Searching for Cancer Biomarkers and Investigating Neural Dynamic Behavior with Globus

Dr. Ioan L. Muntean,

Advanced Computing Application Lab

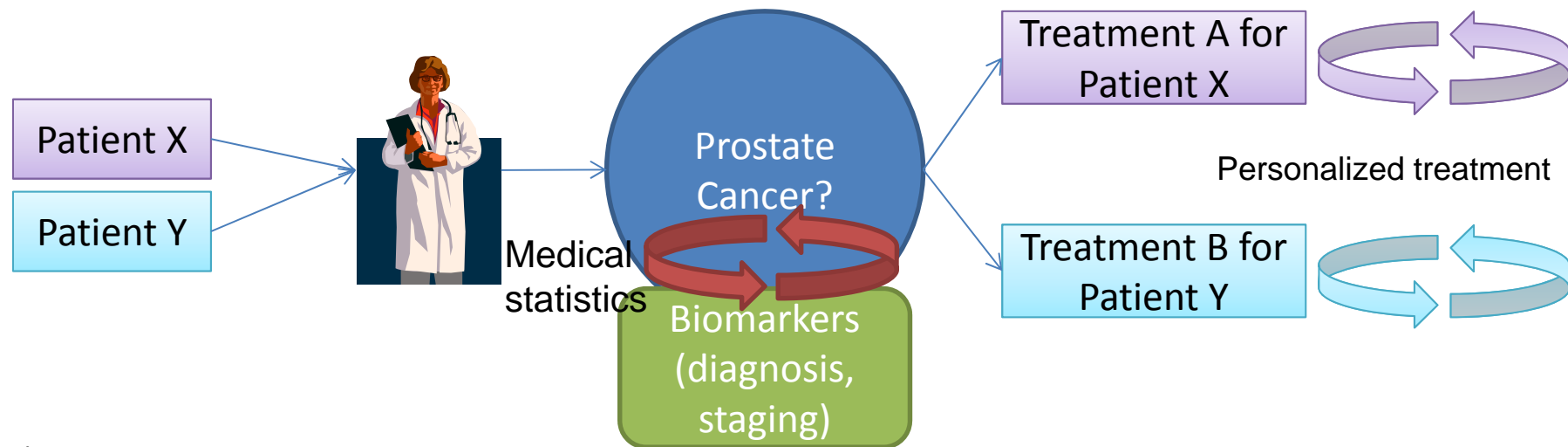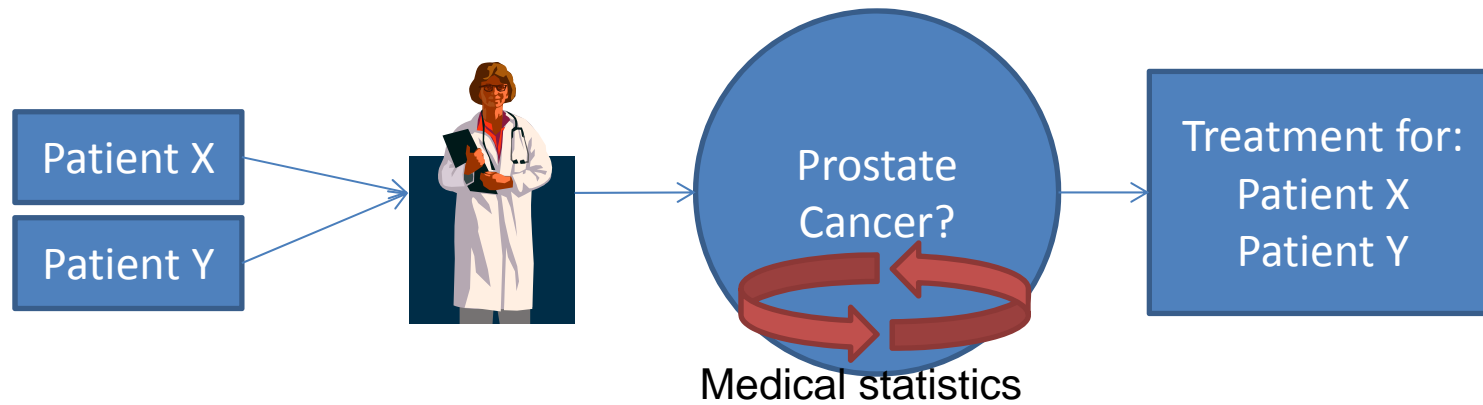Computer Science Department

UTC-N, Cluj-Napoca

Globus World 2013

Chicago, USA

Scenario One

IN-SILICO CANCER BIOMARKER DISCOVERY AND VALIDATION

Motivation: Personalized Medicine – a Vision

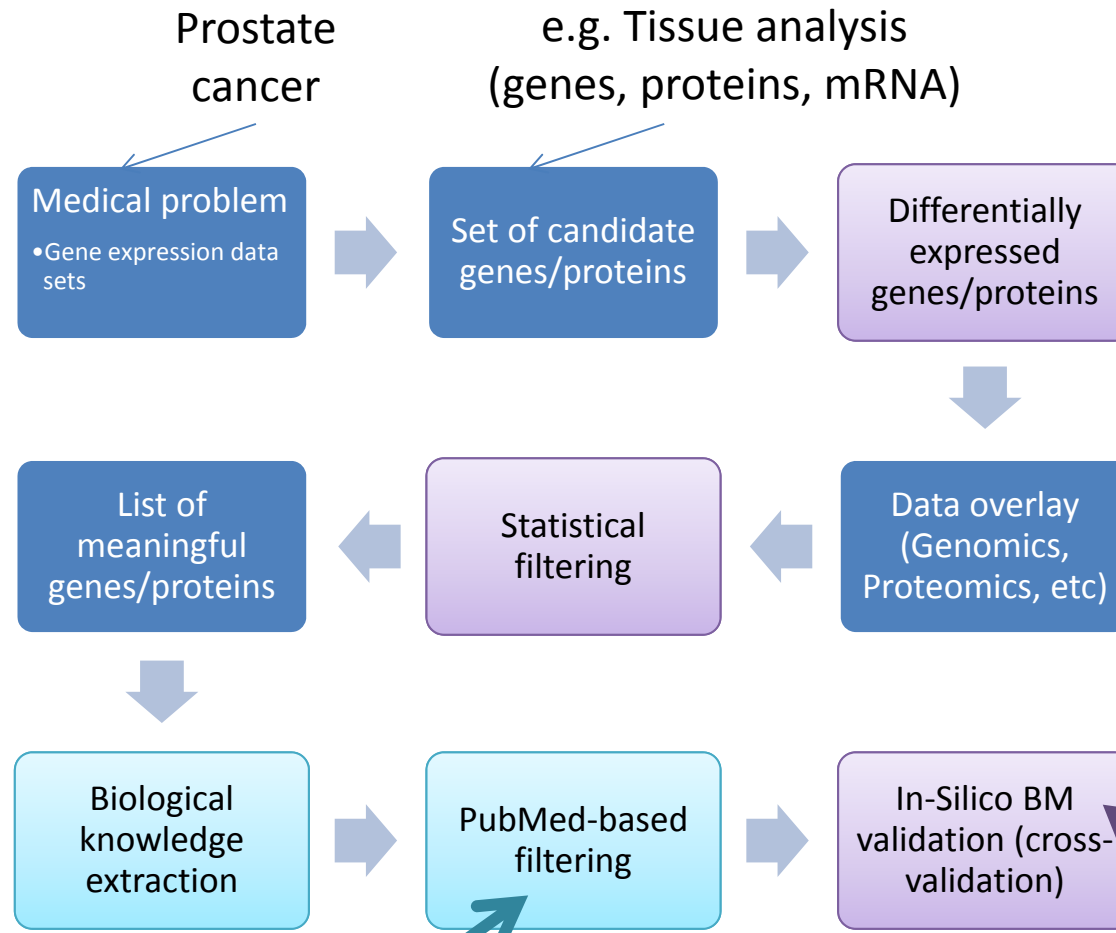# Motivation: Biomarker Discovery and Validation

