Scientific Computing:
An Intellectual Lever for Discovery

Dan Reed
reed@renci.org

Chancellor's Eminent Professor
Vice Chancellor for Information Technology
University of North Carolina at Chapel Hill

Director, Renaissance Computing Institute (RENCI)
Duke, UNC Chapel Hill and North Carolina State University

Chair, Board of Directors
Computing Research Association (CRA)

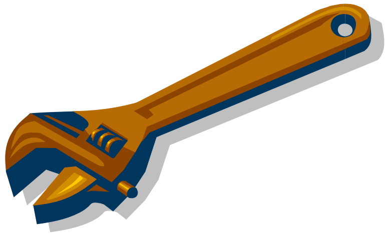
The Instruments of Innovation

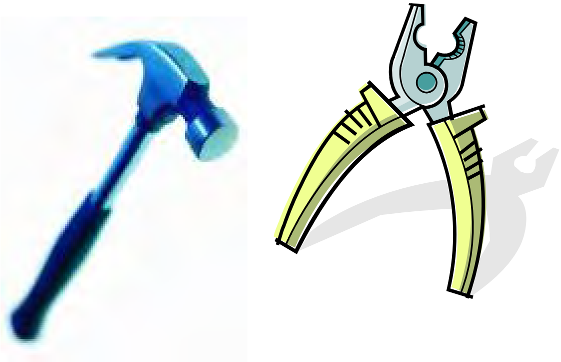
Nothing tends so much to the advancement of knowledge as the application of a new instrument. The native intellectual powers of men in different times are not so much the causes of the different success of their labors, as the peculiar nature of the means and artificial resources in their possession.

Sir Humphrey Davy

The Computer Science Behind Science

- **Computer science (CS) enables science**
  - high-performance computing
  - networks and sensors
  - grids and web services
  - data models and mining
  - scientific and information visualization
  - distributed collaboration tools
  - algorithms, software and tools
  - numerical analysis
  - artificial intelligence



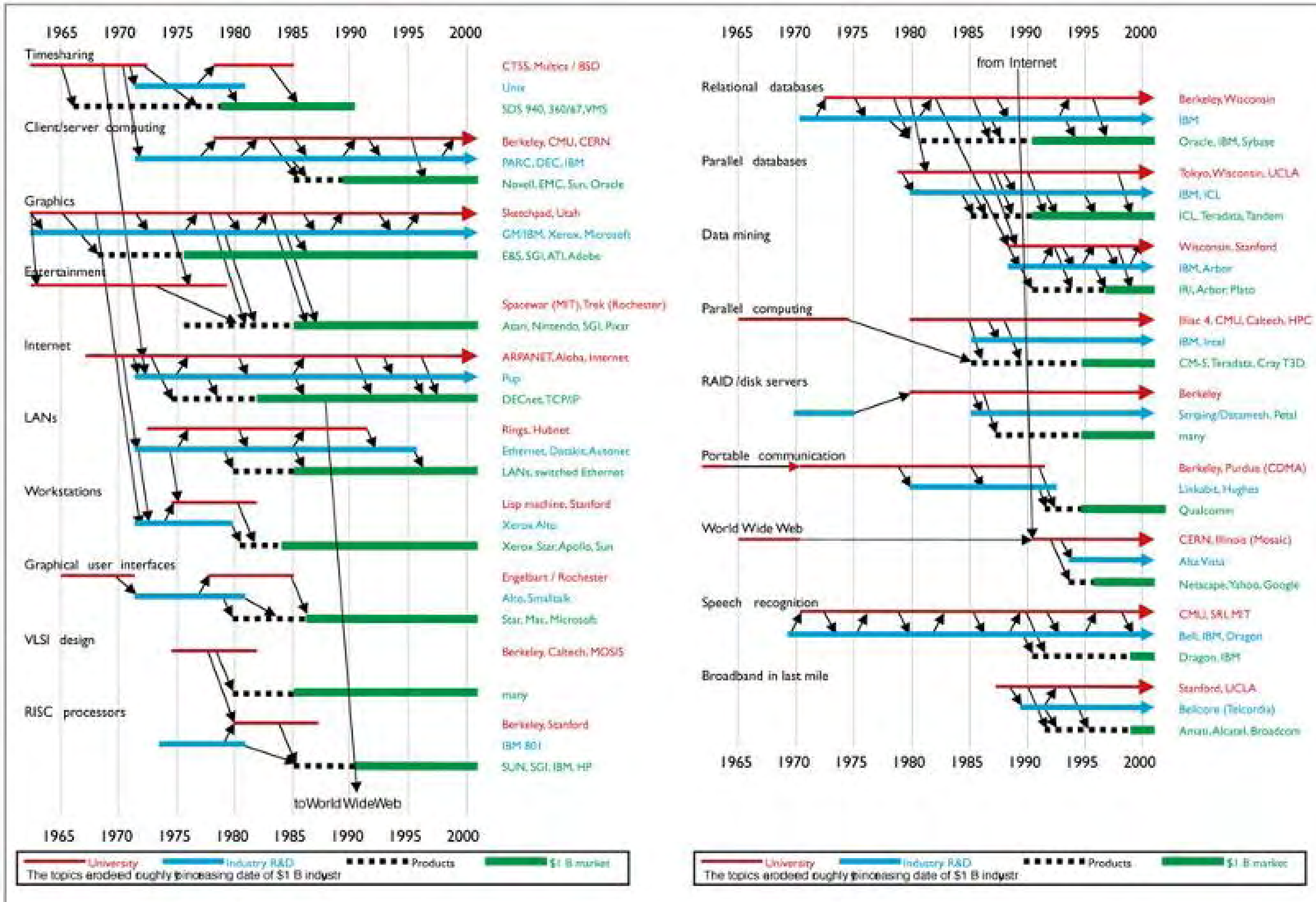


- **Let’s look at some examples …**
  - where CS and science meet as computational science

COMPUTING RESEARCH ASSOCIATION CRA

renci
renaissance computing institute

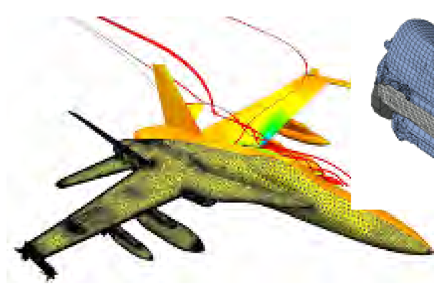
Computer Science Impact

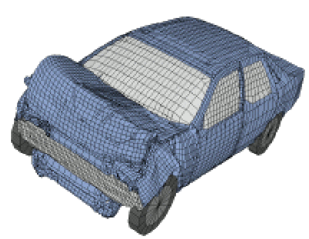


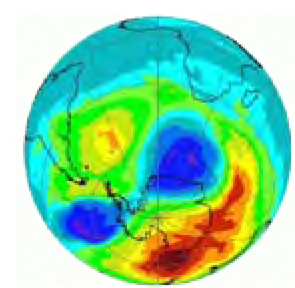
Source: From [6], reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington D.C. © 2003.

