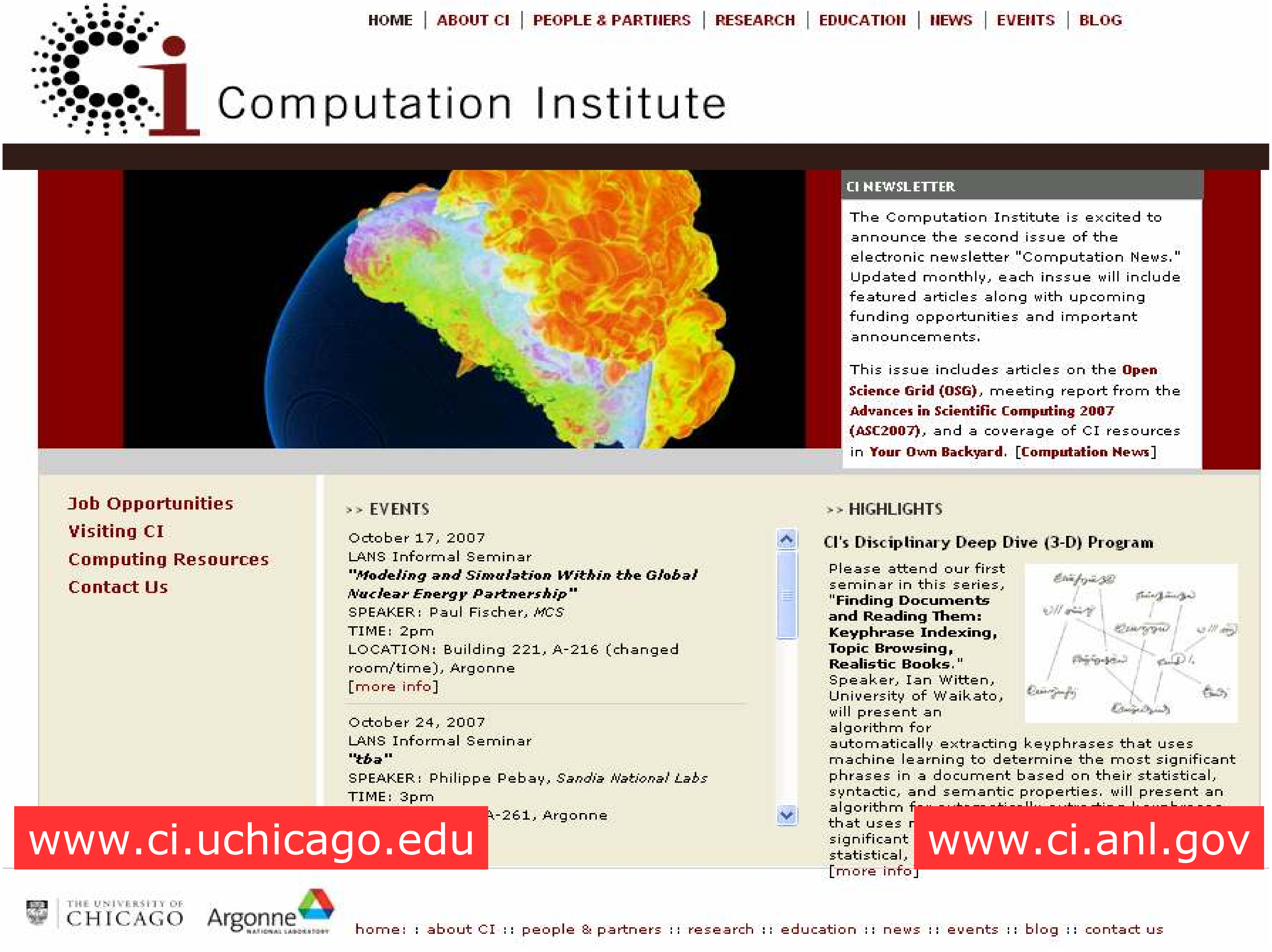
Office of Science
U.S. Department of Energy

Data Mining and Grid

Ian Foster

Computation Institute
Argonne National Lab & University of Chicago
http://ianfoster.typepad.com



HOME | ABOUT CI | PEOPLE & PARTNERS | RESEARCH | EDUCATION | NEWS | EVENTS | BLOG

Computation Institute

CI NEWSLETTER

The Computation Institute is excited to announce the second issue of the electronic newsletter "Computation News." Updated monthly, each inssue will include featured articles along with upcoming funding opportunities and important announcements.

This issue includes articles on the **Open Science Grid (OSG)**, meeting report from the **Advances in Scientific Computing 2007 (ASC2007)**, and a coverage of CI resources in **Your Own Backyard**. [**Computation News**]

>> EVENTS

October 17, 2007
LANS Informal Seminar
"Modeling and Simulation Within the Global Nuclear Energy Partnership"
SPEAKER: Paul Fischer, *MCS*
TIME: 2pm
LOCATION: Building 221, A-216 (changed room/time), Argonne
[more info]

October 24, 2007
LANS Informal Seminar
"tba"
SPEAKER: Philippe Pebay, *Sandia National Labs*
TIME: 3pm
A-261, Argonne

>> HIGHLIGHTS

CI's Disciplinary Deep Dive (3-D) Program

Please attend our first seminar in this series, **"Finding Documents and Reading Them: Keyphrase Indexing, Topic Browsing, Realistic Books."** Speaker, Ian Witten, University of Waikato, will present an algorithm for automatically extracting keyphrases that uses machine learning to determine the most significant phrases in a document based on their statistical, syntactic, and semantic properties. will present an algorithm f

that uses r

significant

statistical,

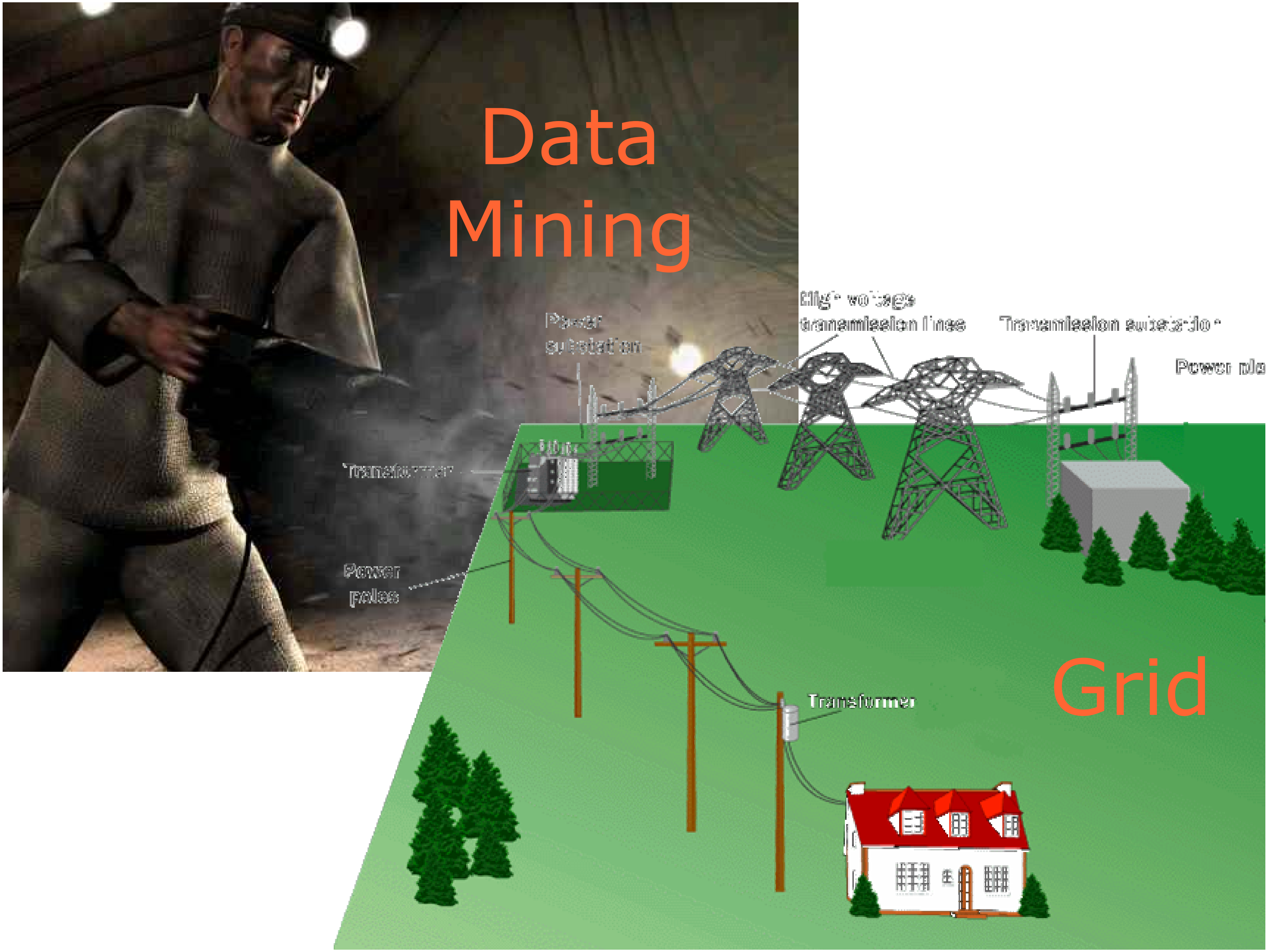
[more info]

home: : about CI :: people & partners :: research :: education :: news :: events :: blog :: contact us

www.ci.uchicago.edu

www.ci.anl.gov

Data Mining



Grid

In the Next 50 Years,
We Must …

- Increase energy production by 5, while reducing GHG emissions by 2 or more

- Mitigate and adapt to climate change

- Address increasingly drug resistant diseases

- Provide meaningful livelihoods for 9B people

→ Innovation

Innovation as a Systems Problem

- Quasi-ubiquitous Internet …
- … connects many potential innovators
  - Millions of scientists, billions of people
- Who need to leverage
  - Enormous data of tremendous complexity
  - Immensely powerful computing
  - Experimental apparatus of great power