- Biomarker – indicator of normal biological or pathogenic processes or pharmacological responses to a therapeutic intervention

- Biomarker discovery and validation (BMVD) process
  - Typically implies large clinical trials (expensive, time consuming)
  - Could be greatly improved by analyzing and correlating findings from multi-omics areas (proteomics, genomics, transcriptomics etc.)

- Challenges of multi-omics data overlay
  - Biological understanding of 'omics' data
  - Processing of large amount of experiments
  - **Complexity of the integration**: significant differences in experimental data/data storage formats, data collection and access policies, scientific review of collected data, interfaces for queries

> **Biomarkers (BM)**
> - different genes expression
> - altered or mutated genes
> - miRNA, transcription factors
> - RNA, proteins, lipids etc.
> Correlated with biological aspects or a clinical outcome

The In-Silico BMDV Process

Prostate cancer

e.g. Tissue analysis (genes, proteins, mRNA)

Medical problem
• Gene expression data sets

→

Set of candidate genes/proteins

→

Differentially expressed genes/proteins

↓

List of meaningful genes/proteins

←

Statistical filtering

←

Data overlay (Genomics, Proteomics, etc)

↓

Biological knowledge extraction

→

PubMed-based filtering

→

In-Silico BM validation (cross-validation)

BMDV process
- Inherently **distributed**
- High degree of **parallel and distributed processing**

Examples
- Search in multiple 'omics' sources
- Parallel search of the gene/protein list
- Parallel statistics
- etc.

Parallel processing

Distributed processing

R. Suharoschi et al. *A proposed curation protocol for discovery of cancer potential biomarker candidates*, AAPS, (81-82),2012

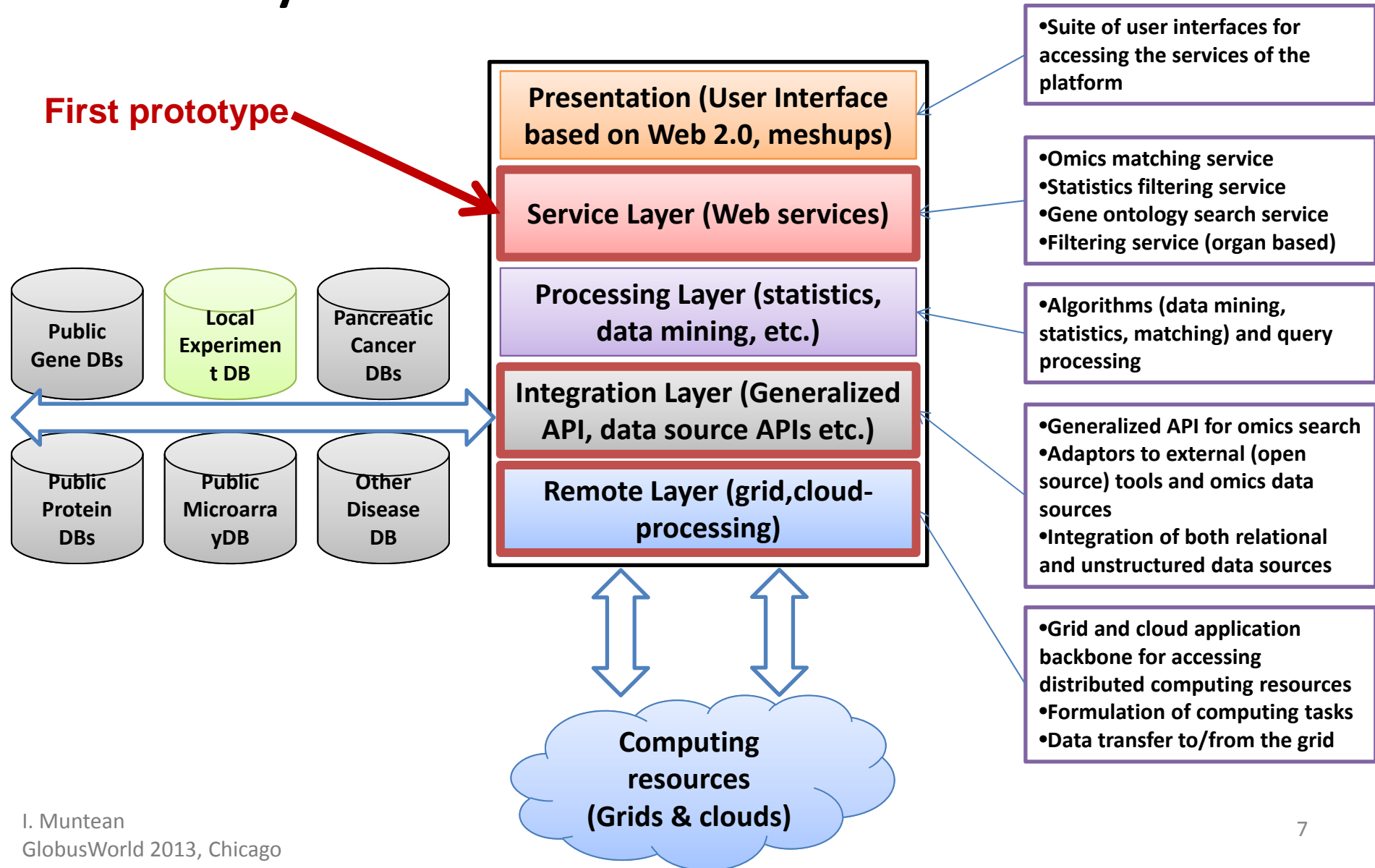I. Muntean
GlobusWorld 2013, Chicago

Integration Realm

- GenBank (NCBI) – sequences and annotations, uniquely identified by the accession number
  - Very powerful search interface
  - Data access: FTP, Web service calls, Entrez suite etc.

