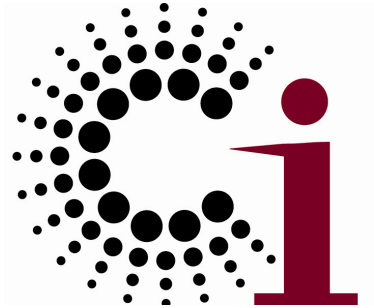
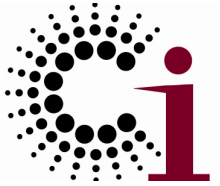
Why Computer Science is Fundamental to Almost Everything

Ian Foster



Computation Institute

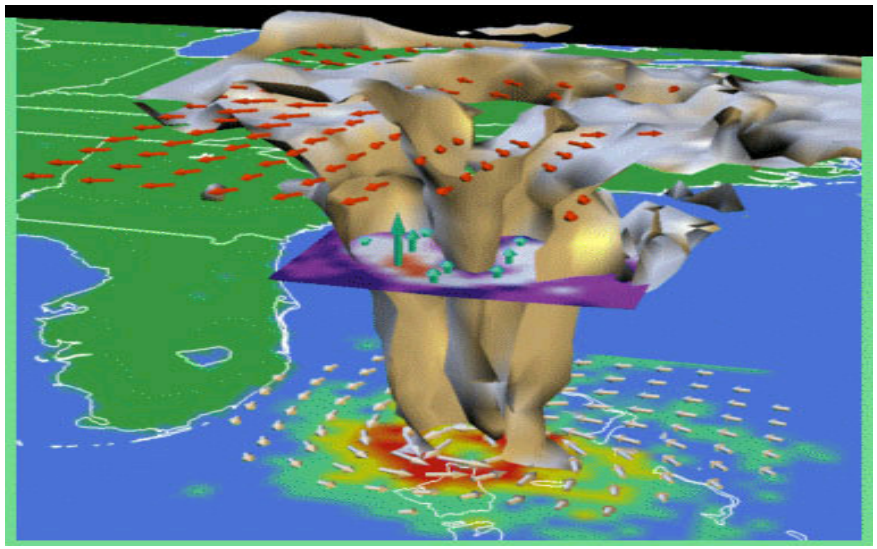
Argonne National Lab & University of Chicago



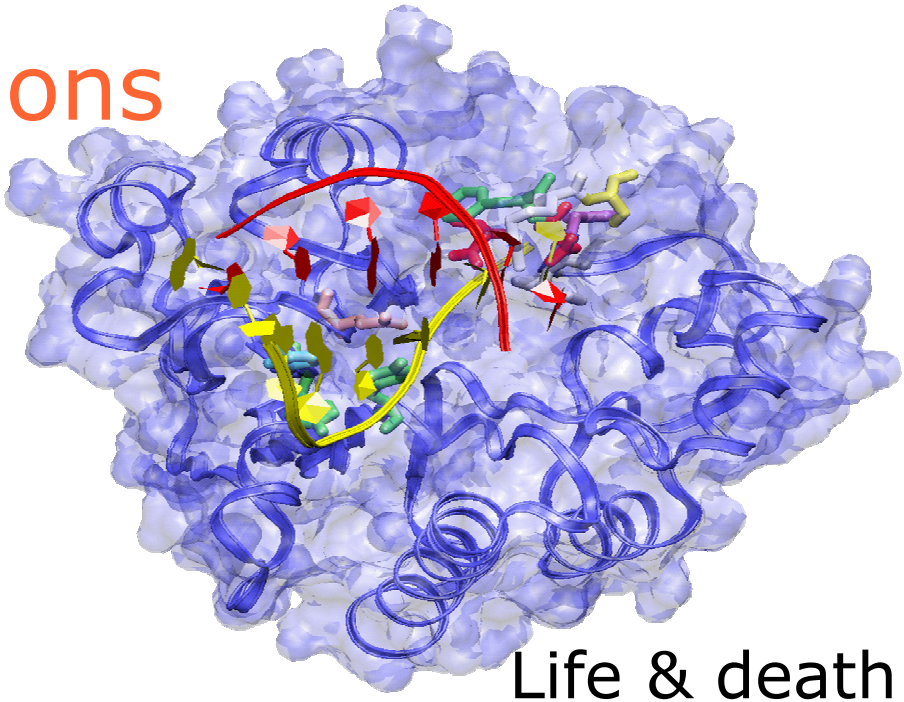
"Applied computer science is now playing the role that mathematics did from the 17th through the 20th centuries: providing an orderly, formal framework & exploratory apparatus for other sciences."