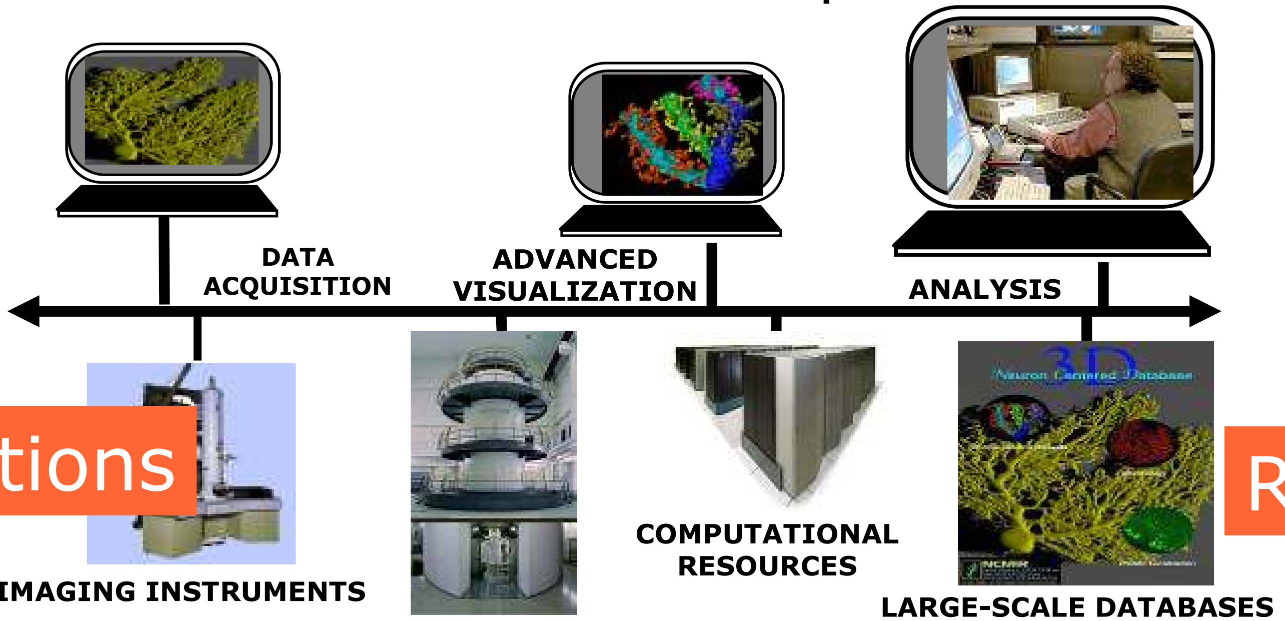
➔ We must address problem solving as an distributed, end-to-end, systems problem

Grid: A Unifying Concept & Technology

Grid enables the **federation** of resources

- Distributed computers, storage, data, people, …
- **Networks** provide connectivity
- **Software** & **standards** provide the “glue”
- **Infrastructure** services facilitate operation

