Applicazione delle tecnologie ICT al settore Life Sciences: casi di studio e dimostrazioni di utilizzo delle soluzioni Nice per l'analisi e gestione dei dati generati dalle analisi biotecnologiche

Livia Torterolo, PhD

livia.torterolo@nice-software.com

Outline

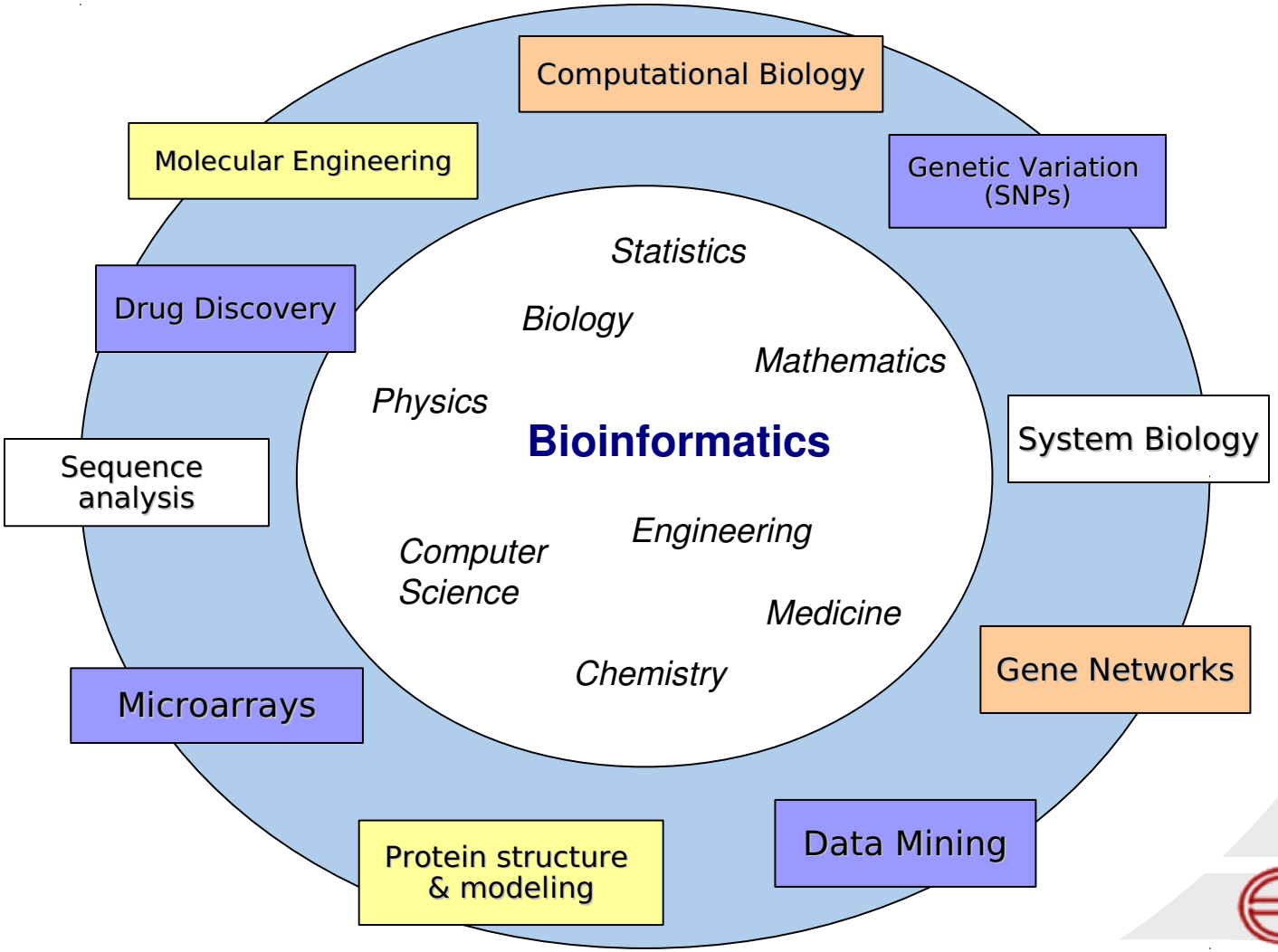
NICE

NICE Products&Solutions

Life Sciences Solutions

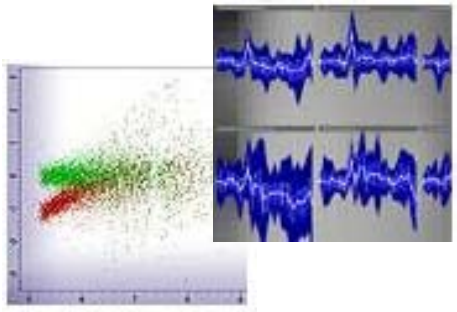
- Case studies
- Demonstrations

Scenario


Computational Biology
Molecular Engineering
Genetic Variation (SNPs)
Drug Discovery
Statistics
Biology
Mathematics
Physics
Bioinformatics
System Biology
Sequence analysis
Engineering
Computer Science
Medicine
Gene Networks
Chemistry
Microarrays
Data Mining
Protein structure & modeling

Challenges

- **Data Analysis**
  - Analyze large datasets
  - Automate procedures – reproducibility