—George Djorgovski

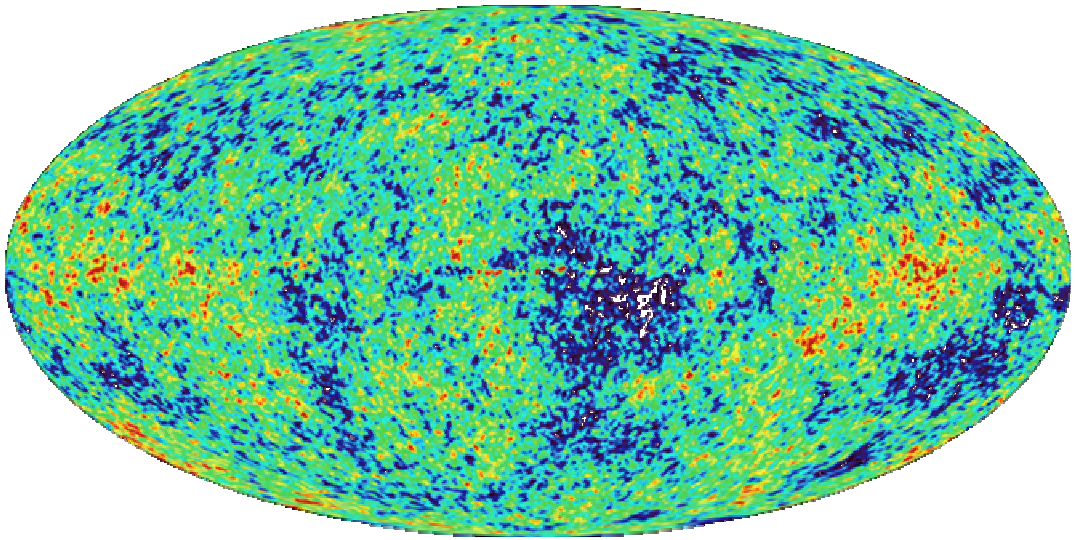
The Big Questions



Future of the planet



Life & death



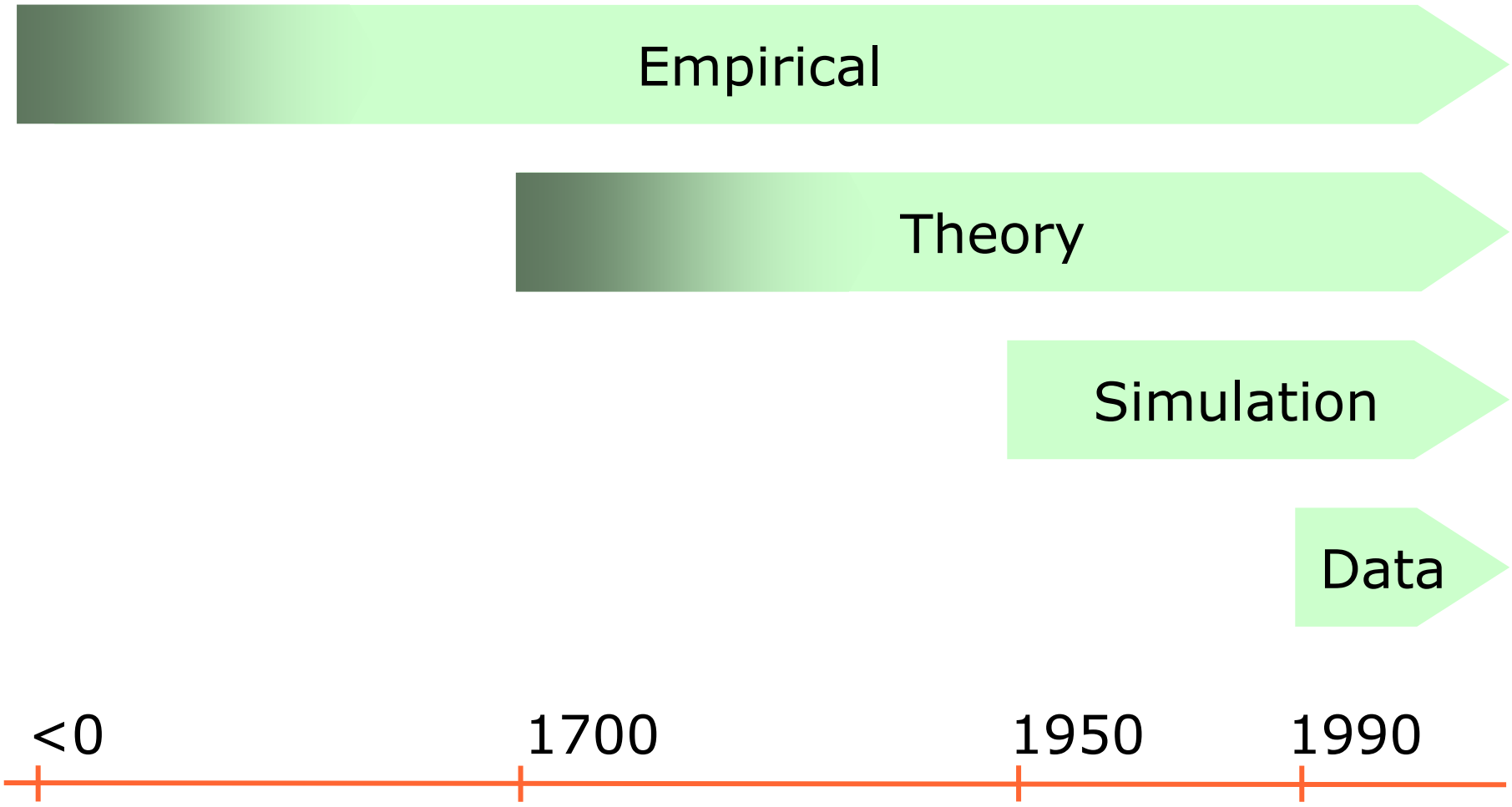
Nature of the universe

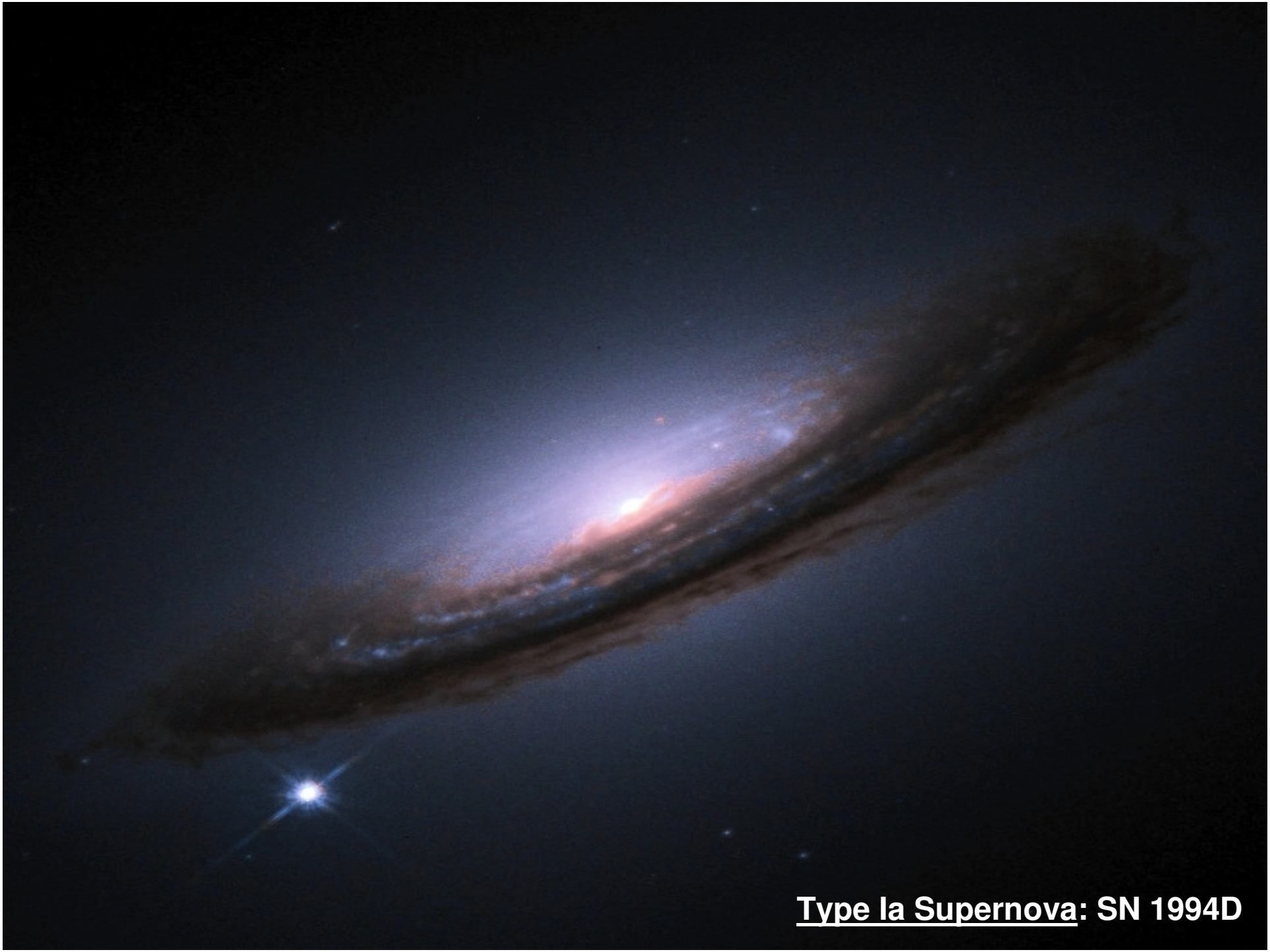
Consciousness

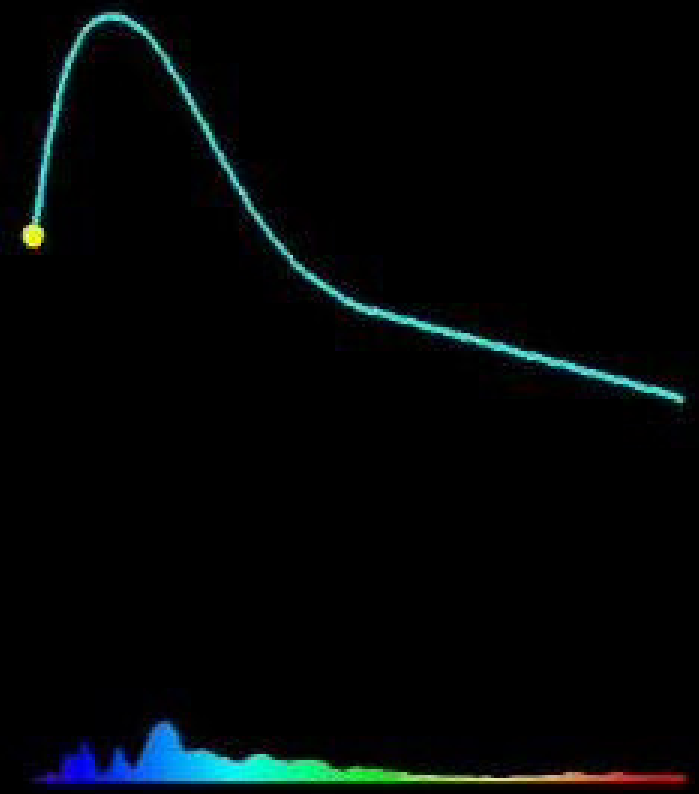
The Little Questions

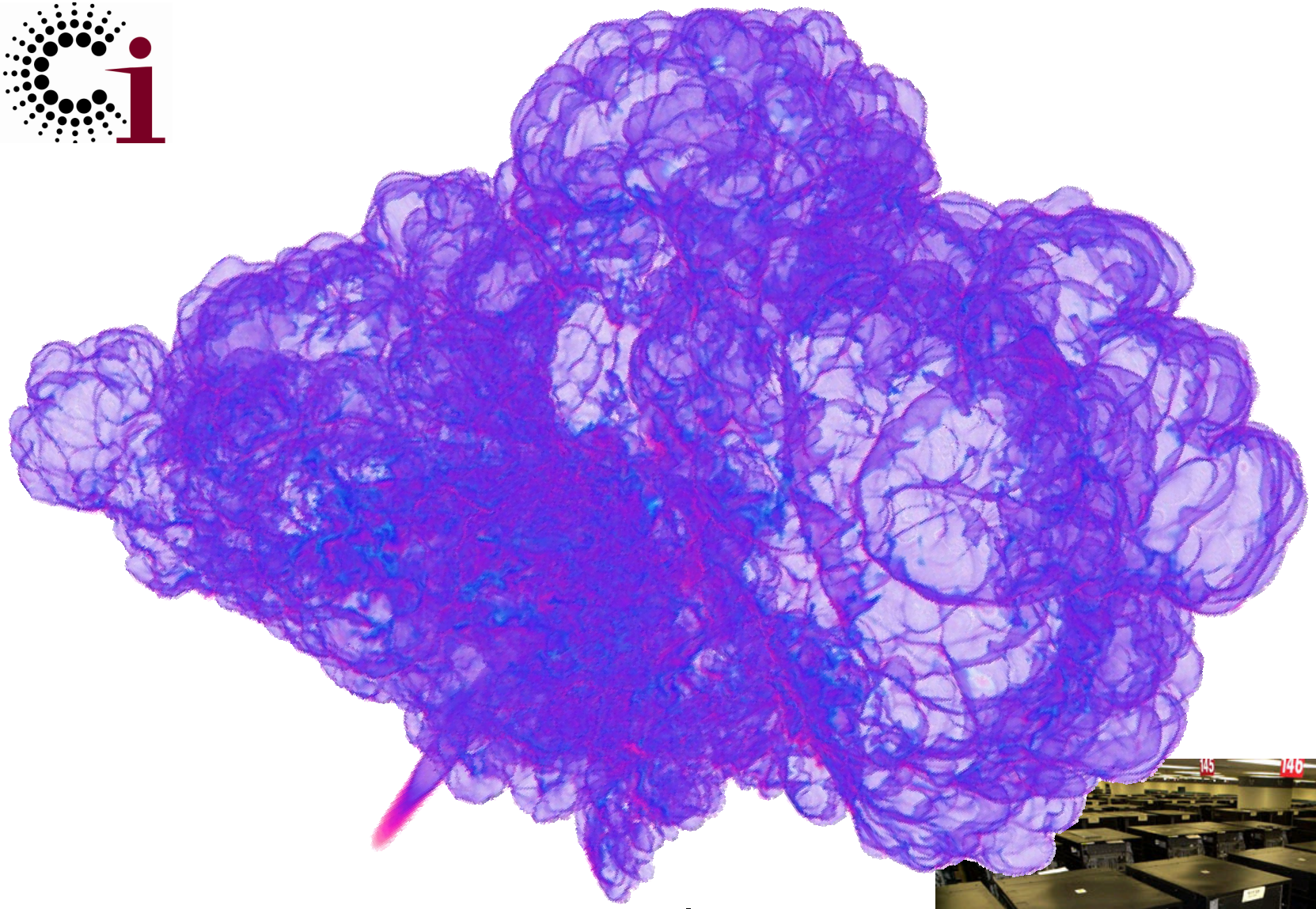
- Friends
- Sales
- Entertainment
- Spelling
- Parking

How Do We Answer Them?

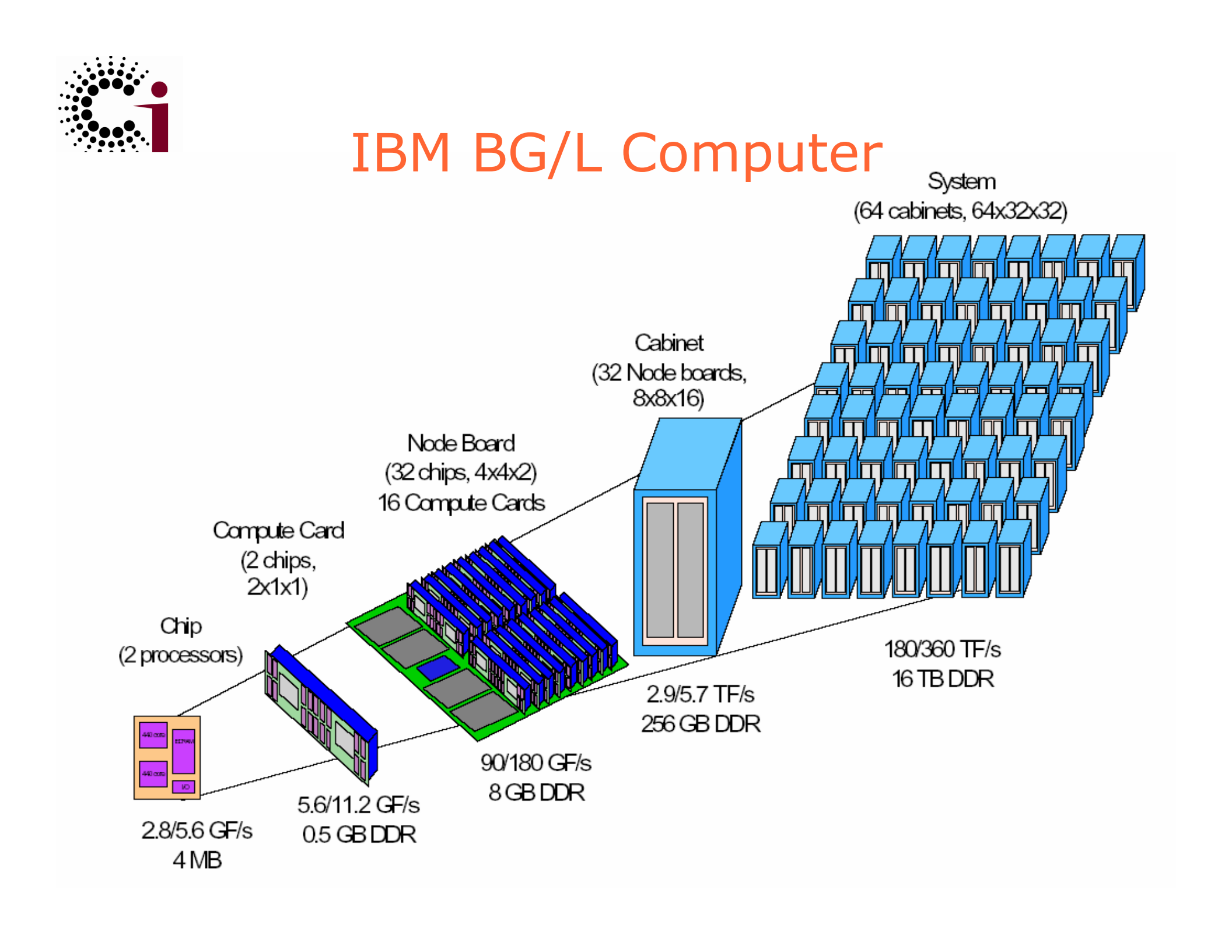



Type Ia Supernova: SN 1994D



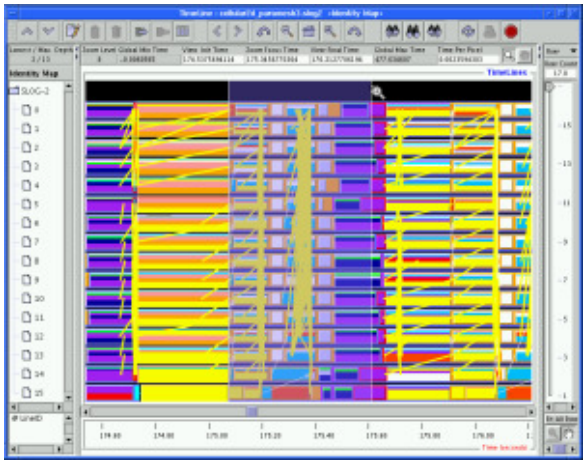


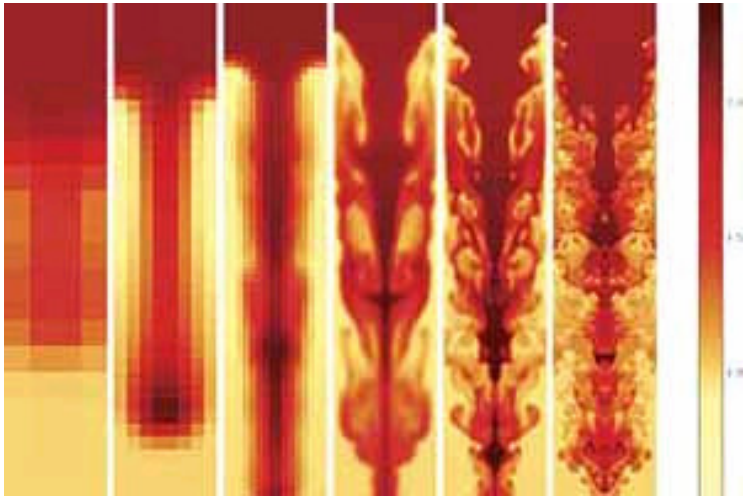
Type Ia Supernova Explosion:
Gravitationally Confined Detonation
(Calder, Plewa, Vladimirova, Lamb, and Truran, 2004)


IBM BG/L Computer
System
(64 cabinets, 64x32x32)
Cabinet
(32 Node boards,
8x8x16)
Node Board
(32 chips, 4x4x2)
16 Compute Cards
Compute Card
(2 chips,
2x1x1)
Chip
(2 processors)
2.8/5.6 GF/s
4 MB
5.6/11.2 GF/s
0.5 GB DDR
90/180 GF/s
8 GB DDR
2.9/5.7 TF/s
256 GB DDR
180/360 TF/s
16 TB DDR

Challenges Include …

- Multi-scale, multi-physics modeling
  - Adaptive mesh refinement
  - Component architectures
- Scaling to 100K+ processors
  - Scalable parallel libraries
  - Parallel operating systems
- Understand & validating results
  - Visualization, data mining
  - Quantifying uncertainty



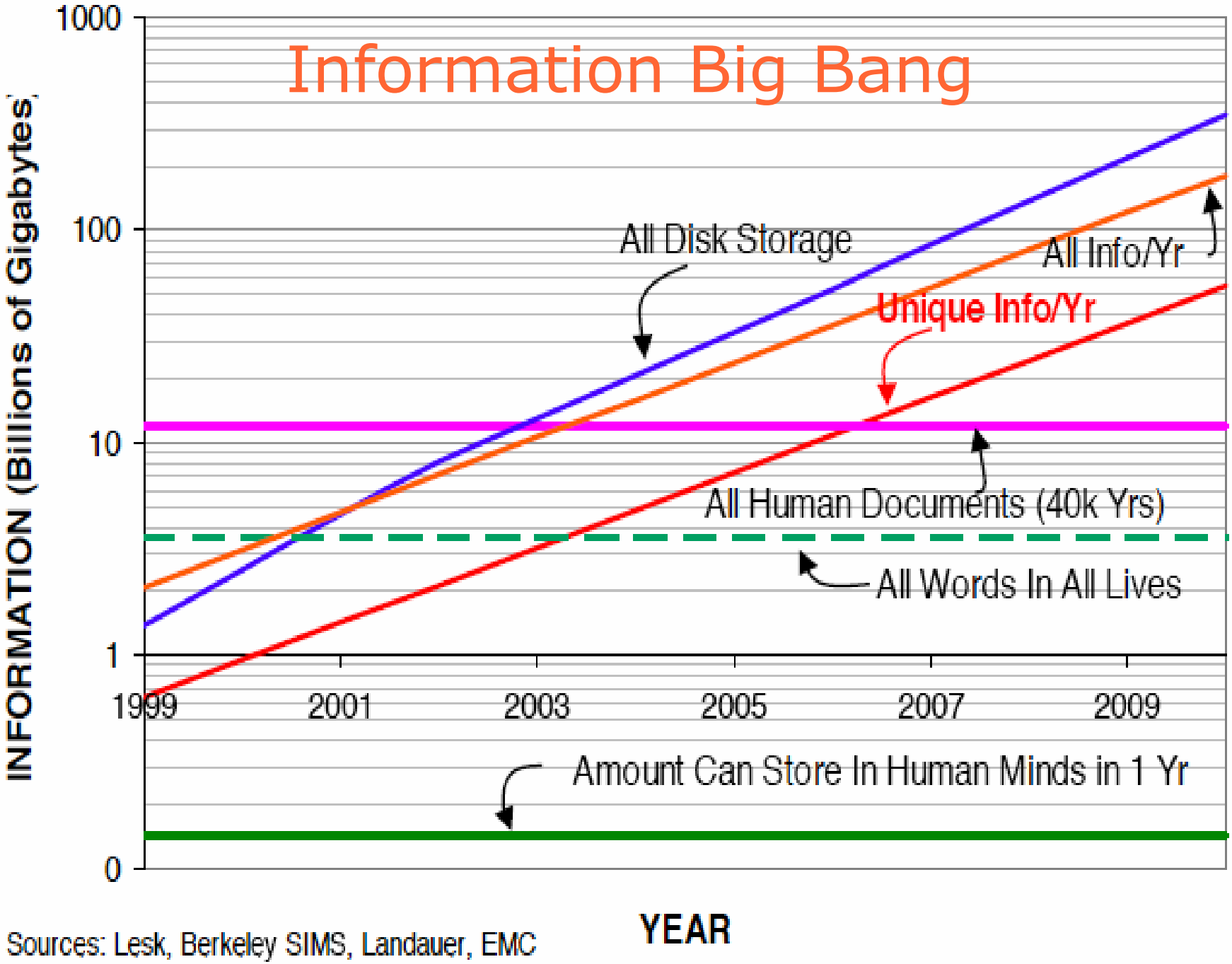


How Much Data?