Infrastructure



DATA ACQUISITION

ADVANCED VISUALIZATION

ANALYSIS

IMAGING INSTRUMENTS

COMPUTATIONAL RESOURCES

LARGE-SCALE DATABASES

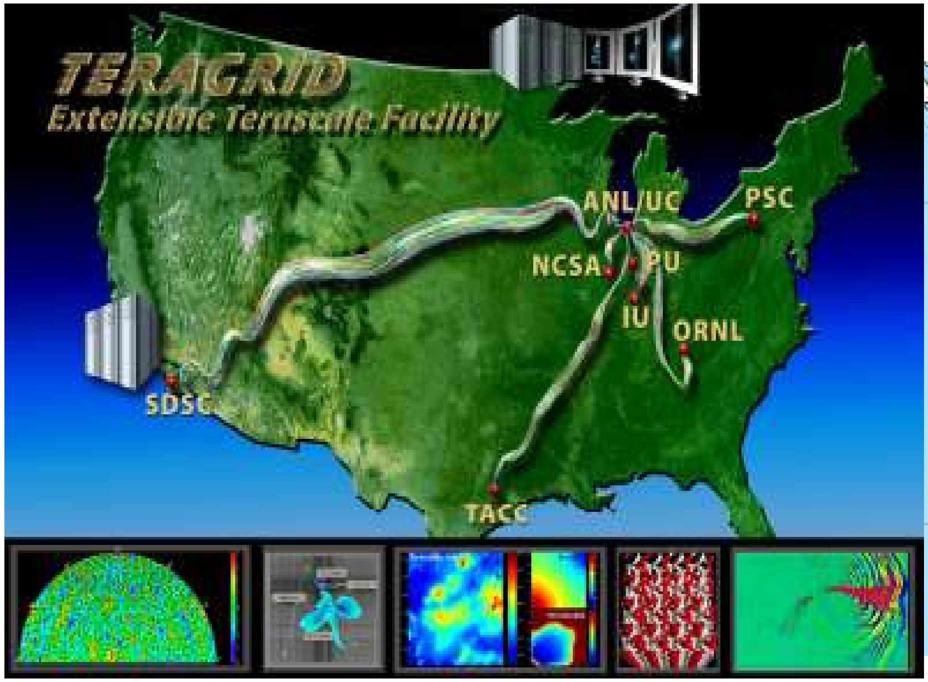
Applications

Research

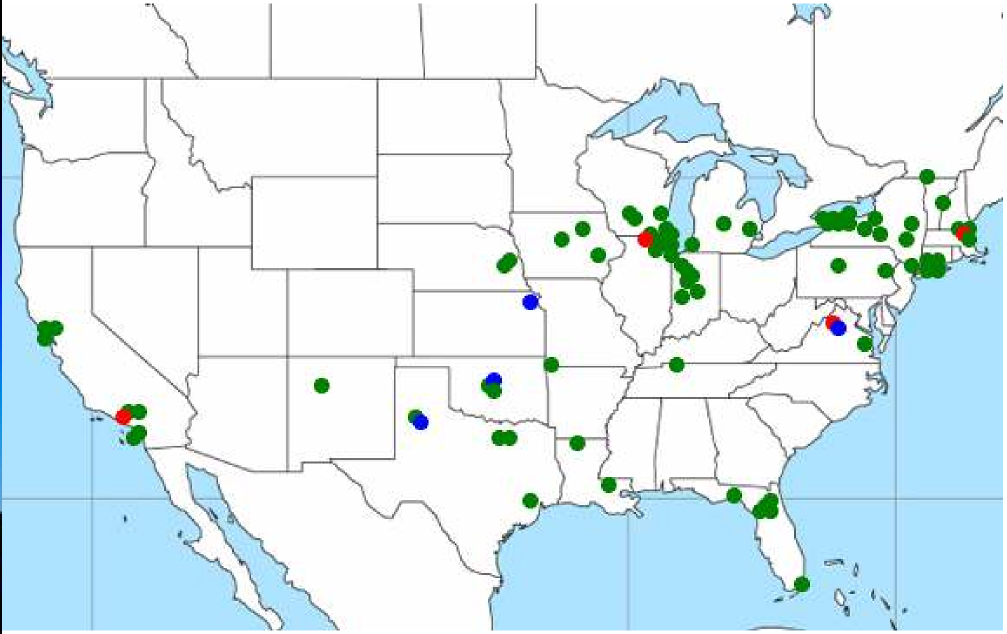
Credit: Mark Ellisman

Grid Infrastructure

- Massive computing and storage
- Service interfaces facilitate access and use

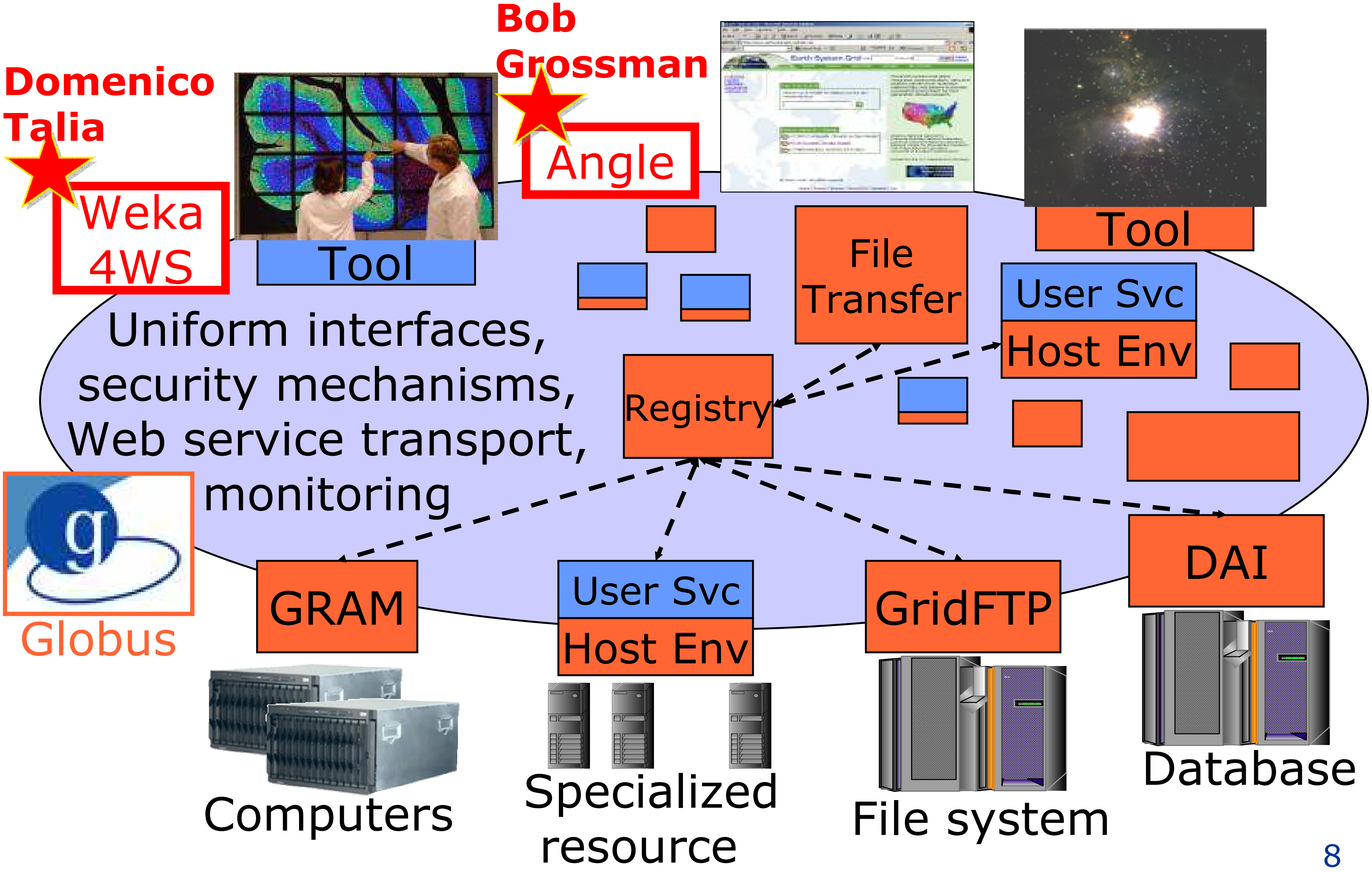


TeraGrid



Open Science Grid

Software and Standards



Bob Grossman

Domenico Talia

Weka 4WS

Tool

Angle

Tool

File Transfer

User Svc

Host Env

Uniform interfaces, security mechanisms, Web service transport, monitoring

Registry

Globus

GRAM

User Svc

Host Env

GridFTP

DAI

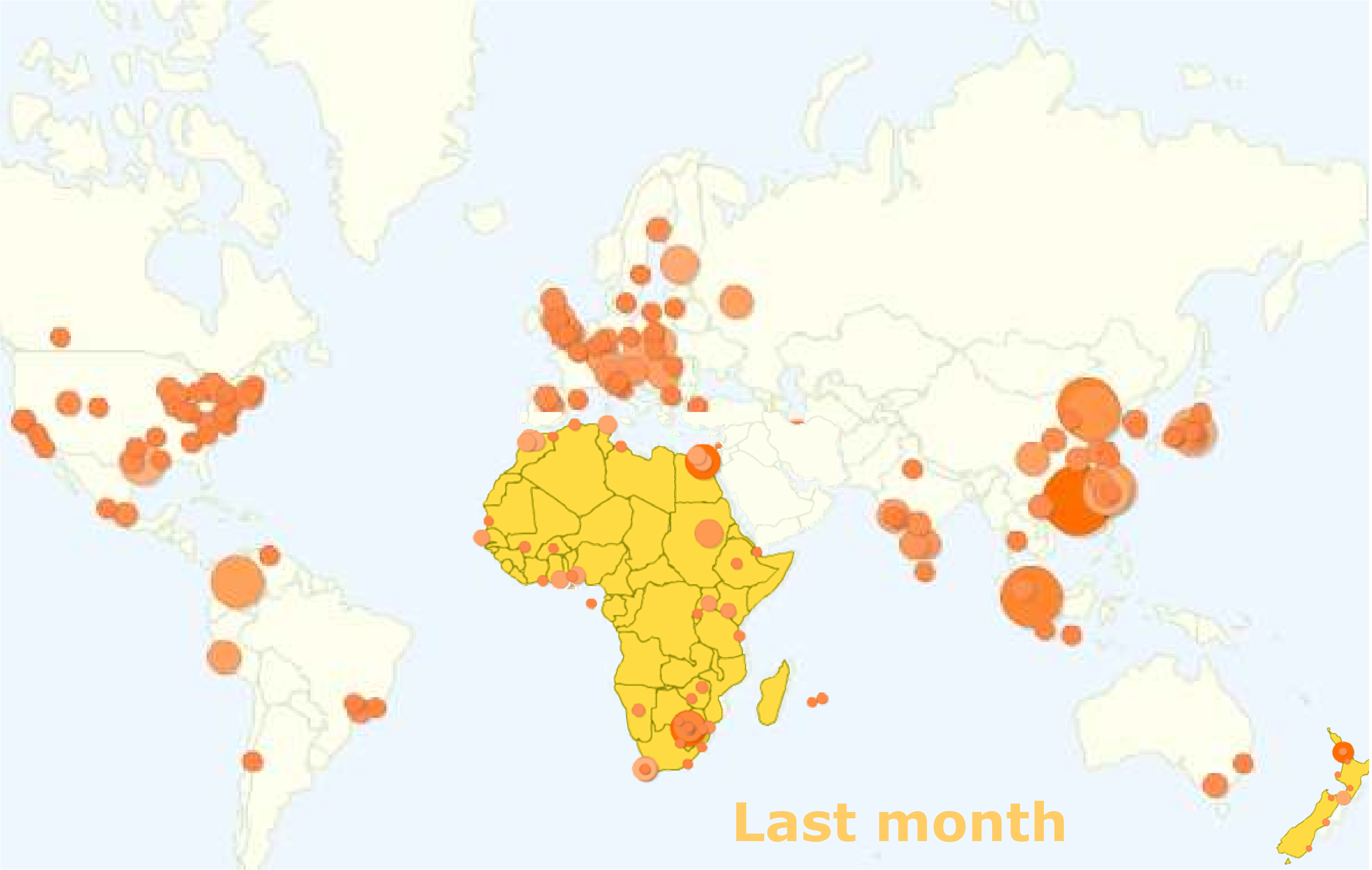
Computers

Specialized resource

File system

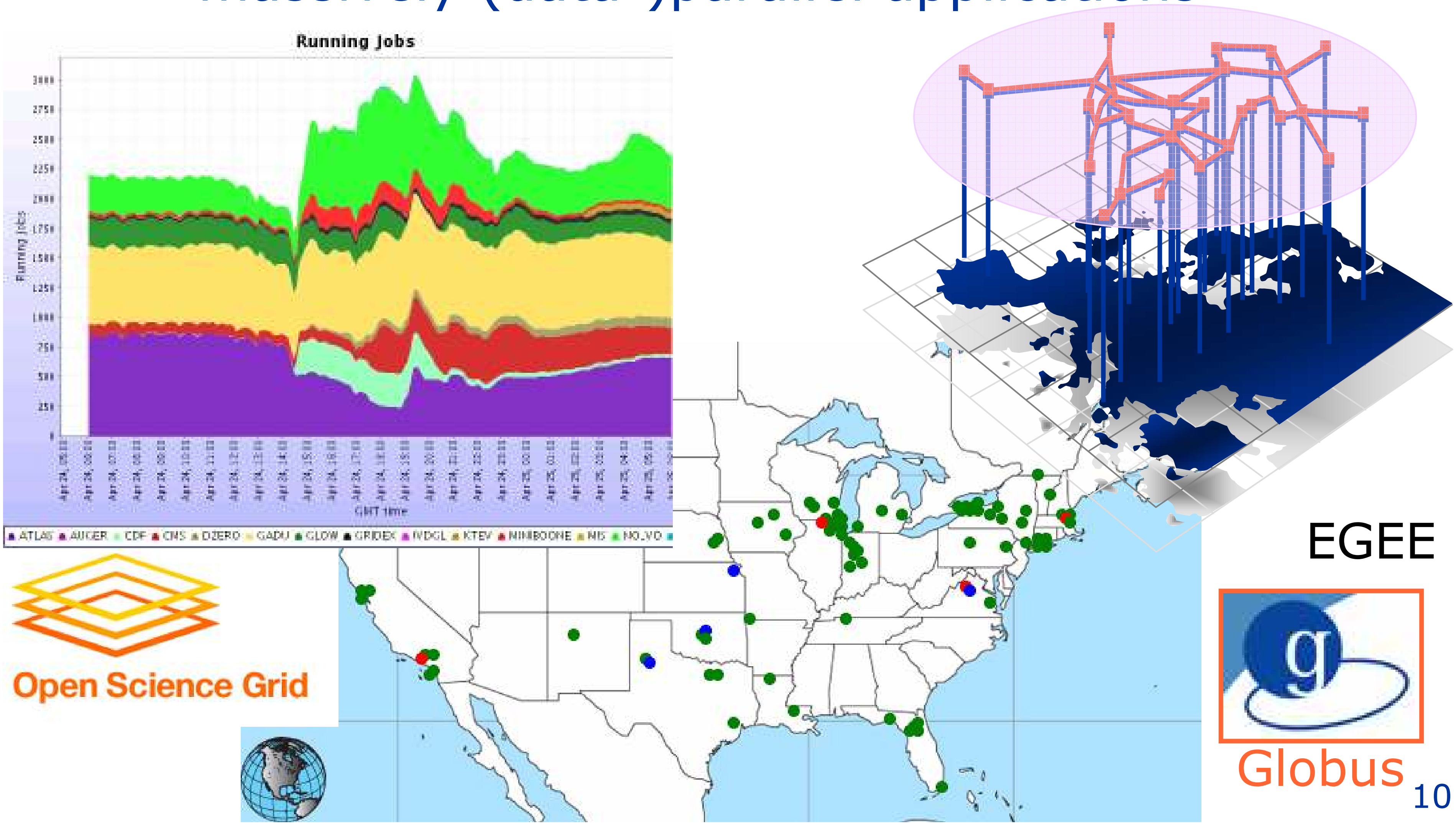
Database

Globus Downloads Last 24 Hours

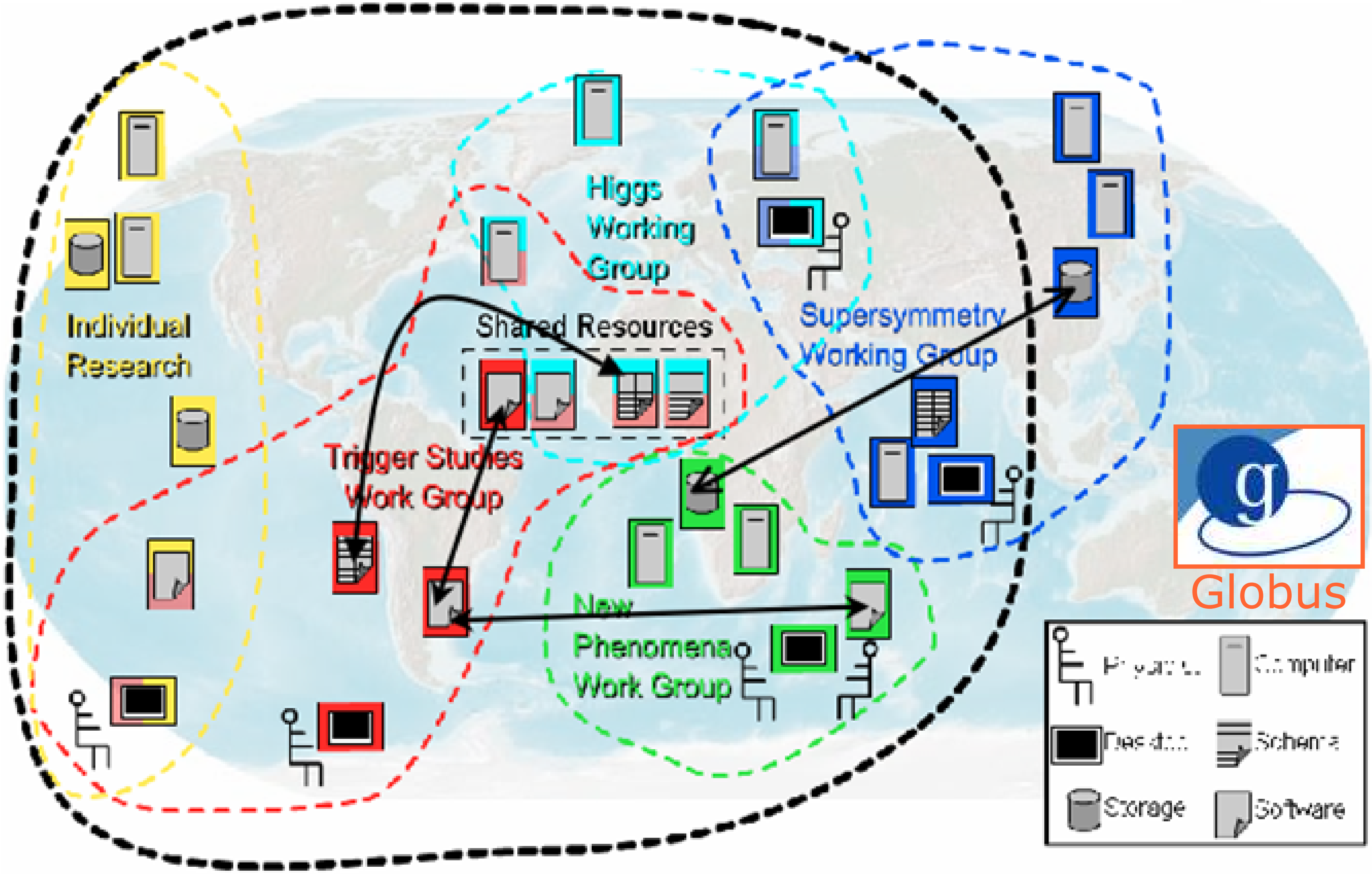


First Generation Grids: On-Demand/Batch Computing

Focus on aggregation of many resources for massively (data-)parallel applications