- ArrayExpress (EBI) – holds raw and annotated microarray experimental data (MIAME reqs)
  - Query interface: Experiments, Protocols and Array designs (by their various attributes, such as species, authors or array platforms)
  - Data access: REST-style queries, JSON web service format etc.

- UniProt – protein sequence and functional annotations
 - Queries using REST services, UniProtJAPI (Java API) etc.

Choice of Technology
- REST web services
- JAXB and Data Transfer Objects
- Application server (Glass Fish) and cloud hosting (Google App Engine)

Multi-Layer Architecture of BioGenProtOMICS

First prototype

Presentation (User Interface based on Web 2.0, meshups)

Service Layer (Web services)

Processing Layer (statistics, data mining, etc.)

Integration Layer (Generalized API, data source APIs etc.)

Remote Layer (grid,cloud-processing)

Public Gene DBs

Local Experiment DB

Pancreatic Cancer DBs

Public Protein DBs

Public MicroarrayDB

Other Disease DB

Computing resources (Grids & clouds)

- Suite of user interfaces for accessing the services of the platform

- Omics matching service
- Statistics filtering service
- Gene ontology search service
- Filtering service (organ based)

- Algorithms (data mining, statistics, matching) and query processing

- Generalized API for omics search
- Adaptors to external (open source) tools and omics data sources
- Integration of both relational and unstructured data sources

- Grid and cloud application backbone for accessing distributed computing resources
- Formulation of computing tasks
- Data transfer to/from the grid

PubMed Filtering Service

- Mines abstracts published medical libraries for evaluating the relevance of biomarker candidates for a given pathology
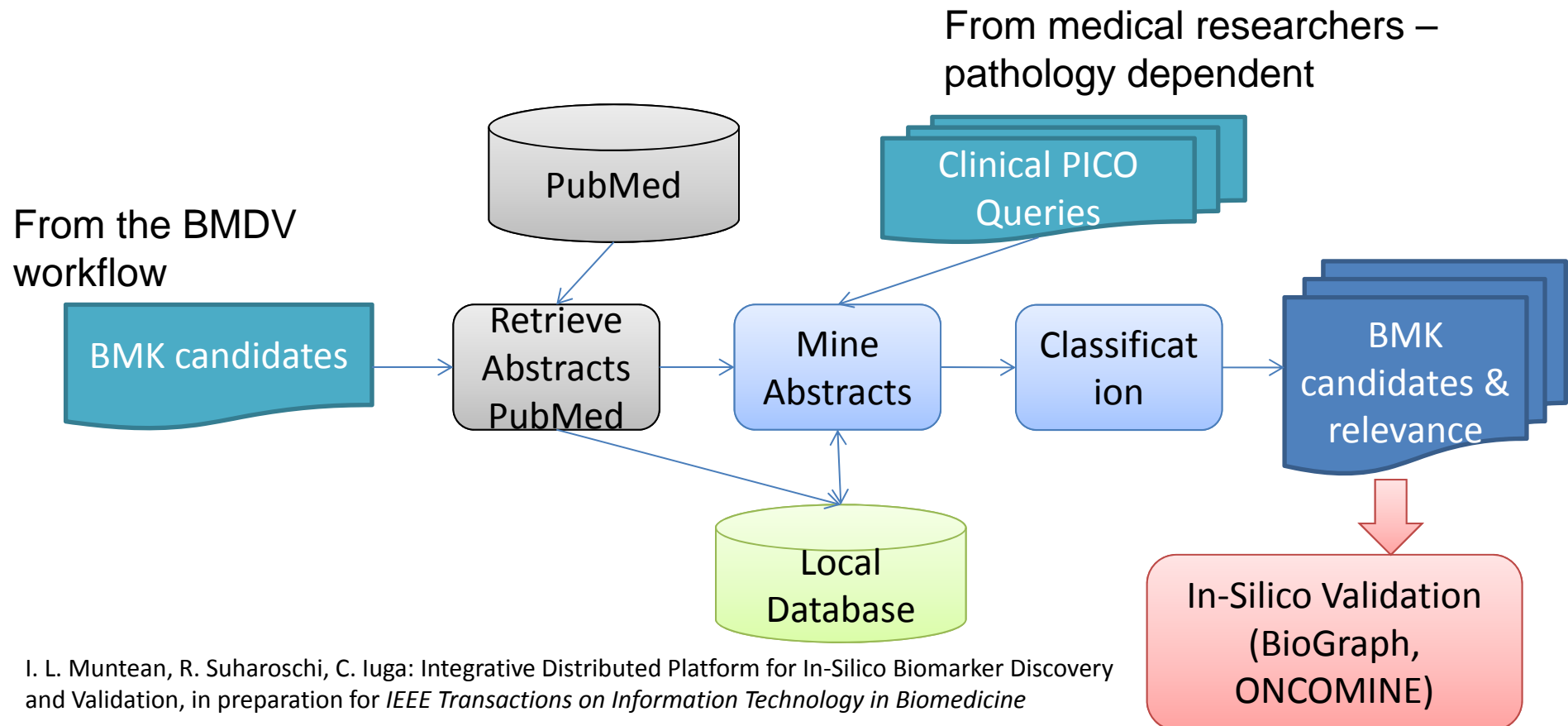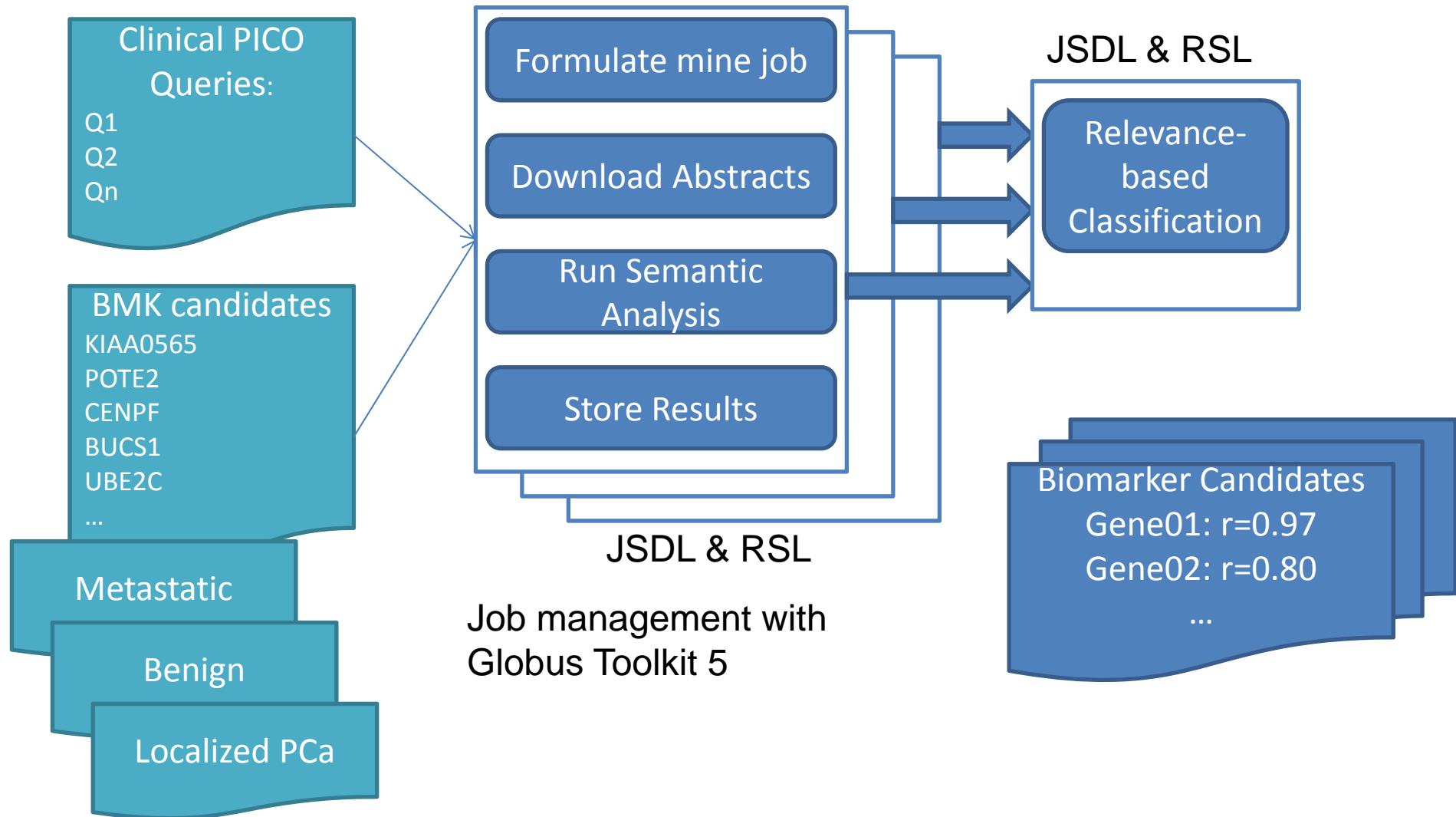

