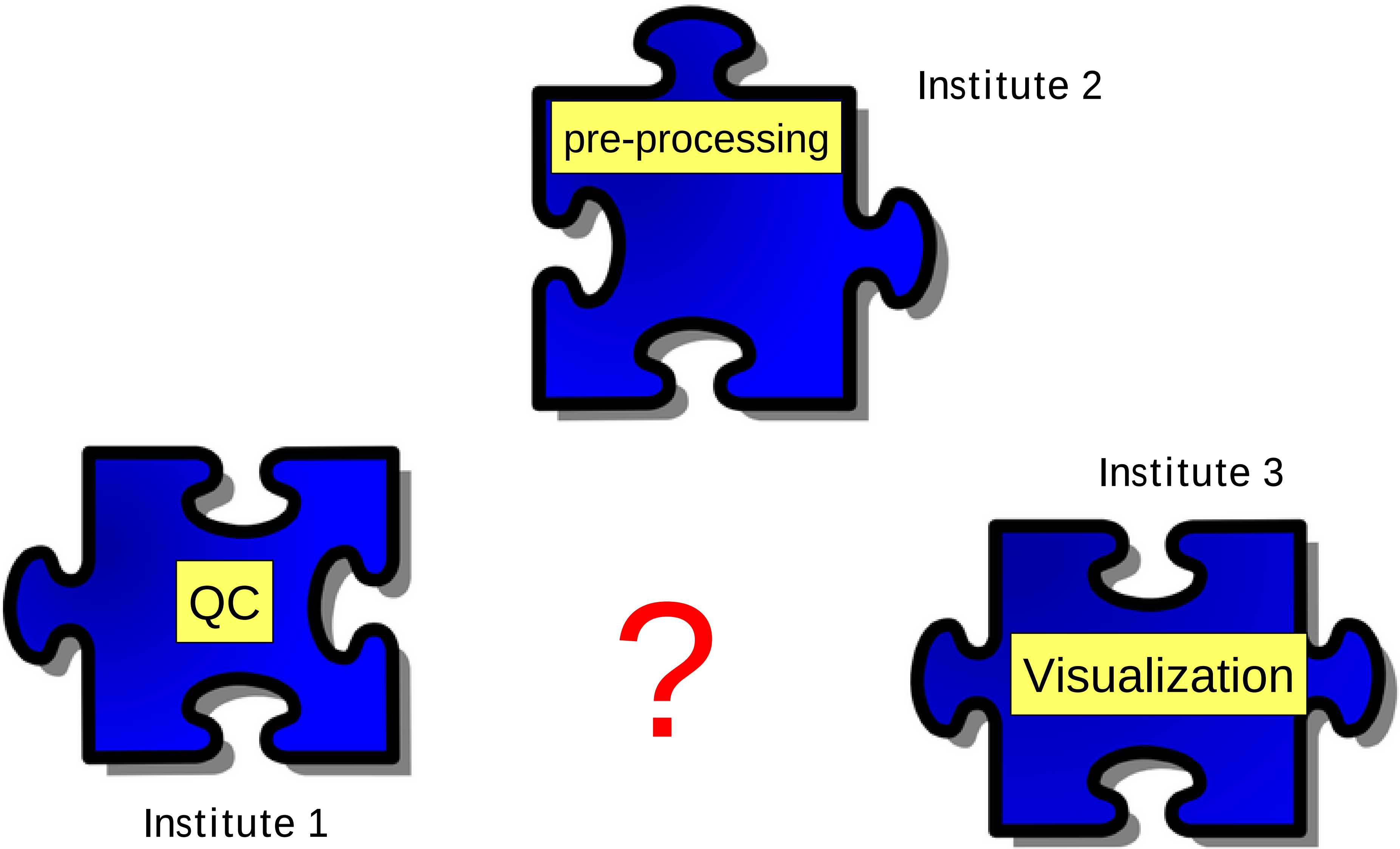
Netherlands Bioinformatics Centre



- Many groups developed their own pipeline
- Large effort
- Development of modules may require specific expertise
- Difficult to use (state-of-the-art) methods of other groups

nbic Netherlands Bioinformatics Centre

....but how to share tools, data, expertise?
....how to jointly solve problems?