Applications: High Energy Physics

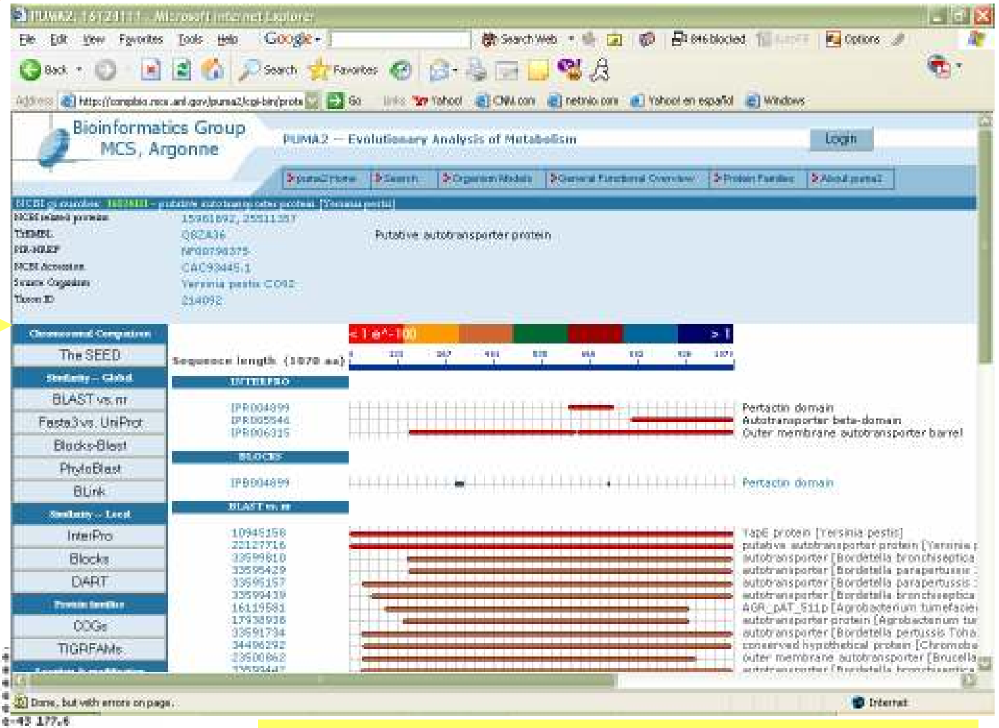


Globus

Integrating Data and Computing, on Demand

Public PUMA Knowledge Base

Information about proteins analyzed against ~2 million gene sequences



PUMA2: 16121111 - Microsoft Internet Explorer

Address http://compbio.mcs.anl.gov/puma2/cgi-bin/prote

Bioinformatics Group MCS, Argonne

PUMA2 — Evolutionary Analysis of Metabolism

NCBI gi number: 16121111 - putative autotransporter protein [Yersinia pestis]

NCBI related proteins	15981892, 22511357	
TrEMBL	Q8ZA36	Putative autotransporter protein
PIR-NREF	NF00798375	
NCBI Accession	CAC93445.1	
Source Organism	Yersinia pestis CO92	
Taxon ID	214092	

Sequence length (1070 aa)

INTERPRO

IPR004899	Pertactin domain
IPR005546	Autotransporter beta-domain
IPR006315	Outer membrane autotransporter barrel

BLOCKS

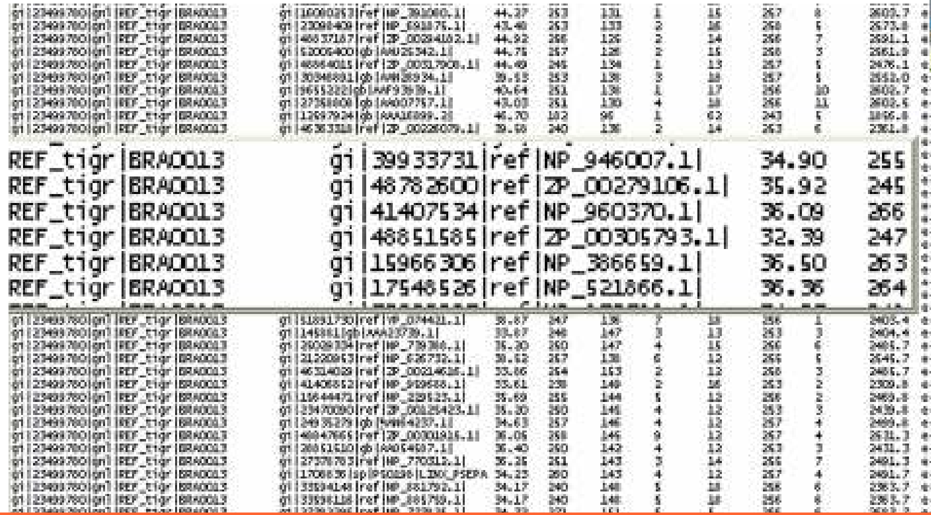
IPB004899	Pertactin domain

BLAST vs. nr

10945158	YapE protein [Yersinia pestis]
22127716	putative autotransporter protein [Yersinia p
33599810	autotransporter [Bordetella bronchiseptica
33595429	autotransporter [Bordetella parapertussis
33595157	autotransporter [Bordetella parapertussis
33599439	autotransporter [Bordetella bronchiseptica
16119581	AGR_pAT_511p [Agrobacterium tumefacie
17938938	autotransporter protein [Agrobacterium tu
33591734	autotransporter [Bordetella pertussis Toha
34496292	conserved hypothetical protein [Chromoba
23500862	outer membrane autotransporter [Brucella

Done, but with errors on page.

File
Edit
View
Favorites
Tools
Help
Google
Search Web
846 blocked
Options
Back
Search
Favorites
Go
Links
Yahoo!
CNN.com
netnio.com
Yahoo! en español
Windows
Login
puma2 home
Search
Organism Models
General Functional Overview
Protein Families
About puma2
Chromosomal Comparison
The SEED
Similarity – Global
BLAST vs. nr
Fasta3 vs. UniProt
Blocks-Blast
PhyloBlast
BLink
Similarity – Local
InterPro
Blocks
DART
Protein families
COGs
TIGRFAMs
< 1 e^-100
> 1
Internet



```
gi|23499780|gnl|REF_tigr|BRA0013   gi|16080253|ref|NP_391080.1|     44.27 253 131 1 15 267  8 2603.7
gi|23499780|gnl|REF_tigr|BRA0013   gi|23098409|ref|NP_691875.1|     43.48 253 133 2 16 268  5 2673.8
gi|23499780|gnl|REF_tigr|BRA0013   gi|48837187|ref|ZP_00294162.1|   44.92 266 126 2 14 266  7 2691.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|52005400|gb|AAU25342.1|       44.75 257 128 2 15 266  3 2561.9
gi|23499780|gnl|REF_tigr|BRA0013   gi|48864015|ref|ZP_00317908.1|   44.49 246 134 1 13 267  5 2476.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|30346931|gb|AAH28934.1|       39.53 253 138 3 16 267  5 2552.0 e-43 177.6
gi|23499780|gnl|REF_tigr|BRA0013   gi|9655222|gb|AAF93878.1|        40.64 251 138 1 17 266 10 2602.7 e-43 177.2
gi|23499780|gnl|REF_tigr|BRA0013   gi|27358008|gb|AAO07757.1|       43.03 251 130 4 18 266 11 2602.5 e-41 170.6
gi|23499780|gnl|REF_tigr|BRA0013   gi|12697924|gb|AAA16899.2|       46.70 182  96 1 62 243  5 1856.8 e-39 162.5
gi|23499780|gnl|REF_tigr|BRA0013   gi|46363318|ref|ZP_00226079.1|   39.58 240 136 2 14 263  6 2361.8 e-36 164.5
                                                                                                  e-33 142.9
REF_tigr|BRA0013        gi|39933731|ref|NP_946007.1|      34.90   255                             e-33 142.1
REF_tigr|BRA0013        gi|48782600|ref|ZP_00279106.1|    35.92   245                             e-32 141.4
REF_tigr|BRA0013        gi|41407534|ref|NP_960370.1|      36.09   266                             e-32 140.2
REF_tigr|BRA0013        gi|48851585|ref|ZP_00305793.1|    32.39   247                             e-32 138.4
REF_tigr|BRA0013        gi|15966306|ref|NP_386659.1|      36.50   263                             e-32 139.0
REF_tigr|BRA0013        gi|17548526|ref|NP_521866.1|      36.36   264                             e-31 137.9
                                                                                                  e-31 137.1
                                                                                                  e-30 135.2
                                                                                                  e-30 134.8
                                                                                                  e-30 134.8
                                                                                                  e-30 134.4
                                                                                                  e-30 134.
                                                                                                  e-30 134.
gi|23499780|gnl|REF_tigr|BRA0013   gi|51891730|ref|YP_074421.1|     38.87 247 138 7 18 256  1 2405.4 e-30 133.7
gi|23499780|gnl|REF_tigr|BRA0013   gi|145881|gb|AAA23739.1|         33.67 246 147 3 13 253  3 2404.4 e-30 133.3
gi|23499780|gnl|REF_tigr|BRA0013   gi|20089334|ref|NP_739386.1|     35.20 250 147 4 15 266  6 2405.7 e-30 132.9
gi|23499780|gnl|REF_tigr|BRA0013   gi|21220953|ref|NP_626732.1|     38.62 257 138 6 12 265  5 2646.7 e-30 132.9
gi|23499780|gnl|REF_tigr|BRA0013   gi|46314028|ref|ZP_00214616.1|   33.86 254 153 2 12 268  3 2465.7 e-30 132.9
gi|23499780|gnl|REF_tigr|BRA0013   gi|41408852|ref|NP_959688.1|     33.61 238 140 2 16 253  2 2309.8 e-30 132.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|15644471|ref|NP_229523.1|     35.69 255 144 5 12 268  2 2469.8 e-30 132.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|23470090|ref|ZP_00125423.1|   35.30 250 141 4 13 253  3 2439.8 e-30 132.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|24935279|gb|AAN64237.1|       34.63 257 146 4 12 257  4 2489.8 e-30 132.1
gi|23499780|gnl|REF_tigr|BRA0013   gi|48047665|ref|ZP_00301816.1|   36.06 258 146 9 12 267  4 2631.3 e-29 131.7
gi|23499780|gnl|REF_tigr|BRA0013   gi|28851610|gb|AAO54687.1|       35.40 250 142 4 12 263  3 2431.3 e-29 131.7
gi|23499780|gnl|REF_tigr|BRA0013   gi|27378703|ref|NP_770312.1|     35.26 251 143 3 14 266  7 2491.3 e-29 131.7
gi|23499780|gnl|REF_tigr|BRA0013   gi|1708636|sp|P50198|LINX_PSEPA  34.23 260 143 4 12 267  4 2491.7 e-29 131.3
gi|23499780|gnl|REF_tigr|BRA0013   gi|33594148|ref|NP_881792.1|     34.17 240 148 5 18 258  6 2363.7 e-29 130.2
gi|23499780|gnl|REF_tigr|BRA0013   gi|33598118|ref|NP_885769.1|     34.17 240 148 5 18 256  6 2363.7 e-29 130.2
```

Back Office Analysis on Grid

Millions of BLAST, BLOCKS, etc., on OSG and TeraGrid

Natalia Maltsev et al., http://compbio.mcs.anl.gov/puma2

Second Generation Grids: **Service-Oriented Science**

- Empower many more users by enabling on-demand access to **services**
- Grids become an enabling technology for **service oriented science** (or business)
  - Grid infrastructures host services
  - Grid technologies used to build services

Science Gateways

TeraGrid™
EMPOWERING DISCOVERY

caBIG

"Service-Oriented Science", *Science*, 2005

Service-Oriented Science

People **create** services (data or functions) …

which I **discover** (& decide whether to use) …

& **compose** to create a new function ...