I. L. Muntean, R. Suharoschi, C. Iuga: Integrative Distributed Platform for In-Silico Biomarker Discovery and Validation, in preparation for *IEEE Transactions on Information Technology in Biomedicine*
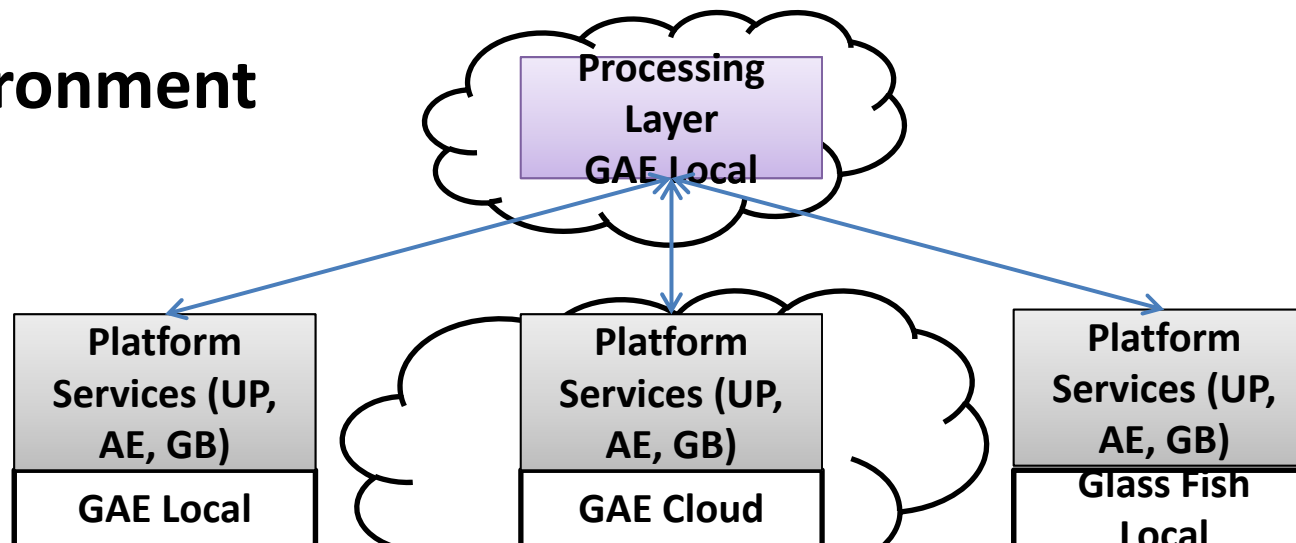
Calculating the Relevance of Biomarker Candidates

Clinical PICO Queries:
Q1
Q2
Qn

BMK candidates
KIAA0565
POTE2
CENPF
BUCS1
UBE2C
…

Metastatic

Benign

Localized PCa

Formulate mine job

Download Abstracts

Run Semantic Analysis

Store Results

JSDL & RSL

Job management with Globus Toolkit 5

JSDL & RSL

Relevance-based Classification

Biomarker Candidates
Gene01: r=0.97
Gene02: r=0.80
…

Evaluation Environment

Clinical queries (prostate cancer)

ID	Query	#UP	#AE	#GB
q1	prostate cancer	990	476	102253
q2	normal prostate	359	176	13017
q3	prostate adenocarcinoma	180	399	12375
q4	aggressive prostate cancer	6	21	2624
q5	radical prostatectomy	2	25	13160
q6	prostate specific antigen	146	10	24038
q7	biochemical recurrence	2	12	4082
q8	seminal vesicle invasion	7	4	914
q9	lymph node invasion	11	22	13403

Legend
UP – UniProt Service
AE – ArrayExpress Service
GB – GenBank Service
GAE – Google App Engine
GF – Glass Fish

Results – UniProt Data Integration – Query Service

Fetch time – time needed by the data retrieval service to extract data from the external data source
•relevant indicator for the responsiveness of the integration solution provided in the platform

