EDG testbed:
An application point of view

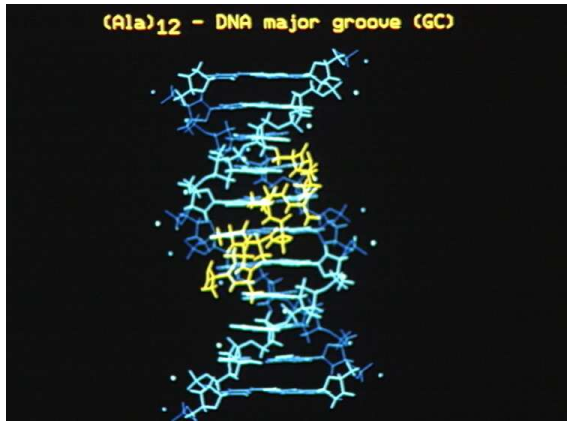
Johan Montagnat

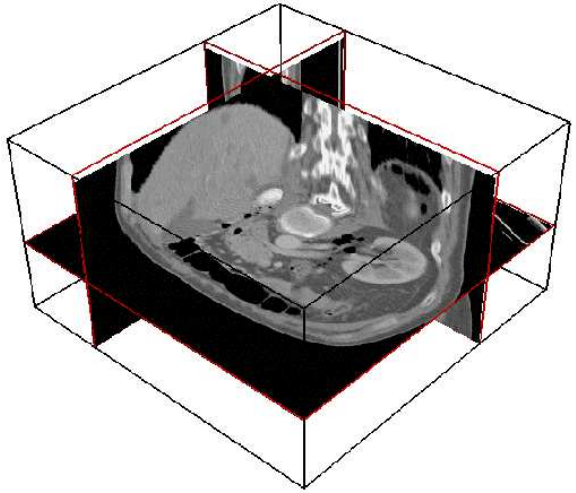
WP10 (Biomedical applications)

EDG tutorial, Feb. 2003

Biomedical applications

- WP10 goals: collect biomedical requirements and deploy first applications on the testbed
  - Biomedical encompass genomics, post-genomics and biomedical applications





- Challenges
  - The biomedical community has no strong center of gravity in Europe
  - Very large user community with little computer science awarness
  - Complex needs (parallel applications, data security, metadata...)

Perspectives for biomedical applications

- Grids open new perspectives in large scale genomics analysis
  - Complete genome annotation
  - Cross-genomes analysis
  - Data mining on distributed databases
  - Pipelining of huge automatic bio-informatics analysis

- Medical image processing
  - Large databases processing
  - Anatomy and physiology modeling
  - Epidemiological studies

Users

- Users
  - Patient: has free access to its medical data.
  - Physician: has complete read access to his/her patients data. Few persons have read/write access.
  - Researchers: may obtain read access to anonymous medical data for research purposes. Nominative data should be blanked before transmission to these users.
  - Biologist: has free access to public databases. Use web portal to access biology server services.
  - Chemical/Pharmacological manufacturer: owns private data. Need to control the possible targets for data storage.

… data …

- Biological data
  - Public and private databases
  - Very fast growth (doubles every 8-12 month)
  - Frequent updates (versionning)
  - Heterogeneous formats
- Medical data
  - Strong semantic
  - Distributed over imaging sites
  - Images and metadata
  - Nominative (critical) and non-nominative (private) data
  - DICOM3 standard compliance for medical images

… and processings

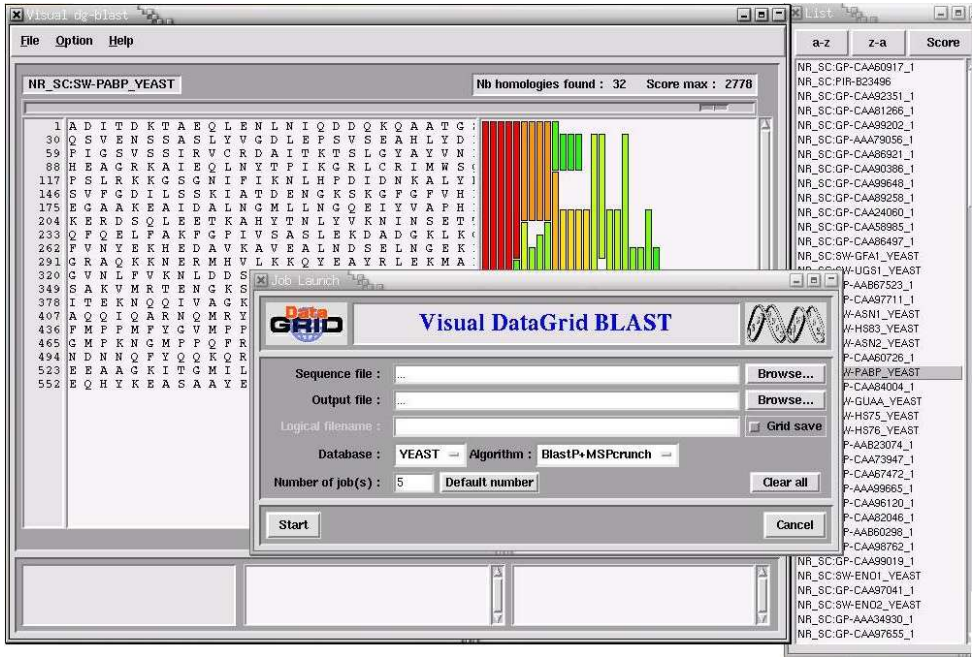
- Processings
  - Are often correlated (pipelines processing)
  - Computation time is often important (to respect clinical applications constraints)
  - Computation time is sometimes critical (e.g. real time simulation)
  - Emergency situation: ambulance jobs
  - Parallel processing is needed to deal with complex algorithms
  - Interactivity may be important

Bioinformatics example: BLAST (Basic Local Alignment Search Tool)