….e-Bioscience

Collaborate to develop experimentation pipeline

Service oriented architecture

Share environment with *de facto* standards; use common approaches

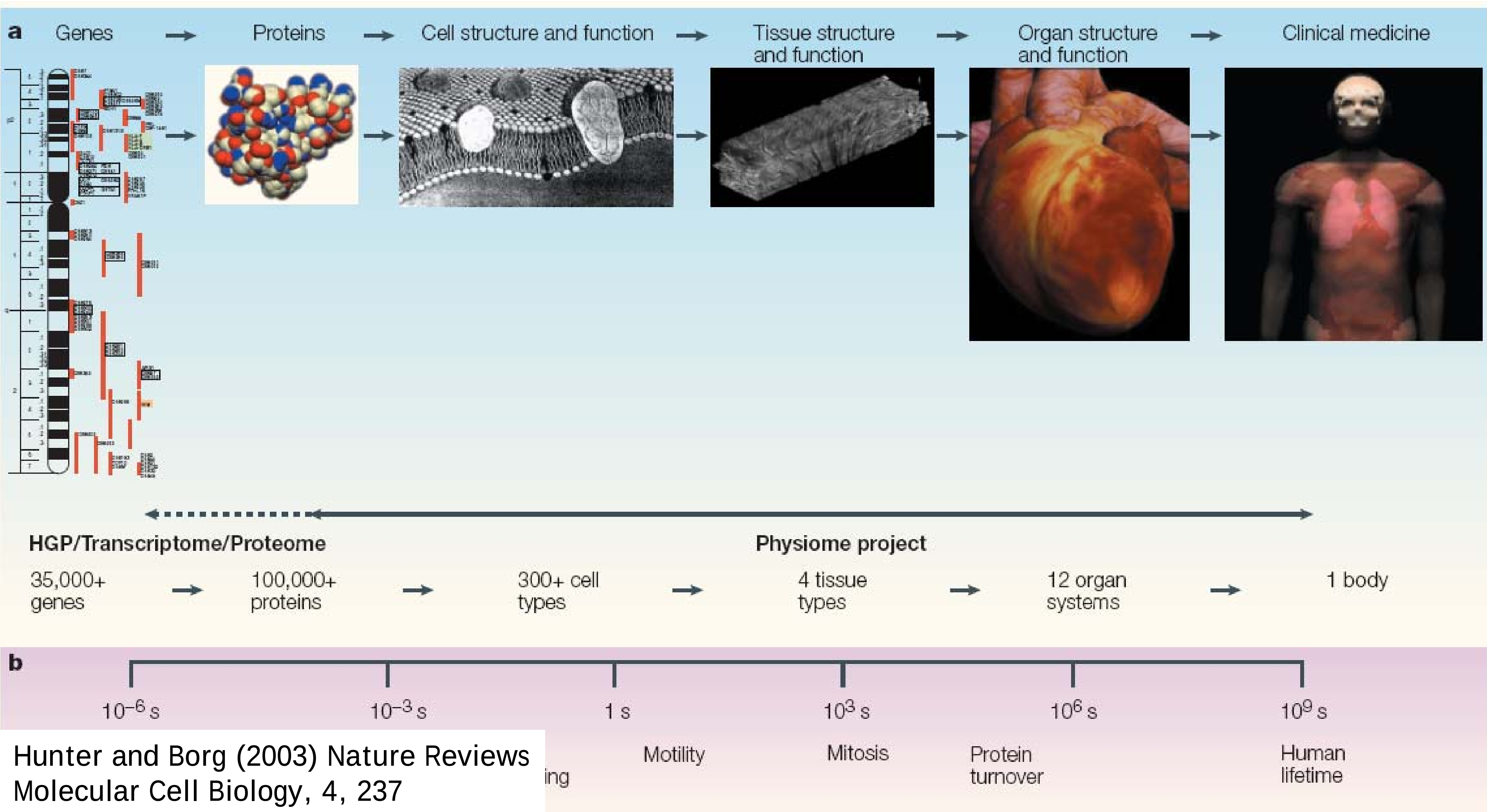
Generic e-Science infrastructure (VLe)