The Third Pillar of 21st Century Science

- **Three pillars**
  - theory, experiment and *computational science*
- **Computational science enables us to**
  - investigate phenomena where
    - economics or constraints preclude experimentation
  - evaluate complex models and manage massive data
    model processes across interdisciplinary boundaries
  - transform business and engineering practices





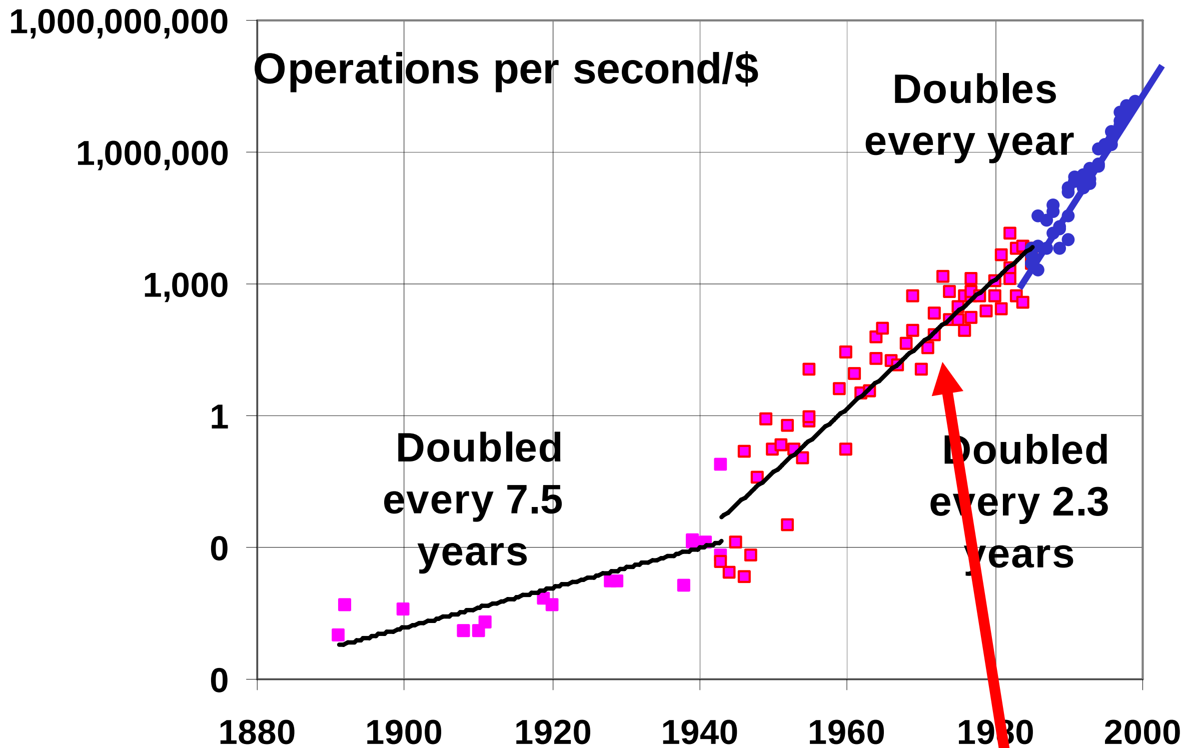


COMPUTING RESEARCH CRA ASSOCIATION

renci
renaissance computing institute

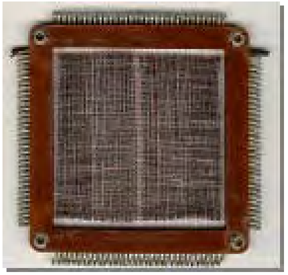
Computing History and Exponentials

- **1890-1945**
  - mechanical, relay
  - 7 year doubling
- **1945-1985**
  - tube, transistor,..
  - 2.3 year doubling
- **1985-2003**
  - microprocessor
  - 1 year doubling



Microprocessor Revolution

- **Exponentials**
  - chip transistor density: 2X in ~18 months
  - graphics: 100X in three years
  - WAN bandwidth: 64X in two years
  - *storage: 7X in two years*



4K bit core plane

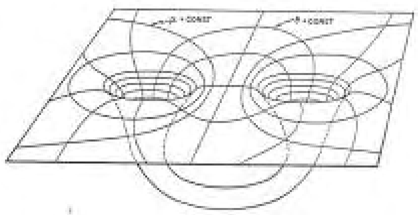
Source: Jim Gray

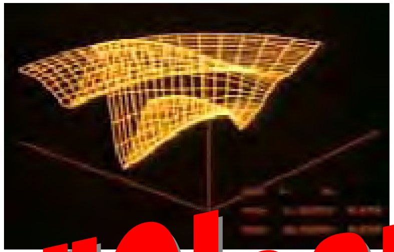
Large Systems

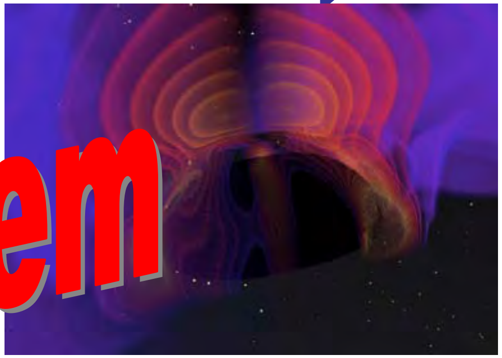
Machine	Processor Count	Teraflops	Year
ASCI Q	8,192	20	’02
ASCI White	8,192	12	’00
NERSC Seaborg	6,656	10	’02
ASCI Red Storm	10,368 (2 MW)	40	’04
ASCI Purple	12,000	100	’05
Blue Gene/L	65,536	180-360	‘05
NASA Columbia	10,240 (2 MW)	60	’04