Fetch Time/Entity	Glass Fish	GAE Local	GAE Cloud
UniProt	26ms	18ms	**6ms**

GAE vs. Glass Fish deployment

Cloud vs. local deployment

GAE Cloud - fastest
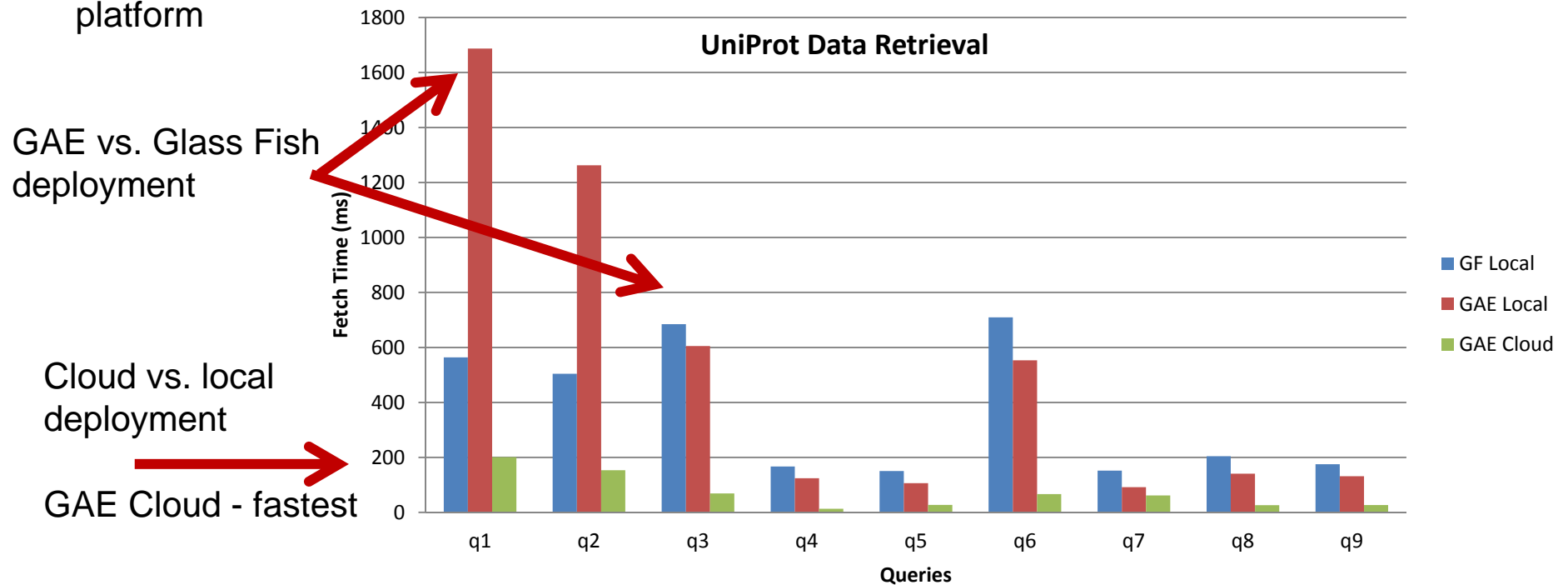
I.L. Muntean et al. *"Experiences in Designing a Distributed Service-Oriented Platform for In-Silico Biomarker Discovery and Validation"*, BIBE 2012, Larnaca

Results – Mining Abstracts with the PubMed Filtering Service

- Prostate cancer investigations

- Microarray experiment E-GEOD-8511 – data set with 98 samples:

  - FC>1.5: →#Genes: metastatic – 680, benign – 73, localized Pca – 48

  - pValue<|0.1|→#Genes: metastatic – 13.089, benign – 14.377, localized Pca – 14.301

  - → More than 42.500 jobs computed

- Average job duration on the IGE Test bed

  - #abstracts <10: < 10 sec

  - 10<#abstracts<100: <30 sec

  - 100<#abstracts: hours

R. Suharoschi et al: *Quencing In-Silico Discovery and Validation of Prostate Cancer Biomarker Candidates from Public Microarray Data Sources*, in preparation for BMC System Biology

Globus Toolkit Experiences

- BMDV activities – distributed nature – grids are the target environment
  - Globus Toolkit computing model – matches our problem & approach
  - Downloading in parallel the abstracts for each job → peak workload for the PubMed library service → computing model and jobs needed to be adjusted
  - Use of the abstracts database conflicts with the typical (firewall) settings of the grid production environments →abstracts as files
- IGE Test bed: ready-to-use, fully functional GT5 environment
  - Most of the jobs ran at the TUDO site – thank you!
- Under evaluation: integration of the platform with Globus Online
- Investigation of the deployment of the BioGenProtOMICS platform on cloud with the IGE Globus Appliances

Conclusions & Outlook

- Aim of our research: perform in-silico biomarker discovery and validation
  - Integration approach of multi-omics data sources, based on mature **web services, grid and cloud technologies**

- Currently – integration backbone of the emerging platform BioGenProtOMICS (GenBank, UniProt, ArrayExpress)
  - **Homogeneous query results** delivered to processing tasks required by BioGenProtOMICS
  - Semantic classification of abstracts

- Further investigations of cloud deployment models for platform services and layers (Amazon EC2, StratusLab)