Life sciences GRID (NCF pilot, BIG GRID)

Basic infrastructure (SURFnet, Gigaport)

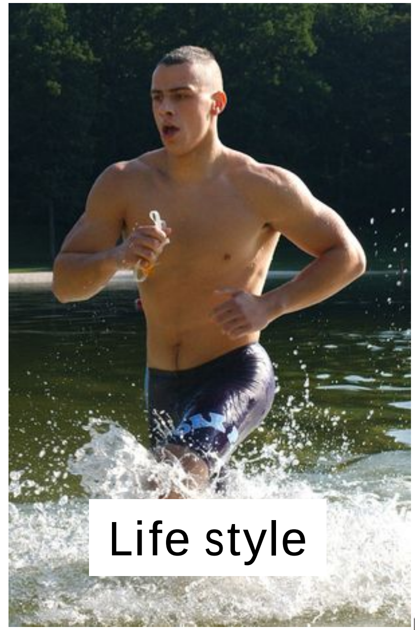
Netherlands Bioinformatics Centre

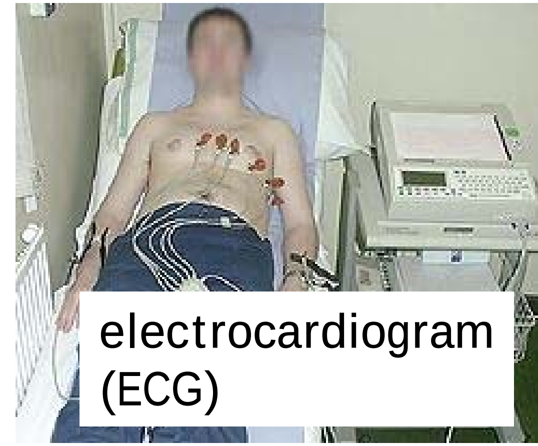
Moreover, science is becoming increasingly complex and multi-disciplinary