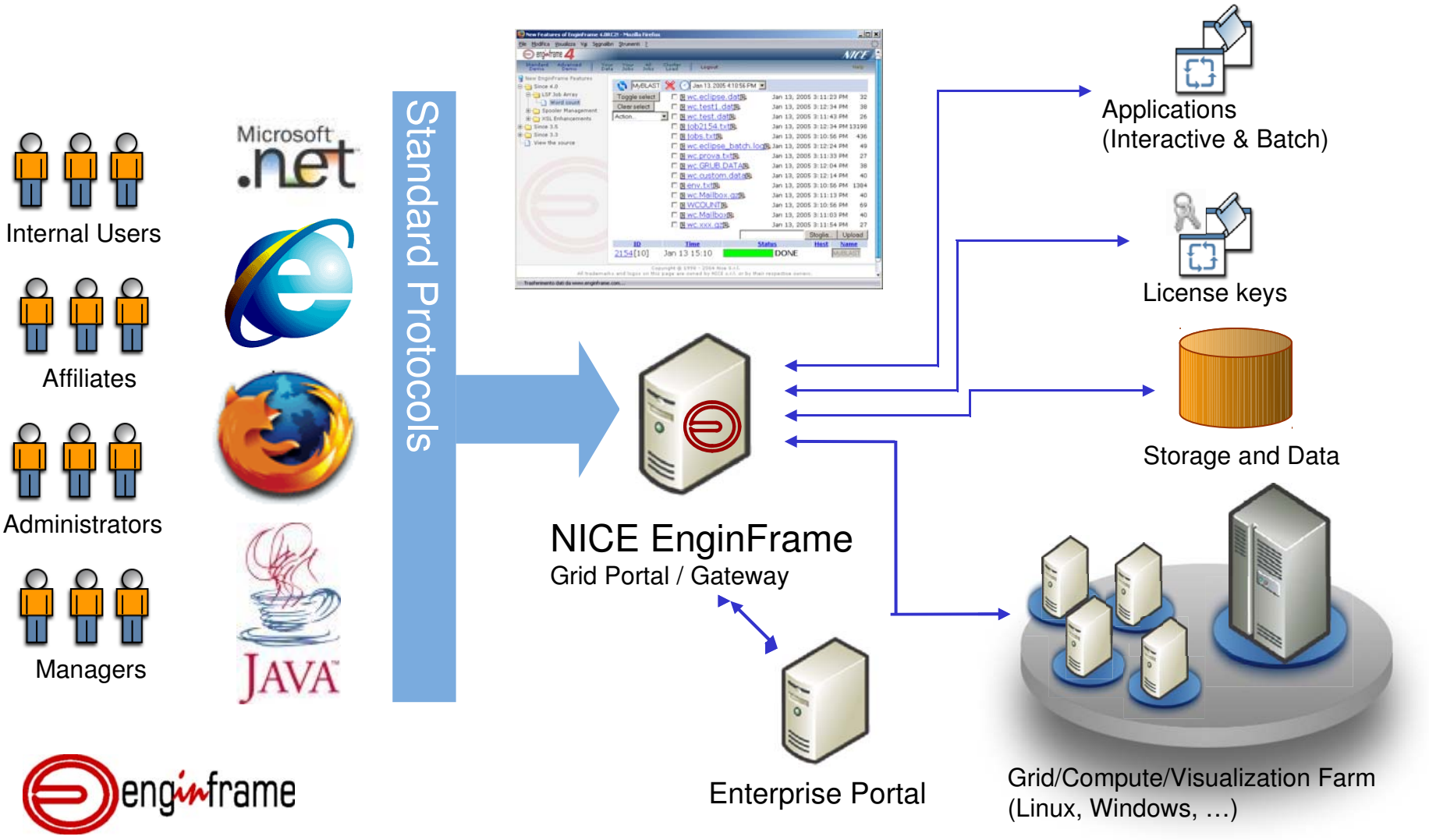
- **Data/Metadata Management**
  - Archive big amount of data/metadata
  - Access and manage distributed data

- **Integration**
  - Integrate heterogeneous tools and data

- **Remote Visualization**
  - Visualize & share data & interactive app remotely

- **Accessibility & Usability**
  - Technology for everyone

NICE EnginFrame Architecture



How EnginFrame can be customized for LS...?

… by using/developing Life Sciences plugins!

- BioInformatics tools (batch)
  - R / BioConductor support
  - Affymetrix Power Tools (parallel)
  - Dchip (Cheng Li Lab) (parallel)
  - Plink commands
  - Amber, AutoDock, BLAST family, BLAT, Charmm, ClustalW, Glue, Gromacs, MrBayes, Namd, Philyp, RaXML, Vienna RNA, X! Tandem, etc.
- Molecular modeling (interactive)
  - Accelerys DS Visualizer
  - MOE
  - Schrodinger/Glide
- Medical Imaging tools (batch&interactive)
  - Statistical Parametric Mapping
  - SeegViewer (3D Viewer)
- Matlab (batch&interactive)
- Workflow tools
  - SOAP/web-service enable
  - Taverna, Moteur
  - A-ware
- Data/metadata mngt framework
  - Irods, gLite

… applications can be provided as services (SaaS)

Outline

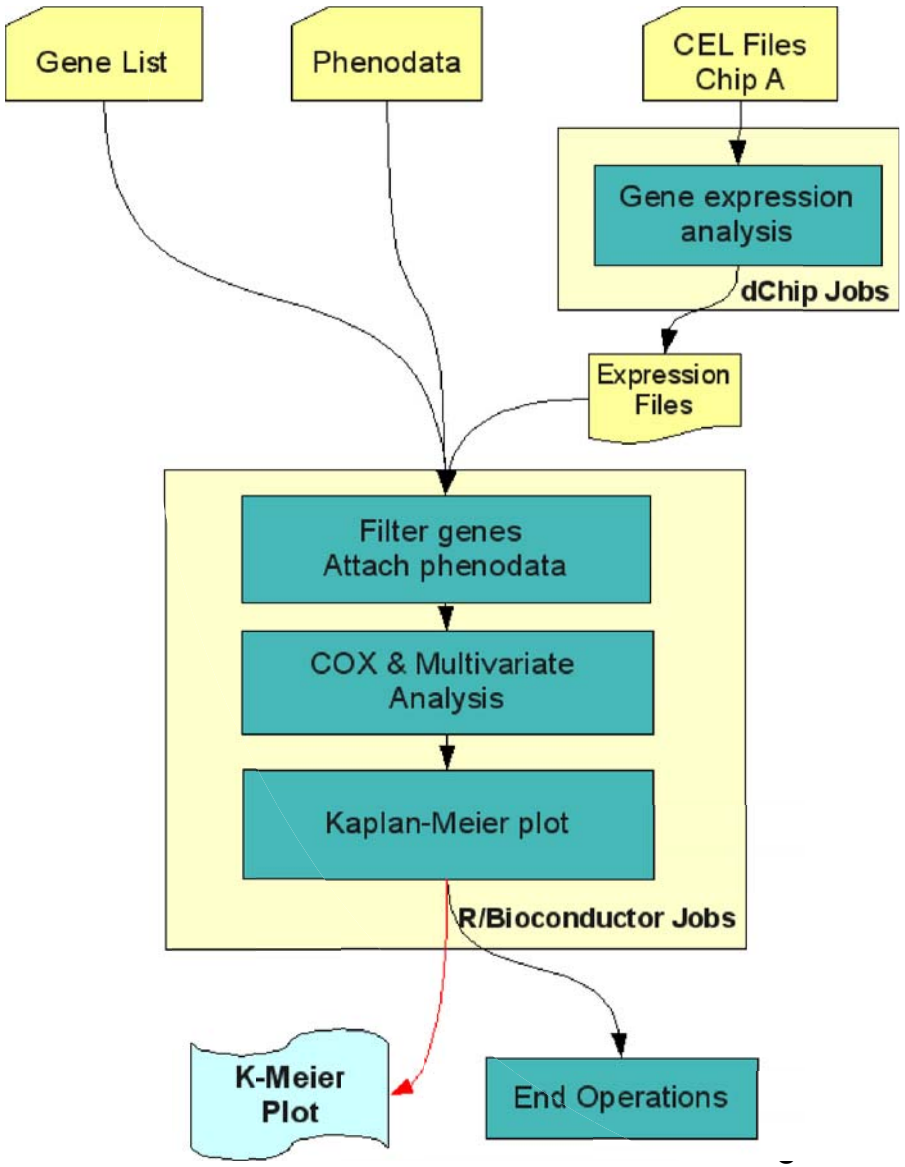
NICE

NICE Products&Solutions

Life Sciences Solutions

- Case studies
- Demonstrations

Case study: Cancer Survival Studies (1)