- In 2006:
  - The world created 161 exabytes (1.6×10^{20} bytes) of digital data
  - There were one billion devices able to capture digital images

- By 2010:
  - Annual data output will reach one zettabyte (1×10^{21} bytes)

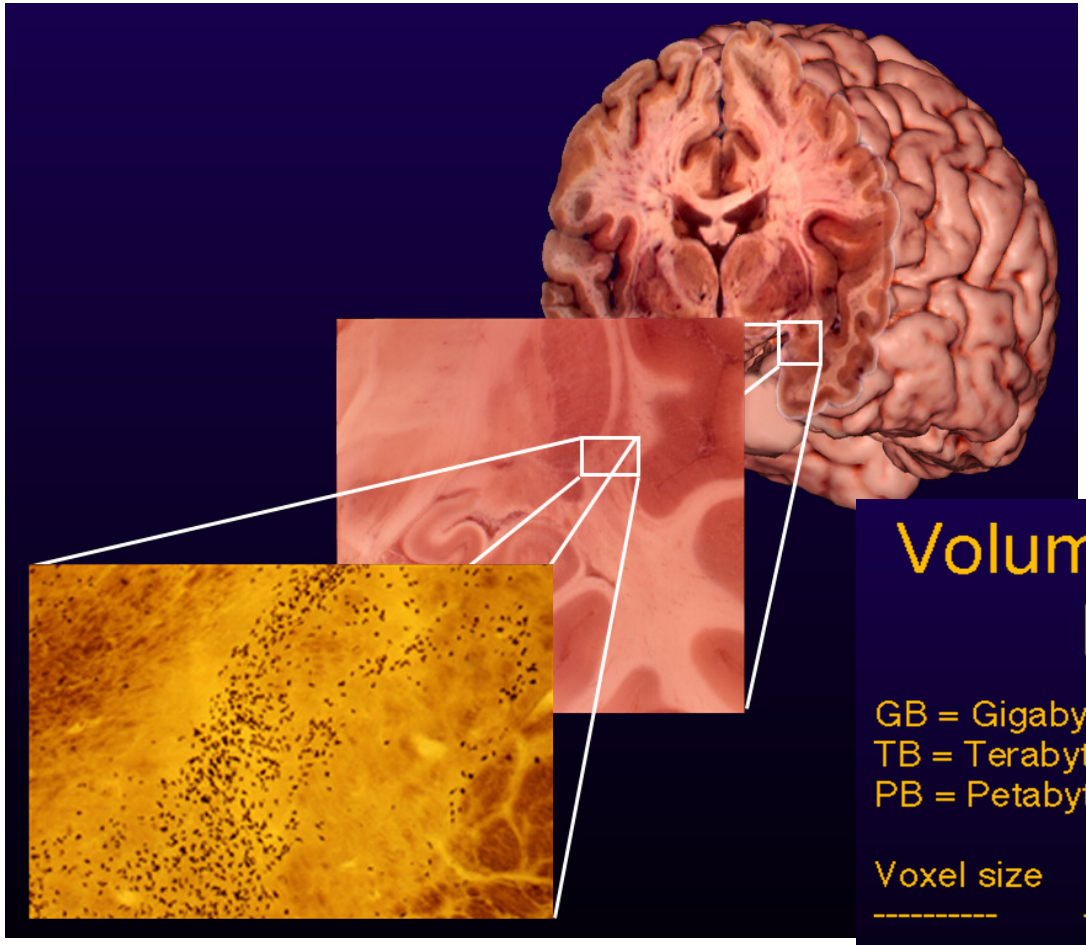
Source: IDC, 2007


Information Big Bang
INFORMATION (Billions of Gigabytes)
1000
100
10
1
0
1999
2001
2003
2005
2007
2009
YEAR
All Disk Storage
All Info/Yr
Unique Info/Yr
All Human Documents (40k Yrs)
All Words In All Lives
Amount Can Store In Human Minds in 1 Yr
Sources: Lesk, Berkeley SIMS, Landauer, EMC

A Data Deluge

A Brain is a Lot of Data!

(Mark Ellisman, UCSD)



And comparisons must be made among many

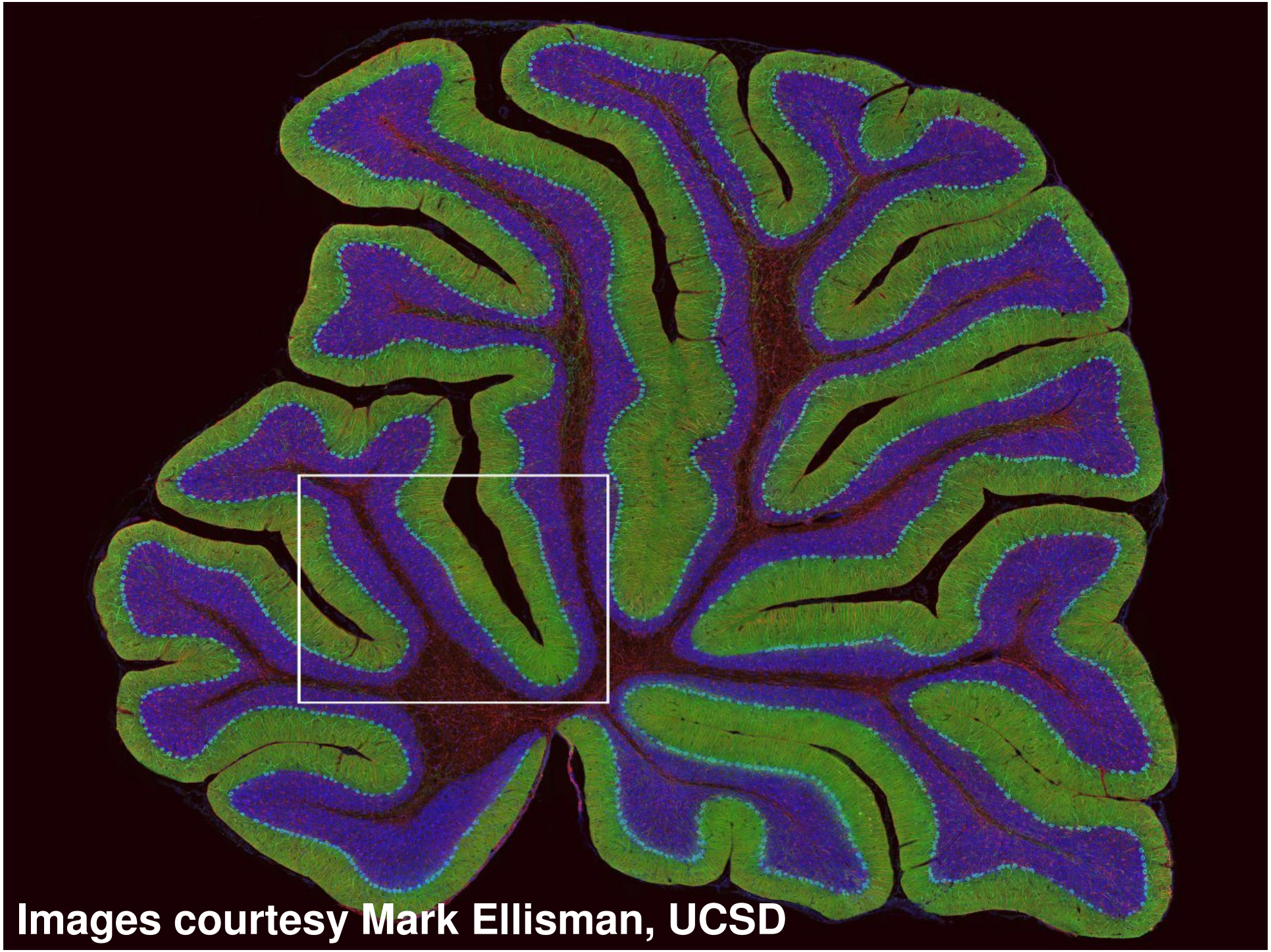
Volume sizes by resolution - brain = 1500 cm^3

GB = Gigabyte = 10^9
TB = Terabyte = 10^{12}
PB = Petabyte = 10^{15}

Voxel size	B&W (1 B/p)	High res (2 B/p)	Color (3 B/p)
cm	1.5 KB	3 KB	4.5 KB
mm	1.5 MB	3 MB	4.5 MB
10 μm	1.5 TB	3 TB	4.5 TB
μm	1.5 PB	3 PB	4.5 PB

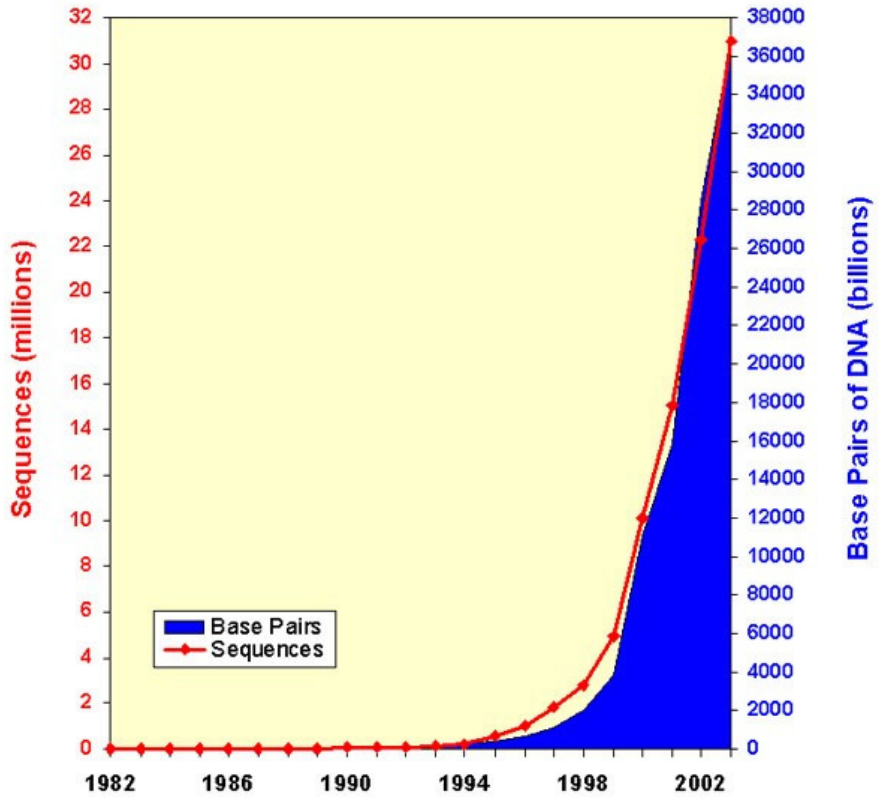
We need to get to one micron to know location of every cell. We are starting to get to 10 microns

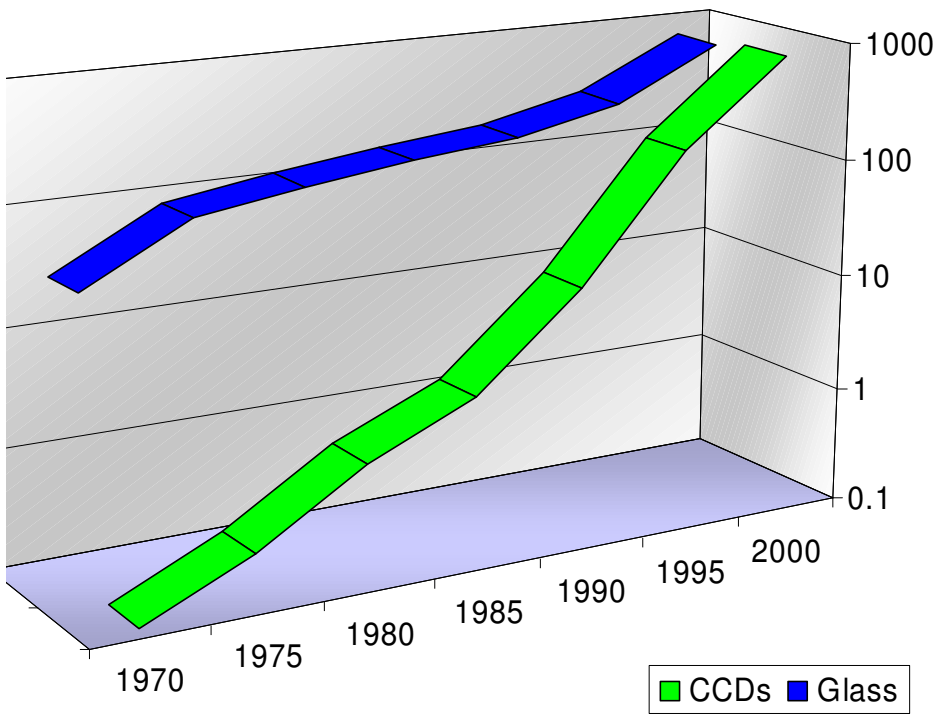
Images courtesy Mark Ellisman, UCSD


Images courtesy Mark Ellisman, UCSD


Images courtesy Mark Ellisman, UCSD


Images courtesy Mark Ellisman, UCSD





- Understanding increases **far** more slowly

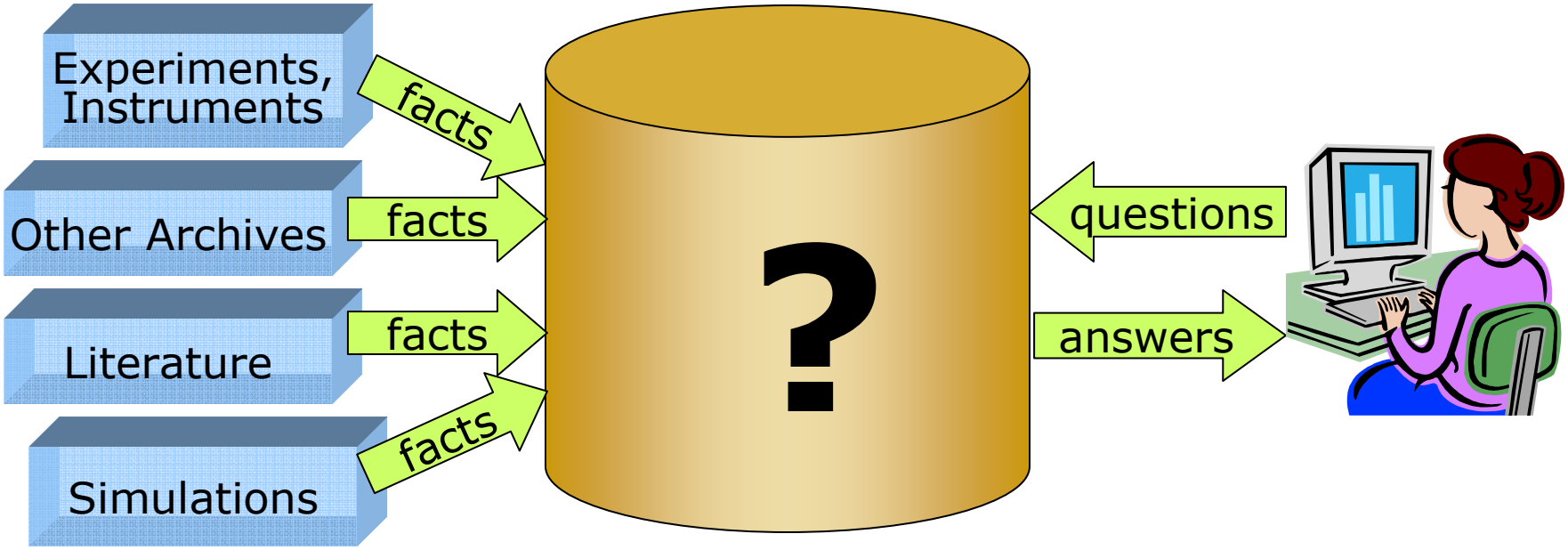
- Methodological bottlenecks?

  ➔ Improved technology

- Human limitations?