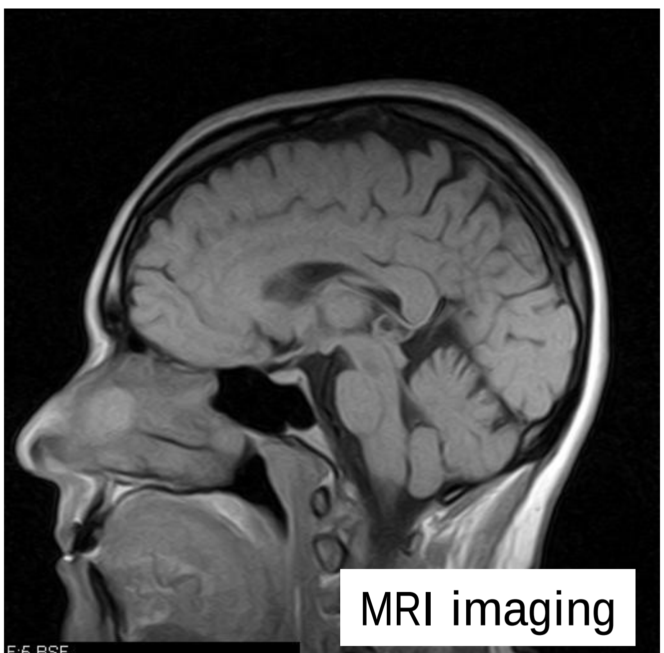


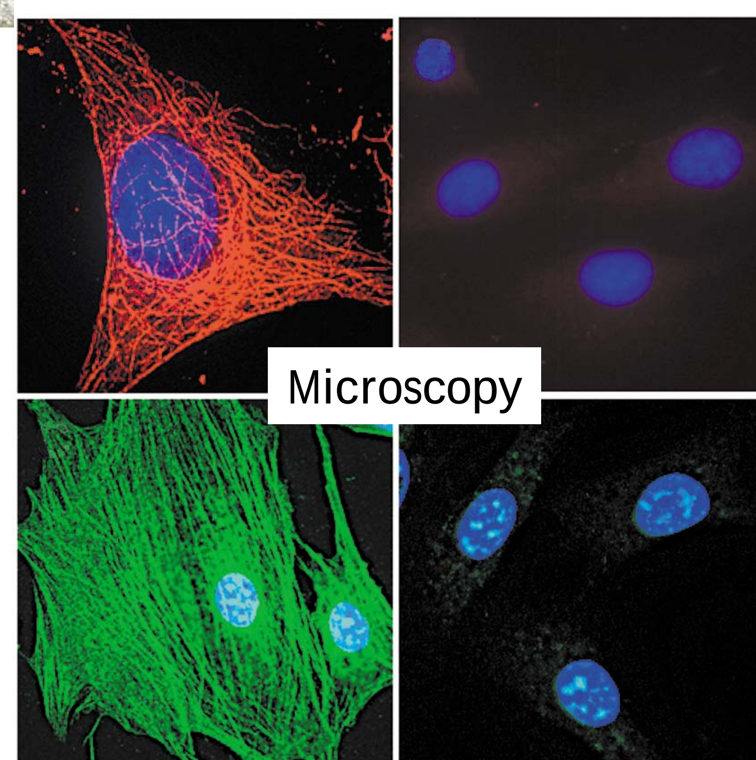
Hunter and Borg (2003) Nature Reviews Molecular Cell Biology, 4, 237