In collaboration with:

biolab

IST

Aim:

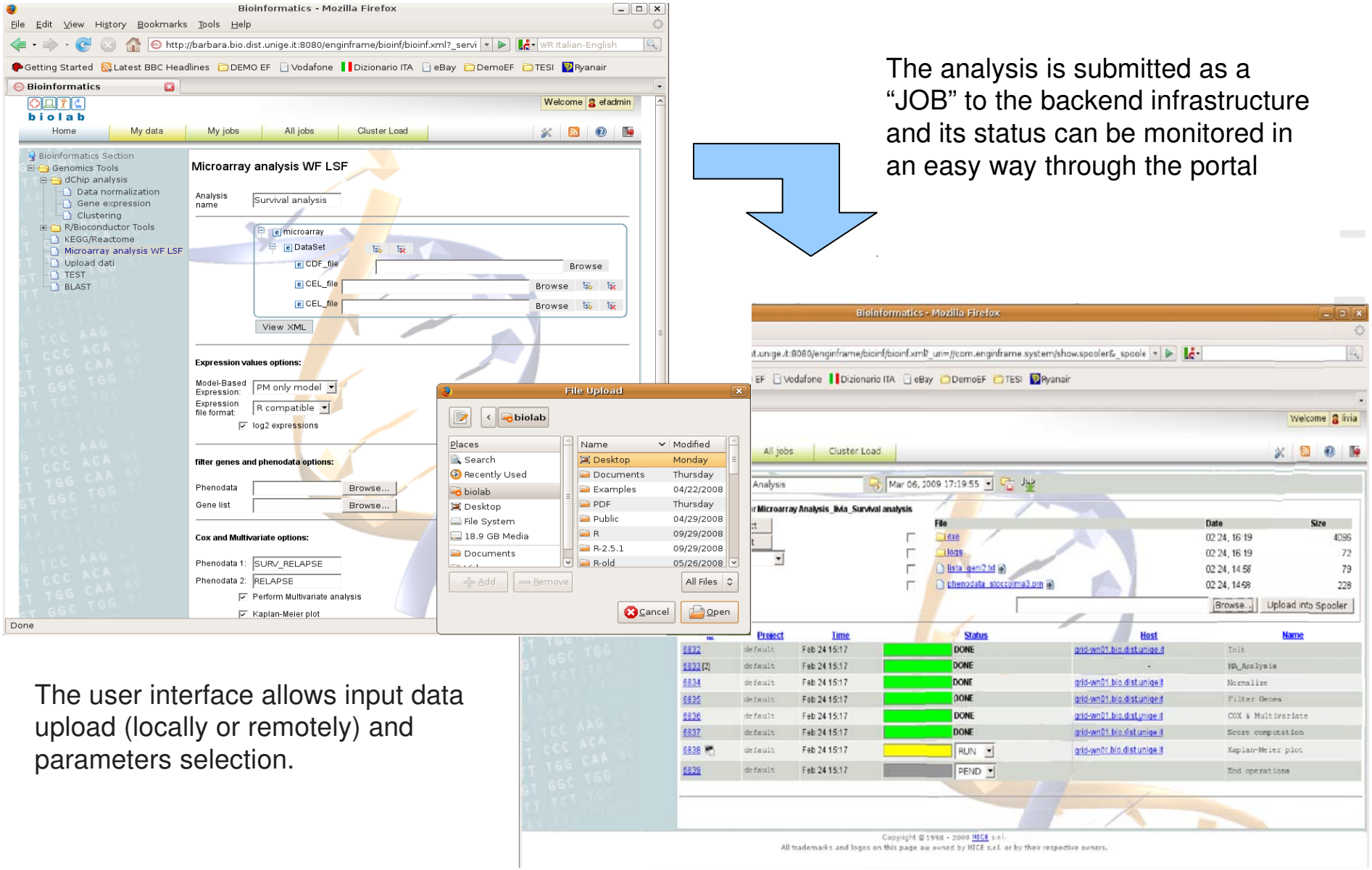
Develop a service for automating functional genomics analyses used in breast cancer survival studies and prognosis assessment.

It allows to study the correlation of clinico-pathological and follow-up data with gene expression both for the definition of signatures and to test hypotheses on the effects of over/under expression of single genes.

The service can be used by bio-medical researchers without specific computation skills to validate potential biomarkers or multi-gene classifiers.

engineframe

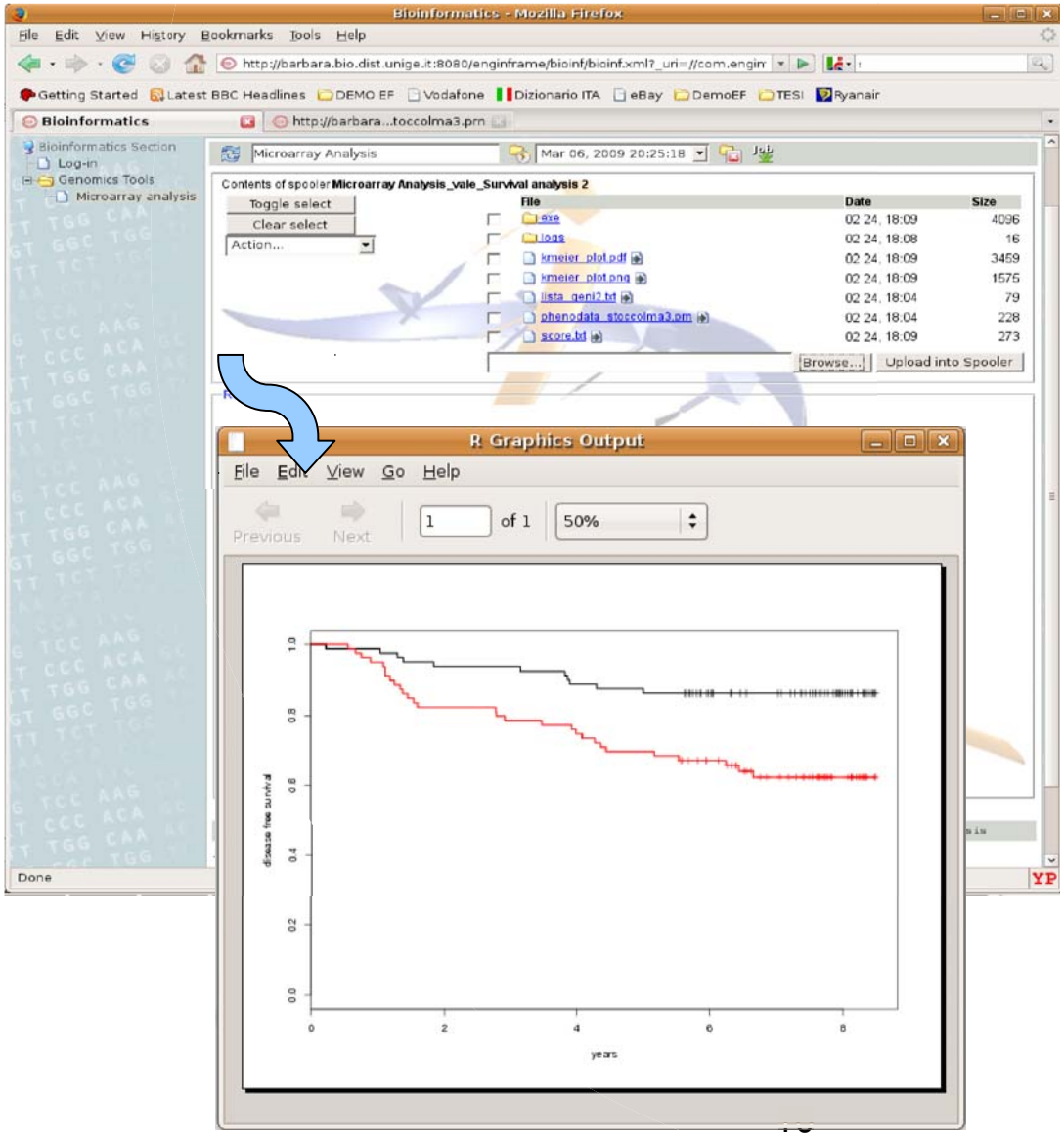
Case study: Cancer Survival Studies (2)