- Interest in using the service from a commercial company and from three omics-research groups/clinics (Romania, Austria, Germany)
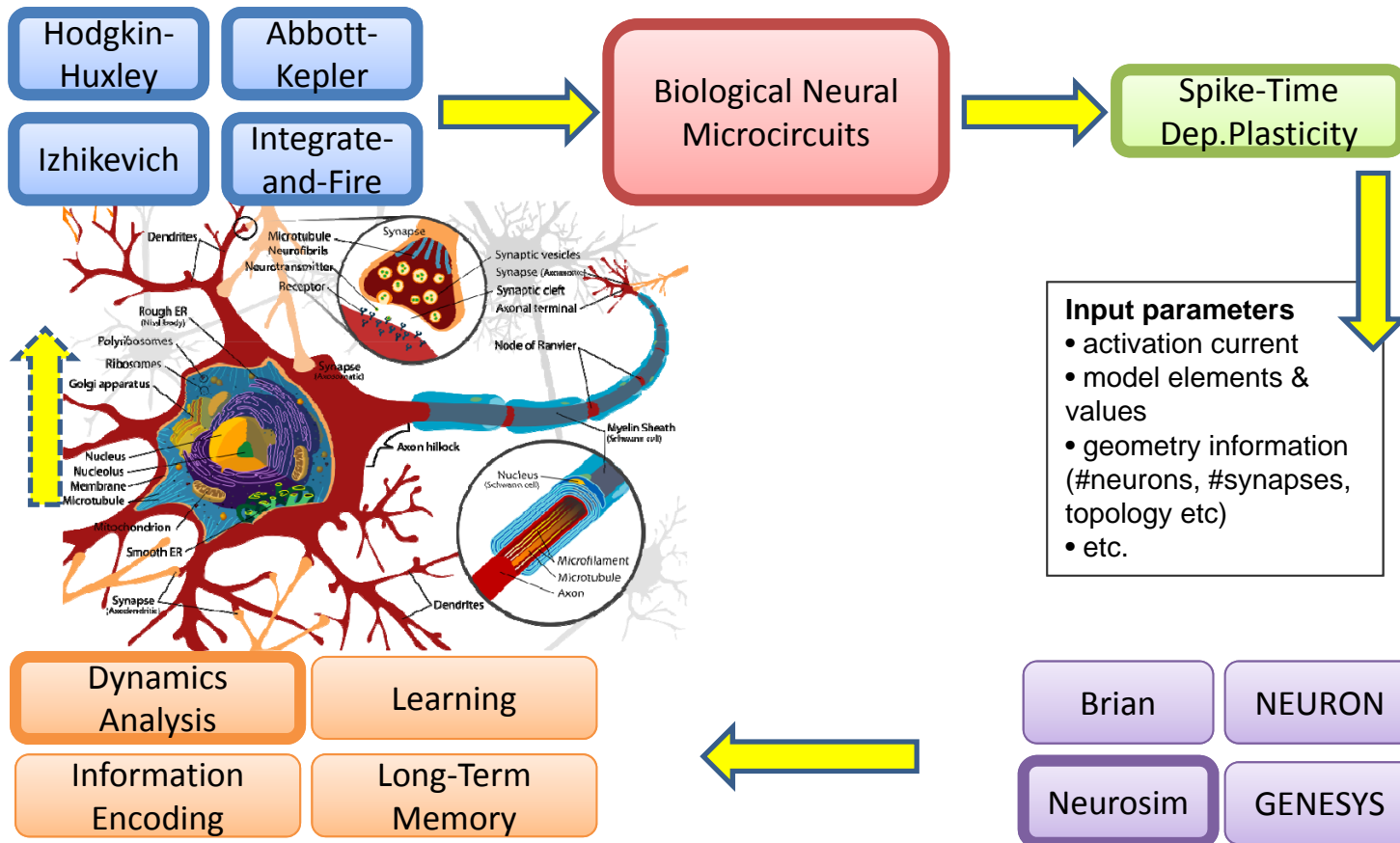
Scenario Two

DYNAMICS BEHAVIOR OF PARALLEL NEURAL MICROCIRCUITS SIMULATION

# Computational Neuroscience Research Context

- Research goal: study the impact of parallelization strategies on the dynamics behavior of neural microcircuits

Neuron image source: *Purves, Neuroscience, 2004.*

Hodgkin-Huxley

Abbott-Kepler

Izhikevich

Integrate-and-Fire

Biological Neural Microcircuits

Spike-Time Dep.Plasticity

Input parameters
- activation current
- model elements & values
- geometry information (#neurons, #synapses, topology etc)
- etc.