& then **publish** as a new service.

→ *I find "someone else" to **host** services, so I don't have to become an expert in operating services & computers!*

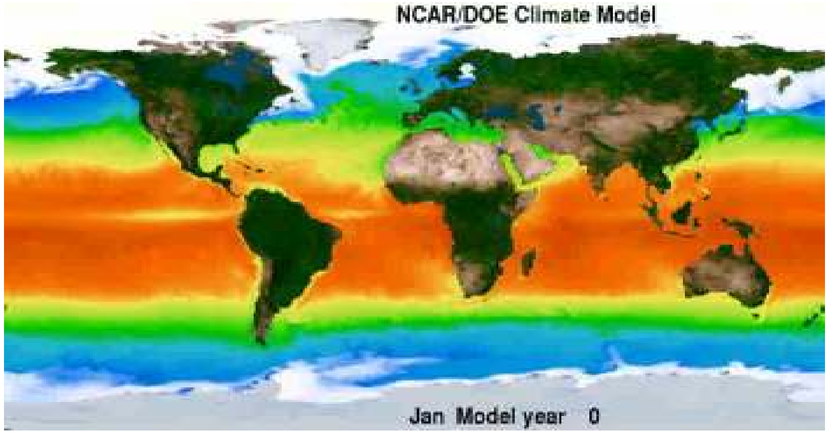
!!

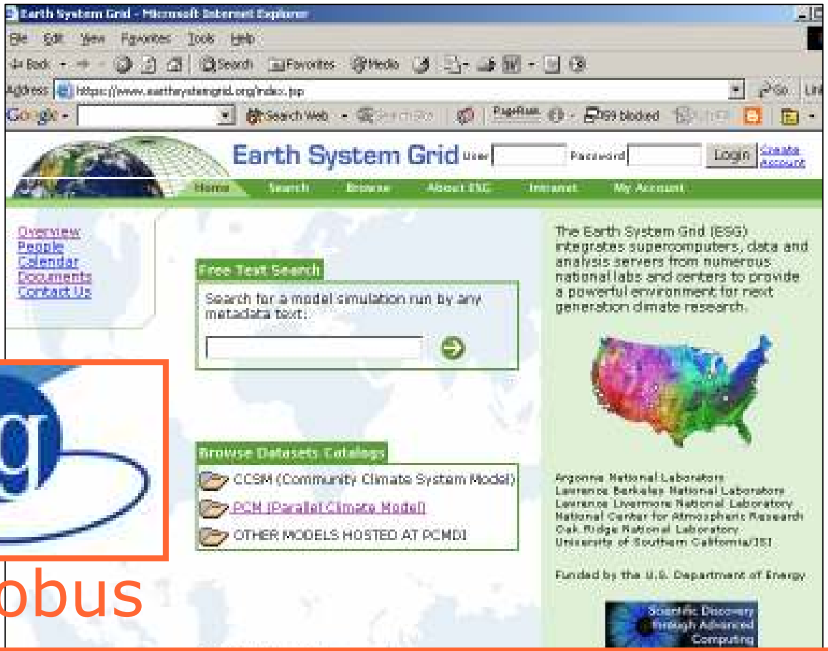
→ *I hope that this "someone else" can **manage** security, reliability, scalability, …*

"Service-Oriented Science", *Science*, 2005

Earth System Grid

- On-demand access to climate simulation data
  - Multiple archives
  - Interactive query
  - Per-collection control
  - Server-side processing
- Major scientific impact
  - >5000 users
  - >200 TB downloaded
  - >300 scientific papers




Earth System Grid - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://www.earthsystemgrid.org/index.jsp
Go
Google Search Web PageRank 69 blocked
Earth System Grid
User
Password
Login
Create Account
Home
Search
Browse
About ESG
Intranet
My Account
Overview
People
Calendar
Documents
Contact Us
Free Text Search
Browse Datasets Catalogs
Scientific Discovery through Advanced Computing

Search for a model simulation run by any metadata text:

CCSM (Community Climate System Model)
PCM (Parallel Climate Model)
OTHER MODELS HOSTED AT PCMDI

The Earth System Grid (ESG) integrates supercomputers, data and analysis servers from numerous national labs and centers to provide a powerful environment for next generation climate research.

Argonne National Laboratory
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
National Center for Atmospheric Research
Oak Ridge National Laboratory
University of Southern California/ISI

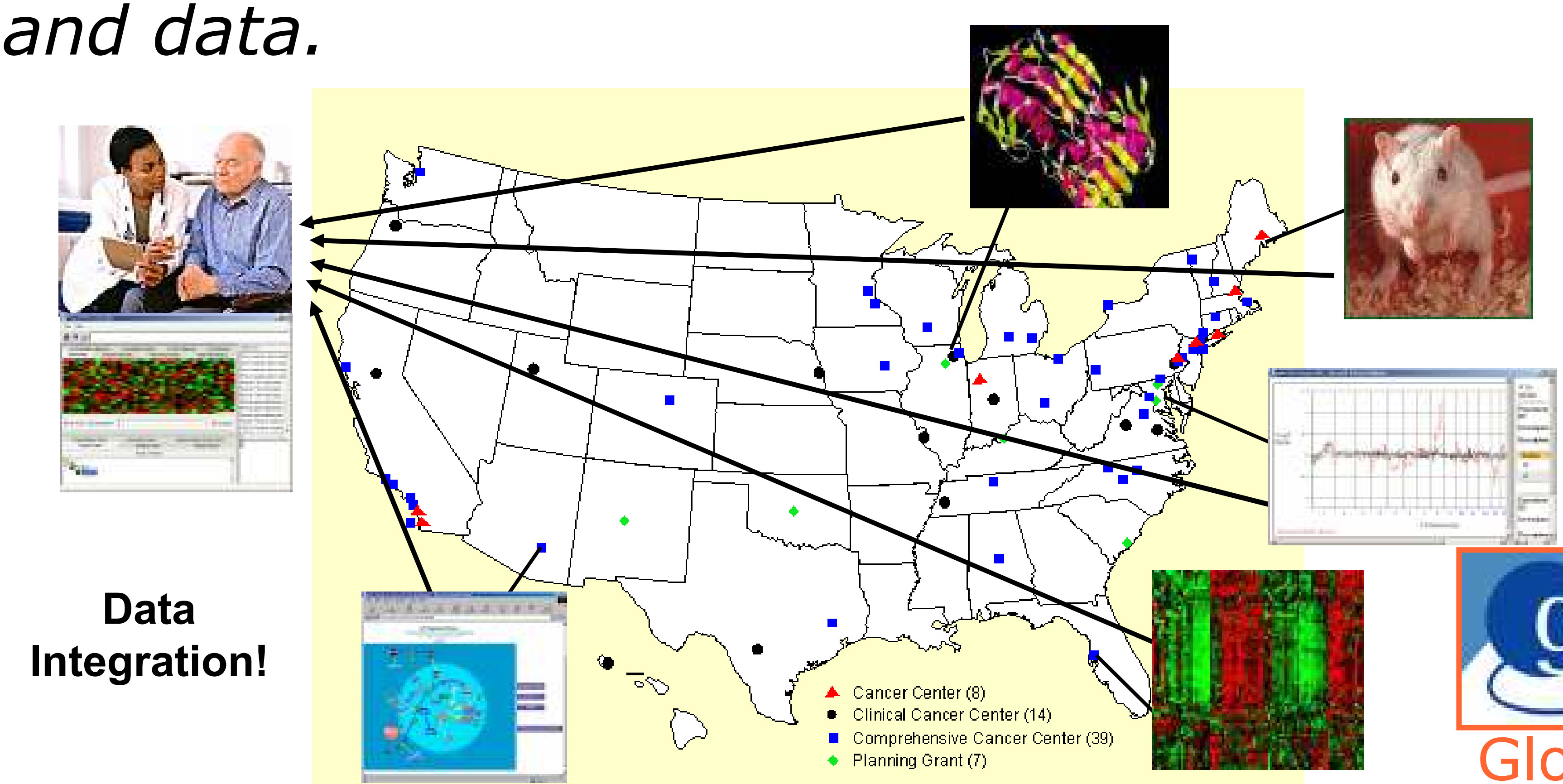
Funded by the U.S. Department of Energy

Globus

www.earthsystemgrid.org — DOE OASCR

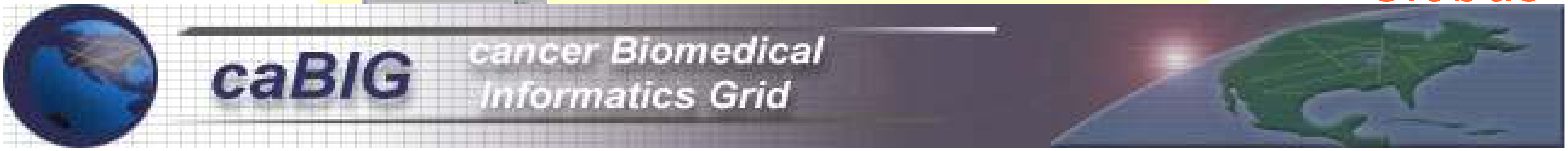
Cancer Biomedical Informatics Grid (caBIG)

caBIG: sharing of infrastructure, applications, and data.