The analysis is submitted as a "JOB" to the backend infrastructure and its status can be monitored in an easy way through the portal

The user interface allows input data upload (locally or remotely) and parameters selection.

Case study: Cancer Survival Studies (3)



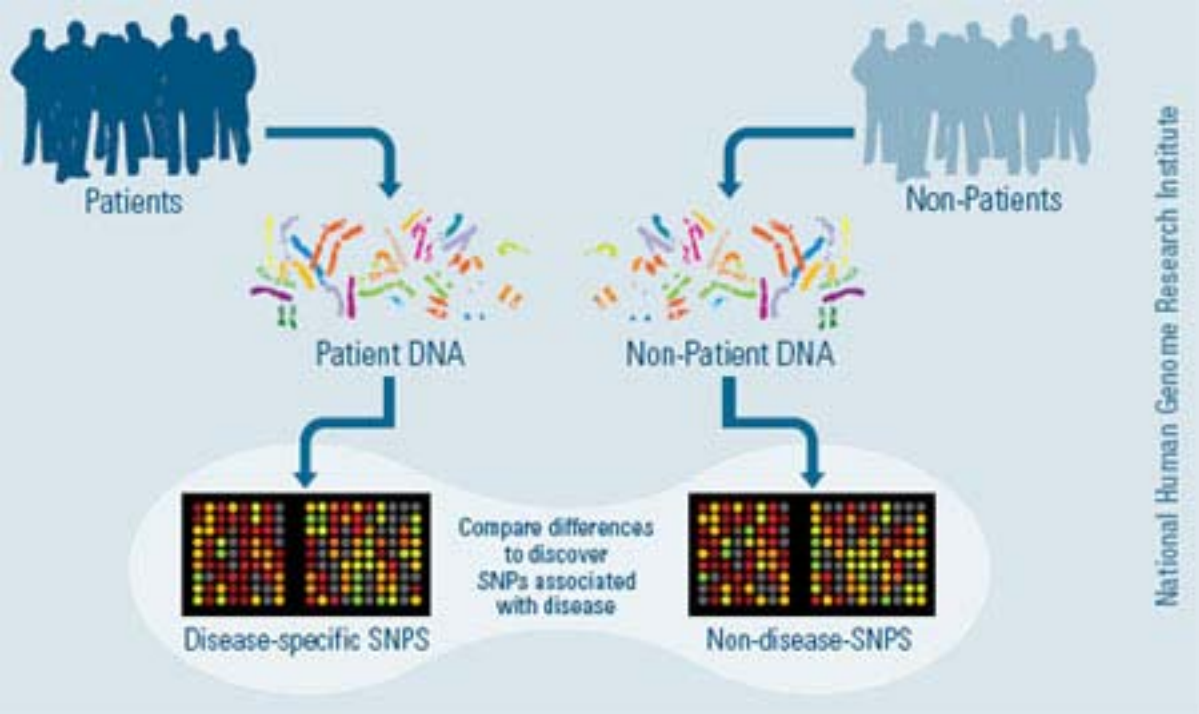
The analysis results (Kaplan-Meier plots) are shown in the portal area.
The plot shows the two groups of patients, classified on the basis of low or high risk of disease relapse.

Advantages obtained:

- Integration of different software tools (dChip, R/Bioc) and data (gene expression, phenodata)
- Automation of procedures → time & cost reduction
- Computational power
- Experiment reproducibility
- Usability

enginframe

Case study: Genome-Wide Association Studies (1)



In collaboration with:

biolab

bmi
BIO-MEDICAL INFORMATICS

Researchers analyze the DNA of two groups of participants: people with the disease (**Case**) being studied and similar people without the disease (**Control**). Each person's complete set of DNA is placed on tiny chips and scanned on automated laboratory machines, which quickly survey each participant's genome for strategically selected markers of **genetic variation**, which are called single nucleotide polymorphisms, or **SNPs**.

enginframe

Case study: Genome-Wide Association Studies (2)

Quality Control of Genome-Wide Association Studies

BIO-MEDICAL INFORMATICS

Aim: Define strategies for setting appropriate genotyping rate thresholds for GWAS quality control based on decision theory.

Approach:

Maximize these parameters = maximize the power of the study

100 genotyping rate filters individual/SNP (90%->99%)

Genotyping rate thresholds

CRITERIA	ALTERNATIVES			
	CR90,SNP90	CR90,SNP91	...	CR99,SNP99
% of Individuals that pass the genotyping filters	99.45	99.24	...	60.66
% of SNPs that pass the genotyping filters	99.68	99.46	...	87.73
% of SNPs with differences in missing rates c/c (p<0.05)	12.88	12.83	...	93.49
% of SNPs with MAF < 1%	0.11	0.11	...	0.11
% of SNPs deviating from HWE (p<0.001)	0.29	0.27	...	0.11
Heterozigosity Rate Standard Deviation (per individual)	0.01	0.01	...	0.00
Genomic Inflation Factor	1.31	1.31	...	1.18

Alternatives: different combinations of Ind and SNP genotyping rate (Campioni>90%,SNPs>90%); (Campioni>90%,SNPs>91%)...