Dynamics Analysis

Learning

Information Encoding

Long-Term Memory

Brian

NEURON

Neurosim

GENESYS

Multi-Core Simulation of Synchronous Neural Microcircuits

Scalability Study (OpenMP): 10000 neurons, 10^7+ synapses, Spike-Response Model (SRM) and Spike Time-Dependent Plasticity (STDP)
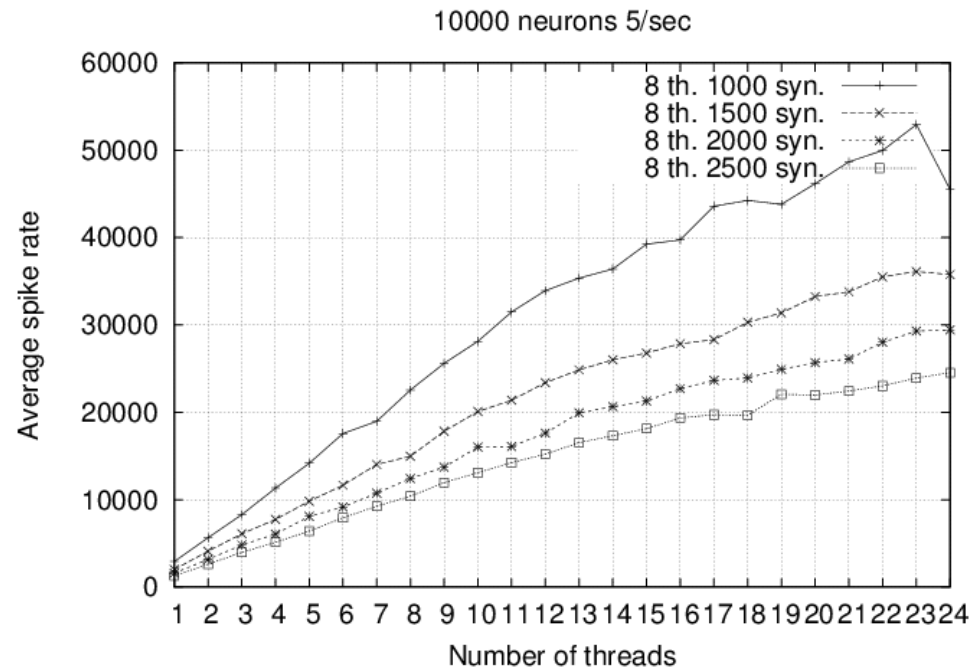
Dynamics analysis (Liapunov exponent, MH Entropy etc.): 800 neurons, 10^4+ synapses, SRM+STDP
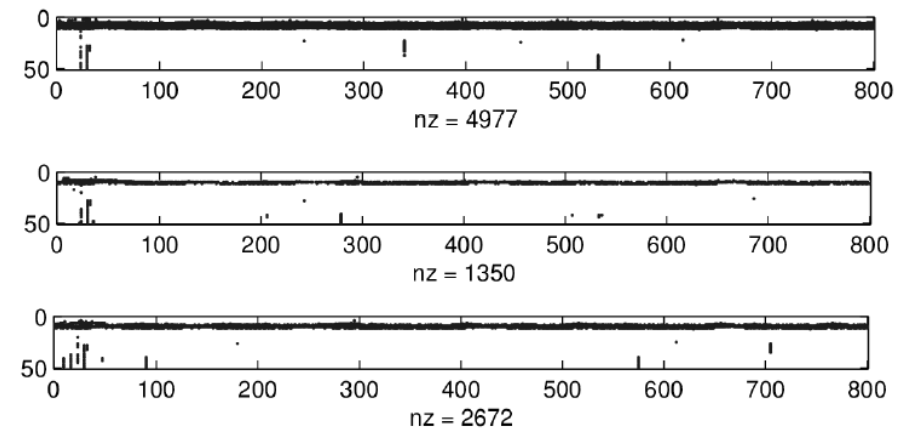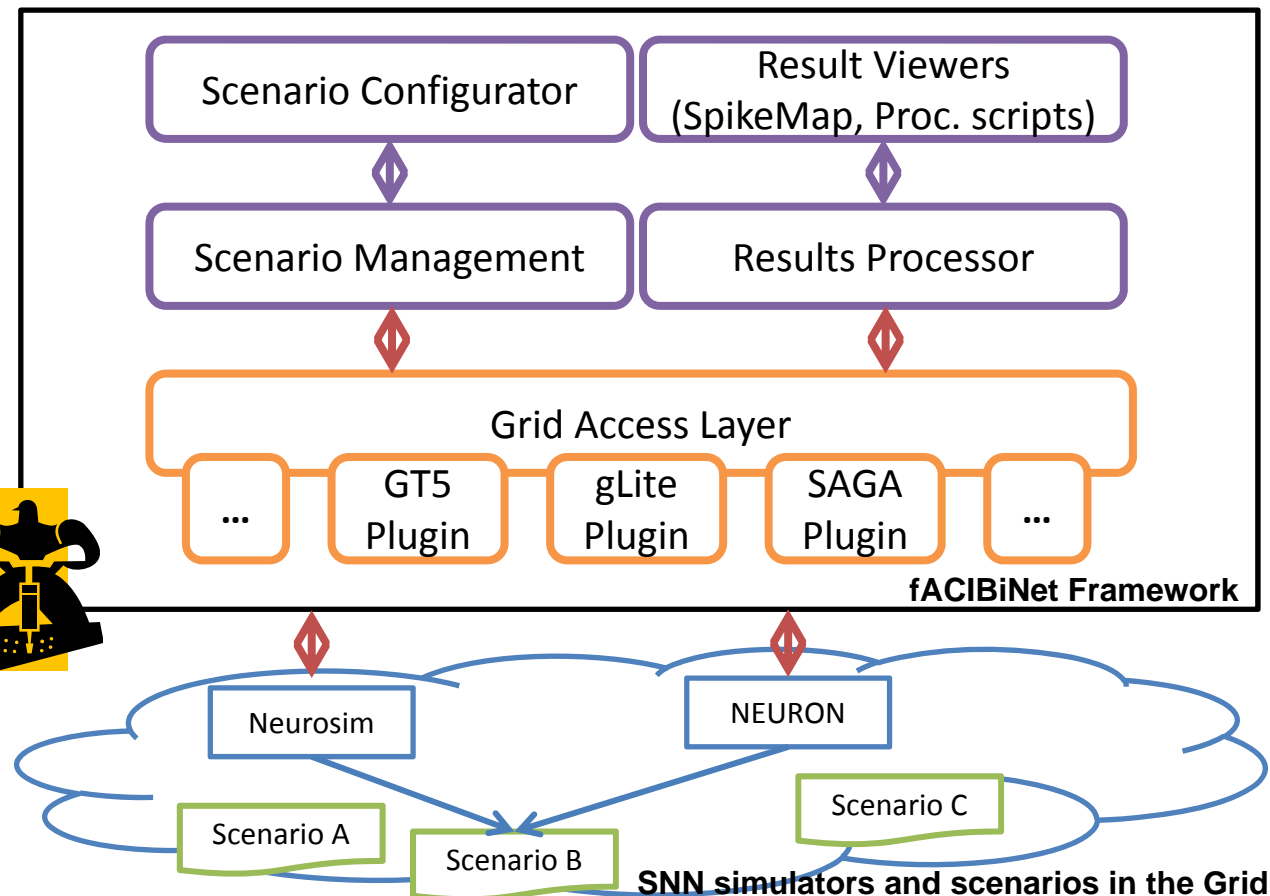

Non-zeros represent variations of spiking times

I.L. Muntean, M. Joldos. Simulation of Biological Neural Microcircuits on Multi-Core Systems, ePaMUS workshop, CISIS, July 4-6, 2012

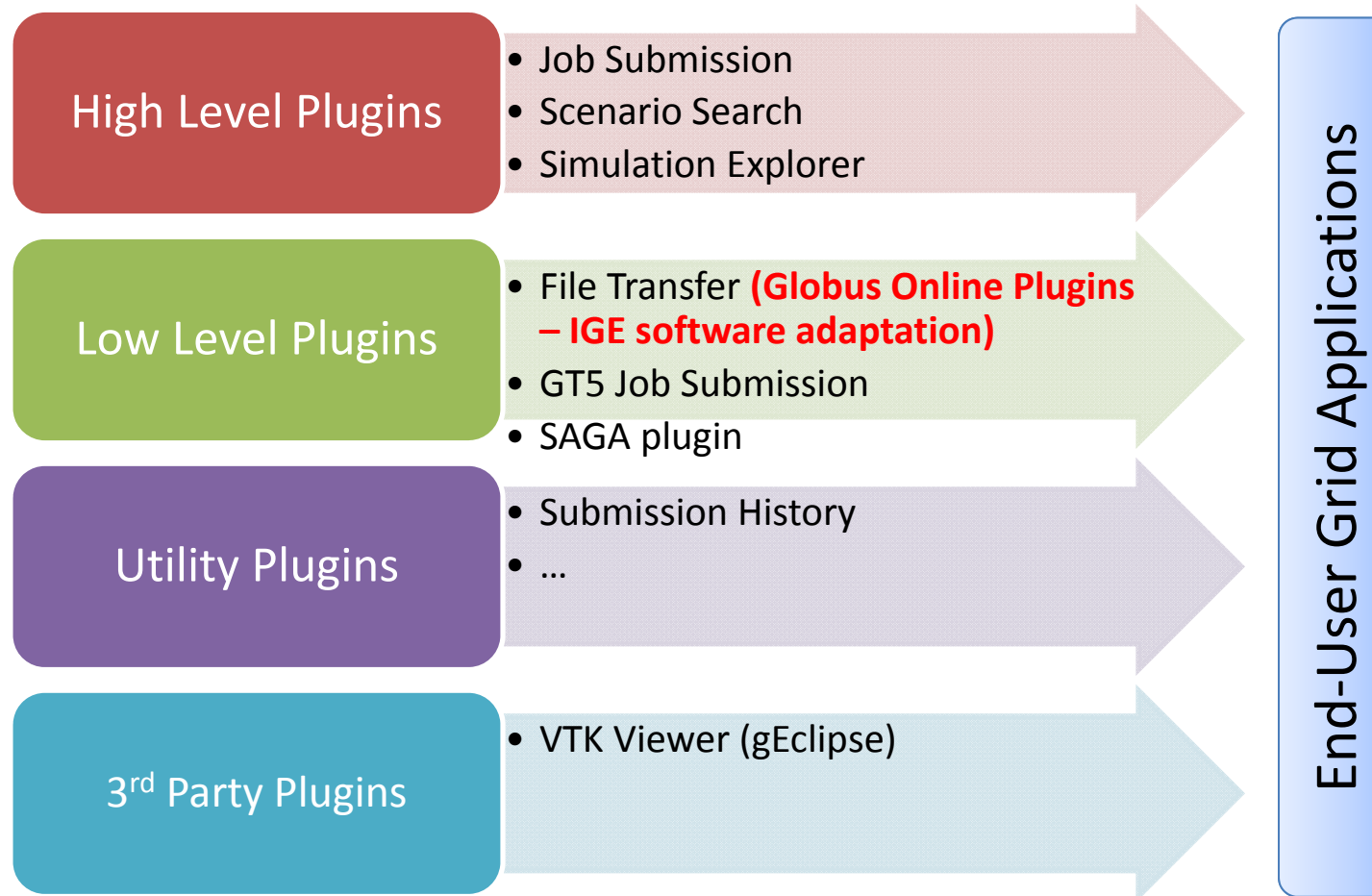# Parallel Simulations & Dynamics Analysis with fACIBiNet