  ➔ AI-assisted discovery

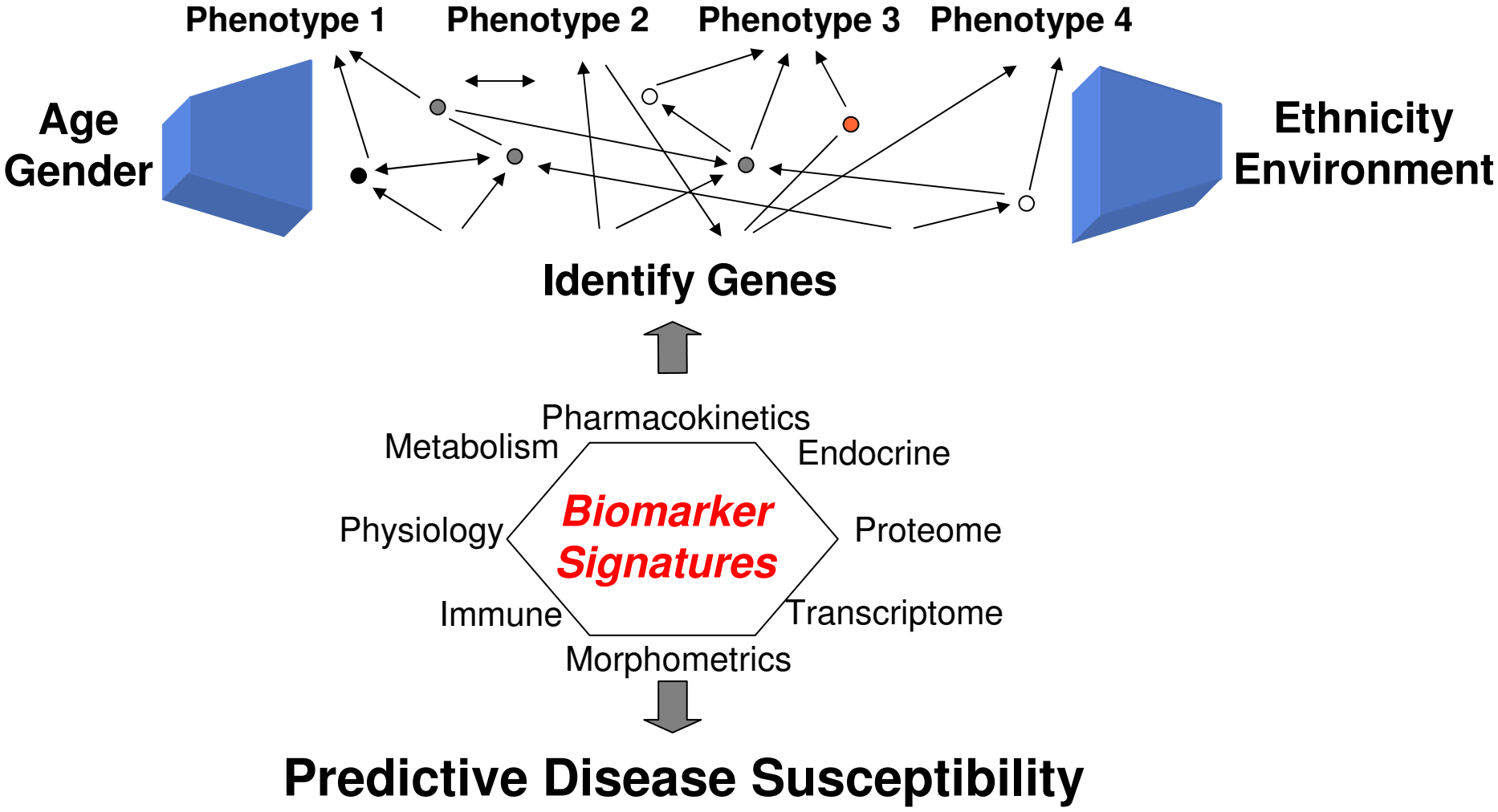
The Problem



- Data ingest
- Managing a petabyte
- Common schema
- How to organize it?
- How to *re*organize it?

- Data query & visualization
- Support/training
- Performance: interactivity, scale in data size, analysis complexity, demand

Evidence Integration: Genetics & Disease Susceptibility



Source: Terry Magnuson

GeneWays as an Info-Grinder

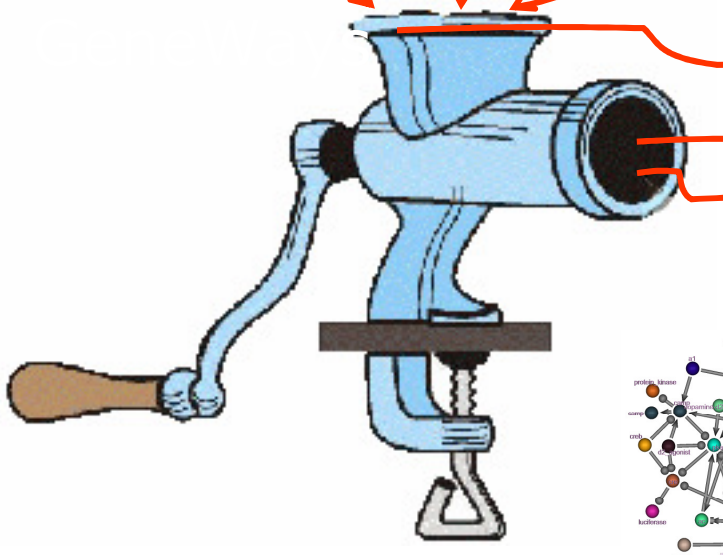
On-line Journals

Andrey Rzhetsky et al.,
U.Chicago

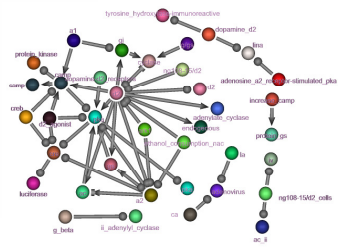
We are currently screening 250,000 journal articles…

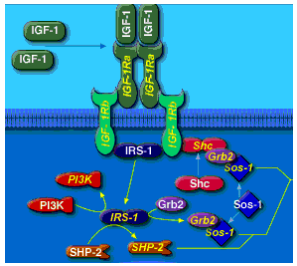
2.5M reasoning chains

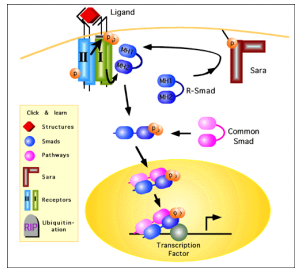
4M statements



Pathways

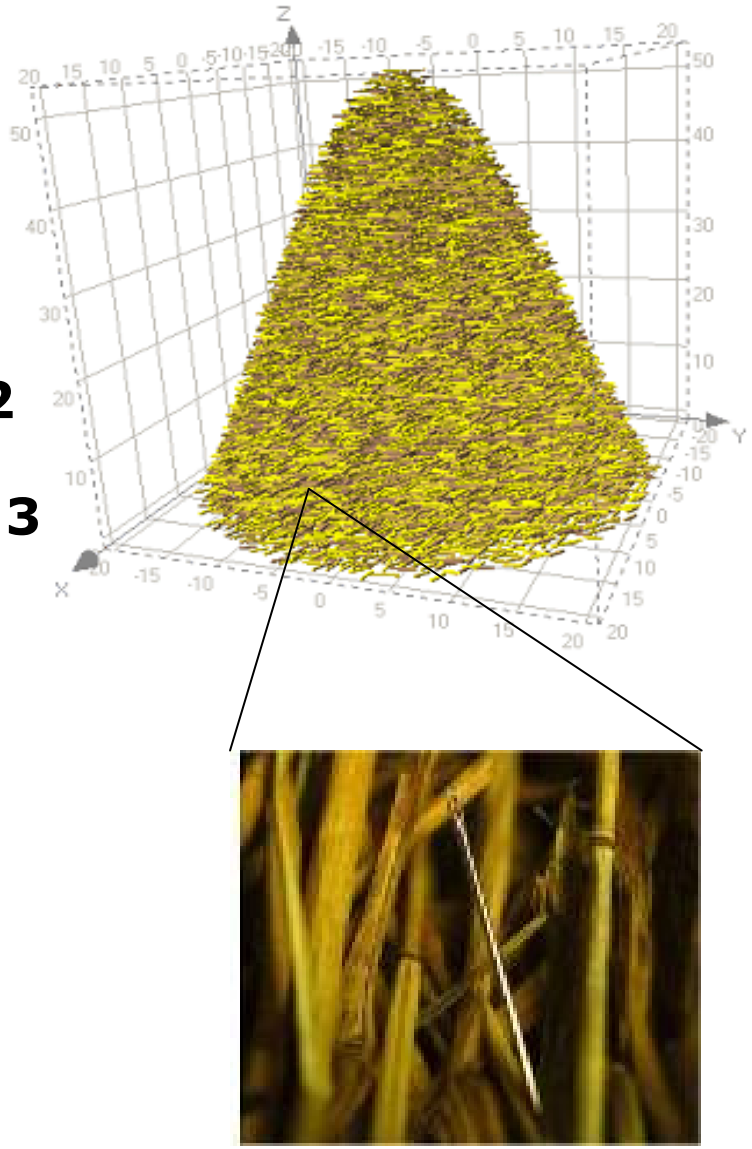






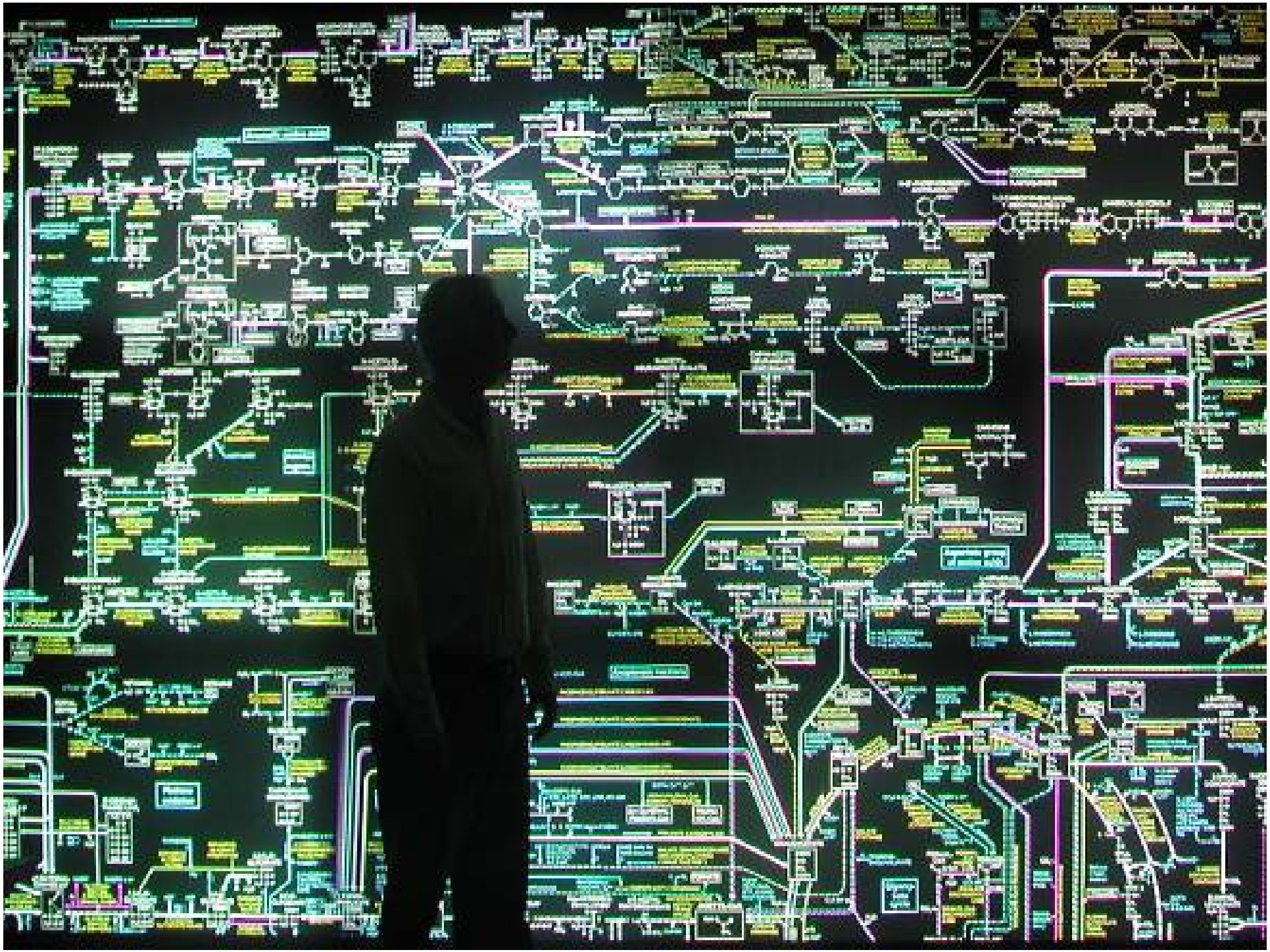
Data Analysis gets Fuzzy

- Global statistics?
  - Correlation functions: **N^2**
  - Likelihood techniques: **N^3**
- Best we can do is **N** or maybe **N logN**

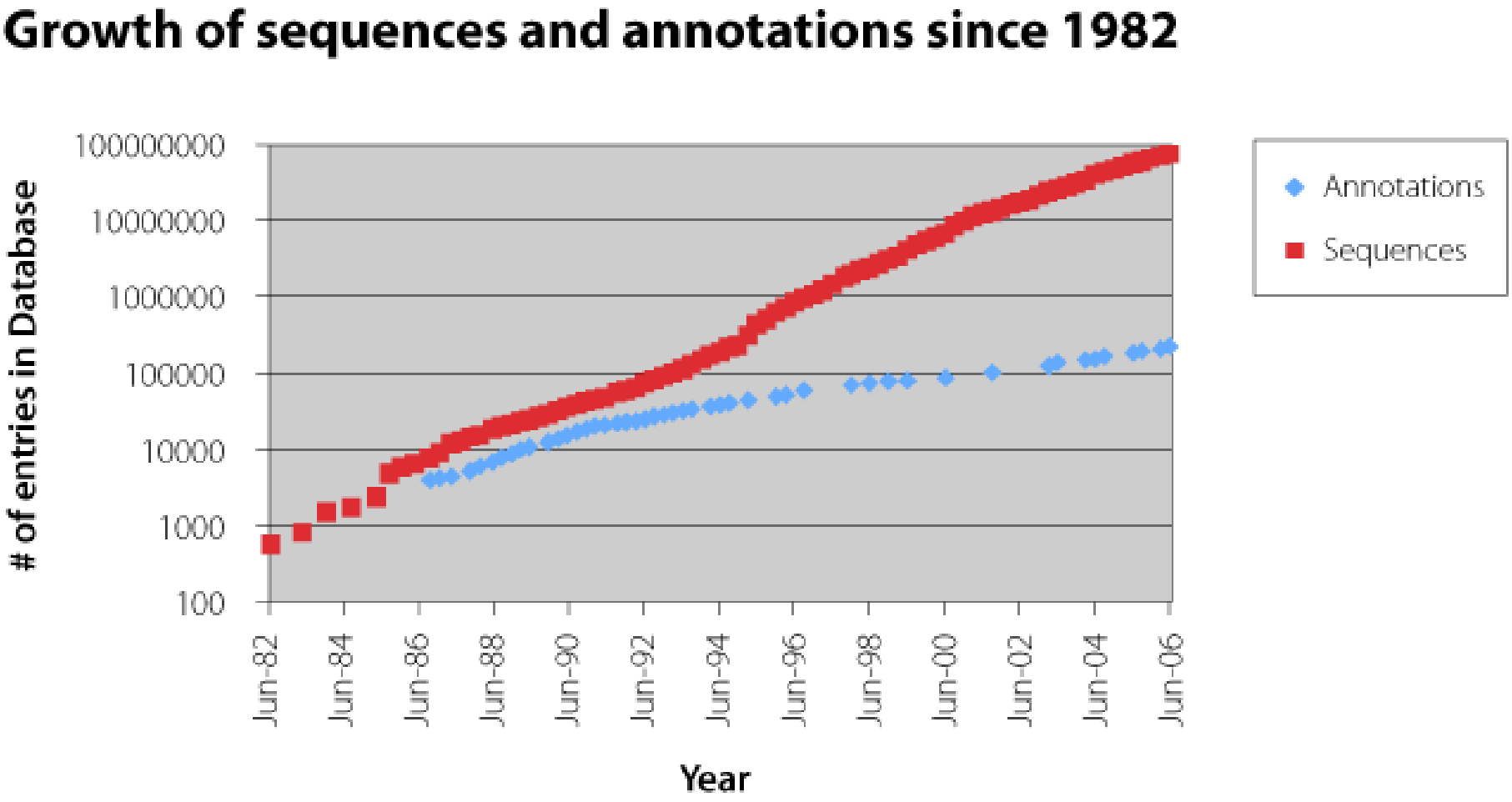


(scale approximate)