Clinical chemistry

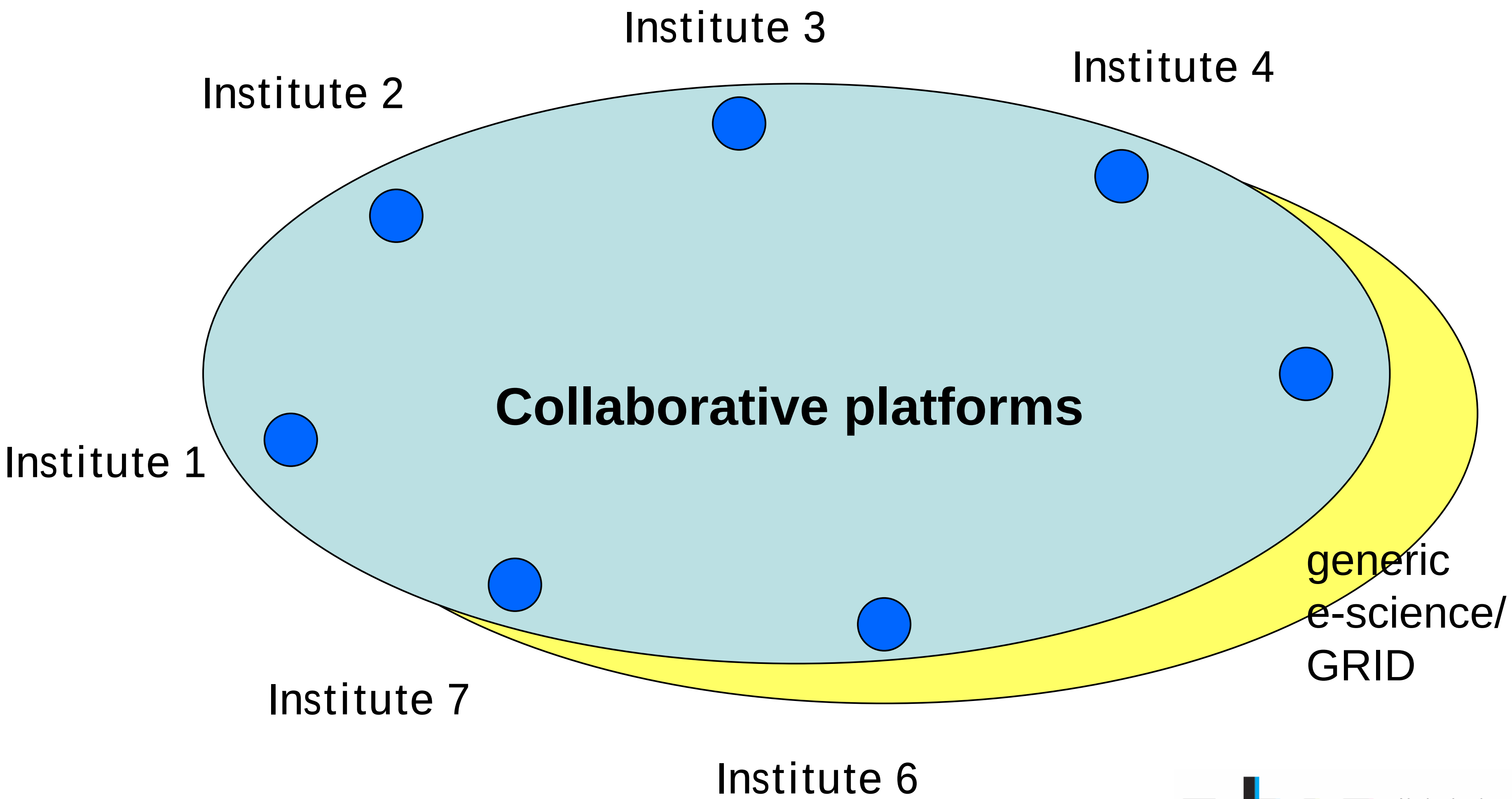

Life style


electrocardiogram
(ECG)