Black Hole Collision Problem

1,800,000,000X





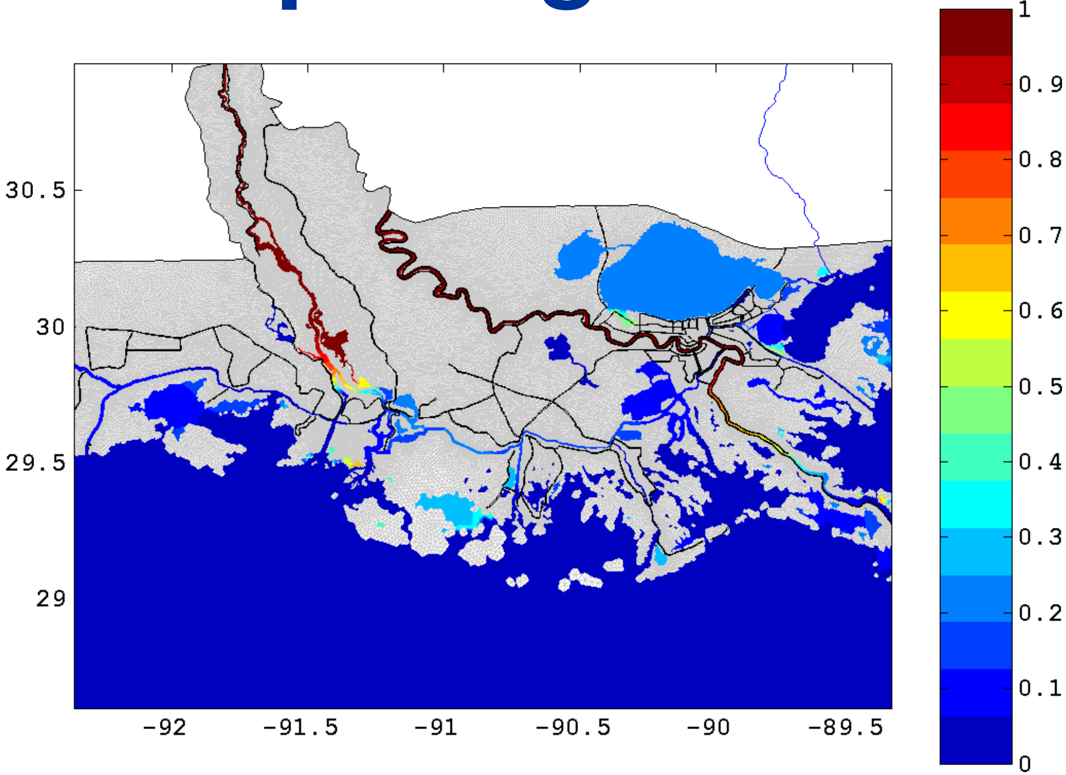


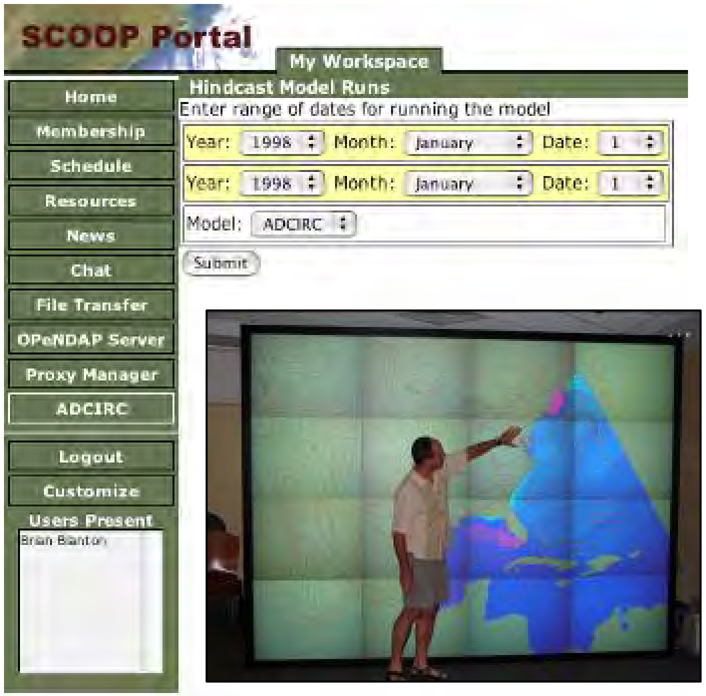
A "Clean" Problem

1963		1999	2001
Hahn and Lindqui	Epple… Smarr	Seidel and Suen, *et al.*	Seidel *et al*
IBM 7090	CDC 7600	NCSA SGI Origin	NCSA Pentium III
One Processor	One Processor	256 Processors	256 Processors
Each 0.2 MF	Each 35 MF	Each 500 MF	Each 1 GF
3 Hours	5 Hours	40 Hours	500,000 Hours total plus 500,000 hours at NERSC

300X → **30,000X** → **~200X**

Computing for Disaster Response



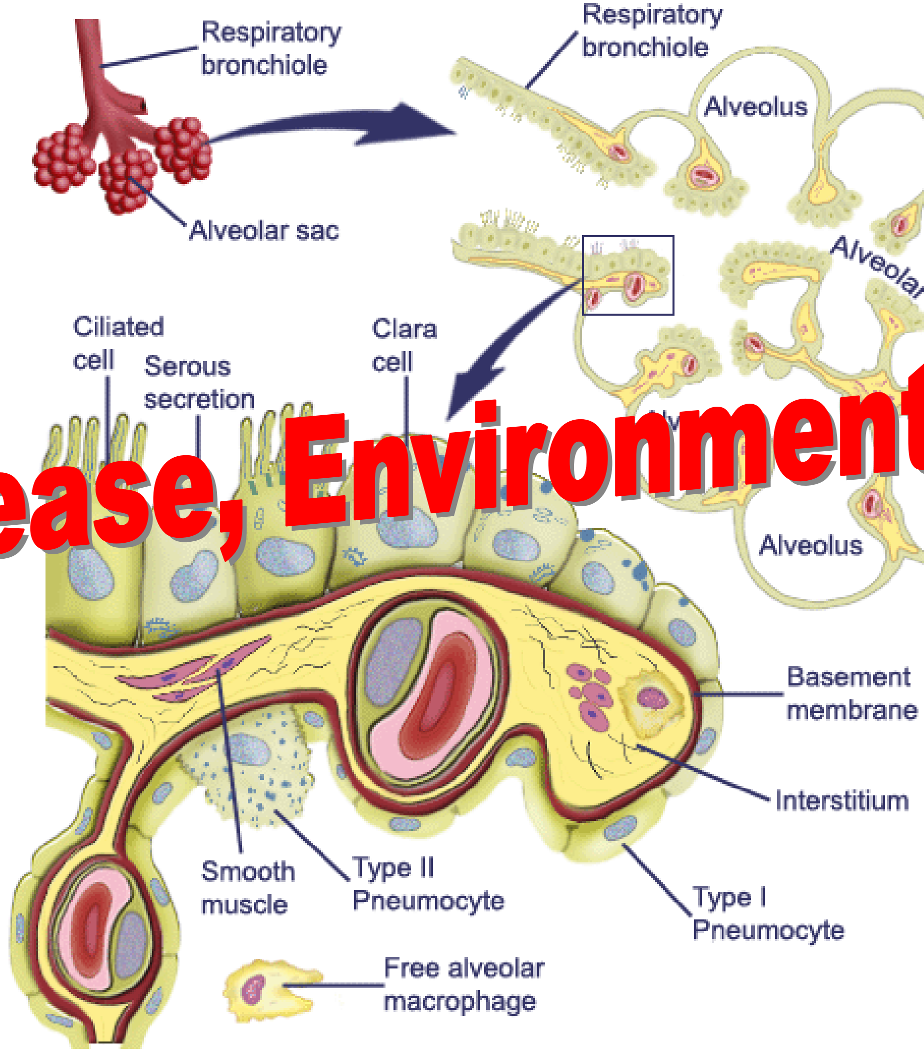


- **Post-Katrina NOAA challenge**
  - petrochemical spillage and remediation
  - water levels determine dissemination
  - answers needed within 48 hours, but inadequate computing capability
- **UNC Marine Sciences and RENCI**
  - ADCIRC storm surge model predicts water levels
  - HPC system yields model predictions

COMPUTING RESEARCH ASSOCIATION CRA

renci renaissance computing institute

Biophysical and Environmental Modeling



Disease, Environment and Medicine

Airway/flow

Mucus

Cilia

Cell biochemistry and structure

Proteomics

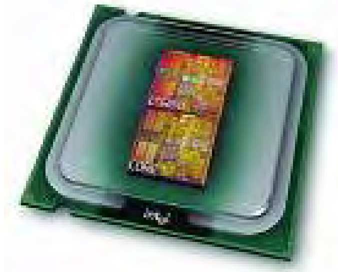
Genomics

Source: Ric Boucher, UNC

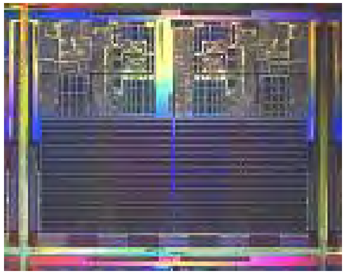
COMPUTING RESEARCH ASSOCIATION CRA

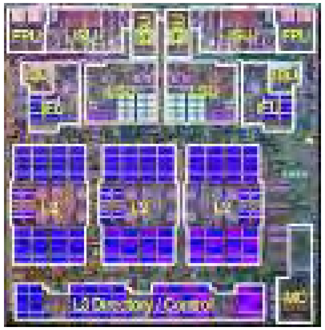
renci renaissance computing institute

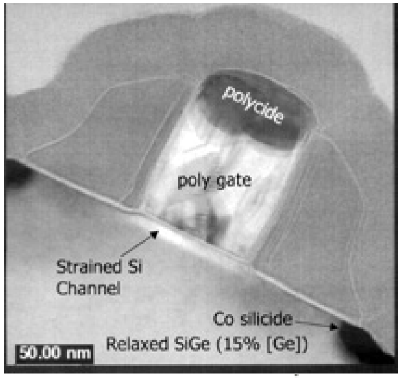
The Coming of Massive Parallelism



- **Technology trends**
  - multicore processors
    - IBM Power5/6 and SUN UltraSPARC IV
    - Intel Xeon and AMD Opteron
    - quad core and beyond are coming
  - reduced power consumption
    - laptop and mobile market drivers
  - greater I/O and memory integration
    - PCI Express, Infiniband, …
  - Justin Ratter (Intel)
    - “100’s of cores on a chip in 2015”
- **Moore’s law isn’t a birthright**
  - CMOS scaling issues are now a challenge
    - power, junction size, fab line costs, …