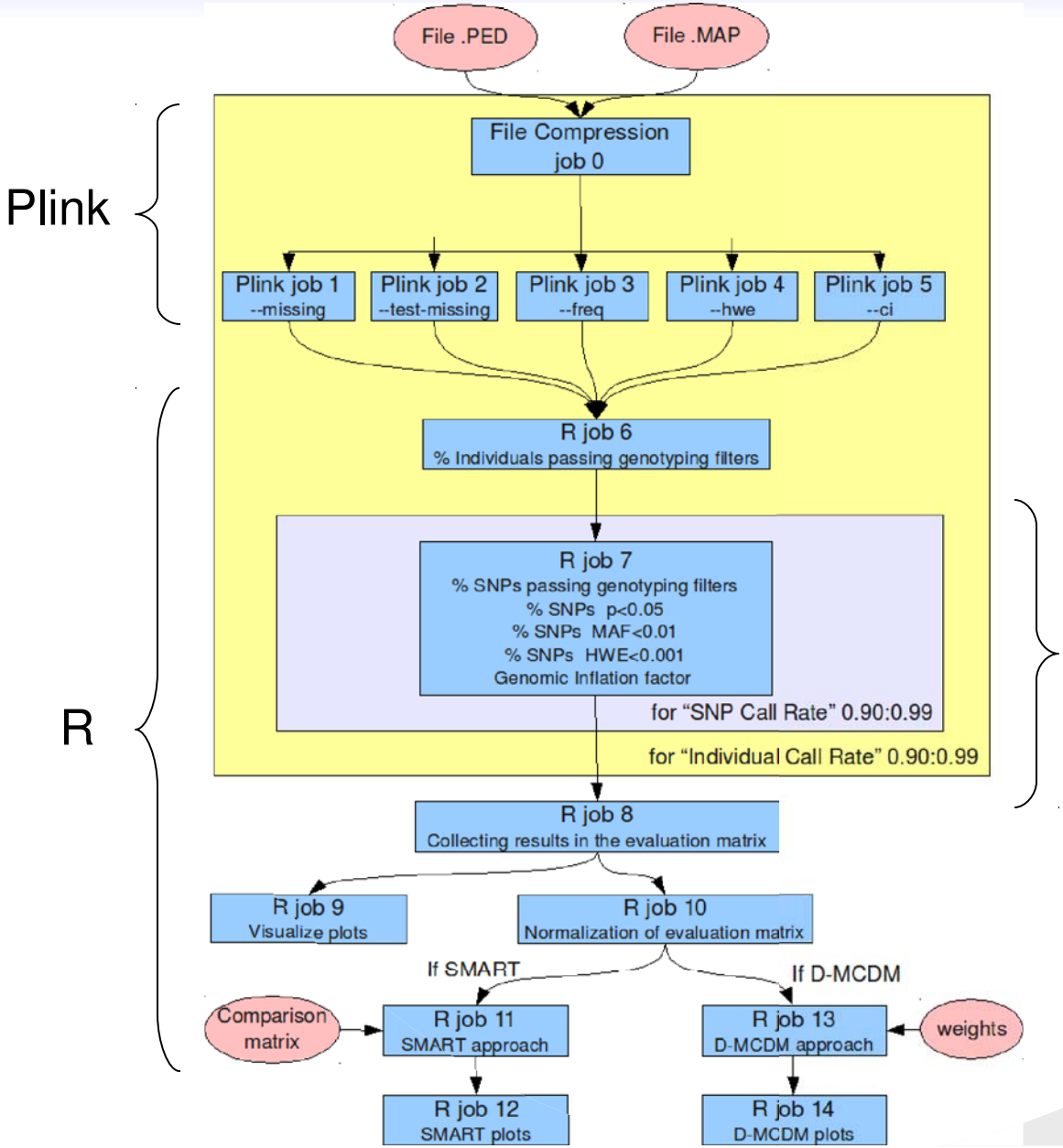
Decision Theory model

Score profile for alternatives

Minimize these parameters = minimize genotyping errors

engineframe

Case study: Genome-Wide Association Studies (3)

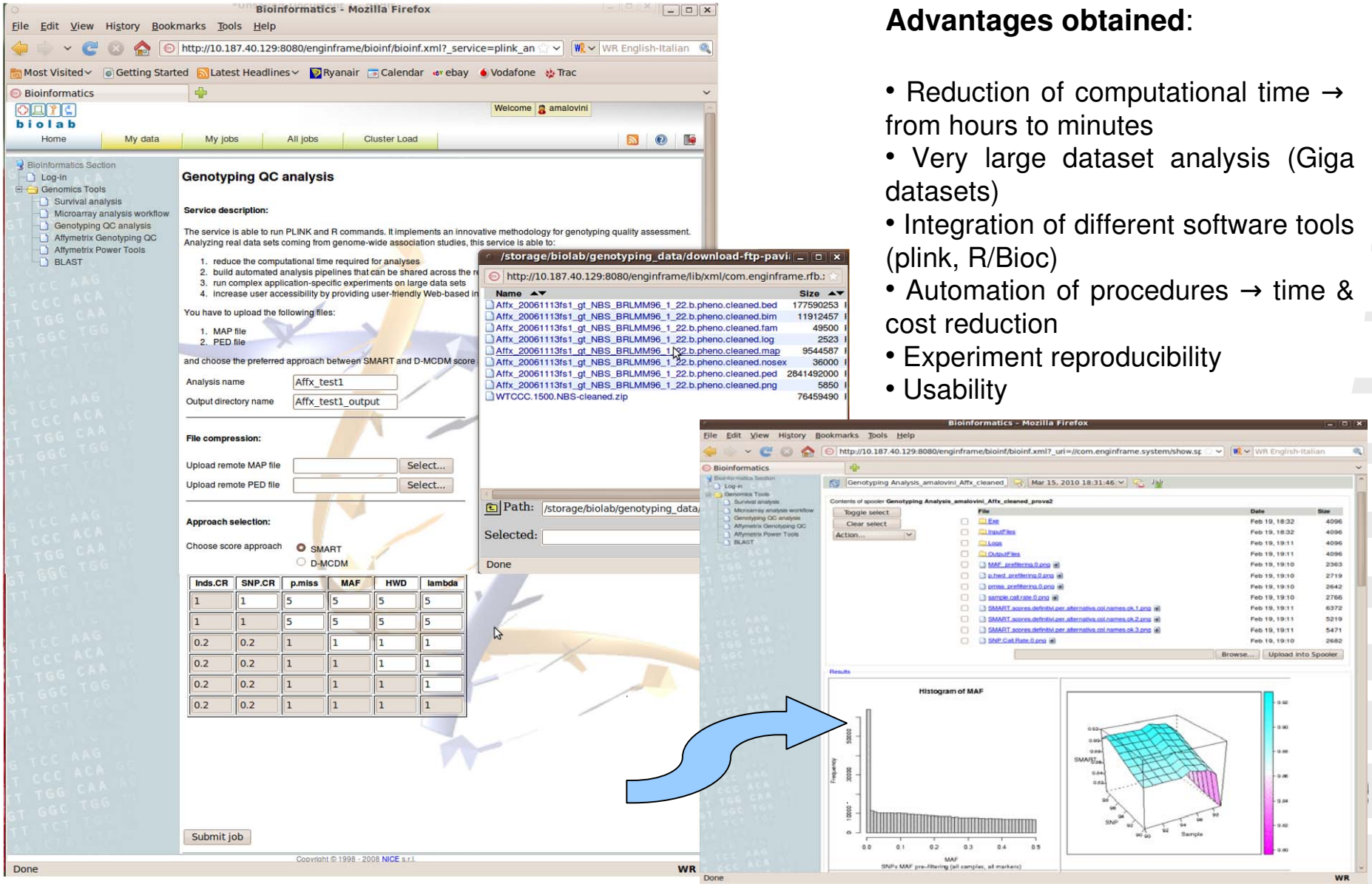


Design the analysis pipeline and study the best computational model to optimize computational effort

… this service submits more than 100 independent jobs, one for any alternative...