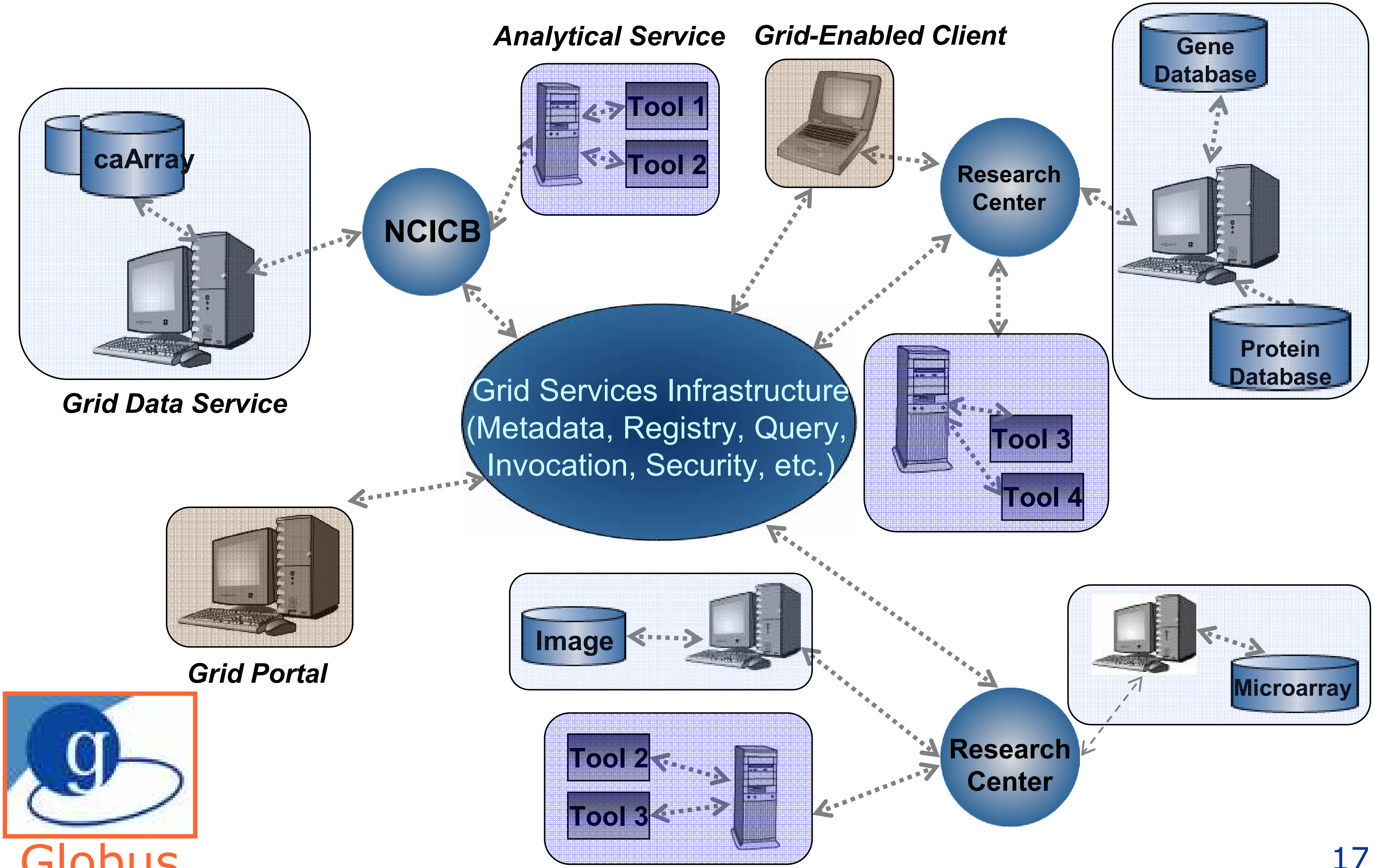

▲ Cancer Center (8)
● Clinical Cancer Center (14)
■ Comprehensive Cancer Center (39)
◆ Planning Grant (7)

Data Integration!

Globus



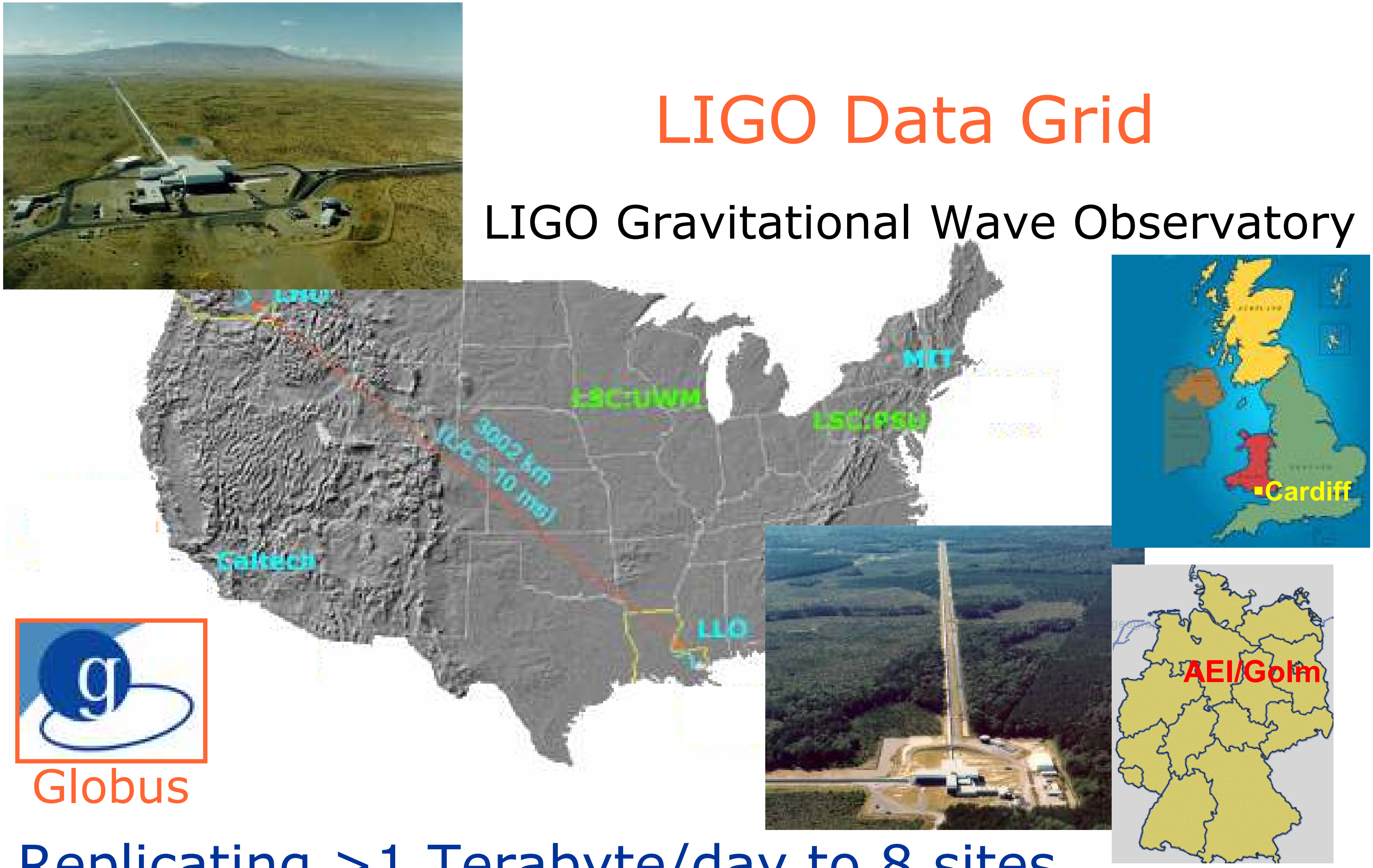
caBIG Under the Covers



Globus

LIGO Data Grid

LIGO Gravitational Wave Observatory



Globus

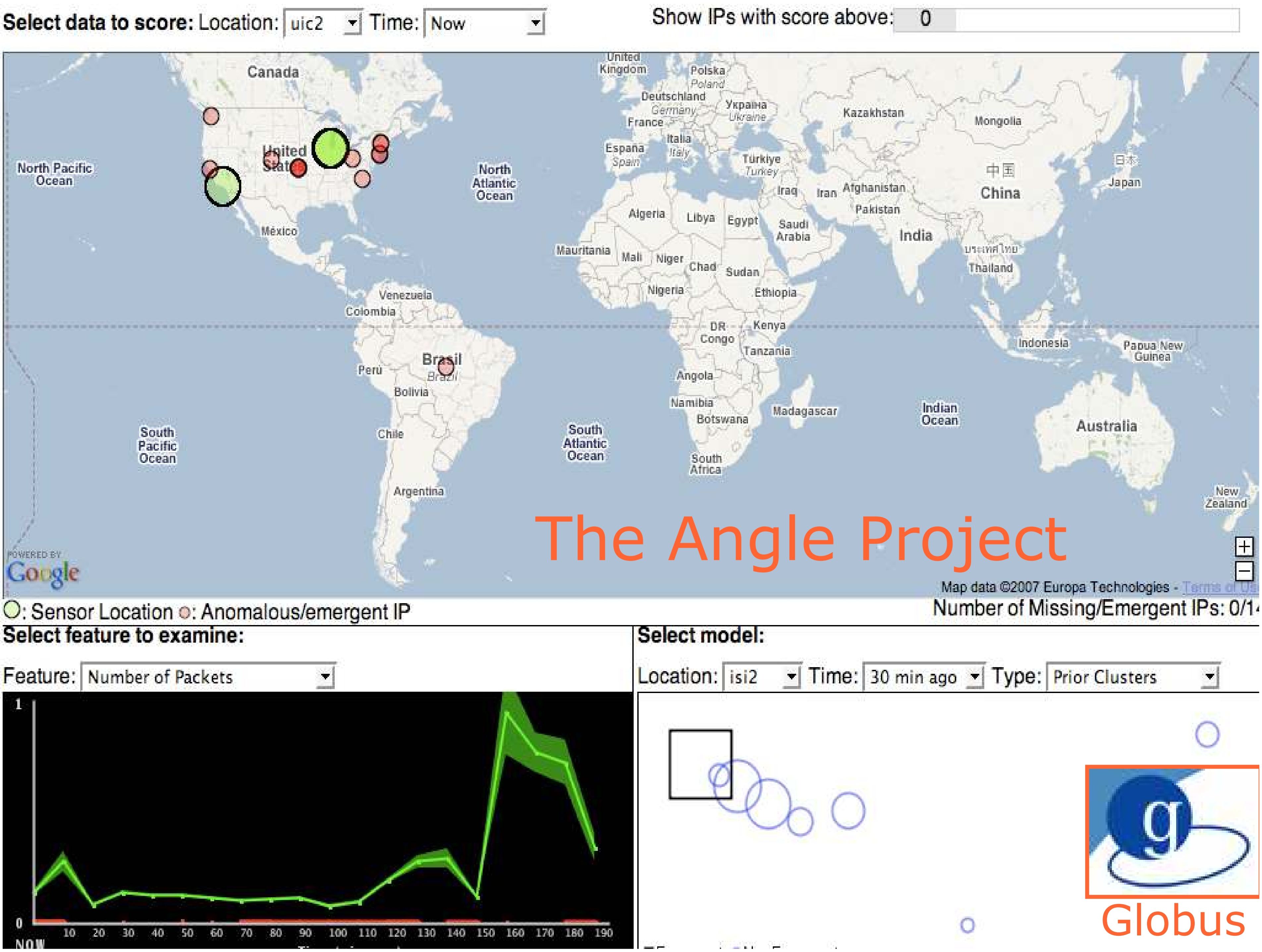
Replicating >1 Terabyte/day to 8 sites

>150 million replicas so far

MTBF = 1 month

www.globus.org/solutions

LSC



Select data to score: Location: uic2 Time: Now

Show IPs with score above: 0

○: Sensor Location ○: Anomalous/emergent IP

Number of Missing/Emergent IPs: 0/1

Select feature to examine:

Feature: Number of Packets

Select model:

Location: isi2 Time: 30 min ago Type: Prior Clusters

The Angle Project

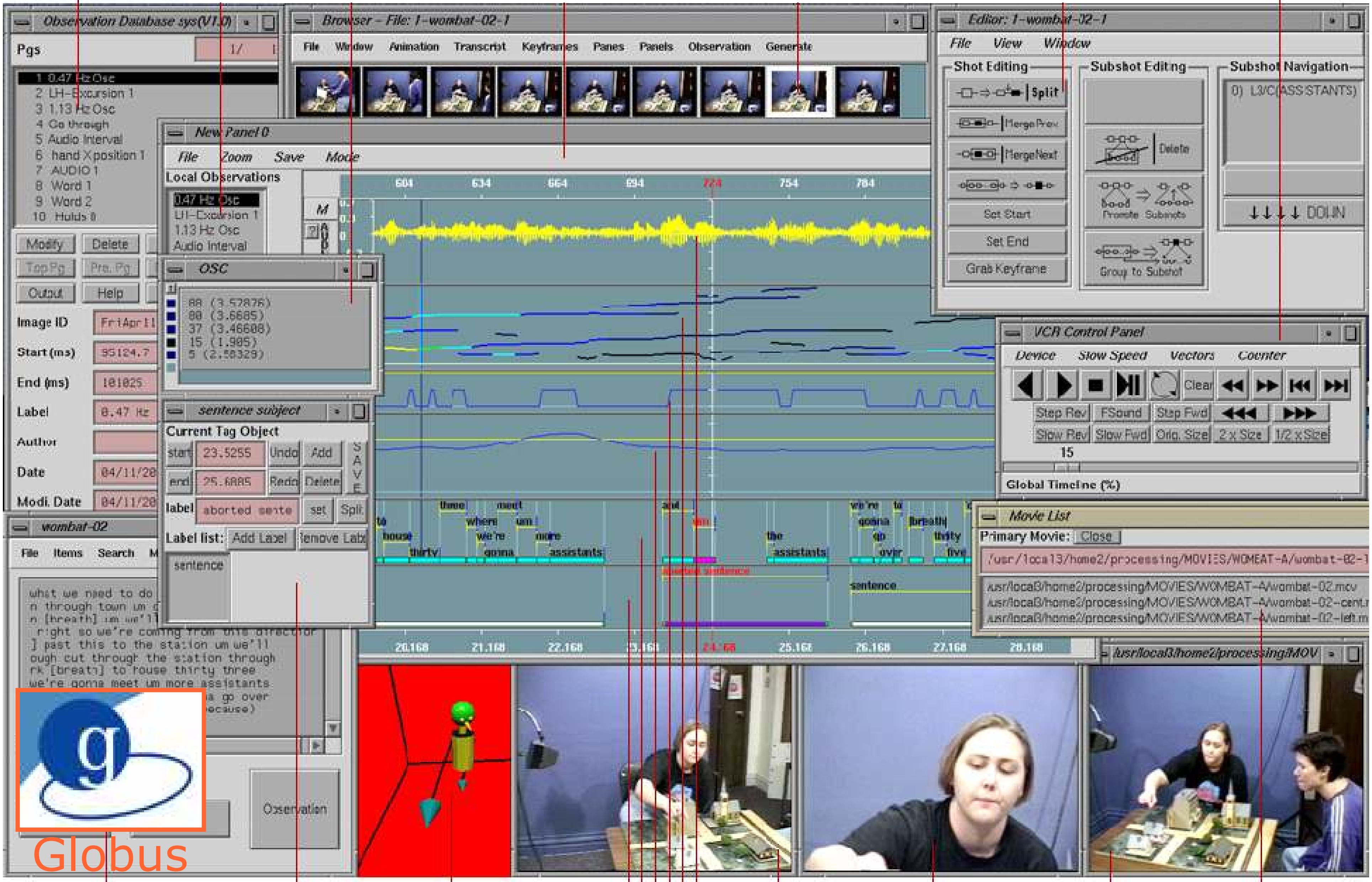
Globus

Social Informatics Data Grid

Global Observation Database (View)

G
C

VCR-Style Control Panel



1 0.47 Hz Osc
2 LH-Excursion 1
3 1.13 Hz Osc
4 Go through
5 Audio Interval
6 hand Xposition 1
7 AUDIO 1
8 Word 1
9 Word 2
10 Holds 8

Image ID: FriApril
Start (ms): 95124.7
End (ms): 101025
Label: 0.47 Hz
Author:
Date: 04/11/20
Modi. Date: 04/11/20

Local Observations:
0.47 Hz Osc
LH-Excursion 1
1.13 Hz Osc
Audio Interval

88 (3.57876)
88 (3.6685)
37 (3.46608)
15 (1.985)
5 (2.58329)

Current Tag Object
start 23.5255
end 25.6885
label aborted sente
Label list:
sentence

```
what we need to do
n through town um c
n [breath] um we'll
 right so we're coming from this director
] past this to the station um we'll
ough cut through the station through
rk [breath] to house thirty three
we're gonna meet um more assistants
a go over
ecause)
```

0) L3/C(ASSISTANTS)

15

Primary Movie: /usr/local3/home2/processing/MOVIES/WOMBAT-A/wombat-02-1

/usr/local3/home2/processing/MOVIES/WOMBAT-A/wombat-02.mov
/usr/local3/home2/processing/MOVIES/WOMBAT-A/wombat-02-cent.
/usr/local3/home2/processing/MOVIES/WOMBAT-A/wombat-02-left.m

/usr/local3/home2/processing/MOV

Globus

Animated Text Transcript (Paragraph Representation)

Tag Transcript Editor

Animated Avatar Representation

Animated Graph Panes

Video Displays

Video List

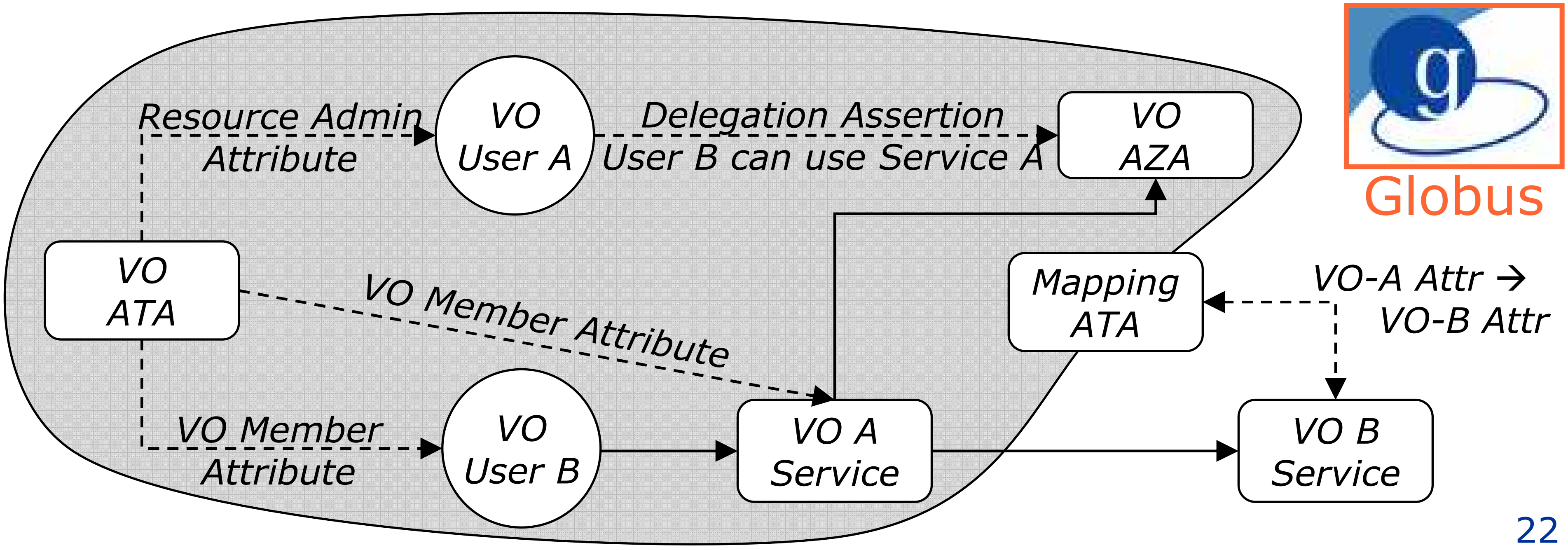
Bennett Berthenthal et al., www.sidgrid.org

A Few Example Research Themes

- Service discovery, composition, provisioning
  - SOA, virtualization, cloud computing, …
- Large-scale (distributed) computation
  - E.g., Swift, Kepler, Taverna
- Provenance
  - E.g., “Provenance Challenge”
- “Virtual organizations”
  - E.g., attribute-based authorization, trust
- Integration of physical systems
  - Optimization of end-to-end workflows