Digital Reality: The Exponentials

1956

1972

2006

- **Megabyte**
  - a small novel
- **Gigabyte**
  - a pickup truck filled with paper or a DVD
- **Terabyte: one thousand gigabytes – ~$1000 today**
  - the text in one million books
  - entire U.S. Library of Congress is ~ten terabytes of text
- **Petabyte: one thousand terabytes**
  - 1-2 petabytes equals all academic research library holdings
    - coming soon to a pocket near you!
  - *soon routinely generated annually by many scientific instruments*
- **Exabyte: one thousand petabytes**
  - 5 exabytes of words spoken in the history of humanity

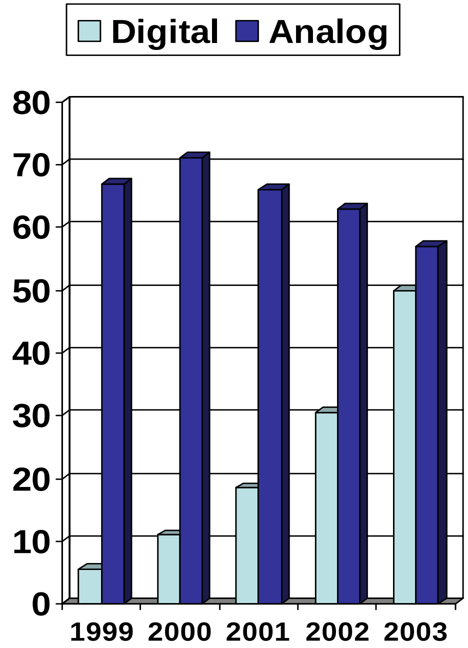
Source: Hal Varian, UC-Berkeley

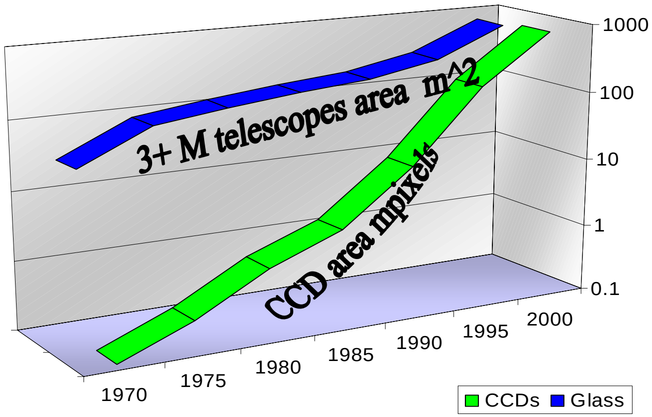
COMPUTING RESEARCH CRA ASSOCIATION

renci
renaissance computing institute

Astronomy and Consumer Cameras

- **Digital camera sales**
  - now exceed analog
  - January 2006
    - Nikon stops film camera production
- **From glass plates to CCDs**
  - detectors follow Moore's law
  - data tsunami
    - data doubles every two years
- **Telescope growth**
  - 30X glass (concentration)
  - 3000X in pixels (resolution)
- **Single astronomy images**
  - 16Kx16K pixels and growing
- **Detector driver**
  - consumer electronics





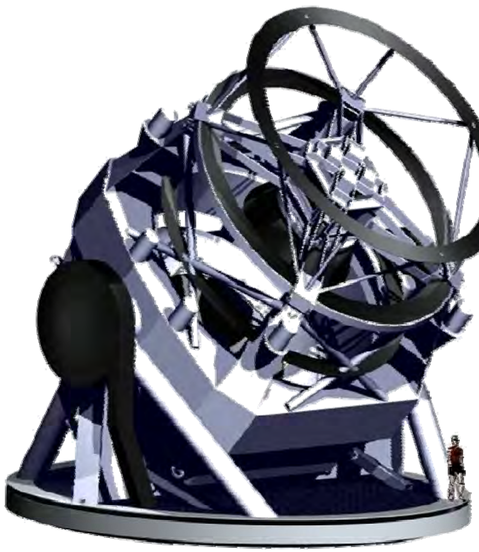
COMPUTING RESEARCH ASSOCIATION CRA

Source: Alex Szalay/Jim Gray

renci
renaissance computing institute

Large Synoptic Survey Telescope (LSST)

- **Top project of the astronomy decadal survey**
- **Celestial cinematography**
  - 3 gigapixel detector for wide field imaging
- **Science**
  - beyond the standard model
    - non-baryonic dark matter
    - non-zero Λ and neutrino oscillations
  - observation targets
    - near Earth object survey
    - weak lensing of wide fields
    - supernovae measurements
- **Features**
  - 9.6 square degree field/6.5 meter effective aperture
  - *~15 TB of data/night, target first light 2012*

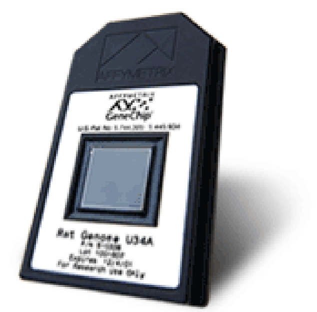


COMPUTING RESEARCH ASSOCIATION CRA

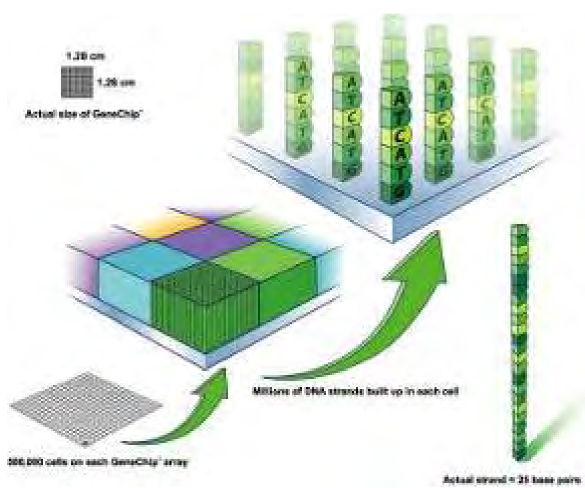
renci renaissance computing institute

Gene Expression and Microarrays

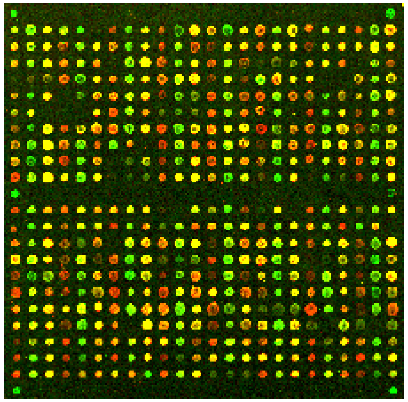
- **Concurrent evaluation**
  - expression levels for thousands of genes
- **Photolithography**
  - up to 500K 10-20 micron cells
    - each containing millions of identical DNA molecules
- **Image capture and analysis**
  - laser scanning and intensity calculation