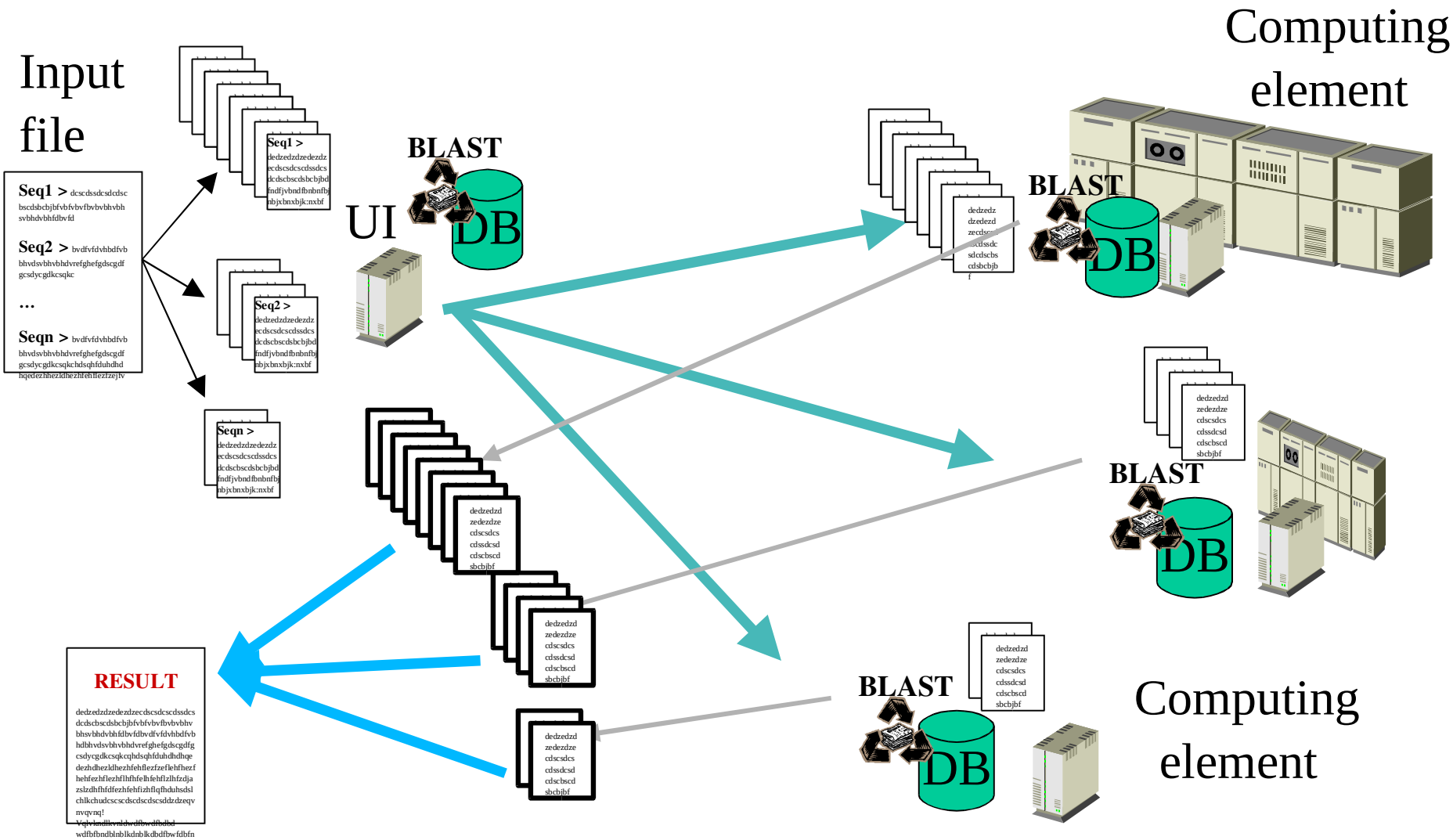
- BLAST is the first step for analysing new sequences: to compare DNA or protein sequences to other ones stored in personal or public databases.
- BLAST is costly and a good candidate for gridification:
  - Requires equipment to store databases and run algorithms
  - Requires manpower for system & network maintenance and frequent update of databases
- More and more biologists…
  - … compare larger and larger sequences (whole genomes)…
  - … to more and more genomes…
  - … with fancier and fancier algorithms !

The visual DataGrid BLAST

- Most biologists use integrated web portals for their genomics comparative analysis: no need to worry about the biological file format and the method arguments.
- The vDG BLAST includes:
  - a graphical interface to enter query sequences and select the reference database
  - A script to execute the BLAST algorithm on the grid
  - A graphical interface to analyze results