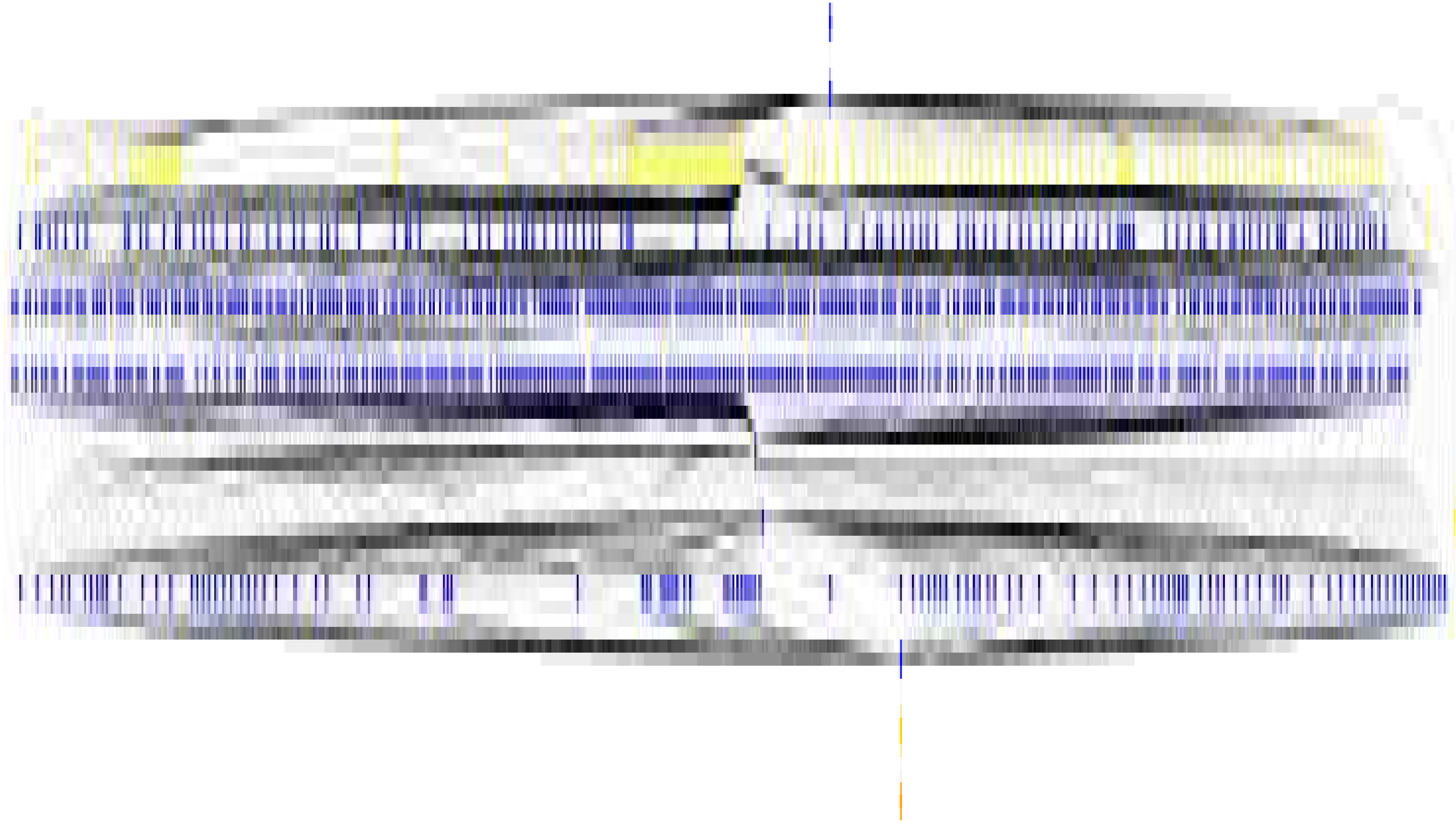
Security Services for Virtual Organization Policy

- Attribute Authority (ATA)
  - Issue signed attribute assertions (incl. identity, delegation & mapping)
- Authorization Authority (AZA)
  - Decisions based on assertions & policy

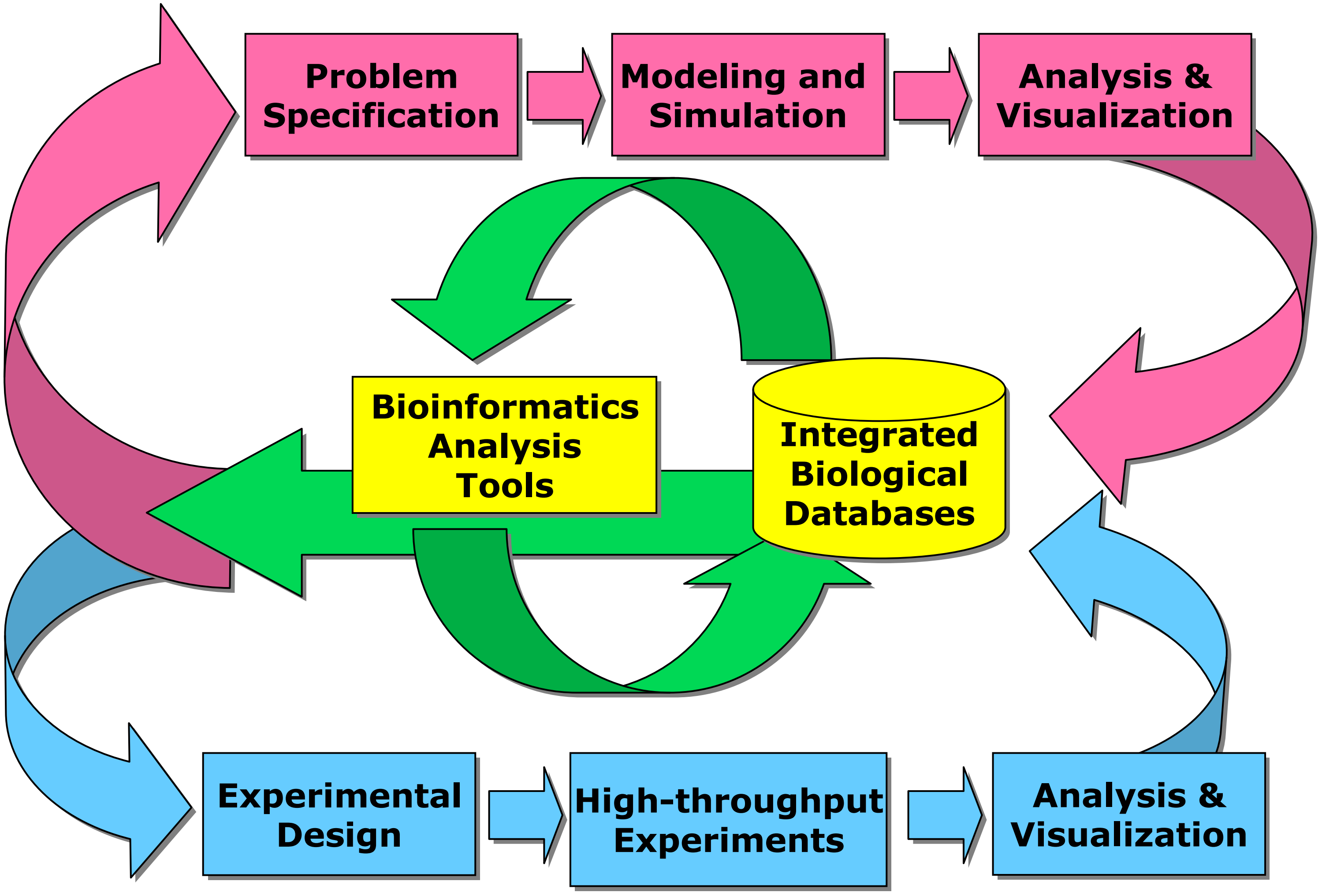


Globus

Swift (www.ci.uchicago.edu/swift)



An Integrated View of Modeling, Simulation, Experiment, & Informatics



Problem Specification

Modeling and Simulation

Analysis & Visualization

Bioinformatics Analysis Tools

Integrated Biological Databases

Experimental Design

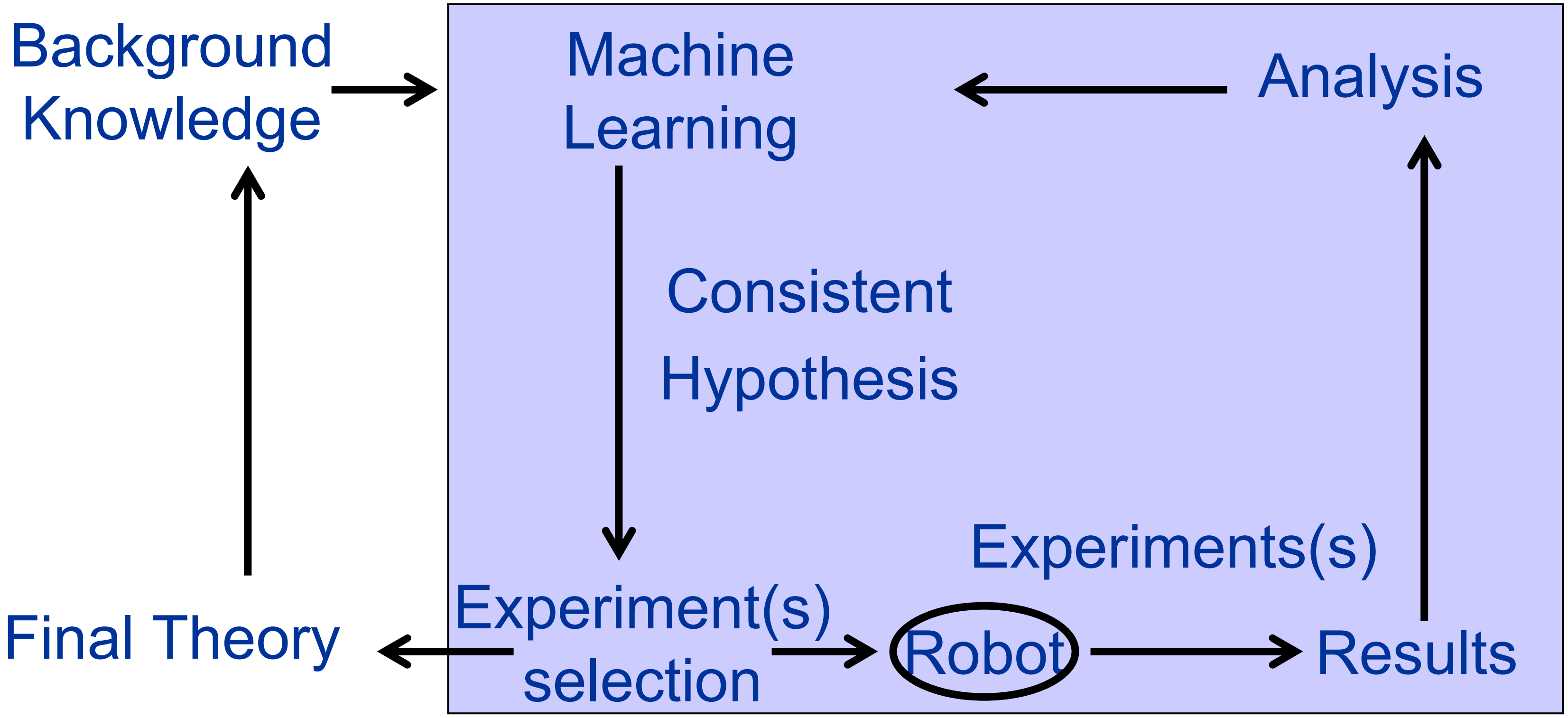
High-throughput Experiments

Analysis & Visualization

Robot Scientist

Biomek 200

"The robot scientist project aims to develop a computer system capable of originating its own experiments, physically doing them, interpreting the results, & then repeating the cycle."



Stephen Muggleton, Ross King et al., UK

Team Science meets Data Deluge