Haystack: Jim Gray/Alex Szalay



Growth of Sequences and Annotations since 1982

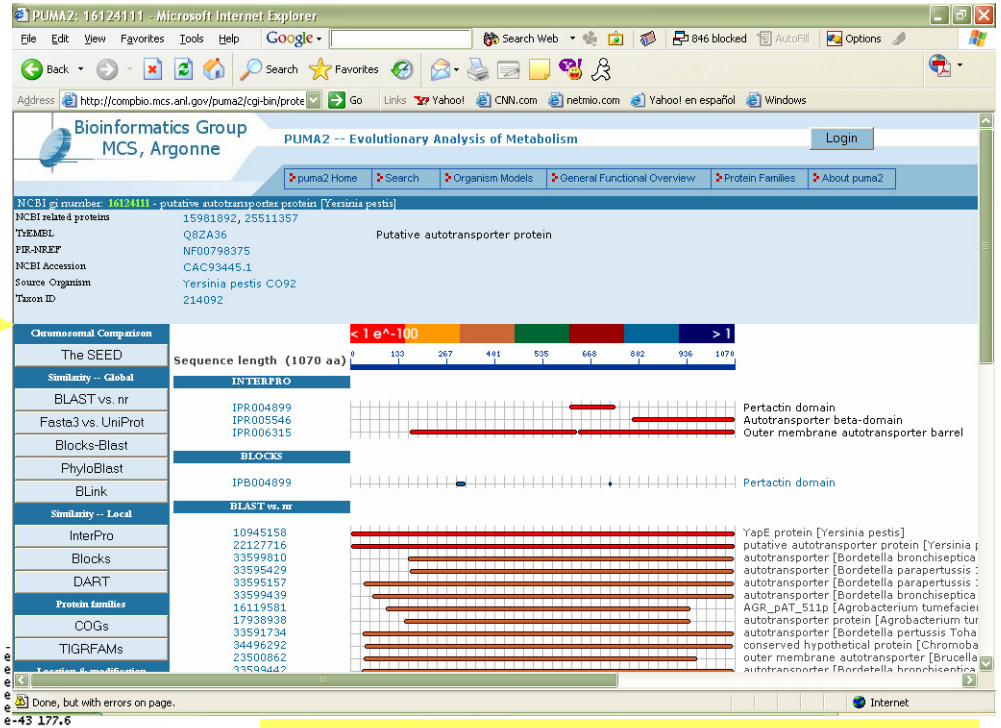


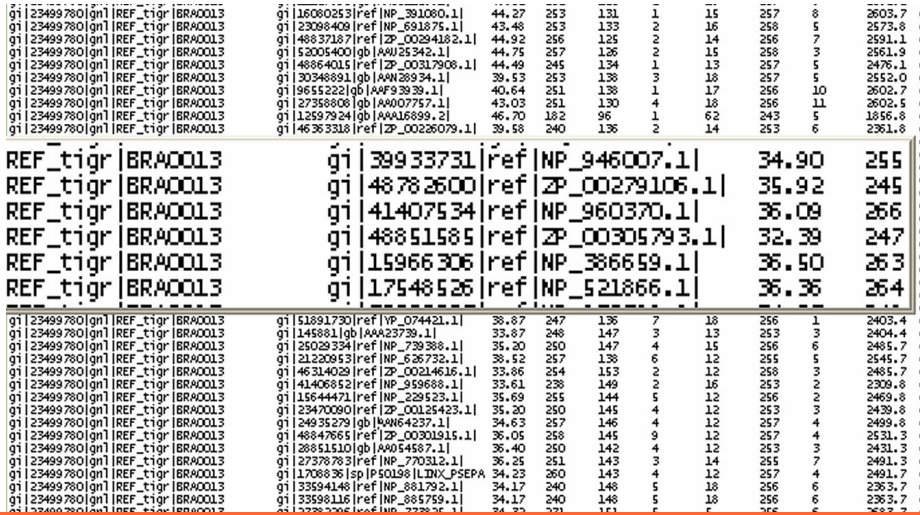
Folker Meyer, Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade, **CTWatch**, August 2006.

Production Science: Biology

Public PUMA Knowledge Base

Information about proteins analyzed against ~2 million gene sequences

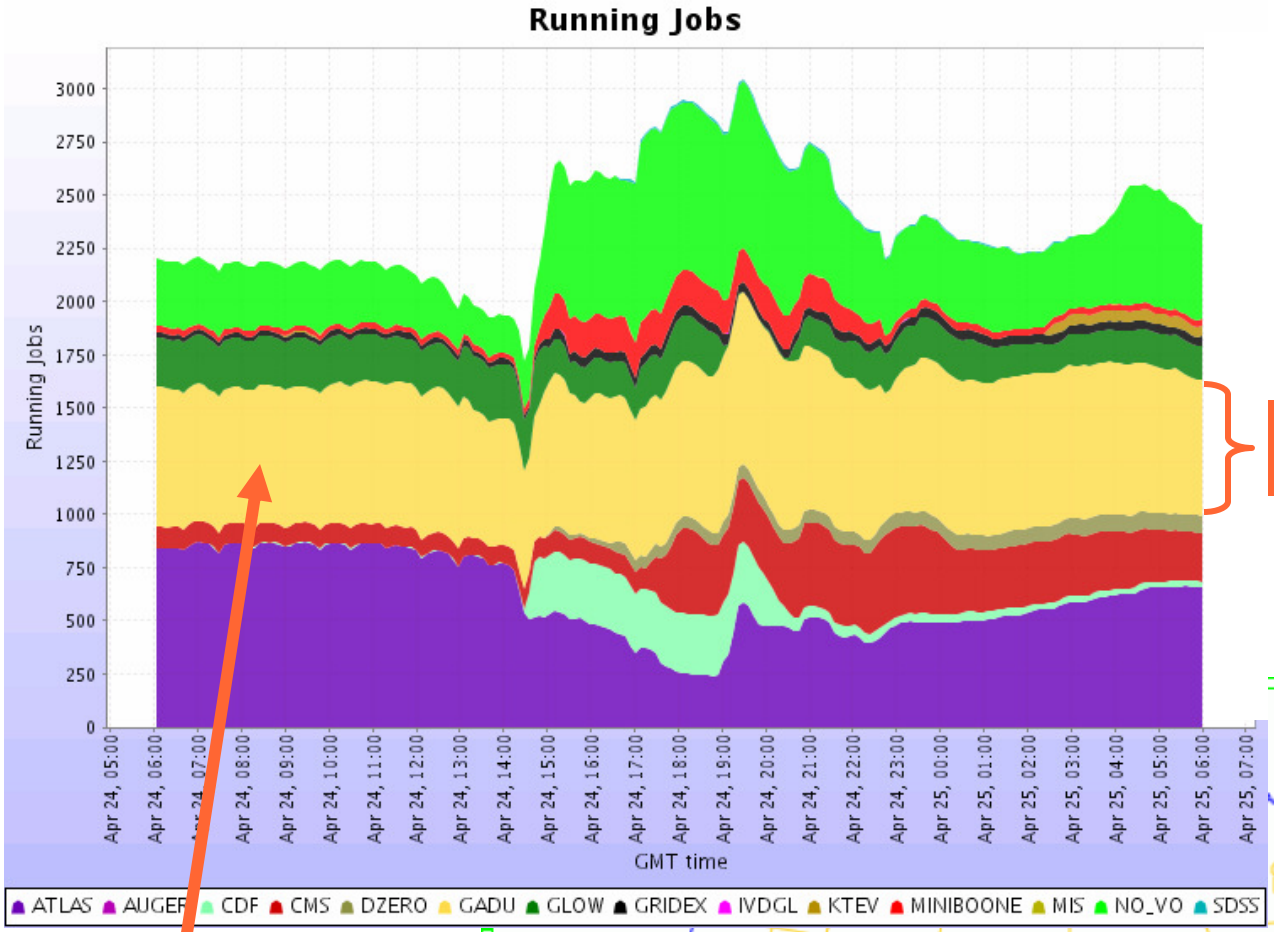




Back Office Analysis on Grid

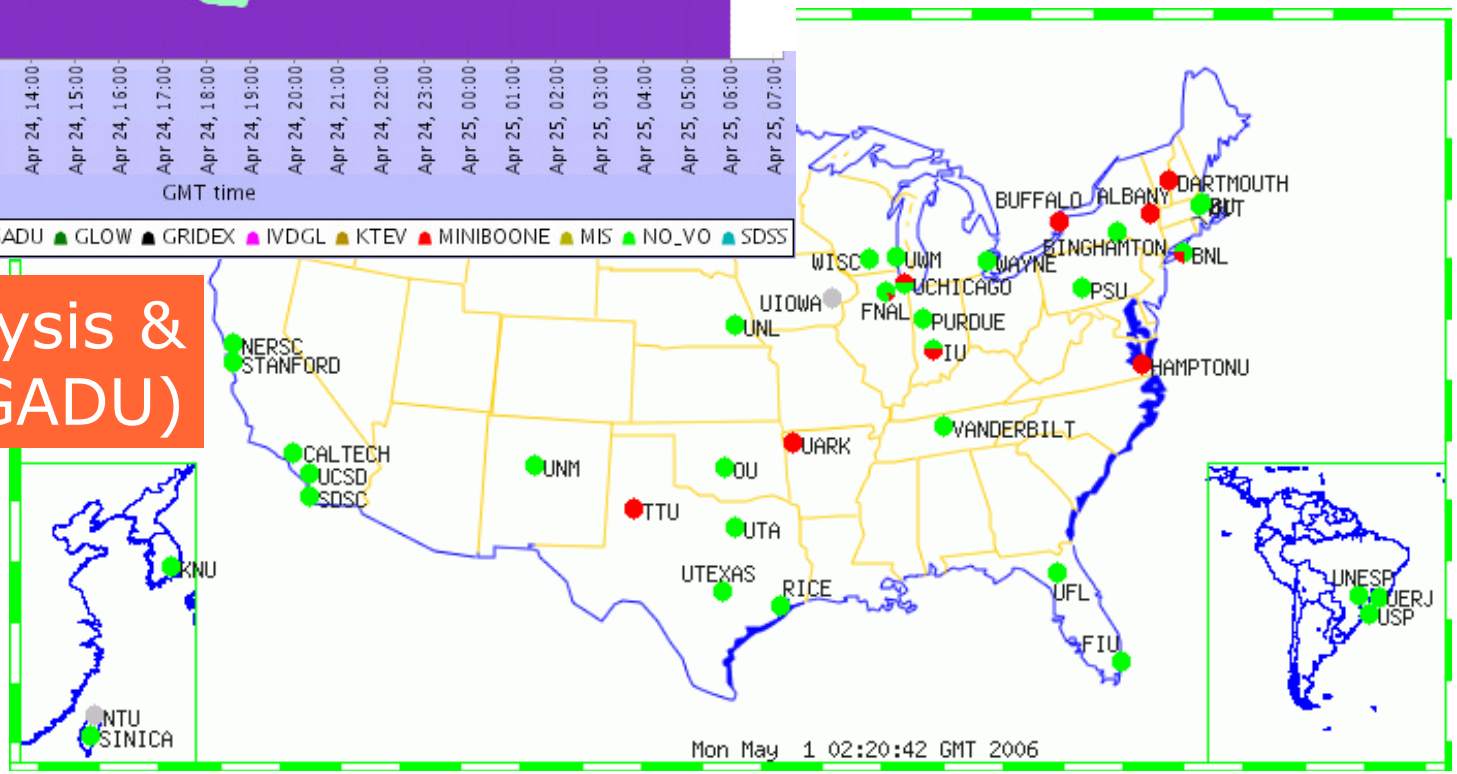
Millions of BLAST, BLOCKS, etc., on OSG and TeraGrid