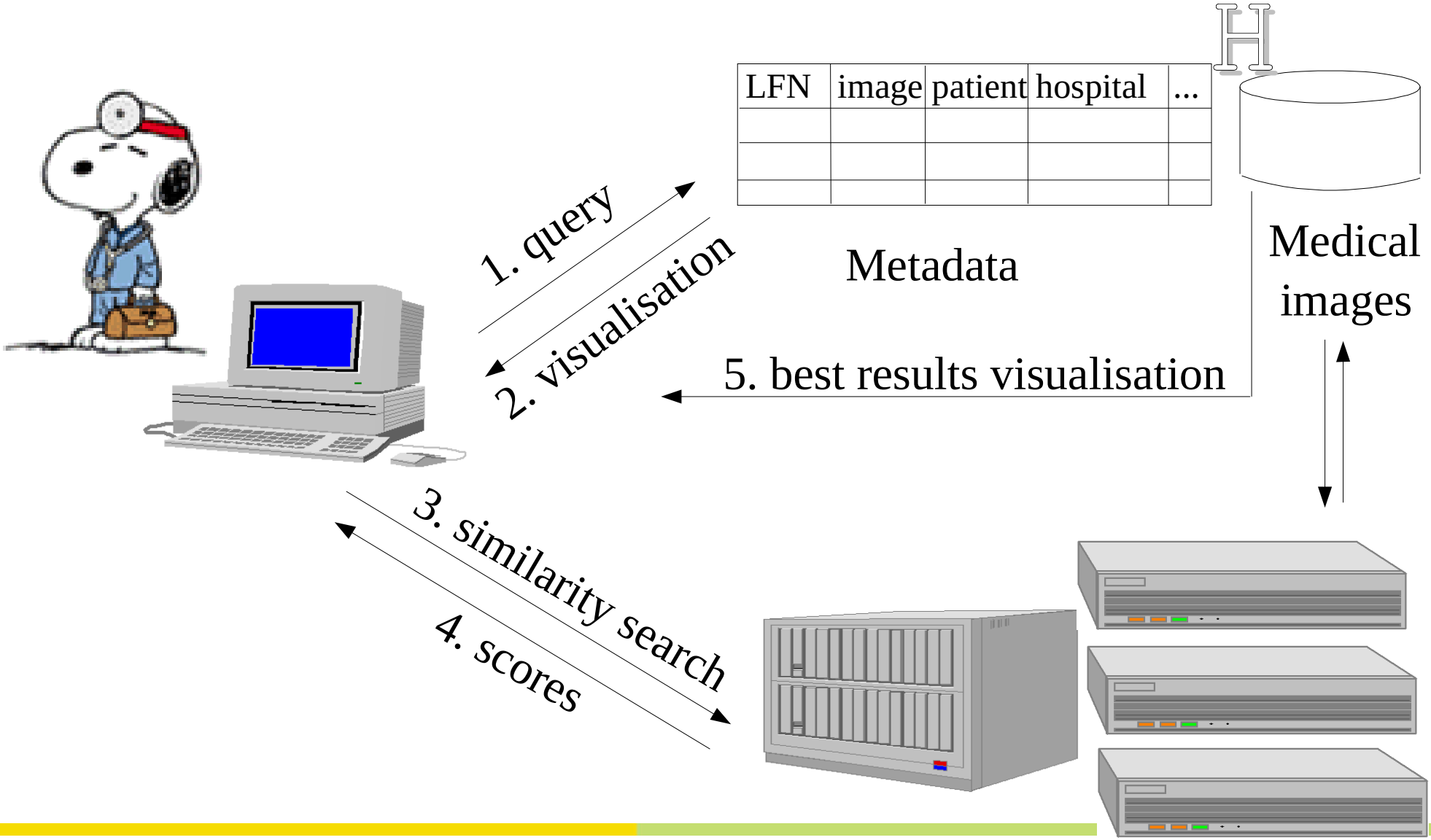
Running BLAST on the grid



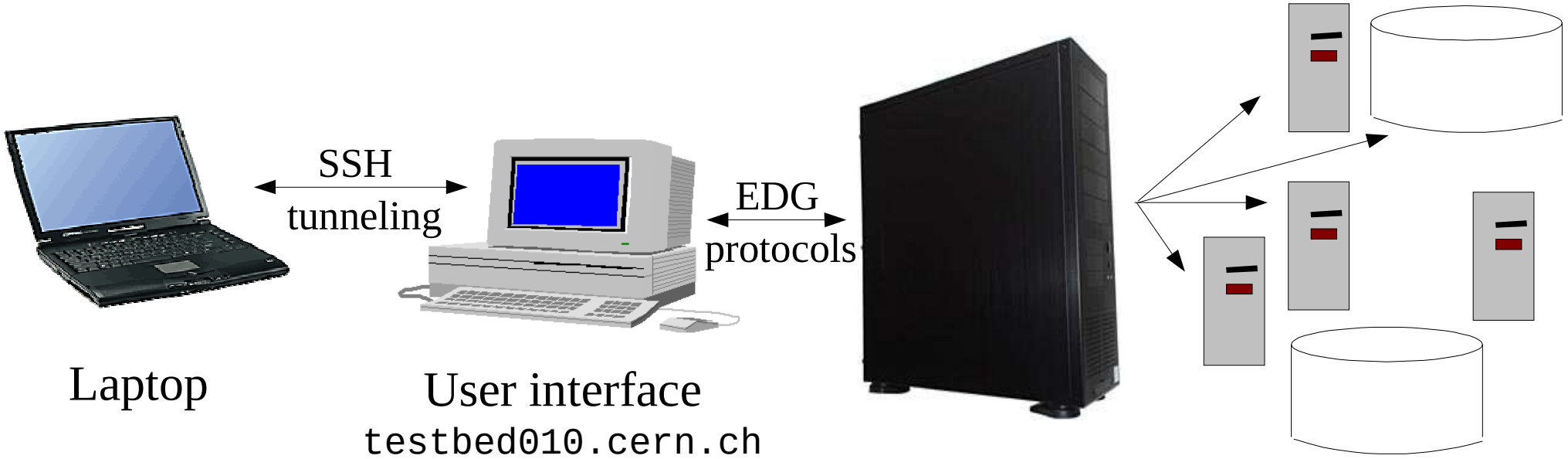
Computation time

- Swissprot vs Swissprot (100000 sequences)
  - Running time on one CPU : 228 hours
  - Tests at Institut de Biologie et Chimie des Protéines (quadripro) : 49 hours
  - Tests on DataGrid testbed1 (cc-in2p3) : 3 hours
- Impacts :
  - Reduced pressure on local computing
  - Ability to handle very large jobs

Medical imaging example: Image content-based query



Development on top of the EDG middleware

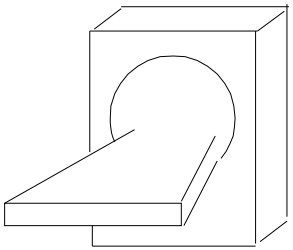


3. Application layer
2. Graphic layer
1. C++ API

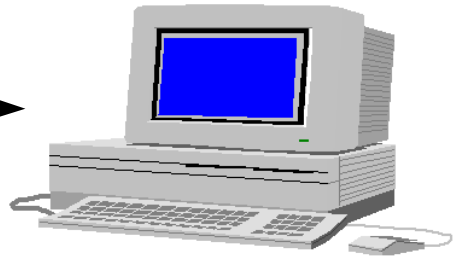
SSH
system calls

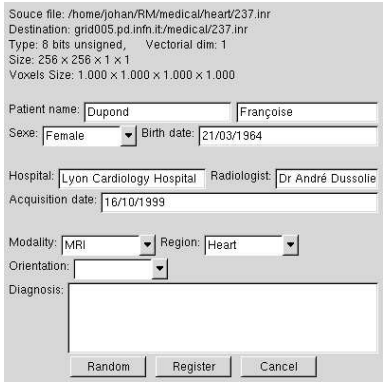
EDG command line interface
EDG middleware
Globus

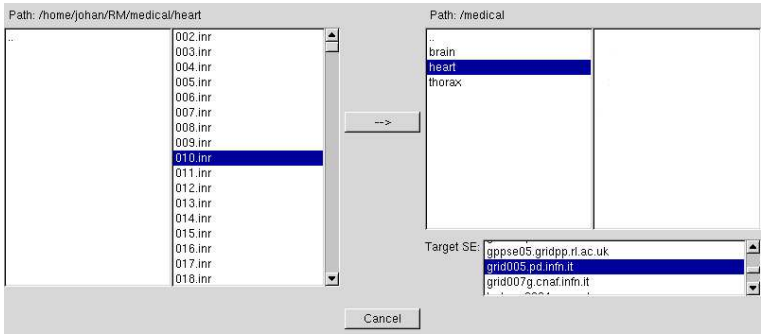
Data and metadata registration



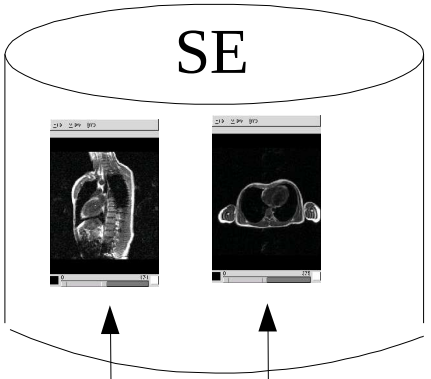
Imager







LFN	image	patient	hospital	...