100 alternatives

Case study: Genome-Wide Association Studies (4)



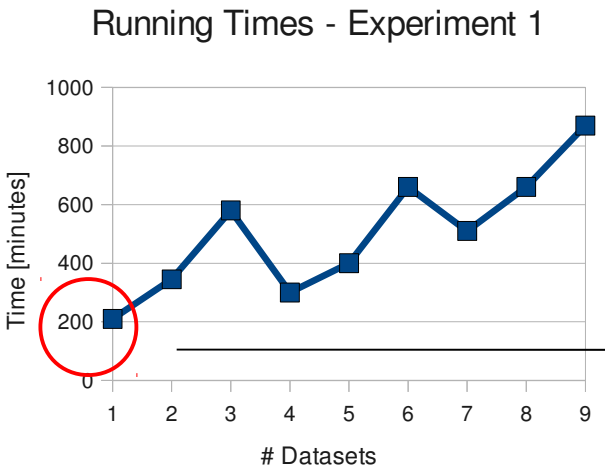
Advantages obtained:

- Reduction of computational time → from hours to minutes
- Very large dataset analysis (Giga datasets)
- Integration of different software tools (plink, R/Bioc)
- Automation of procedures → time & cost reduction
- Experiment reproducibility
- Usability

Case study: Genome-Wide Association Studies (5)

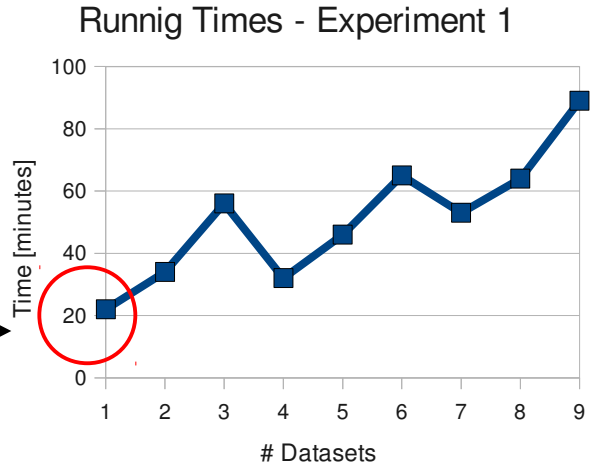
Computational tests

Execution on PC/Workstation



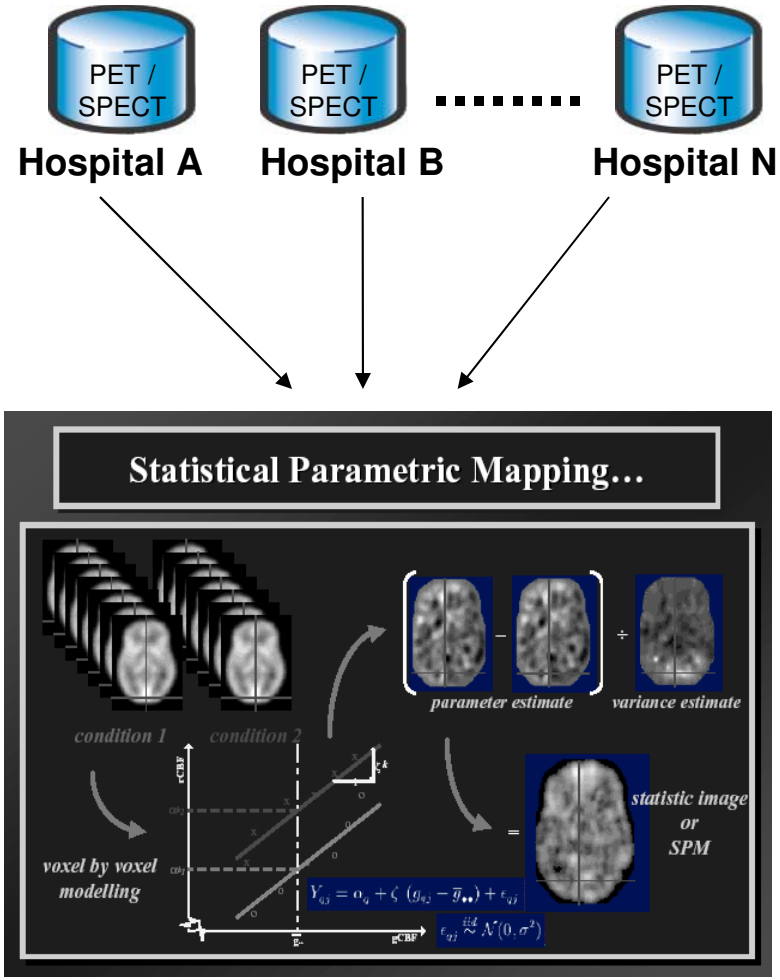
Time reduction!

Execution on HPC/Grid



Dataset 1	cases	controls	SNPs	time PC	time Grid (minuti)
sim-500-samples-300k	500	500	300k	210	22
sim-500-samples-500k	500	500	500k	345	34
sim-500-samples-1000k	500	500	1000k	580	56
sim-1000-samples-300k	1000	1000	300k	300	32
sim-1000-samples-500k	1000	1000	500k	400	46
sim-1000-samples-1000k	1000	1000	1000k	660	65
sim-2000-samples-300k	2000	2000	300k	510	53
sim-2000-samples-500k	2000	2000	500k	660	64
sim-2000-samples-1000k	2000	2000	1000k	870	89

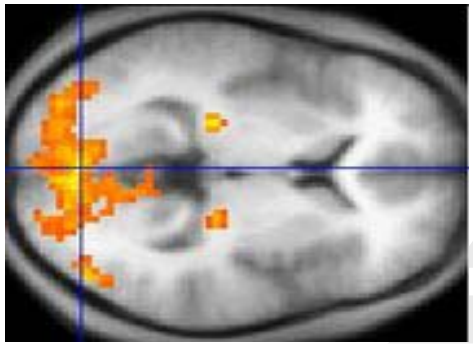
Case study: Data management, Analysis and Visualization in Neuroinformatics (1)



Aim: perform the early diagnosis of Alzheimer through a statistical comparison of PET/SPECT patient image with a database of reference patient images.