AFFYMETRIX GENECHIP®



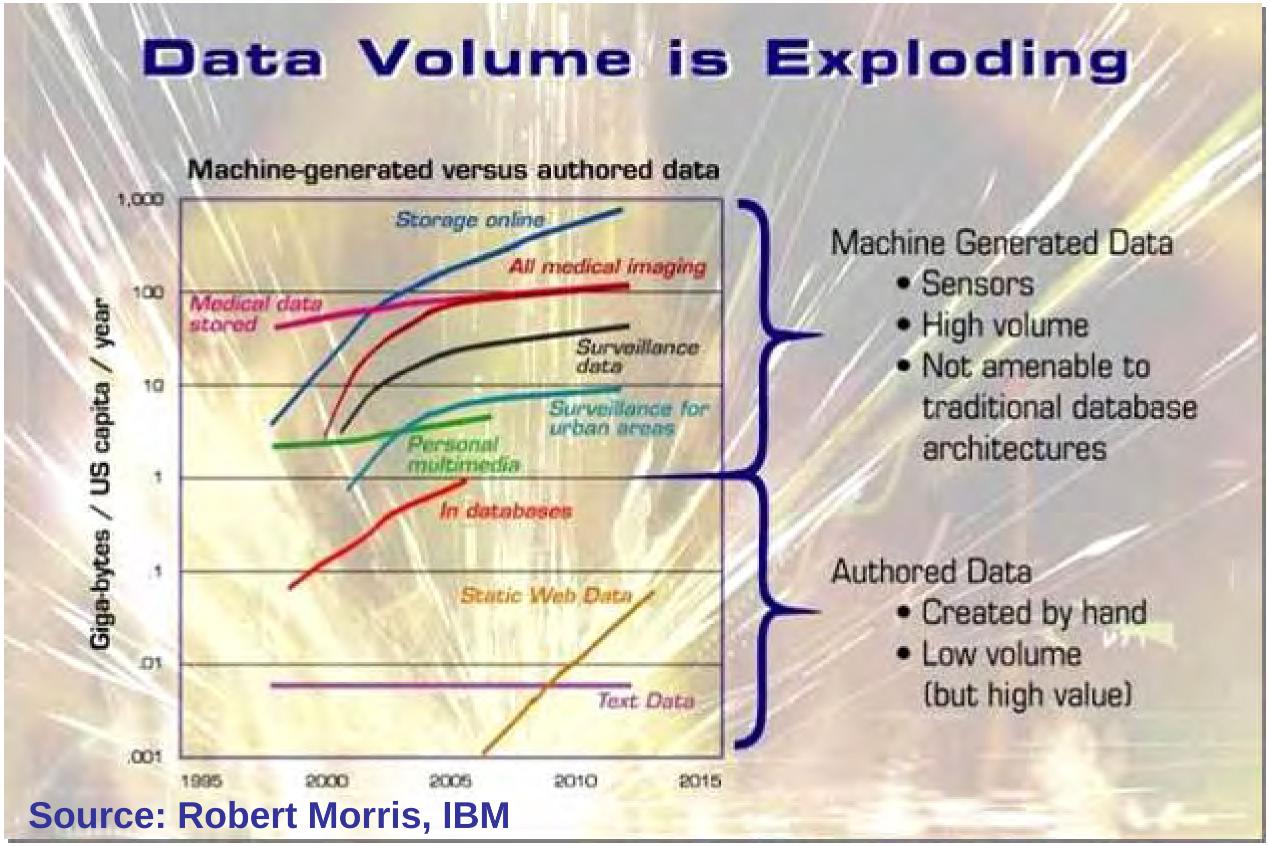
Source: Affymetrix

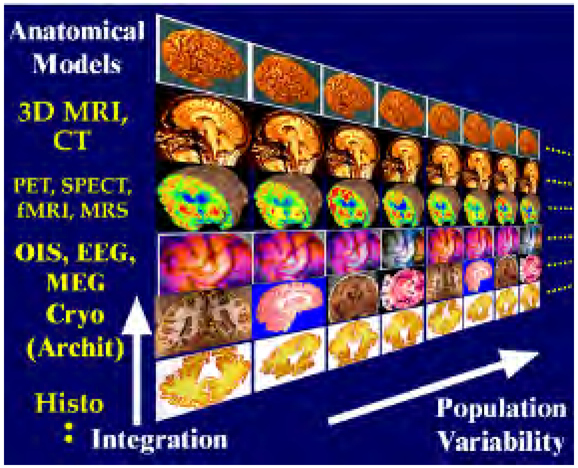


COMPUTING RESEARCH ASSOCIATION CRA

renci
renaissance computing institute

Sensor Data Overload



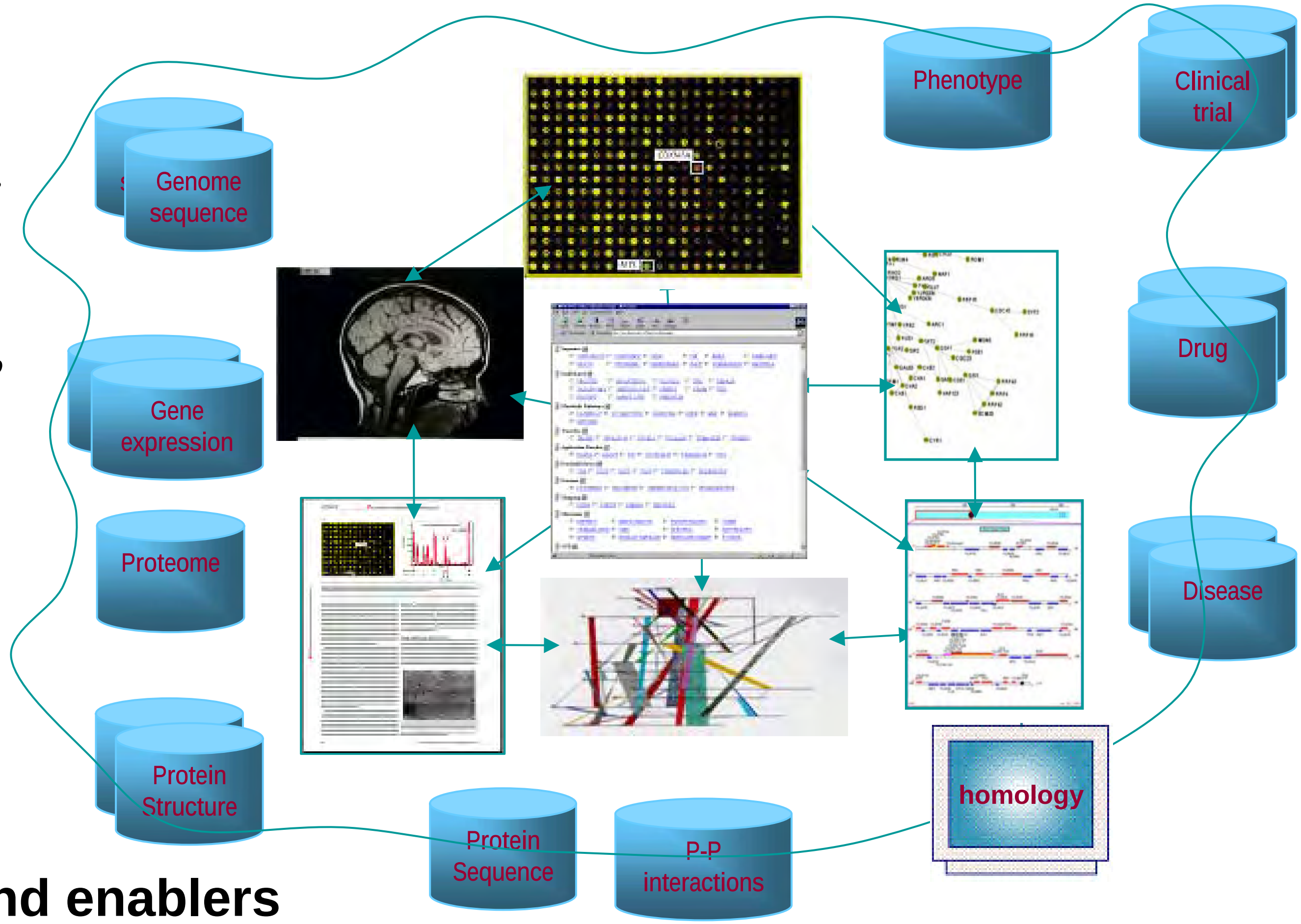


Source: Chris Johnson, Utah
Art Toga, UCLA

- **High resolution brain imaging**
  - 4.5 petabytes (PB) per brain

Data Heterogeneity and Complexity

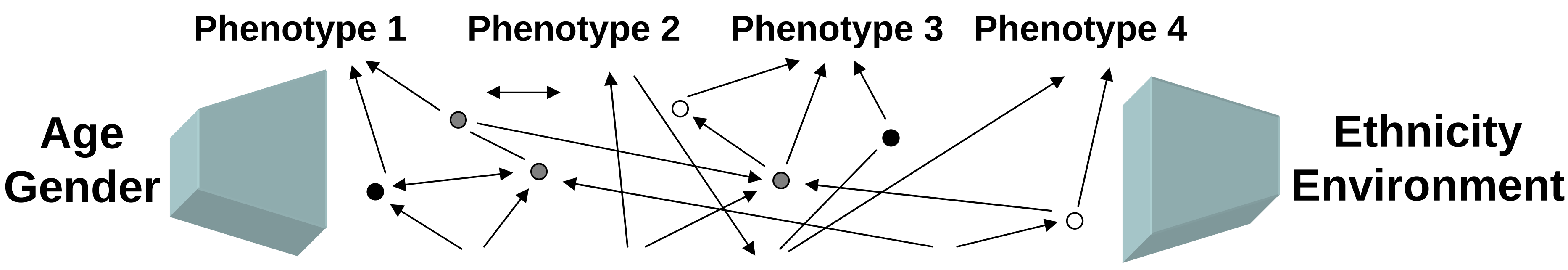