MRI imaging
F:5 BSF

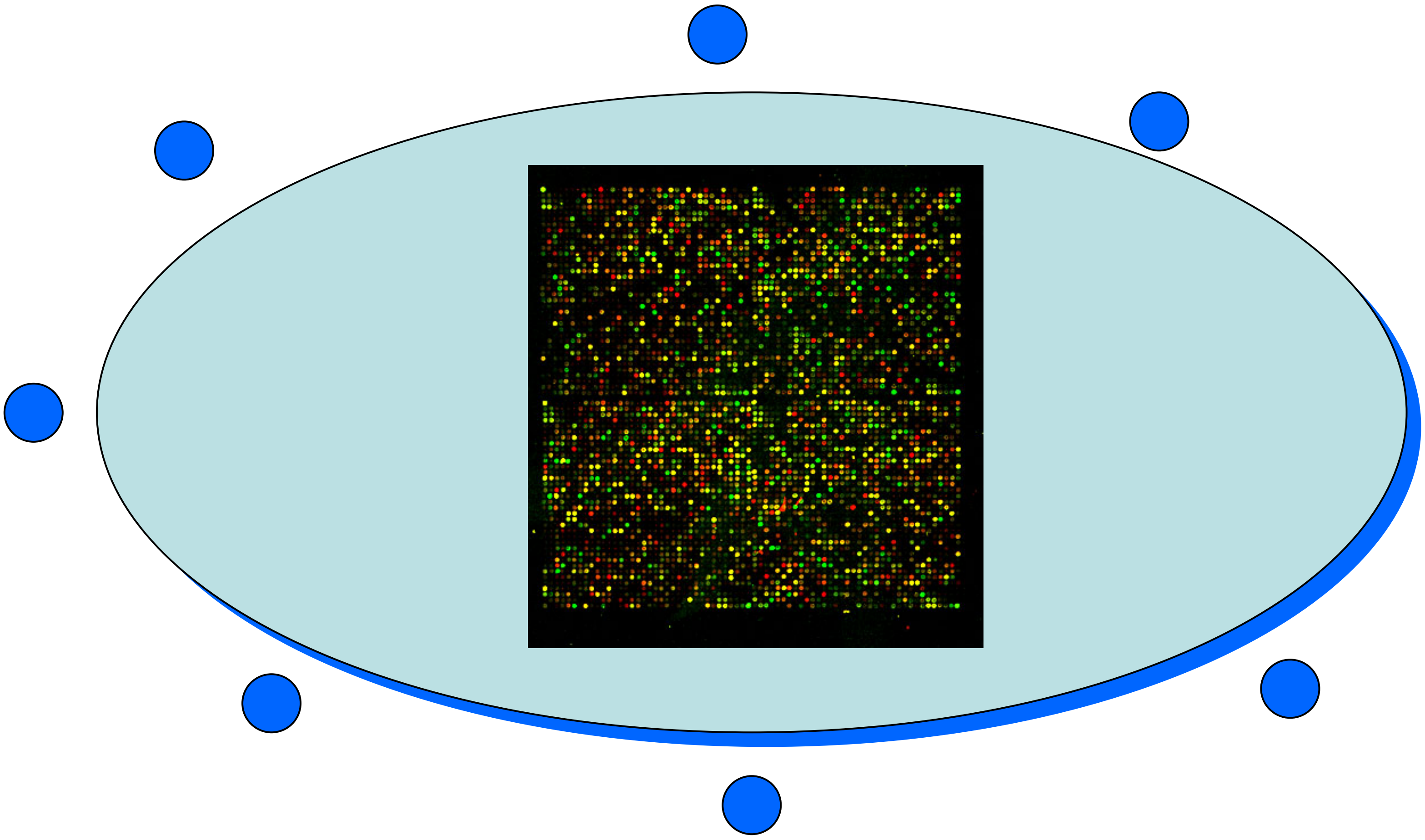

Microscopy

Collaborative platform to address research questions



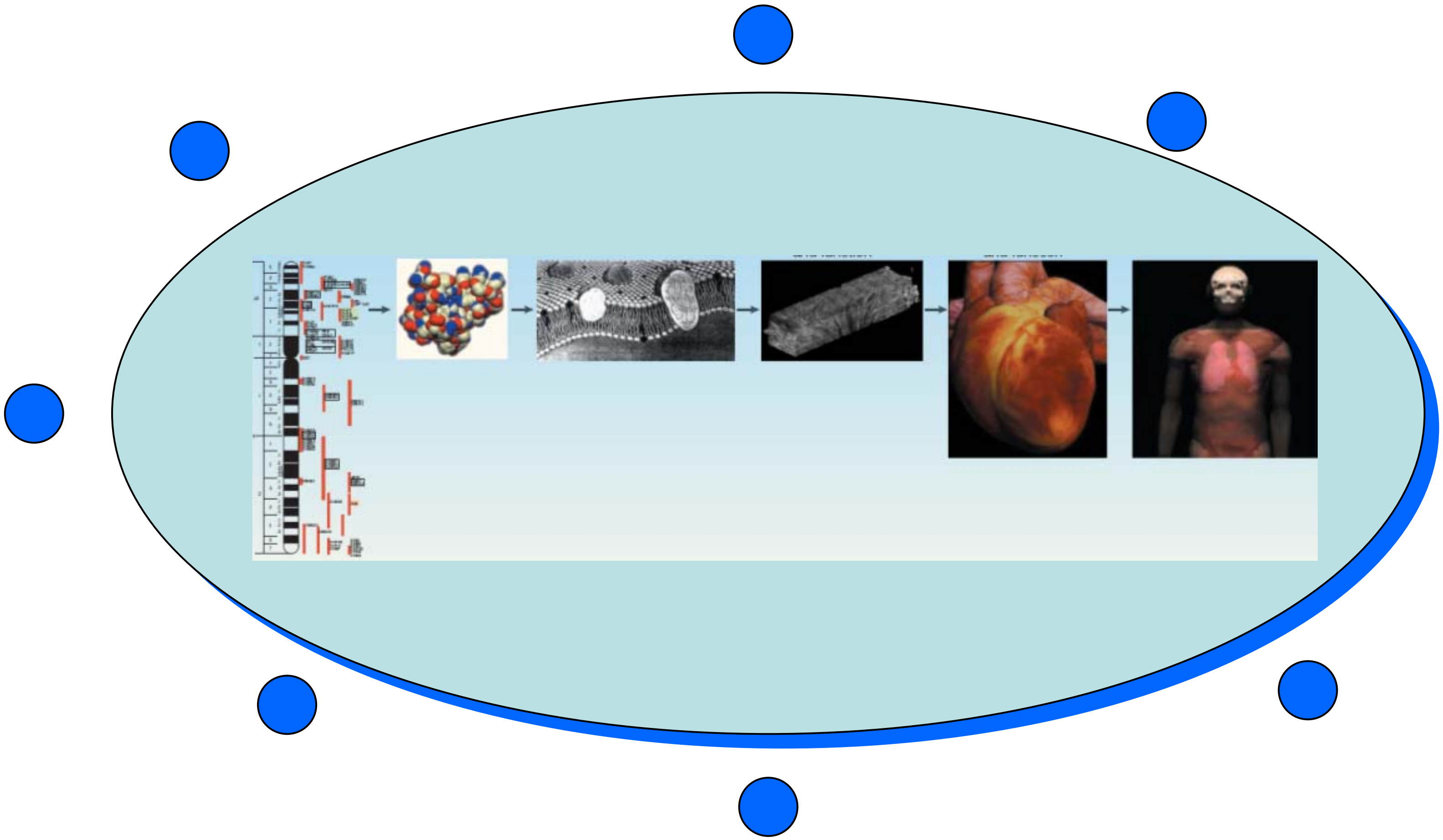
Netherlands Bioinformatics Centre

Collaborative platform for microarray research



Collaborative platform for systems biology

Truly multi-disciplinary!