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

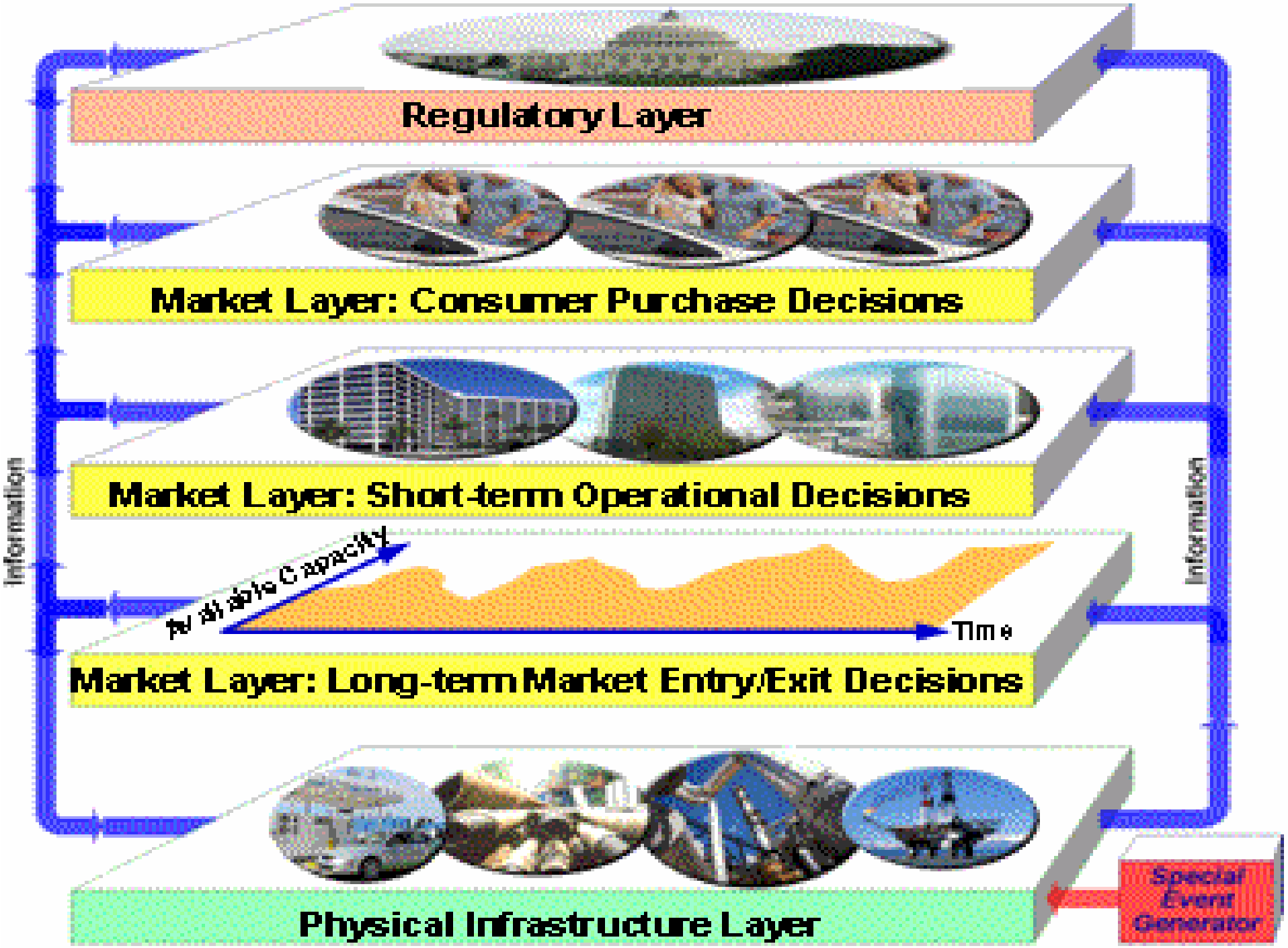


600-1000+ CPUs

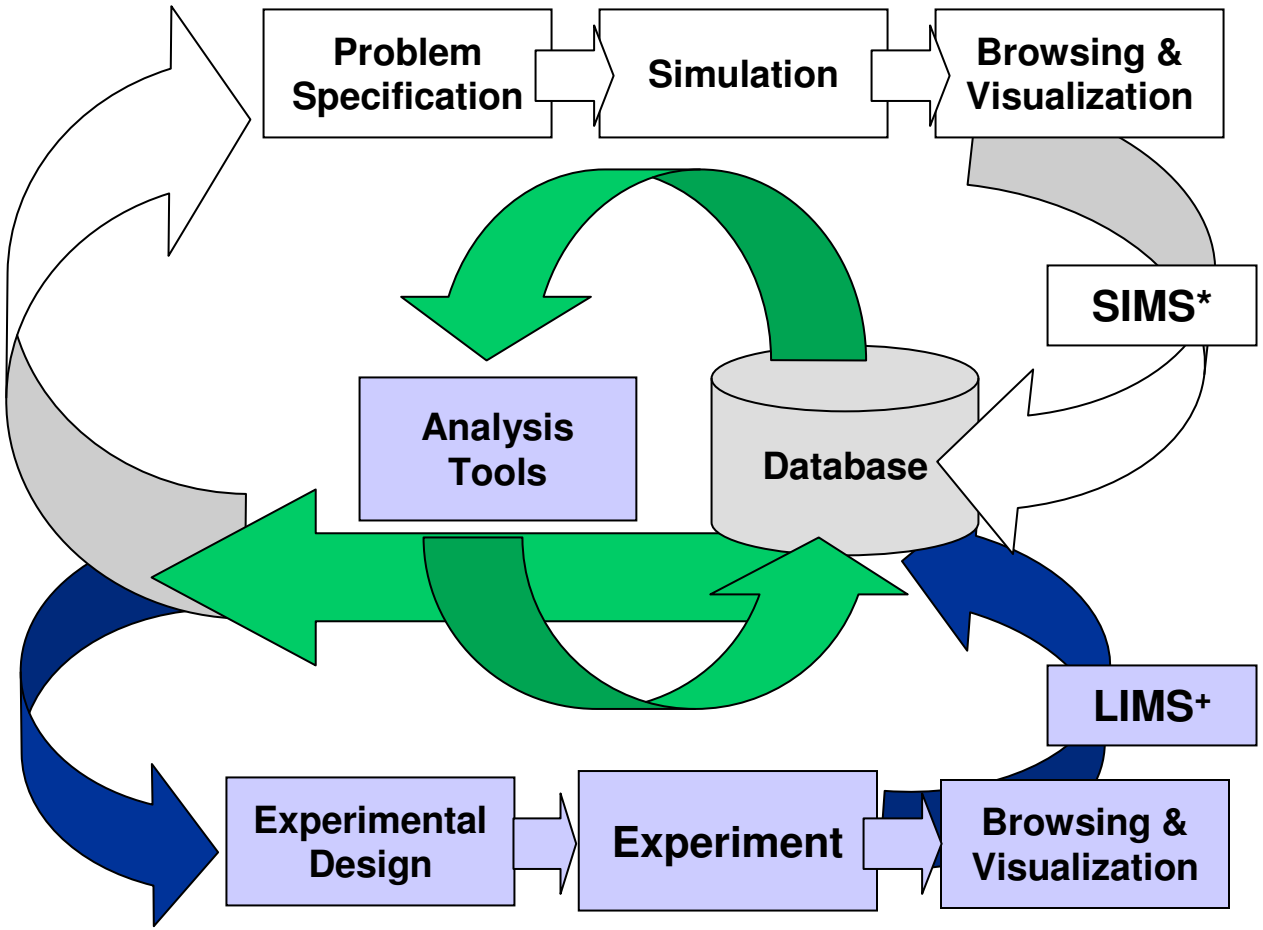
Genome Analysis & DB Update (GADU)



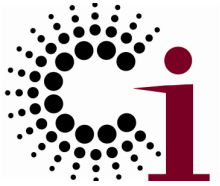
Open Science Grid


Regulatory Layer
Market Layer: Consumer Purchase Decisions
Market Layer: Short-term Operational Decisions
Available Capacity
Time
Market Layer: Long-term Market Entry/Exit Decisions
Physical Infrastructure Layer
Special Event Generator
Information
Information

Integrated View of Simulation, Experiment, & Bioinformatics



*Simulation Information Management System
+Laboratory Information Management System



eScience

Computational science

\+ **Informatics**

= **eScience** [John Taylor, UK EPSRC]

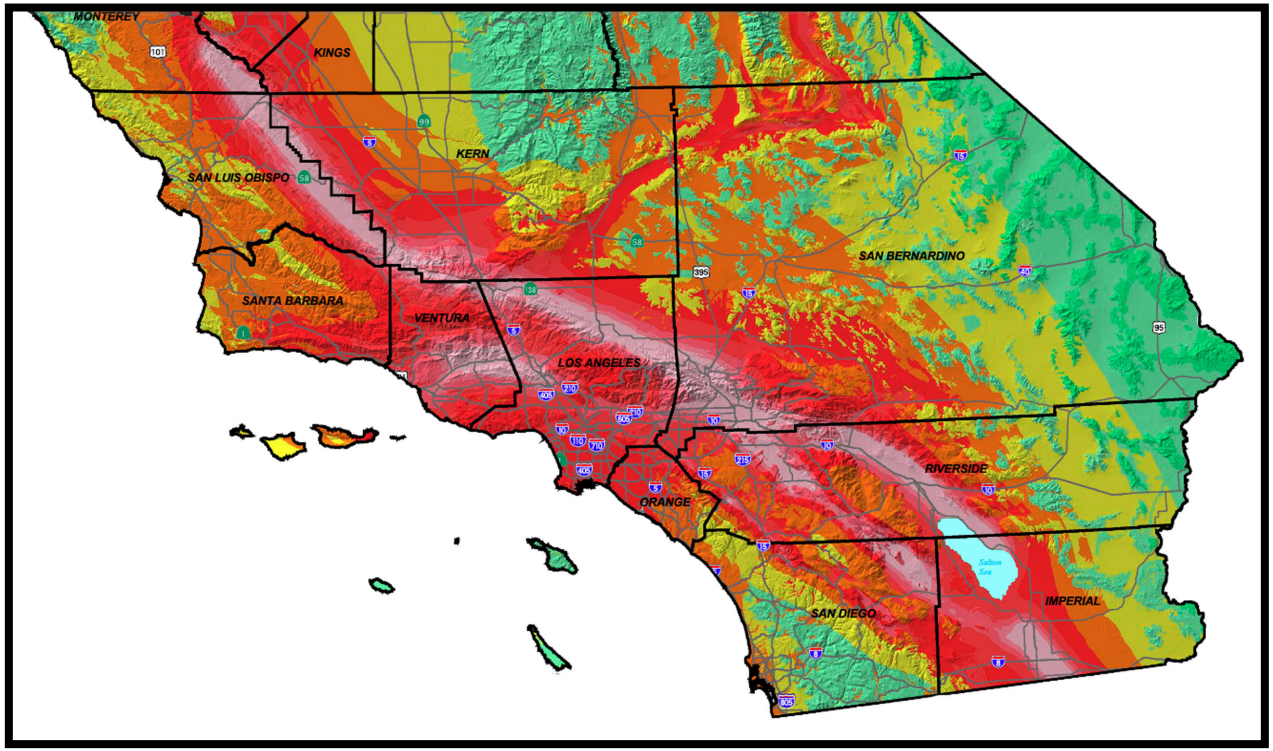
- "Large-scale science carried out through distributed collaborations—often leveraging access to large-scale data & computing"

Seismic Hazard Analysis

Defn: Max. intensity of shaking expected at a site during a fixed time interval

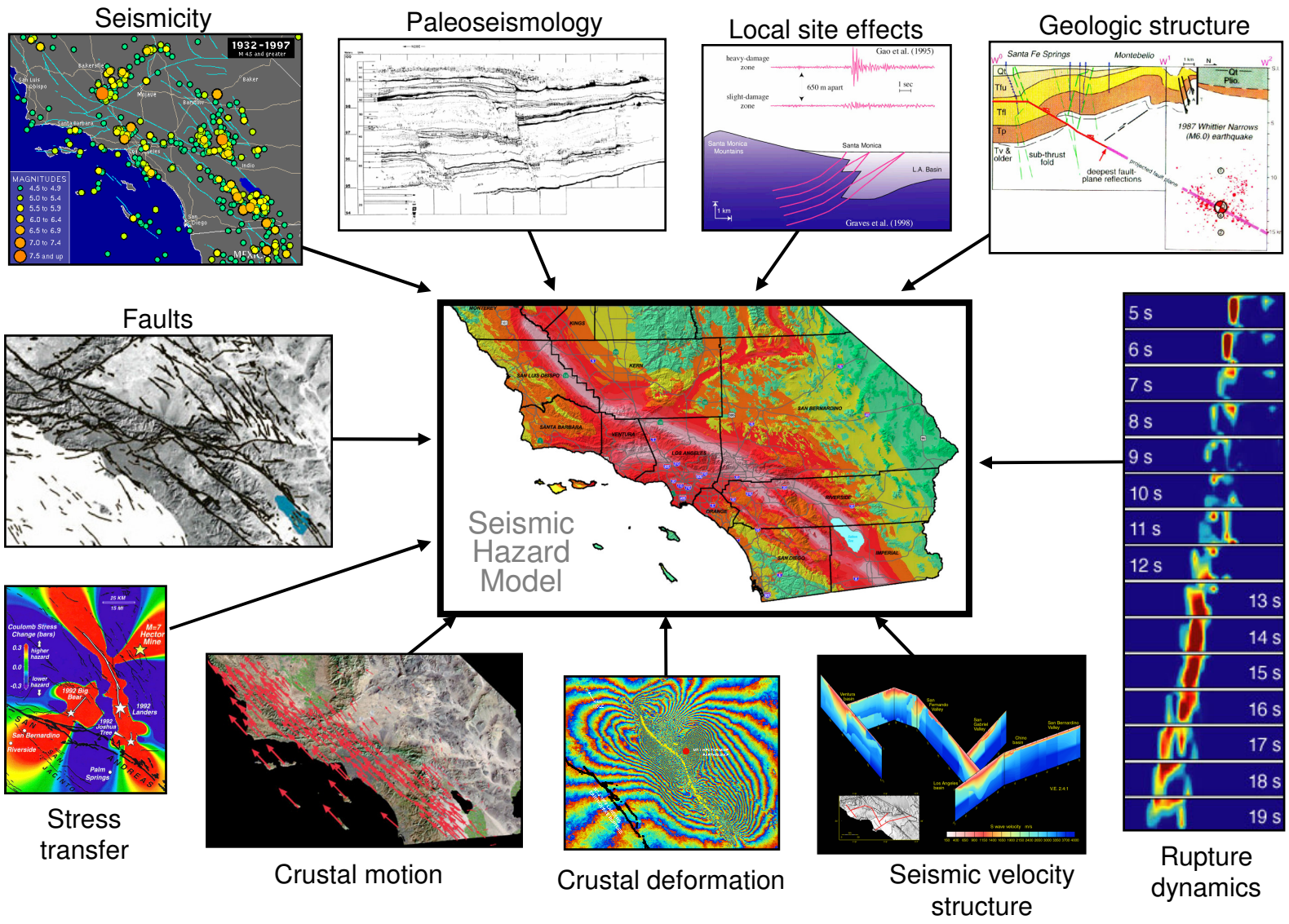
Example: National seismic hazard maps

- Intensity measure: peak ground acceleration
- Interval: 50 yrs
- Probability of exceedance: 2%