Genomic, proteomic, transcriptomic, metabalomic, protein-protein interactions, regulatory bio-networks, alignments, disease, patterns and motifs, protein structure, protein classifications, specialist proteins (enzymes, receptors)

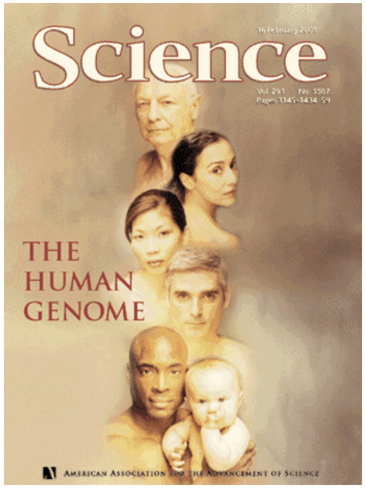


- **Many causes and enablers**
  - increased instrument resolution
  - increased storage capability
- **The challenge: *extracting insight!***

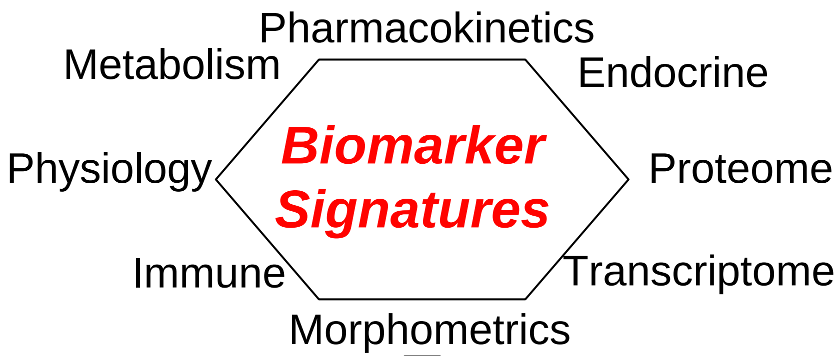
Source: Carole Goble (Manchester)