Advantages obtained: remote analysis of distributed images

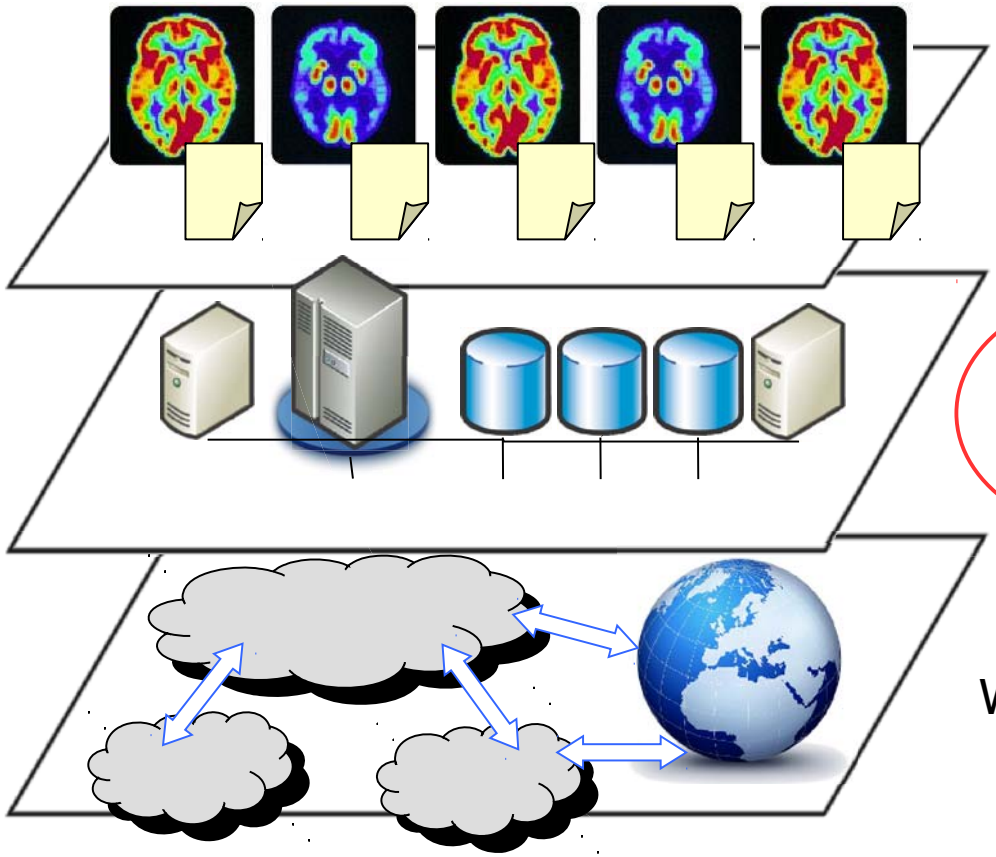
Results



Patient image where suspected brain areas are visualized

Case study: Data management, Analysis and Visualization in Neuroinformatics (2)

How to access to distributed images? How to manage metadata?



Neuro Data/Images

Patient/Data information

Data/Metadata Management Framework

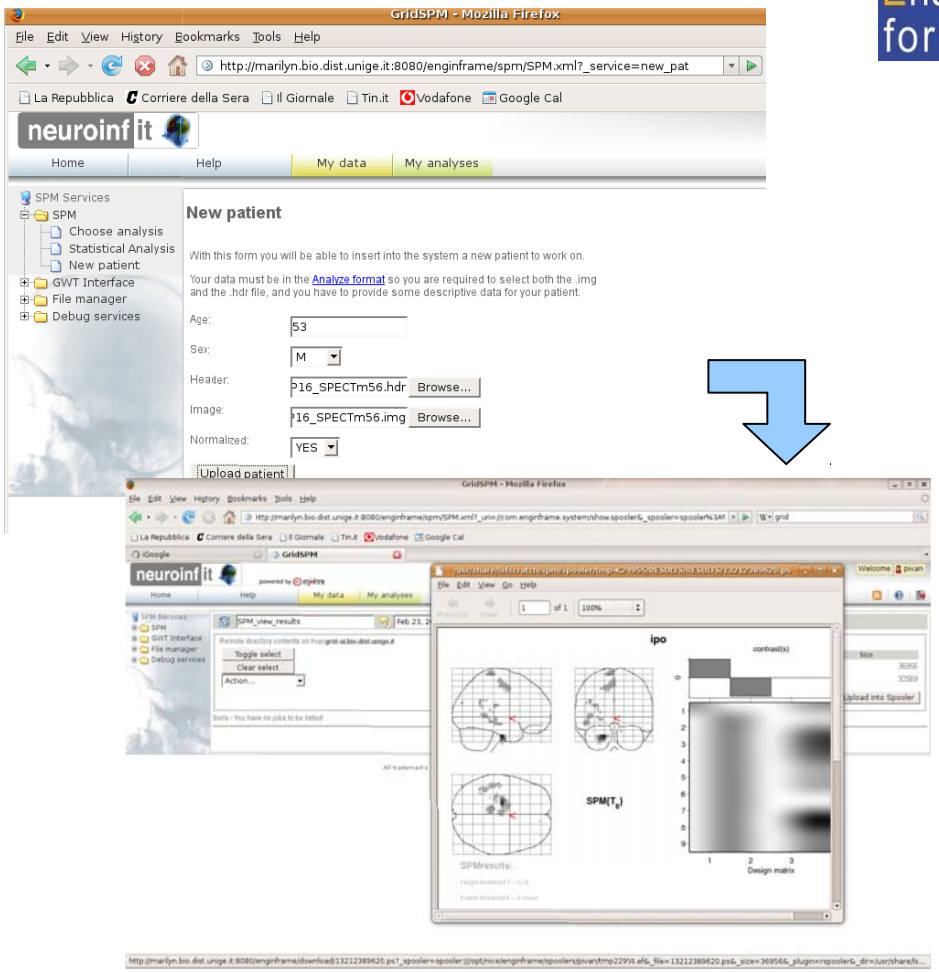
gLite

iRods

Wide Area Network

Case study: Data management, Analysis and Visualization in Neuroinformatics (3)