T. Jordan et al., Southern California Earthquake Center

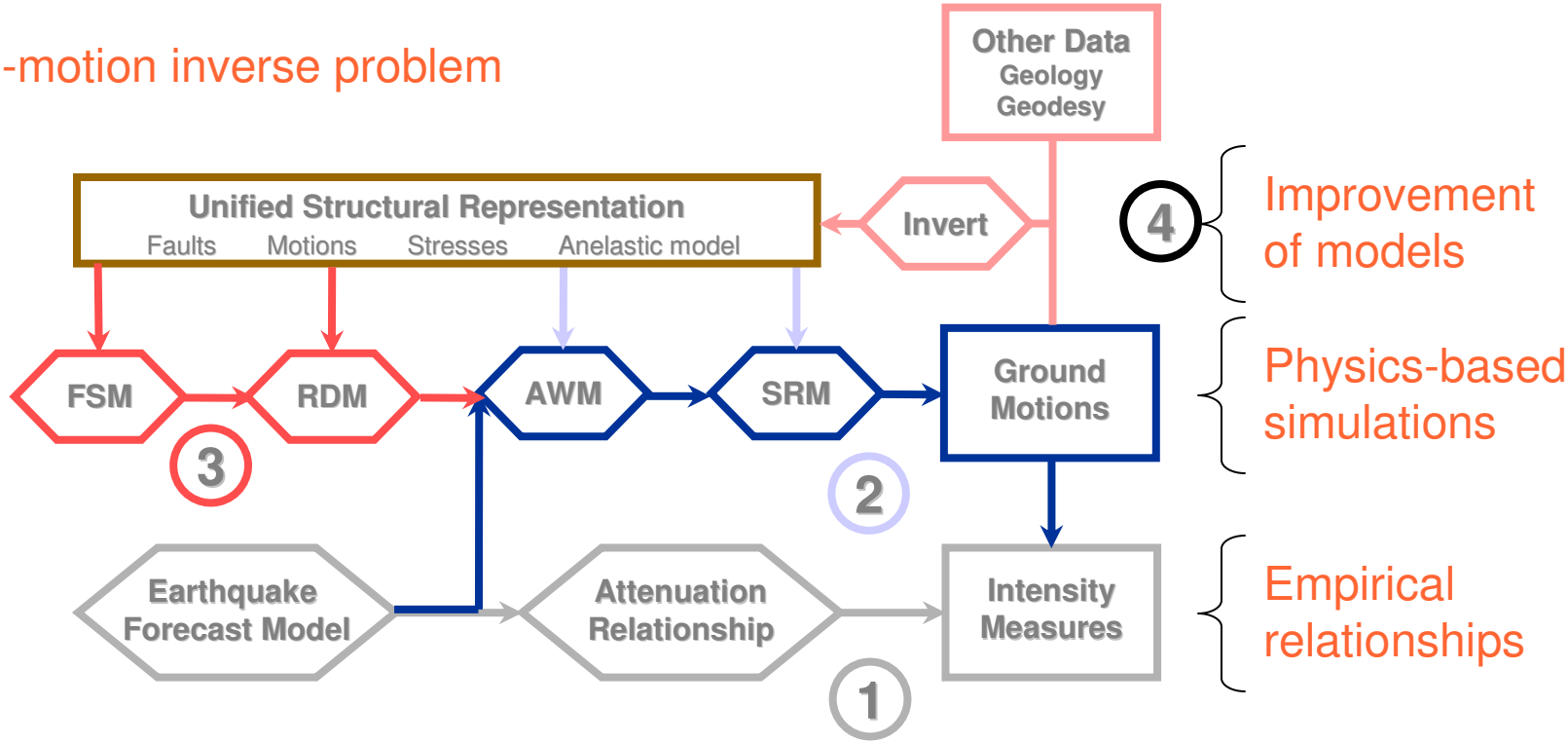
Seismic Hazard Analysis



T. Jordan et al., Southern California Earthquake Center

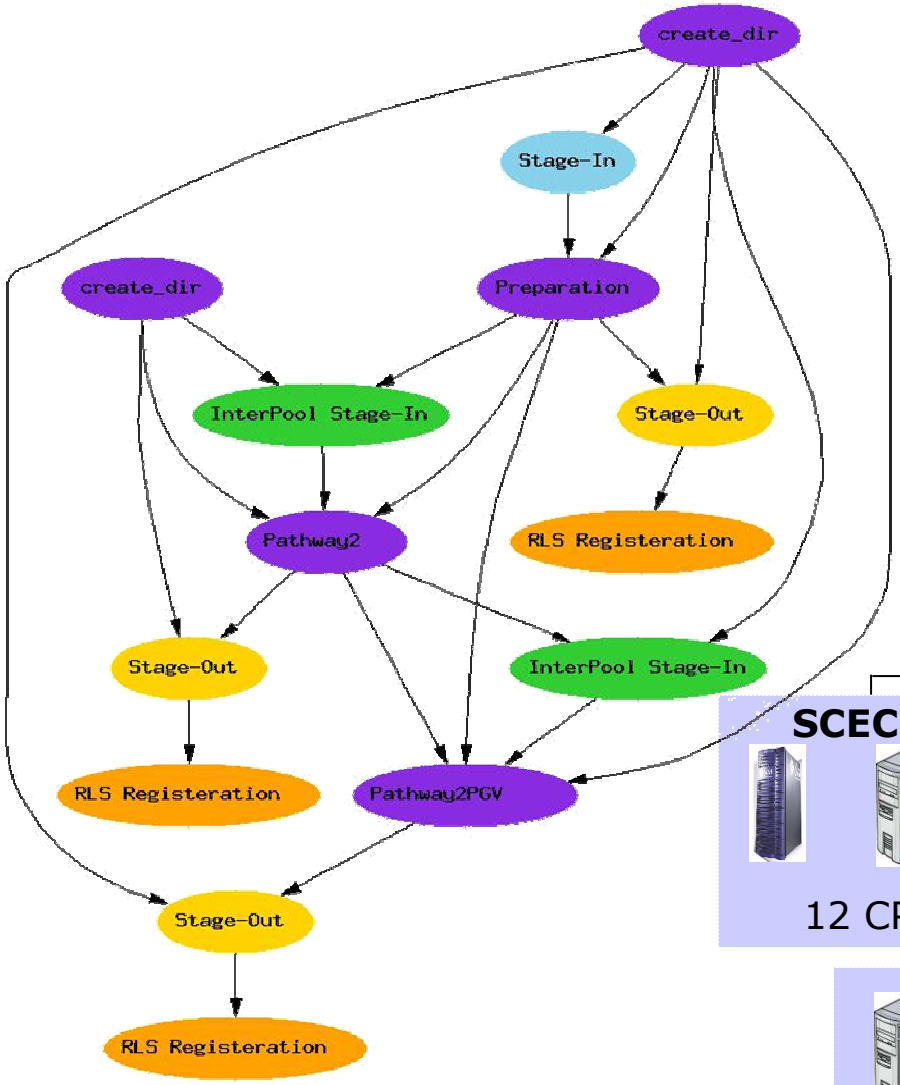
1. Standardized Seismic Hazard Analysis
2. Ground motion simulation
3. Physics-based earthquake forecasting
4. Ground-motion inverse problem

SHA Computational Pathways

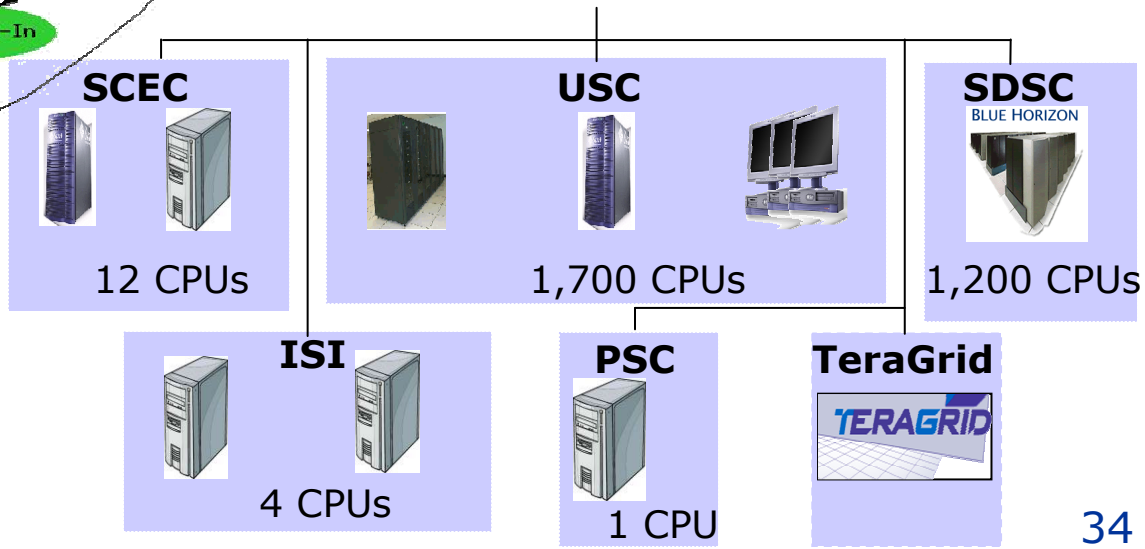


FSM = Fault System Model
RDM = Rupture Dynamics Model
AWP = Anelastic Wave Propagation
SRM = Site Response Model

Access to National Cyberinfrastructure



- Prepare input to Pathway2 wave propagation code
- Pathway2PGV converts output into hazard map
- Map is visualized

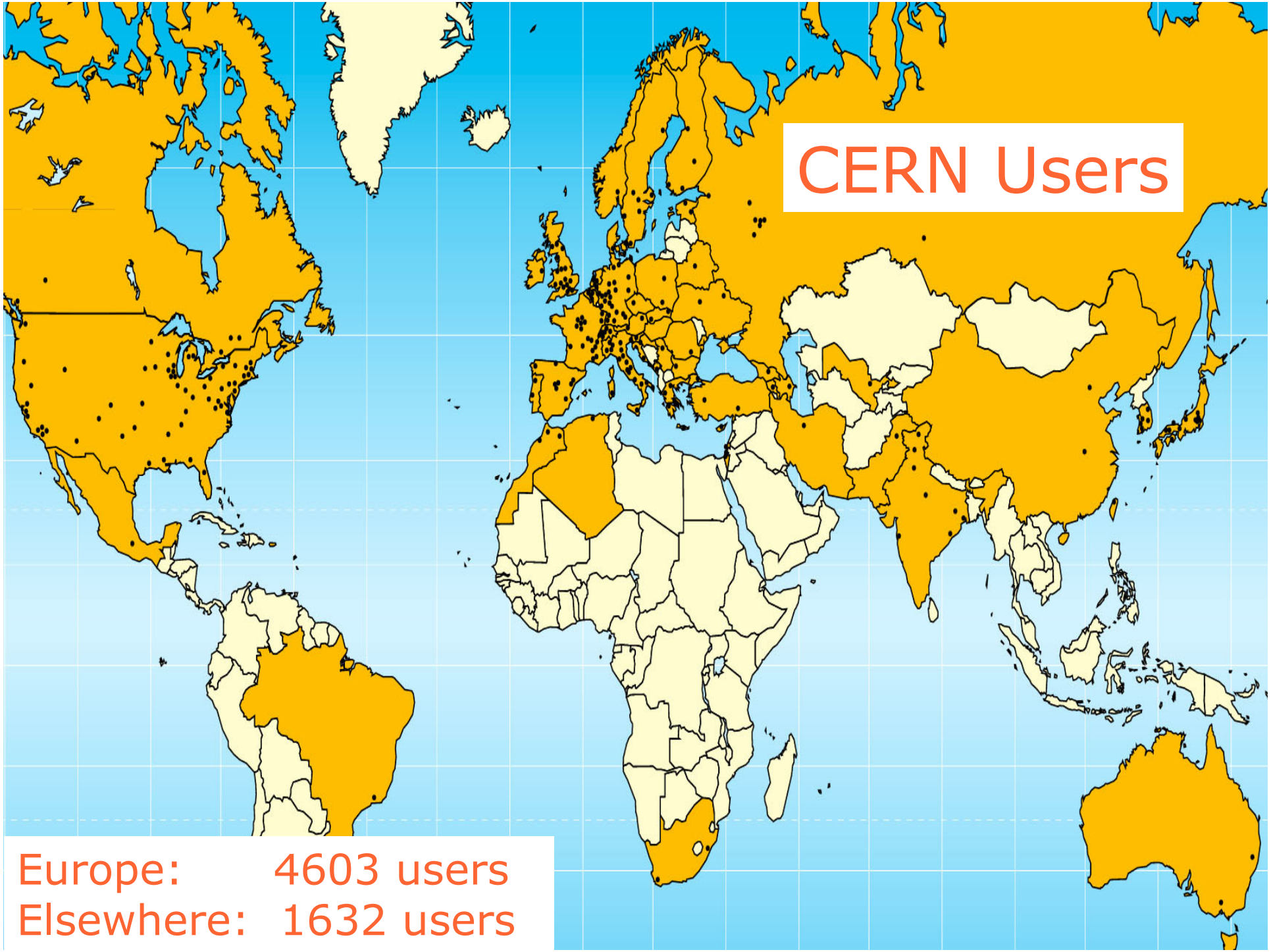


Slide 34

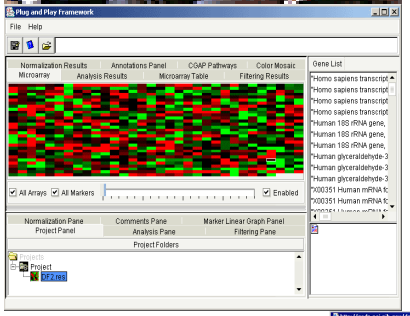
CFK11 This shows only pathway two, where the other pathways involved as well.

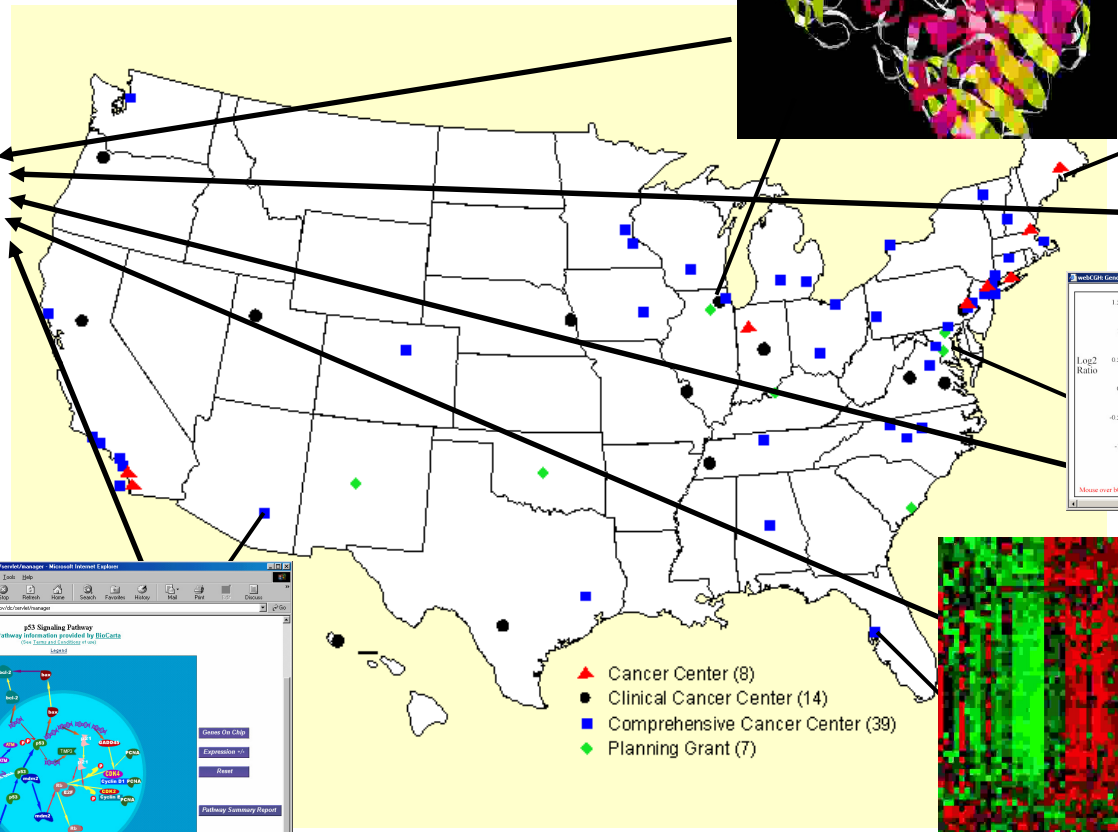
Do you have a visualization of the output of wone of these runs?