- Technical goal: organize simulations, manage results and post-processing
- Framework for Advanced Computing Investigations of Biological Neural Microcircuits (fACIBiNet)

Why grid?
- Parameter studies with multi/manycore codes
- Independent tasks (network generation, simulation)
- High-dimensionality of parameter space (>15)
- High-dimensionality of results (>10^7)

fACIBiNet Framework

| Scenario Configurator | Result Viewers (SpikeMap, Proc. scripts) |

| Scenario Management | Results Processor |

Grid Access Layer

| ... | GT5 Plugin | gLite Plugin | SAGA Plugin | ... |

Neurosim   NEURON

Scenario A   Scenario B   Scenario C

SNN simulators and scenarios in the Grid

Grid Access Layer – Plugin-based Realization*

High Level Plugins
- Job Submission
- Scenario Search
- Simulation Explorer

Low Level Plugins
- File Transfer **(Globus Online Plugins – IGE software adaptation)**
- GT5 Job Submission
- SAGA plugin

Utility Plugins
- Submission History
- …

3rd Party Plugins
- VTK Viewer (gEclipse)

End-User Grid Applications

* Most of the plugins have been initially developed in the research projects "Grid Applications for Numerical Simulation" and "Grid Centre of UTCN".

Instead of Conclusions: IGE GO Client (plugin-based)

I. Muntean
GlobusWorld 2013, Chicago
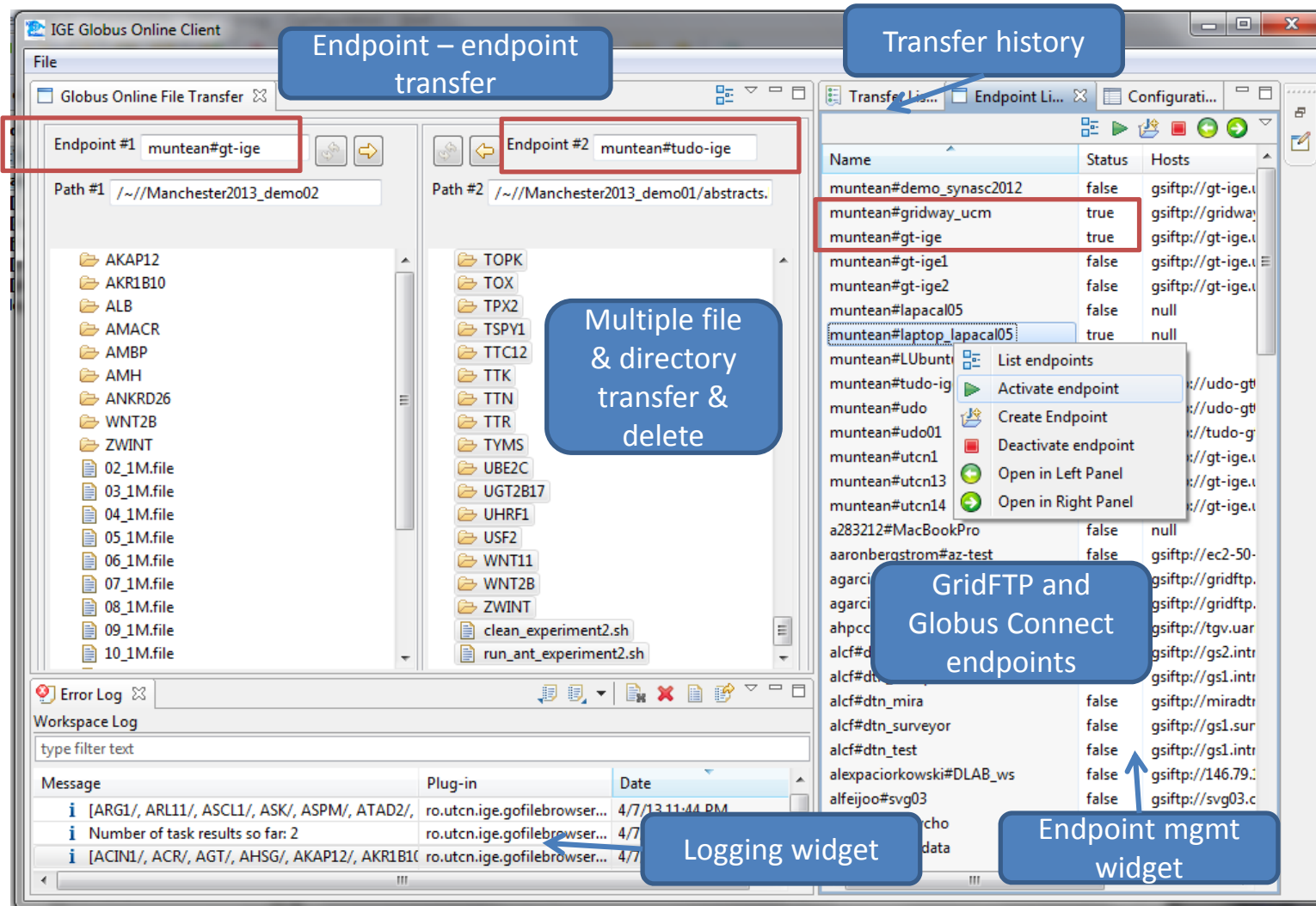
THANK YOU!

QUESTIONS?

Our Team:
Medical doctors:
Cristina Iuga
Ramona Suharoschi
Neuroscientists:
Raul Muresan
Computer Scientists:
Emil Chifu
Adrian Colesa
Gabriel Dragomir
Marius Joldos
Radu Peter
Students:
Carmen Negreanu
Bogan Rotaru
Delia Sumalan
et al.