gLite implementation



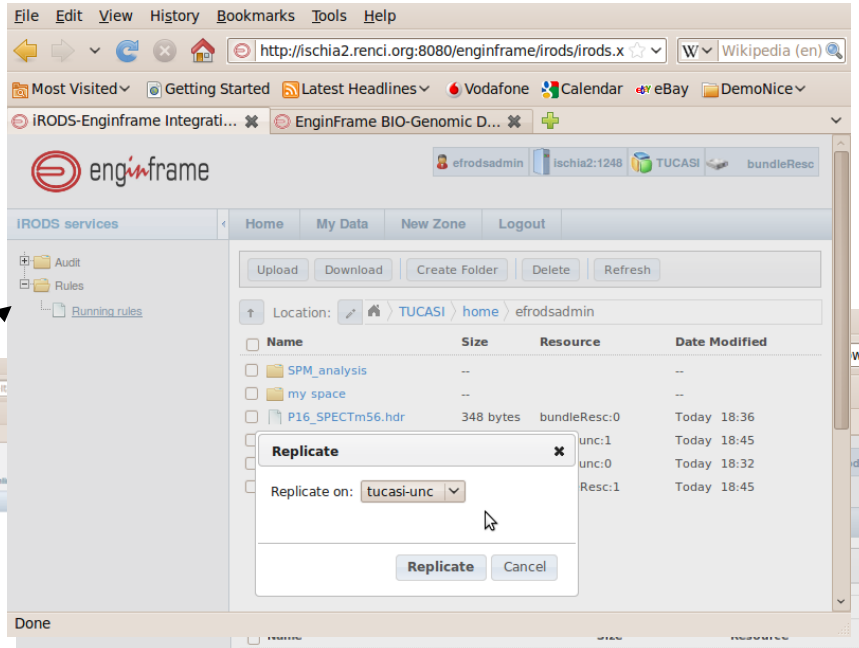
Aim: design, implement, and validate a **service** for the computer-aided extraction of diagnostic markers for Alzheimer's disease and schizophrenia from medical images on a GRID-based e-Infrastructure.

NICE is involved in Front-end development and services provision by integrating EnginFrame with Liferay technology.

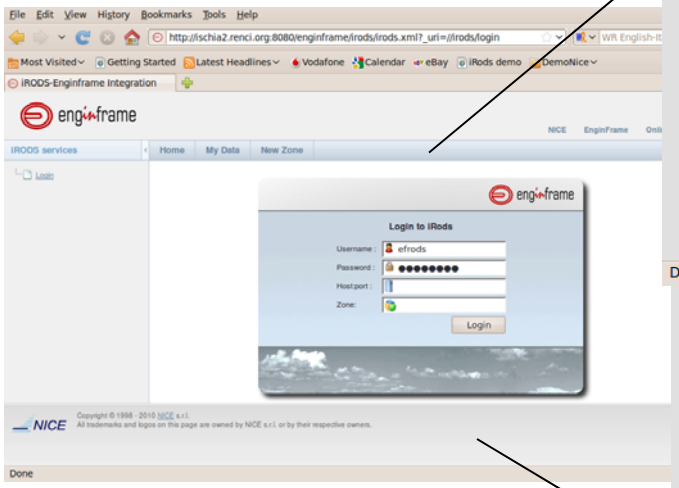
Case study: Data management, Analysis and Visualization in Neuroinformatics (4)

iRODS implementation

Solution for distributed data management and organization of data into sharable logical collections

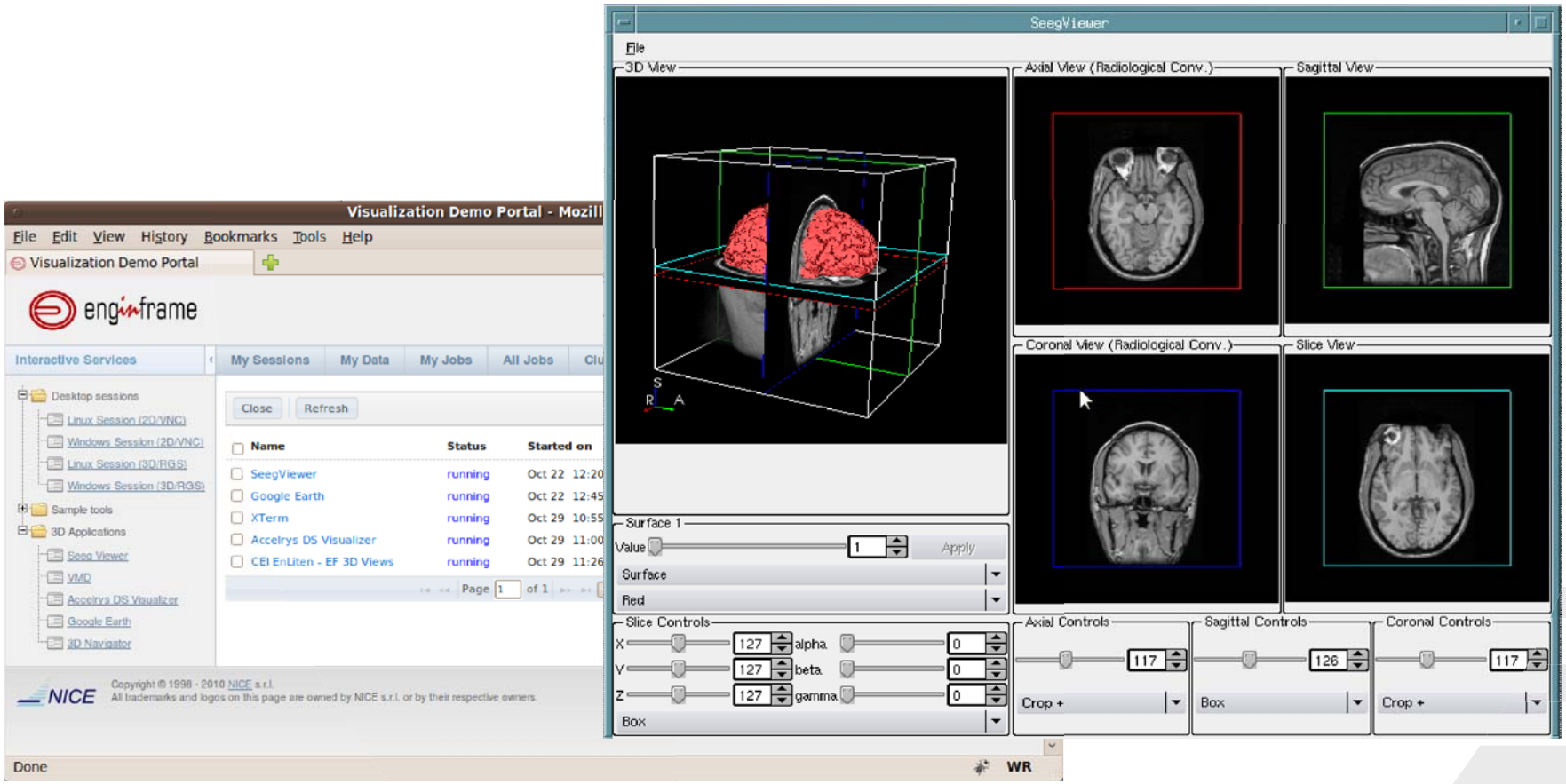


Replication of data on different physical resources



Visualization & setting of metadata

Case study: Data management, Analysis and Visualization in Neuroinformatics (5)



Advantages obtained: access to interactive 2D/3D applications remotely, in a collaborative environment, through a common web browser

Outline

NICE

NICE Products&Solutions

Life Sciences Solutions

Case studies

Demonstrations