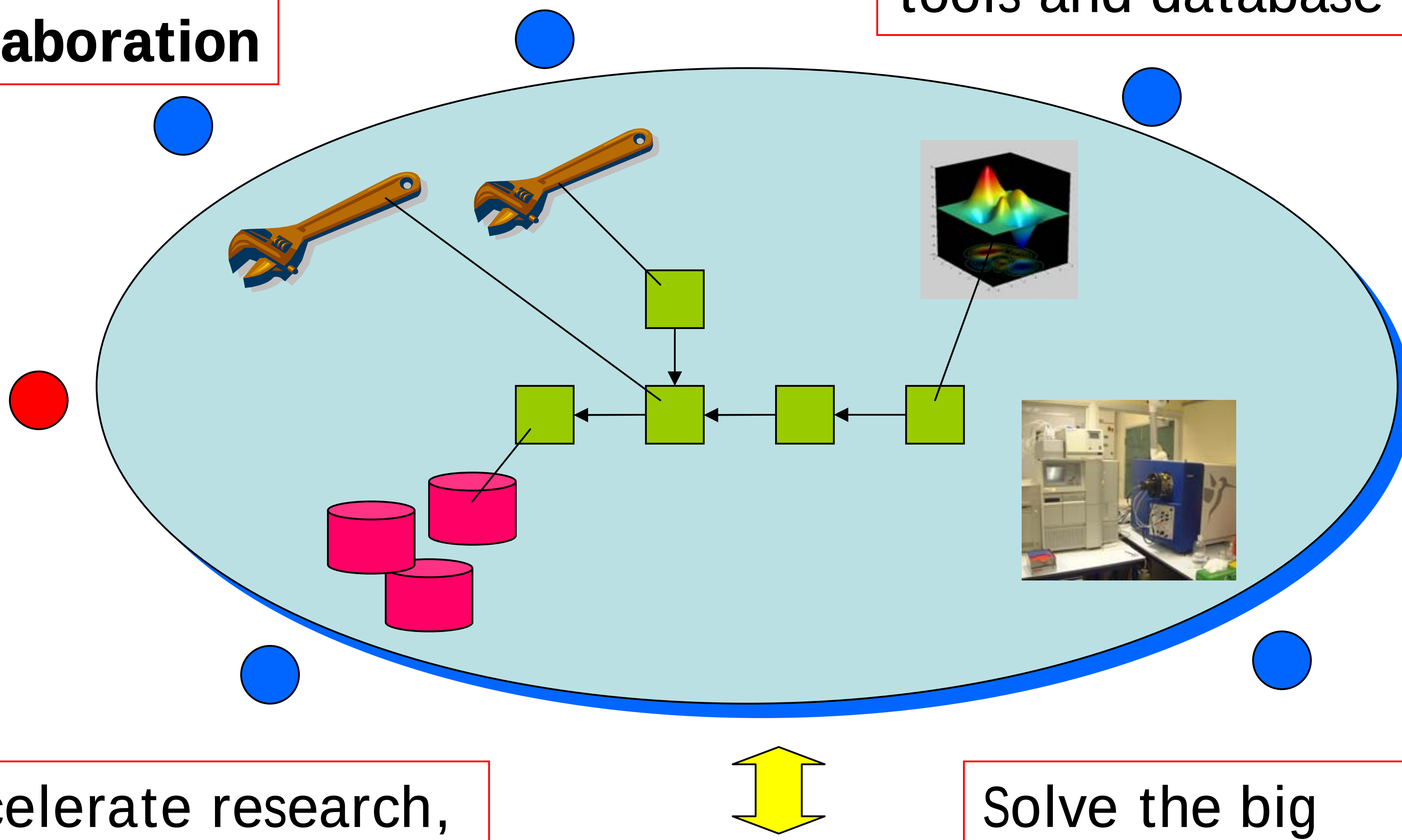
nbic Netherlands Bioinformatics Centre

e-Bioscience

Exchange of data & tools & expertise

Define standardized workflows that connects tools and database

Collaboration



Users

Accelerate research, avoid redundancy, reduce costs

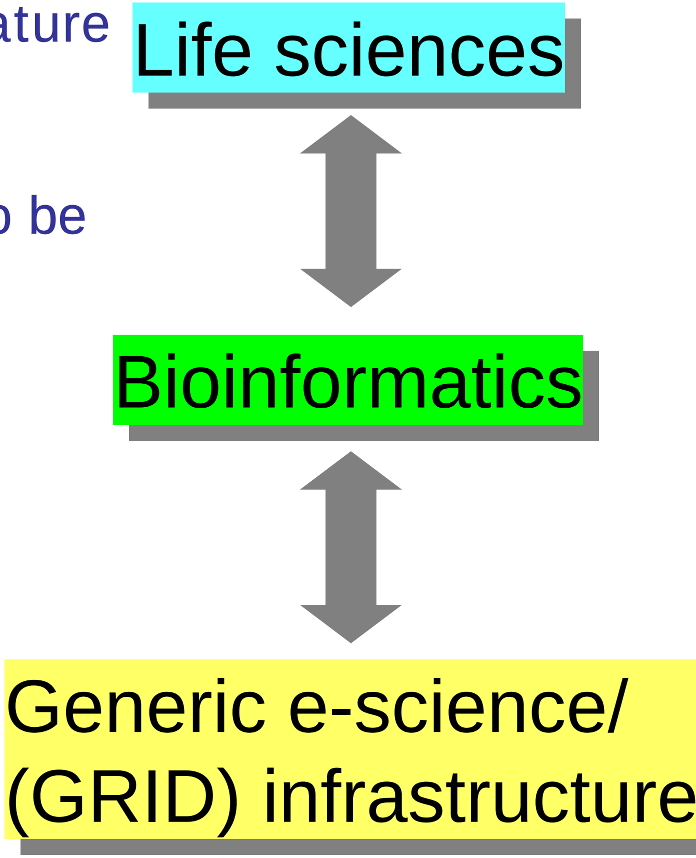
Solve the big scientific questions

Bioinformatics
Centre

The e-Bioscience challenge

E-(bio)science/GRID are not production systems, instead

- developments/research on e-(bio)science and GRID is ongoing.
- Experience from current and future cases will mature this approach
- Collaborative platforms require sufficient time to be designed and implemented
- Requires specific expertise
- Investments (hardware, software, personnel)
- **Willingness to collaborate**



Acknowledgements

University of Amsterdam

- Prof. dr. L.O. Hertzberger
- Dr. T. Breit
- Dr. M. Bouwhuis

CERN

- Dr. C. Jones

Virtual Laboratory for e-Science (VLe; www.vl-e.nl)

BIG GRID (www.nikhef.nl/grid/BIG)

NBIC (www.nbic.nl)

Netherlands Bioinformatics Centre