Running similarity measurements

Similarity computation

Job	Status	Target
27499 (similarity)	Terminated	localhost:0/noqueue
27503 (similarity)	Terminated	localhost:0/noqueue
27507 (similarity)	Terminated	localhost:0/noqueue
27511 (similarity)	Terminated	localhost:0/noqueue
27515 (similarity)	Terminated	localhost:0/noqueue
27520 (similarity)	Terminated	localhost:0/noqueue
27524 (similarity)	Terminated	localhost:0/noqueue
27528 (similarity)	Terminated	localhost:0/noqueue
27532 (similarity)	Terminated	localhost:0/noqueue
27536 (similarity)	Terminated	localhost:0/noqueue
27540 (similarity)	Terminated	localhost:0/noqueue
27544 (similarity)	Terminated	localhost:0/noqueue
27548 (similarity)	Terminated	localhost:0/noqueue
27552 (similarity)	Terminated	localhost:0/noqueue
27556 (similarity)	Terminated	localhost:0/noqueue
27560 (similarity)	Output ready	localhost:0/noqueue
27564 (similarity)	Running	localhost:0/noqueue
27568 (similarity)	Submitted	localhost:0/noqueue
27572 (similarity)	Submitted	
New similarity	Sending to UI	

Dismiss

Job monitoring

File Similarity About

Source image:

Jones Jean	Cardiology Center of Monaco	Dr Jina Carlson	1997-11-18

Results:

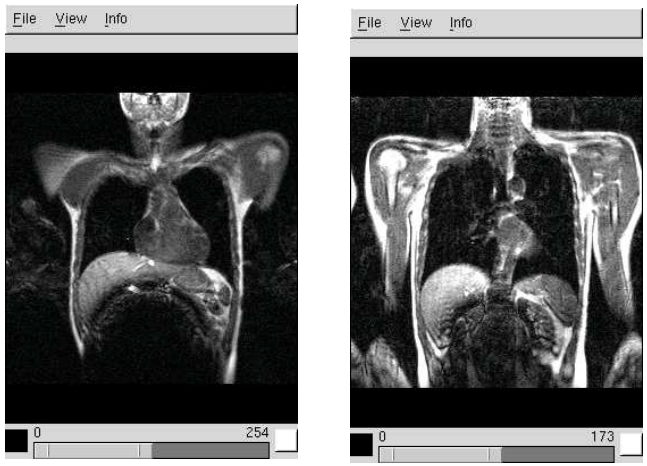
0.904684	Durand Jean	Lyon Cardiology Hospital	Dr Alain Deloin	2002-02-21
0.743148	Dupont Marc	Cardiology Center of Monaco	Dr Francis Black	1998-01-18
0.219426	Durand Jean	Cardiology Center of Monaco	Dr Jina Carlson	2000-10-08
0.217490	Jones Linda	Montreal Neurological Institut	Dr Fany Anderson	2000-12-21
0.193847	Jones Sandra	Cardiology Center of Monaco	Dr Francis Black	2000-12-25
0.003237	Dupont Denise	Montreal Neurological Institut	Dr Norbert White	1998-10-22
0.003084	Dupont John	Montreal Neurological Institut	Dr Norbert White	1998-04-22
0.002636	Smith Marc	Cardiology Center of Monaco	Dr Jina Carlson	1997-04-04
0.001778	Durand Sylvie	Lyon Neurology Hospital	Dr Martine Foliet	2001-02-14
0.001515	Smith Marc	Montreal Neurological Institut	Dr Norbert White	2001-02-09
0.001023	Durand Jean	Cardiology Center of Monaco	Dr Jina Carlson	2000-02-24