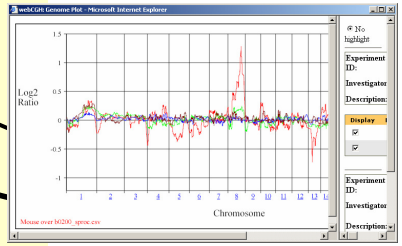
Are the CPU pictures accurate to what you ran?
Carl Kesselman, 9/27/2004

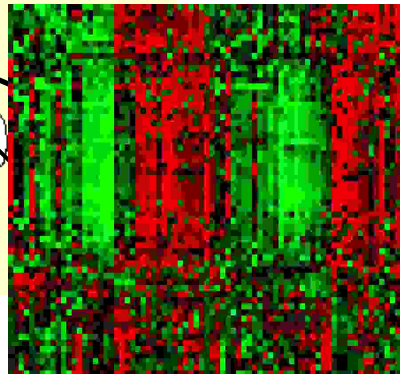

CERN Users
Europe: 4603 users
Elsewhere: 1632 users

Cancer Biology

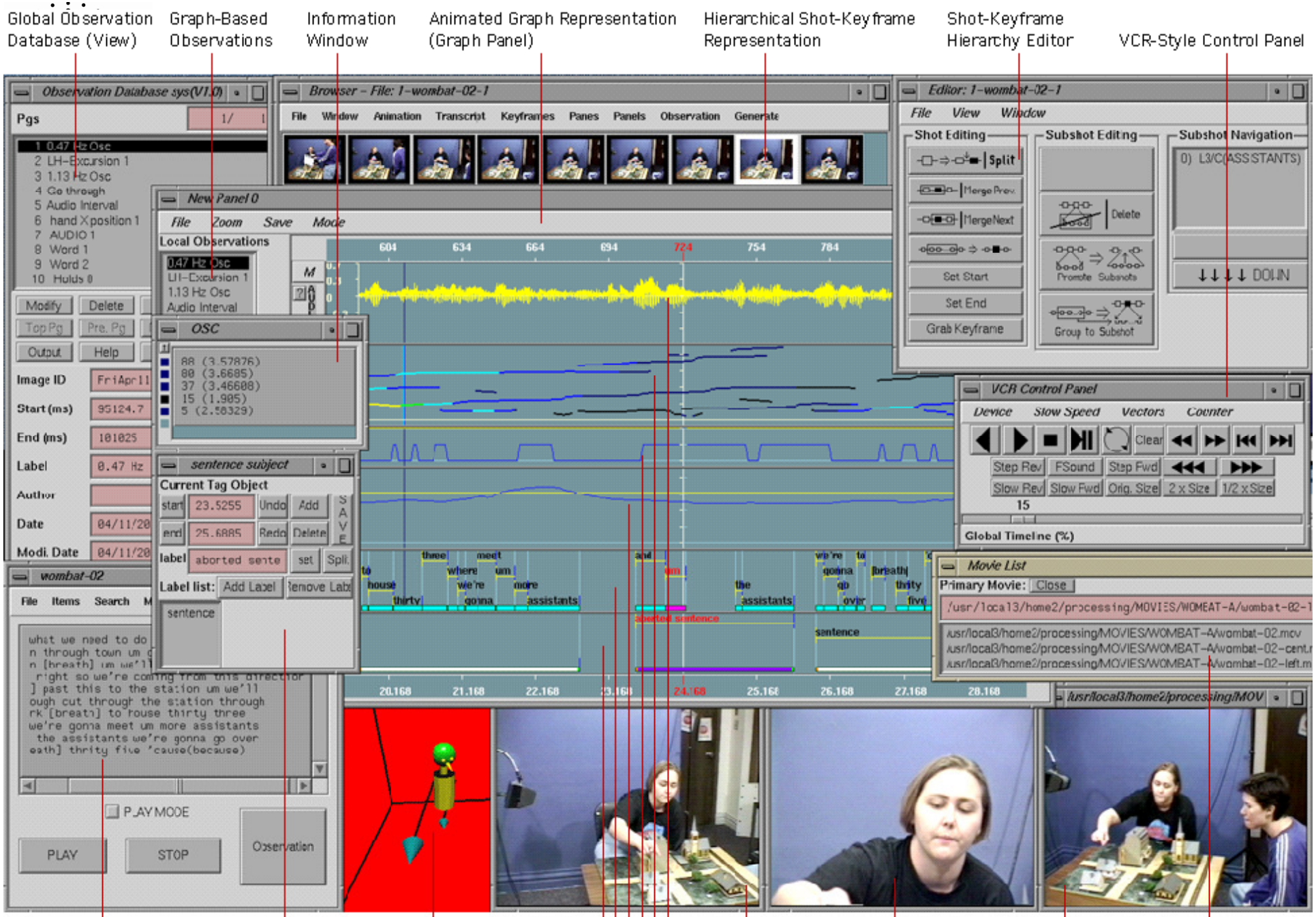


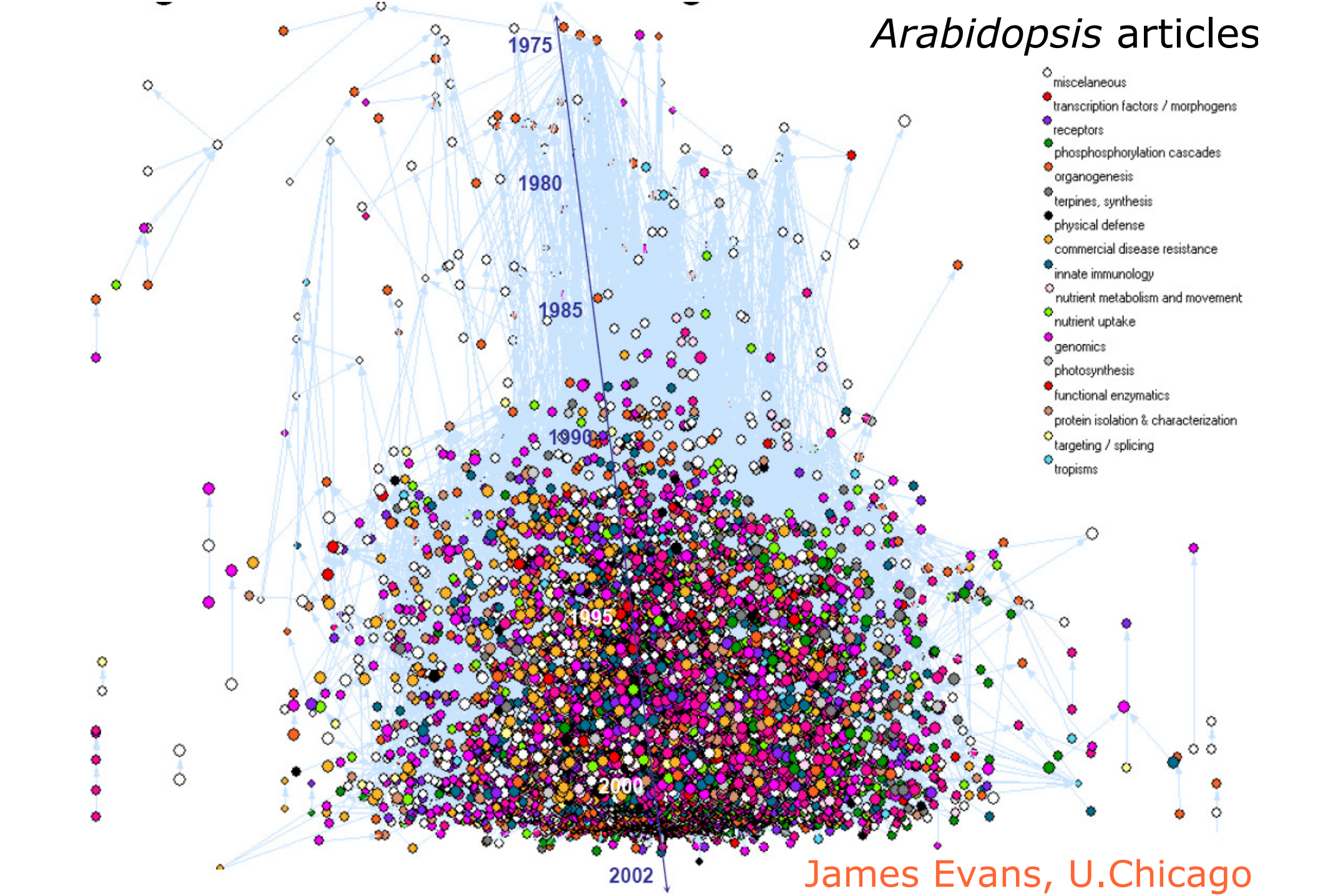


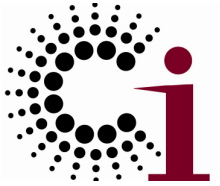




caBIG cancer Biomedical Informatics Grid

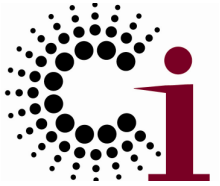

Global Observation Database (View)
Graph-Based Observations
Information Window
Animated Graph Representation (Graph Panel)
Hierarchical Shot-Keyframe Representation
Shot-Keyframe Hierarchy Editor
VCR-Style Control Panel
Animated Text Transcript (Paragraph Representation)
Tag Transcript Editor
Animated Avatar Representation
Animated Graph Panes
Video Displays
Video List
Bennett Berthenthal et al., www.sidgrid.org


Arabidopsis articles
1975
1980
1985
1990
1995
2000
2002
miscelaneous
transcription factors / morphogens
receptors
phosphosphorylation cascades
organogenesis
terpines, synthesis
physical defense
commercial disease resistance
innate immunology
nutrient metabolism and movement
nutrient uptake
genomics
photosynthesis
functional enzymatics
protein isolation & characterization
targeting / splicing
tropisms
James Evans, U.Chicago



eScience Challenges

- Simulate complex, multi-component systems
- Evaluate accuracy of such simulations
- Integrate evidence to draw conclusions
- Evaluate strength of conclusions
- Automate “experimental” workflows
- Document basis for conclusions (provenance)
- Allow these problems to be tackled by distributed teams using federated resources



What is Fundamental?

Bits

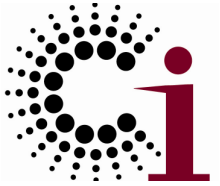
+

Algorithms

+

Complex systems

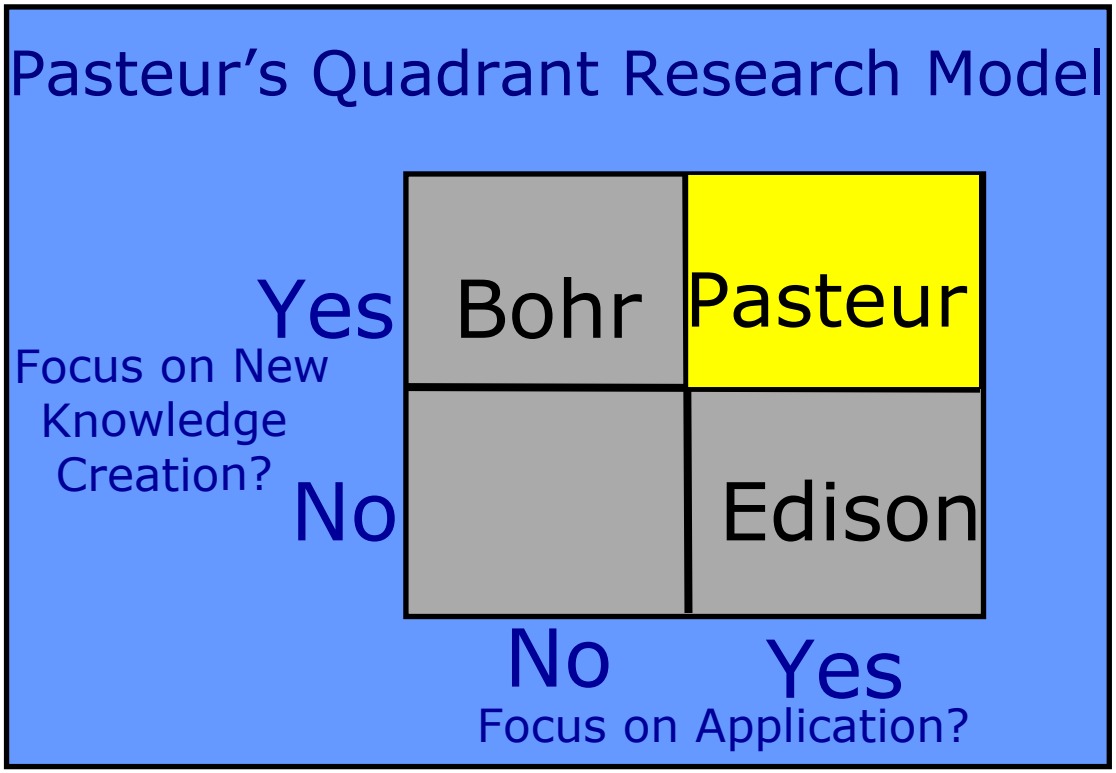
First two, at least © CS

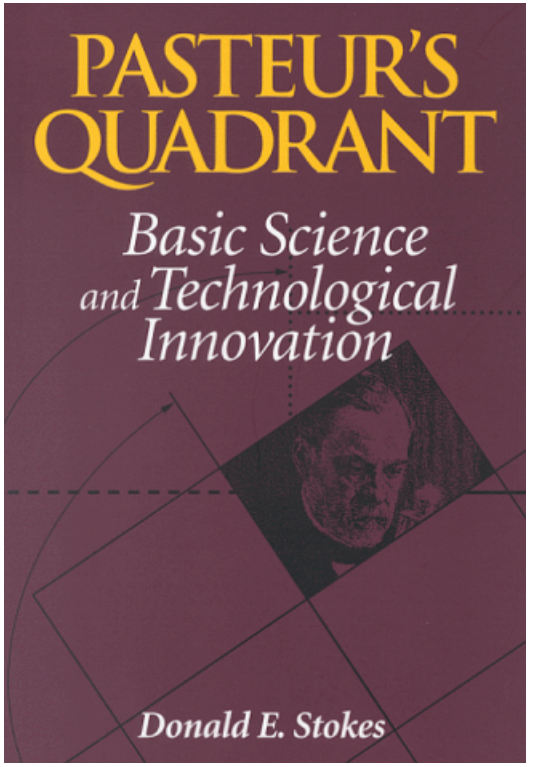


Computer Science: A Narrow or Broad View?

- Narrow
  - CS is programming
  - ➔ Other aspects of information are the domain of "statistics," "bioinformatics", etc., etc.
- Broad
  - CS is "the systematic study of algorithmic processes that describe and transform information, their theory, analysis, design, efficiency, implementation, and application" (Denning et al., CACM, 1989)
  - ➔ Statistics & bioinformatics are subdisciplines of computer science

Effective eScience **requires** PQ research models



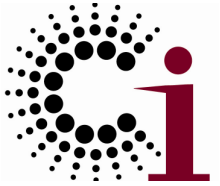


Creation of knowledge: basic, curiosity-driven research

Application of knowledge

Classic Linear Research Model

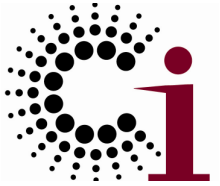
Slide courtesy Dan Atkins, U.Michigan



"**Applied computer science** is now playing the role that mathematics did from the 17th through the 20th centuries: providing an orderly, formal framework & exploratory apparatus for other sciences."

—George Djorgovski

"… the branch of computer science that concerns itself with the application of computing knowledge to other domains"**?**

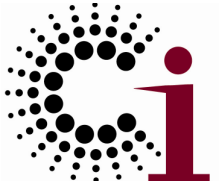


Computation Institute

A joint institute of Argonne and the University of Chicago, focused on advancing **system-level science**

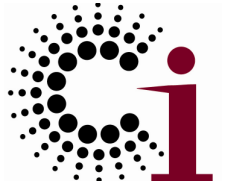
Solutions to many grand challenges facing science and society today are dependent upon the analysis and understanding of entire systems, not just individual components. They require not reductionist approaches but the synthesis of knowledge from multiple levels of a system, whether biological, physical, or social (or all three).

http://www.ci.uchicago.edu



Thanks!

- foster@mcs.anl.gov
- http://www.ci.uchicago.edu
- http://ianfoster.typepad.com



In Memoriam: Jim Gray (1944-2007?)

Turing Award, 1998

"for seminal contributions to database & transaction processing and technical leadership in system implementation"