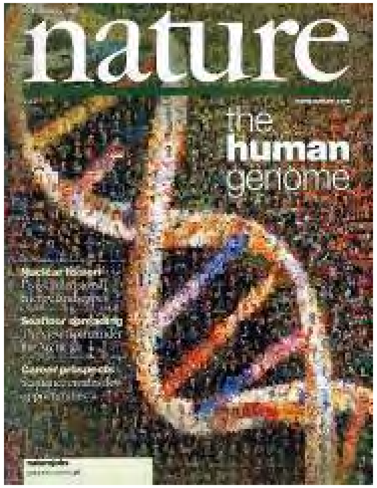
Genetics and Advanced Data Mining





Identify Genes

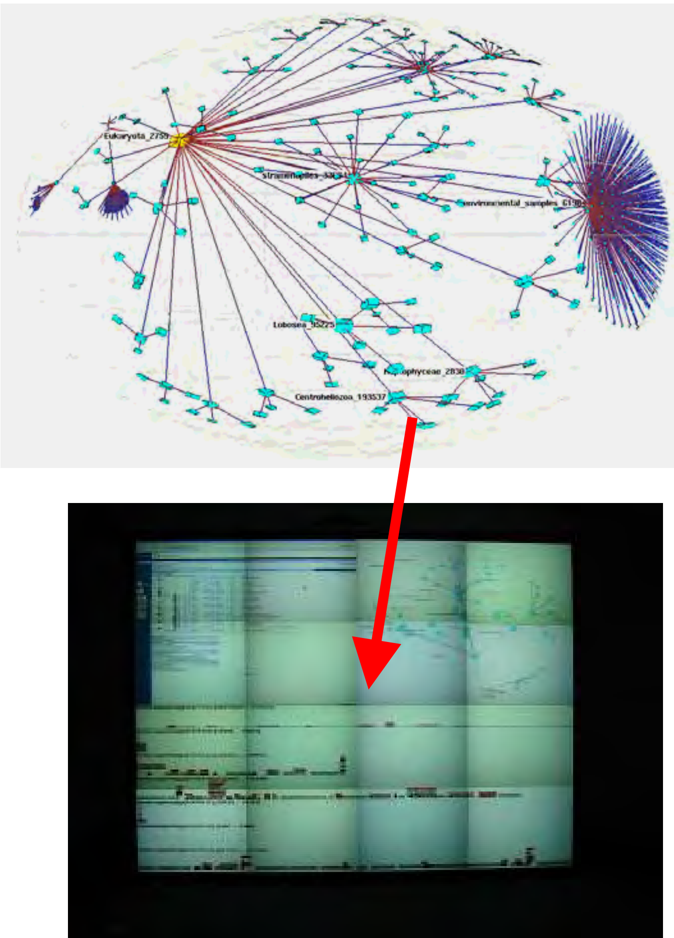


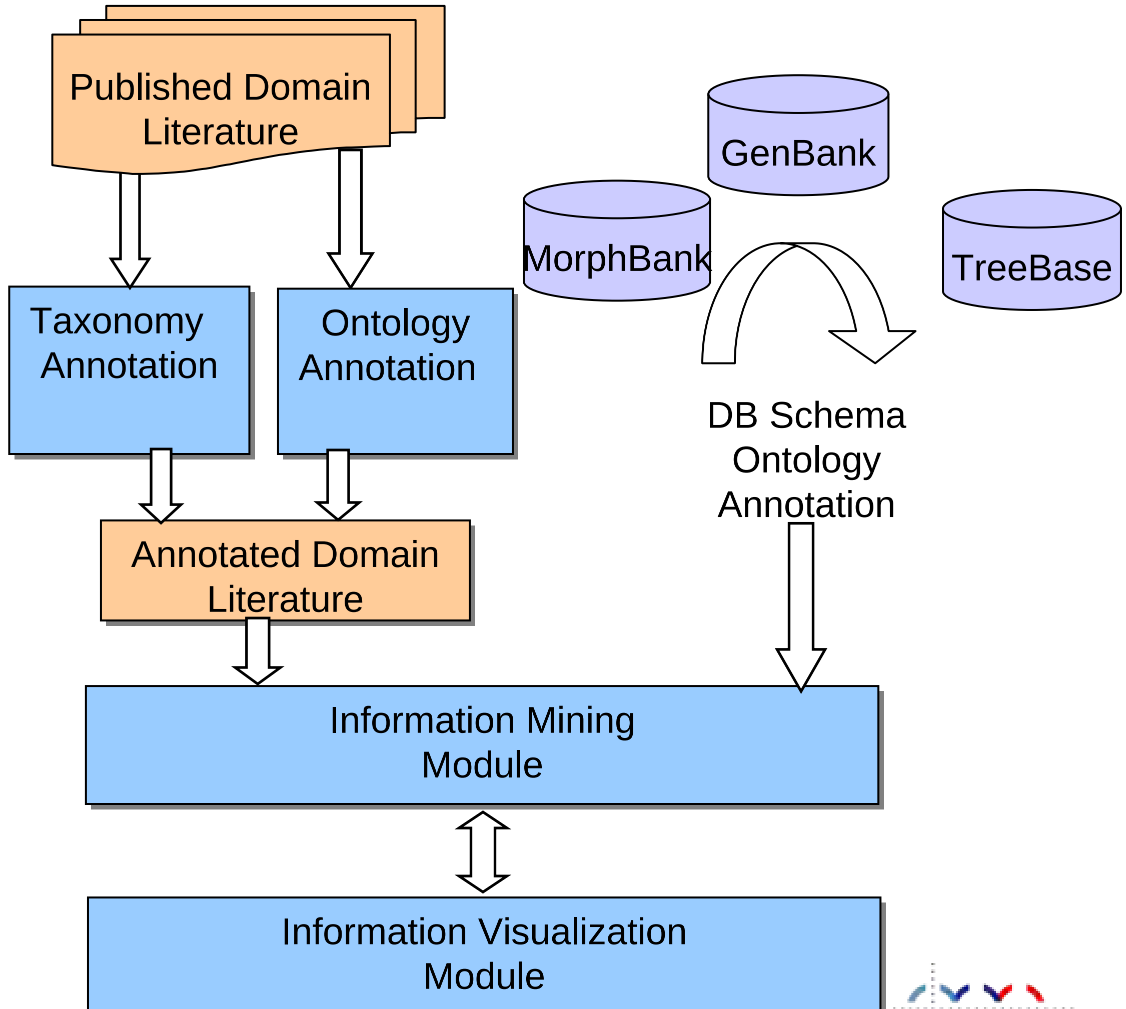


Predictive Disease Susceptibility

Source: David Threadgill/Terry Magnuson, UNC

Data Federation and Info Viz



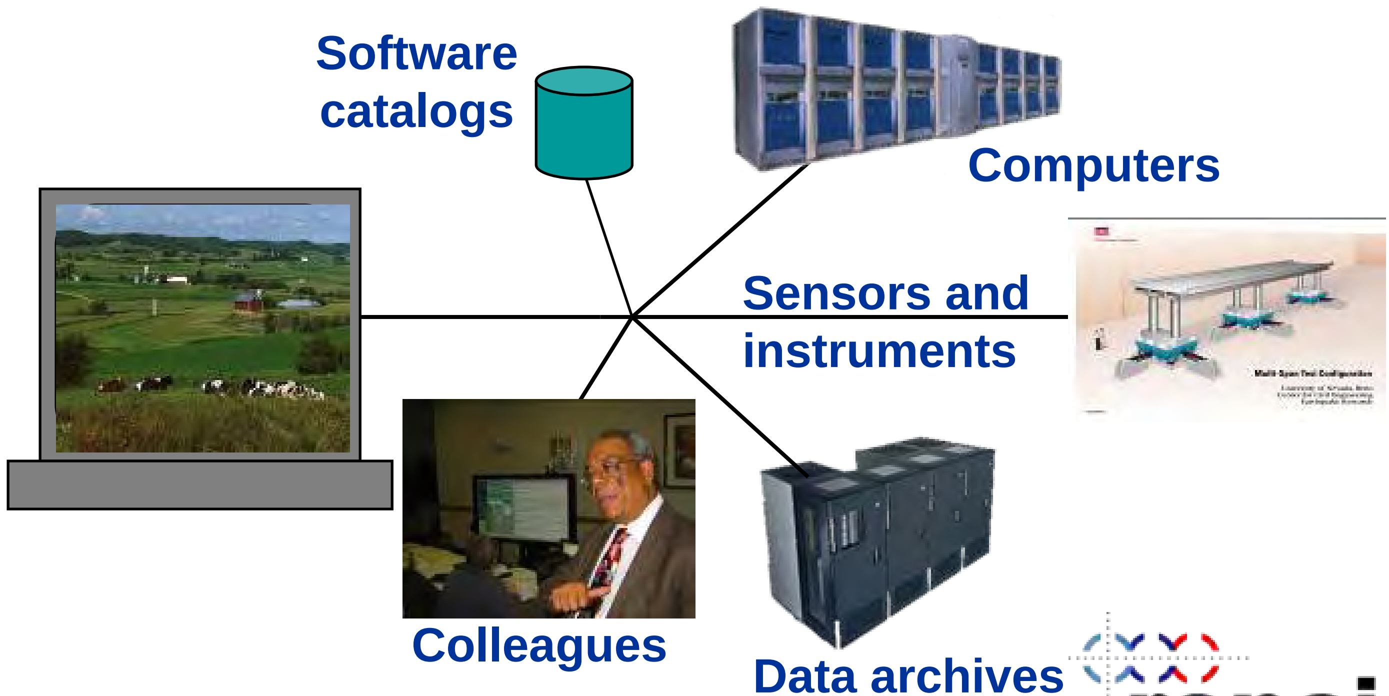


Source: Anita Lungu, Duke

What's A Grid?