View

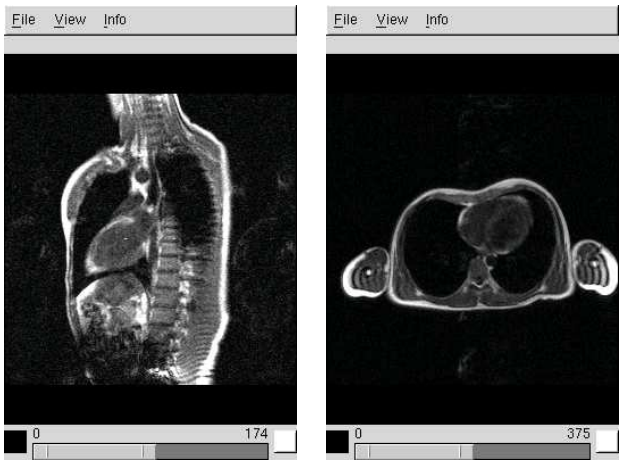
Ranked list of images

Results visualization

Source image



Most similar images



Low score images

Other medical applications

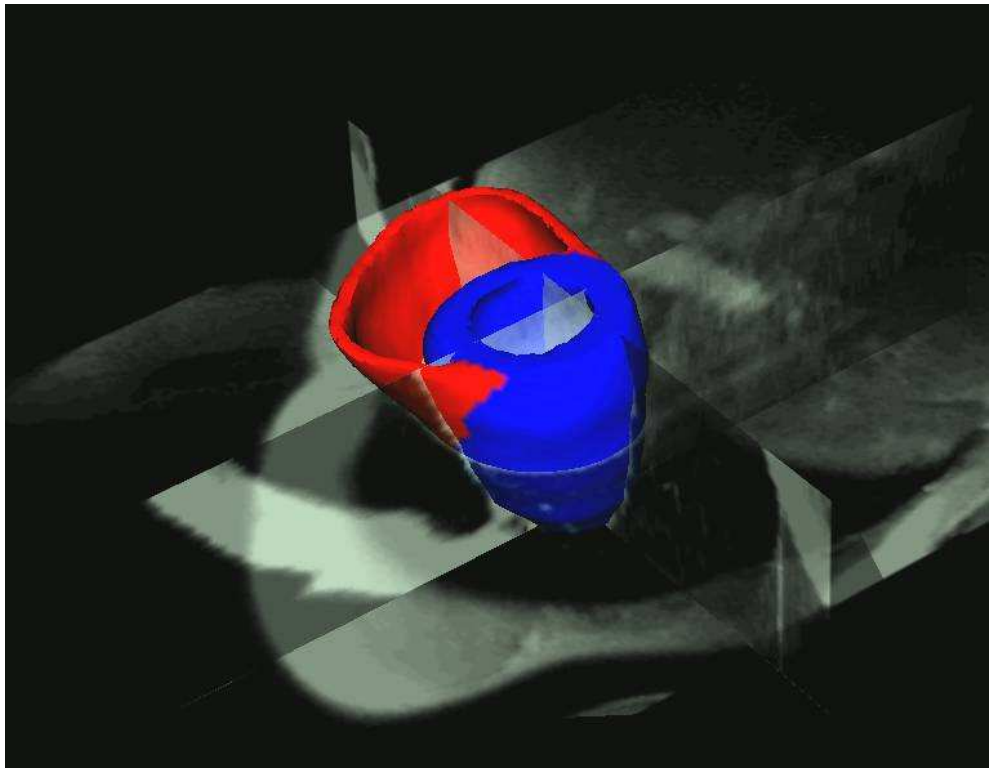
- 1. Complex modeling of anatomical structures
  - anatomical and functional models, parallelization
- 2. Surgery simulation
  - Realistic models, real-time constraints
- 3. Simulation of MRIs
  - MRI modeling, artifacts modeling, parallel simulation
- 4. Mammographies analysis
  - Automatic pathologies detection
- 5. Shared and distributed data management
  - data hierarchy, dynamic indices, optimization, caching

1. Complex modeling of anatomical structures

- Anatomical modeling for:
  - Segmentation
  - Quantitative analysis
- Linear Finite Element

  Modeling of biomechanics
  - Parallelization of large

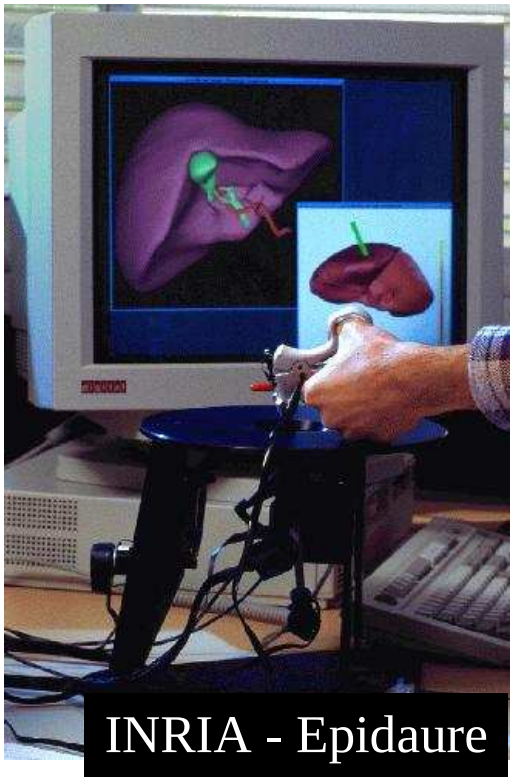
    linear systems

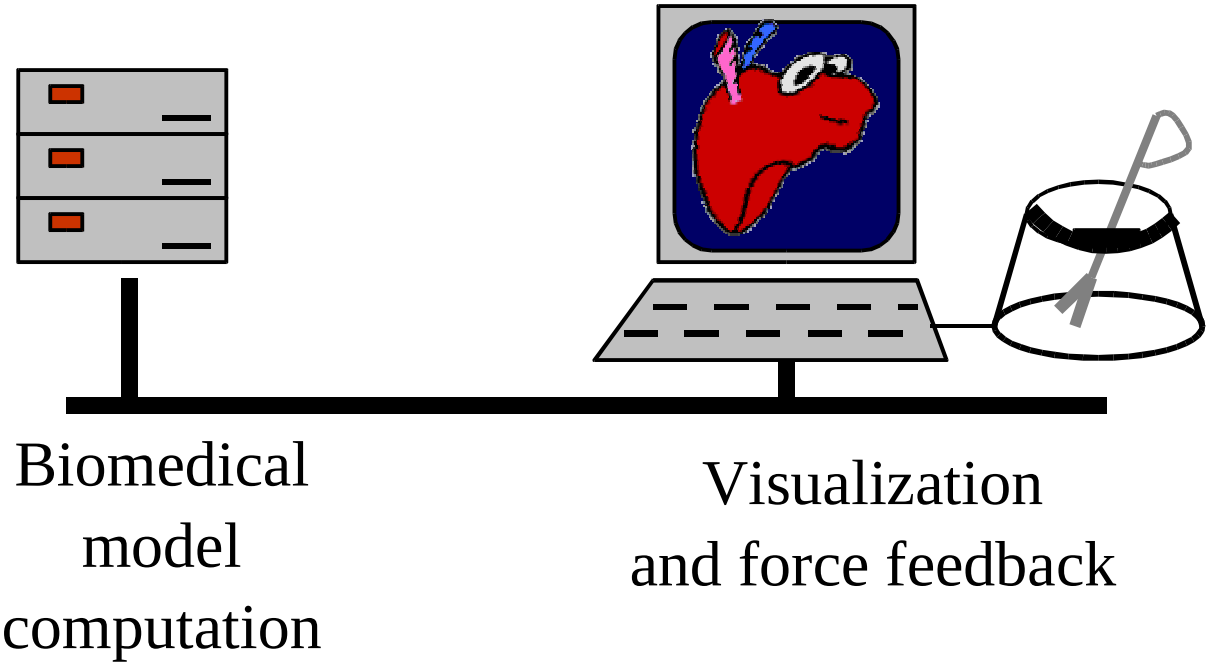


- Modeling / segmentation of 3D+T cardiac sequences in a reasonnably short amount of time (few minutes)

2. Surgery simulation

- **Surgery simulation**

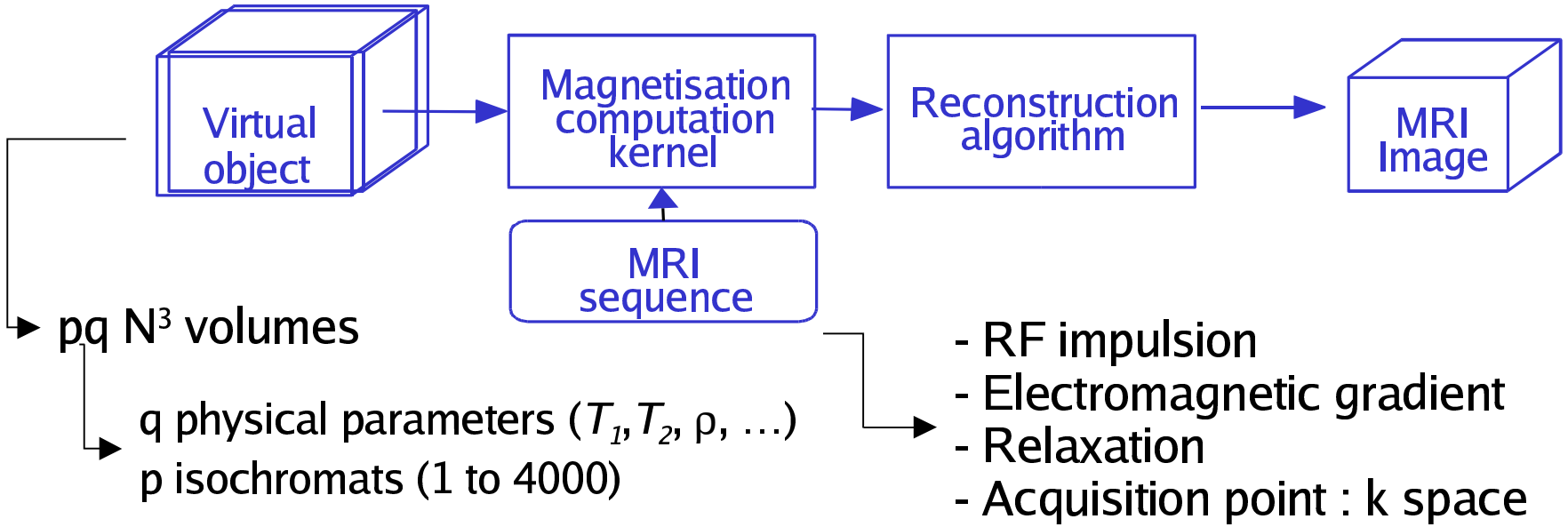




- **Biomedical model deformation on CE**
- **Real time visual (25 Hz) and force (300 Hz) feedback on user interface**

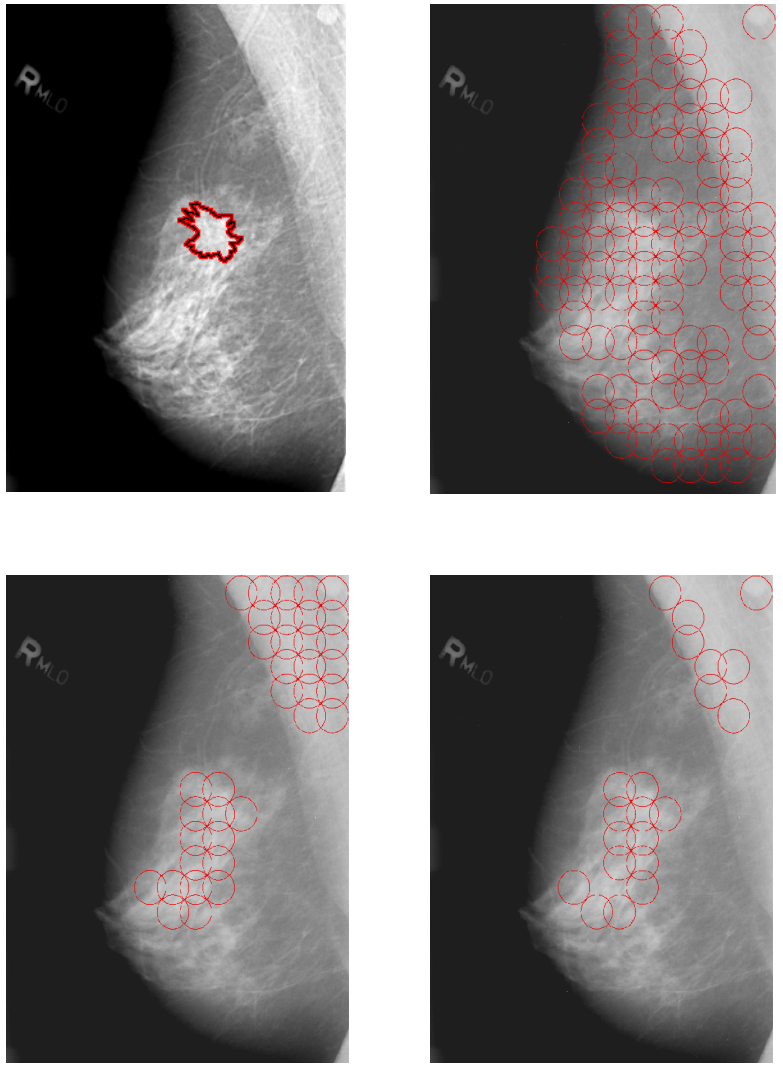
3. Simulation of MRIs

- Produce simulated images from a perfectly known model for:
  - Artifacts study and correction
  - Image processing evaluation
  - MRI sequences testing and design
- Overview



4. Mammographies analysis

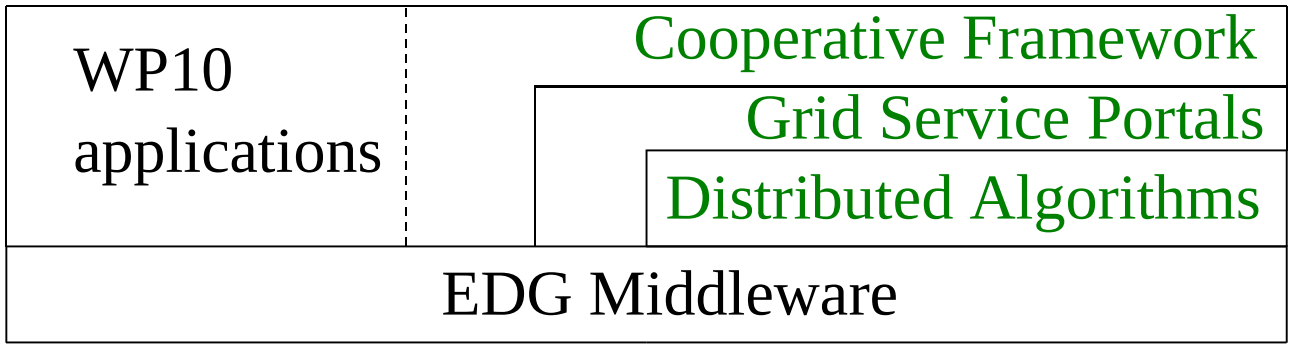
- More than 10000 images, 450 Gbytes
- 400 sub regions (e.g.) per image
- About 250 variables extracted on each region for training and for CBIR
  - Texture, gray-levels and shape analysis
  - Image indexation
- Indexing requires about 30 minutes of computations per image (Sun Ultra-10, 440 MHz), no optimization



5. Shared and distributed data management

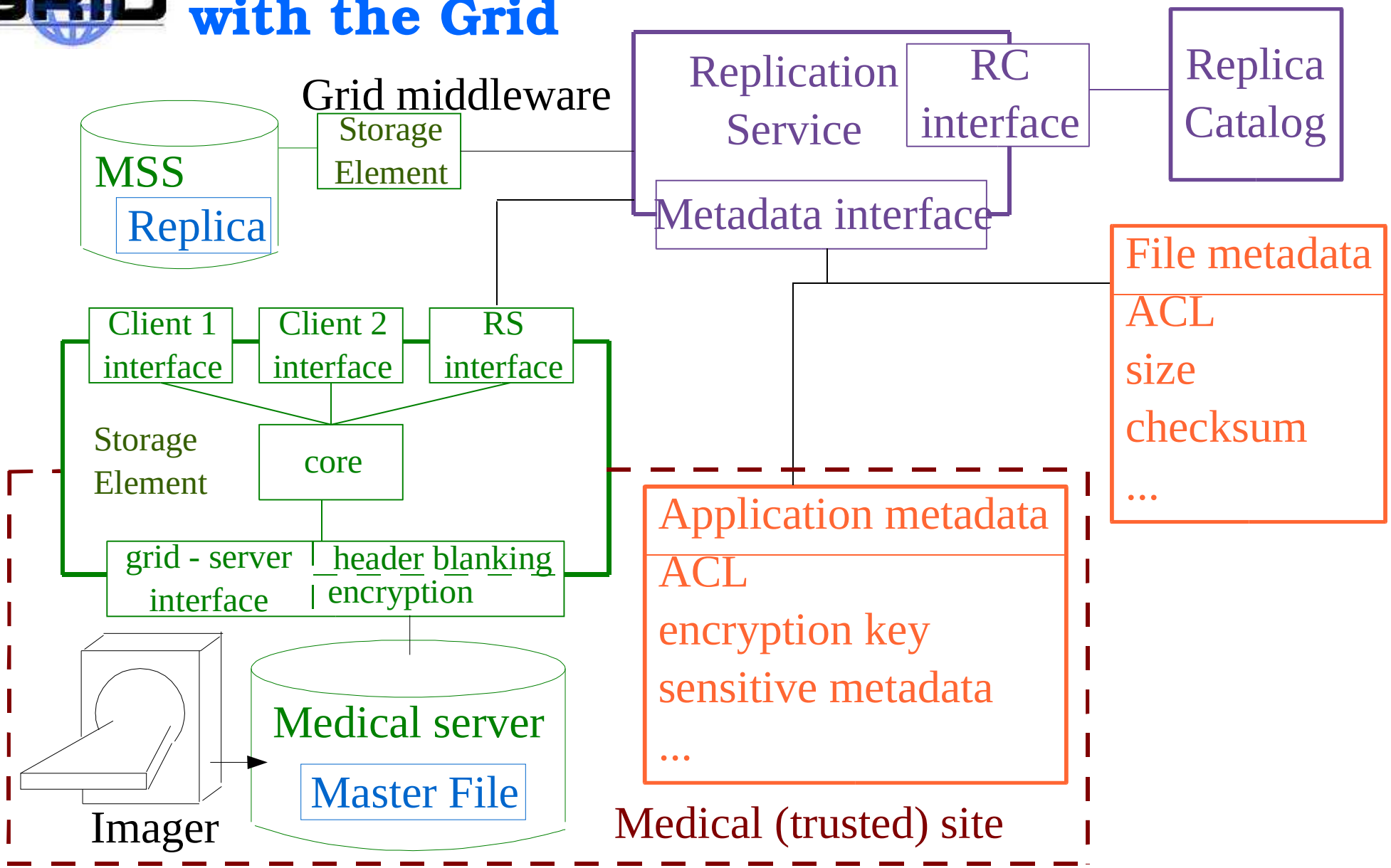
- Distributed data and distributed metadata
  - Metadata Distribution/Location Service (similar to GRID replication services for metadata)
  - Metadata and data should be synchronized (same lifetime, access authorization…)
  - Data traceability (How was data B produced? Which result was obtained from data A?)
- High level layer
  - Intelligent proxy hierarchy
  - Distributed dynamic indices for queries
  - Optimisation / caching of search requests

The future from biomedical applications point of view

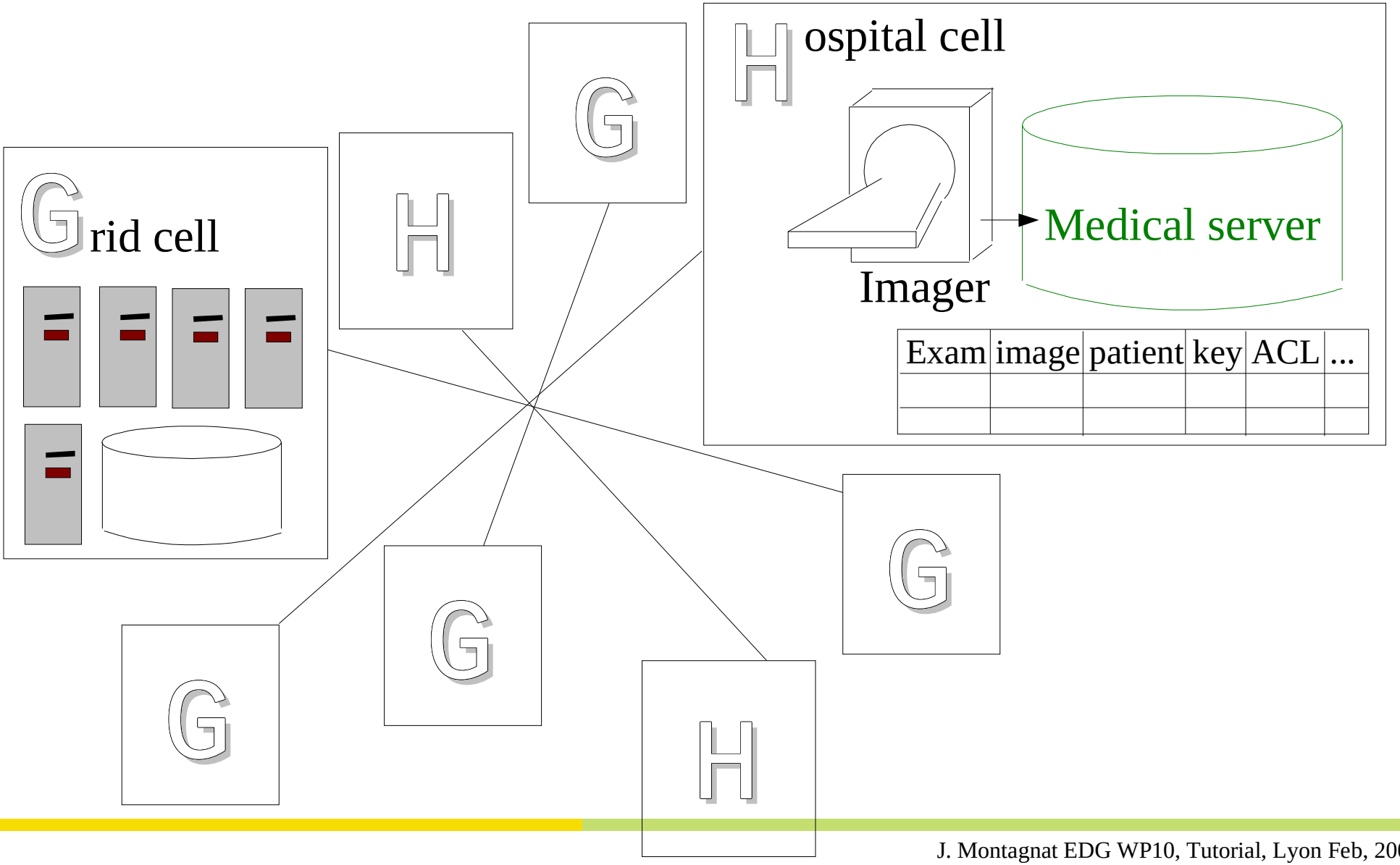


- **Distributed Algorithms**. New distributed "grid-aware" algorithms (c.f. this afternoon's demonstration).
- **Grid Service Portals**. Service providers taking advantage of the DataGrid computational power and storage capacity.
- **Cooperative Framework**. Use the DataGrid as a cooperative framework for sharing resources, algorithms, and organize experiments in a cooperative manner.

Future: Interfacing medical data with the Grid



Future: distributed medical data in a Grid environment



Algorithm repository

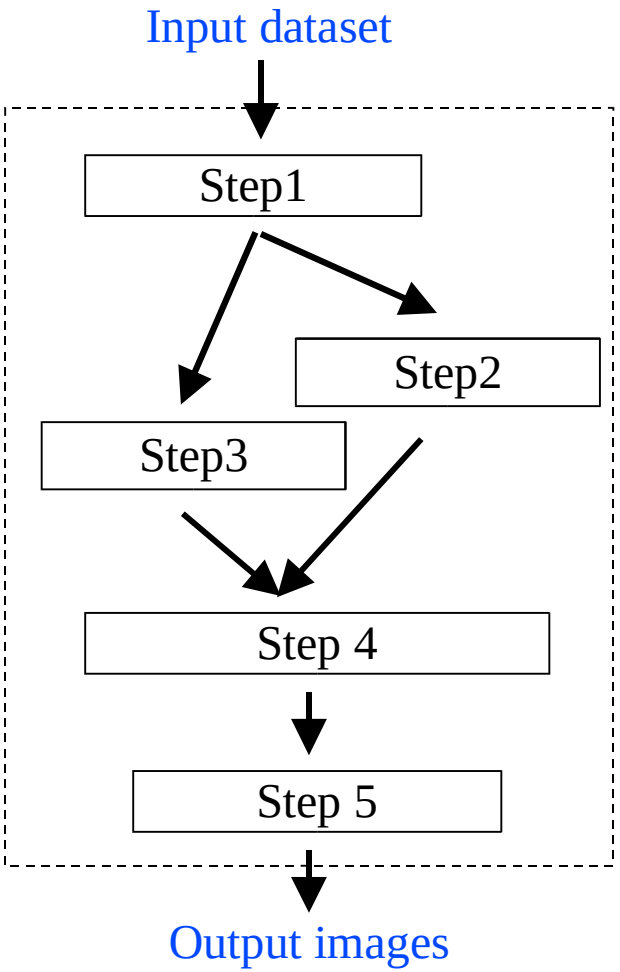
- Algorithms registration
- Versioning
- Formal description of algorithms input/output
  - Formal pipelines
- Tracking
  - Re-executions
  - Replace data by processing description (on-the-fly reconstruction)
- Data / algorithms synchronization
  - find processed data source
  - find processed data

GRID metadata manager

- Distributed metadata
  - The RM is a distributed file system for the GRID
  - Spitfire only deals with local data
  - A Metadata Location Service would be the counterpart of the RM for metadata (Should this involve replication? Not necessarily although it may be a good idea for scalability).
- Metadata / data synchronisation
  - Metadata are attached to data and have the same lifetime
  - This involves an 'LFN / metadata' table
- JSS / metadata synchronisation
  - A Job should be able to process a set of files corresponding to some metadata

Pipeline processing

- Pipeline issues
  - one pipeline for multiple inputs
  - load balancing / synchronization
  - failure / retrial paradigm, logging
  - dynamic extension of the processed image set
- Algorithm DB
- Log data transformations
- Rebuild transformed data from raw data



Technical requirements

Critical / mid-term / long-term

- 1. Large user community
  - anonymous/group login
- 2. Data management
  - data updates and data versioning
- 3. Security
  - disk / network encryption
- 4. Limited response time
  - fast queues
- 5. High priority jobs
  - privileged users
- 6. Interactivity
  - communication between user interface and CE's
- 7. Parallelization
  - MPI site-wide / grid-wide
- 8. Pipeline processing
  - pipeline description language / scheduling

Conclusions

- Increased computing power
  - Large scale applications (genome data mining, epidemiological studies…)
  - Fast response time (ambulance jobs)
- Distributed computing and data storage
  - Data replication and versioning
  - Large data sets processing
- Collaborative environment
  - Sharing resources
  - Sharing data
  - Sharing algorithms
  - Security issues