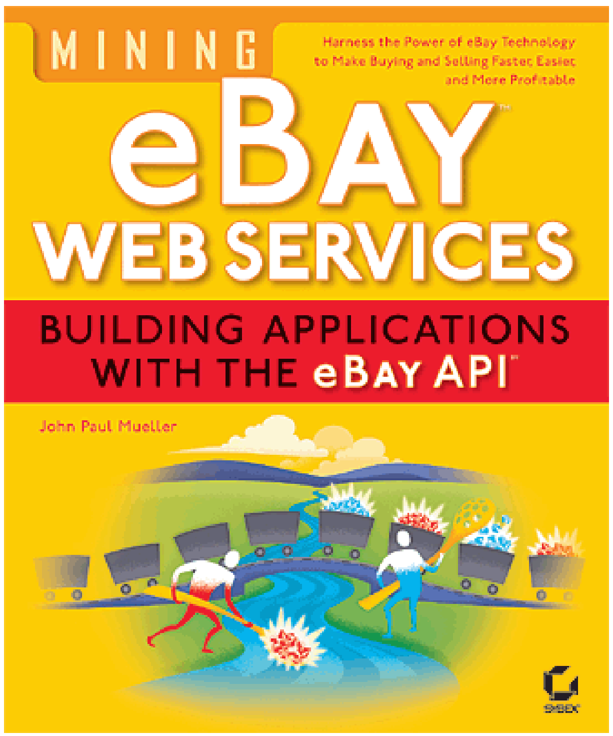
Web: Uniform access to documents

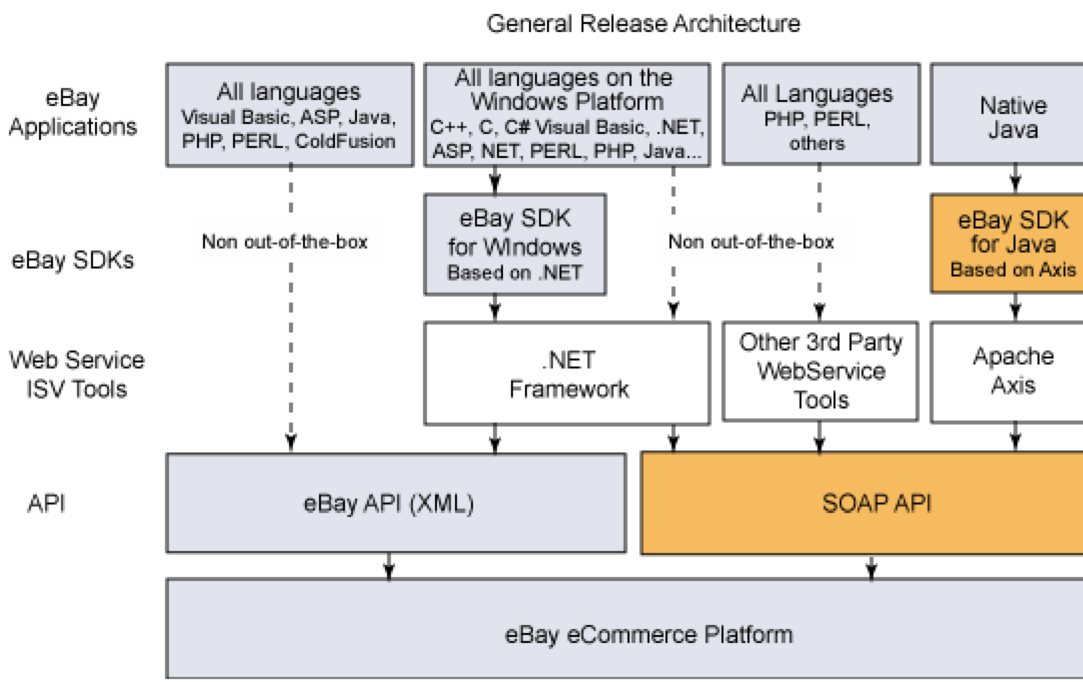
Grid: Flexible, high-performance access to resources and services for *distributed communities*



eBay Web Services Architecture

- Over 40% of eBay's listings are now via API calls





Source: IBM

The World's Online Marketplace

Grids/Web Services of All Flavors

Open Science Grid

The UK GRID

D-GRID

eGee Enabling Grids for E-sciencE

AUSTRIAN GRID

FusionGRID www.fusiongrid.org

PPDG

BIOGRID

myGrid

GridLab

FhRG Fraunhofer Resource Grid

caBIG cancer Biomedical Informatics Grid

HEALTHGRID

NVO NATIONAL VIRTUAL OBSERVATORY

IVOA

nees

Astro Grid

COMPUTING RESEARCH ASSOCIATION CRA

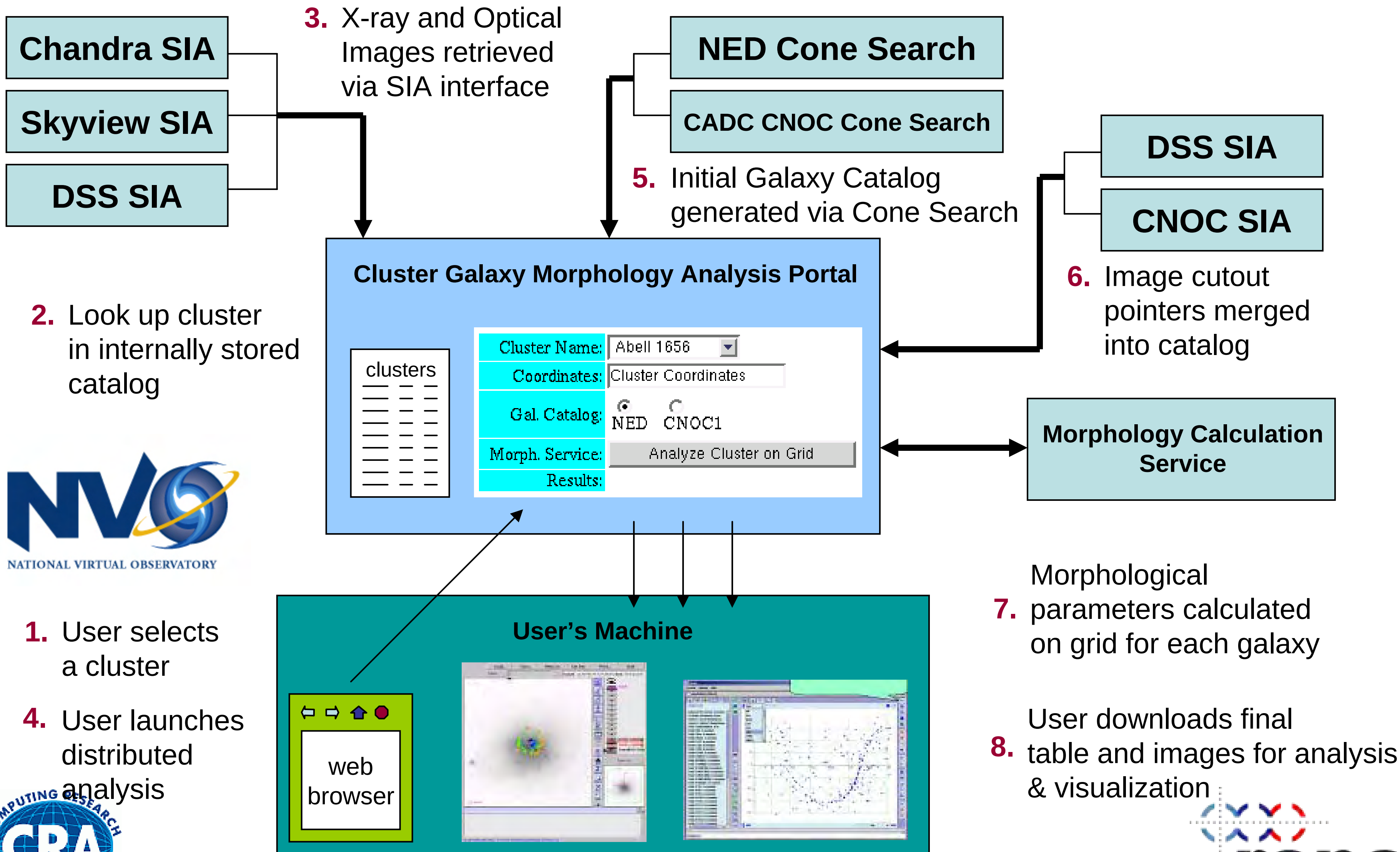
TeraGrid

GRIDS

renci renaissance computing institute

GEON THE GEOSCIENCES NETWORK

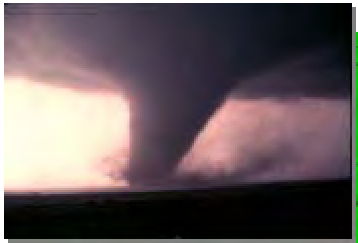
{Inter}national Virtual Observatory

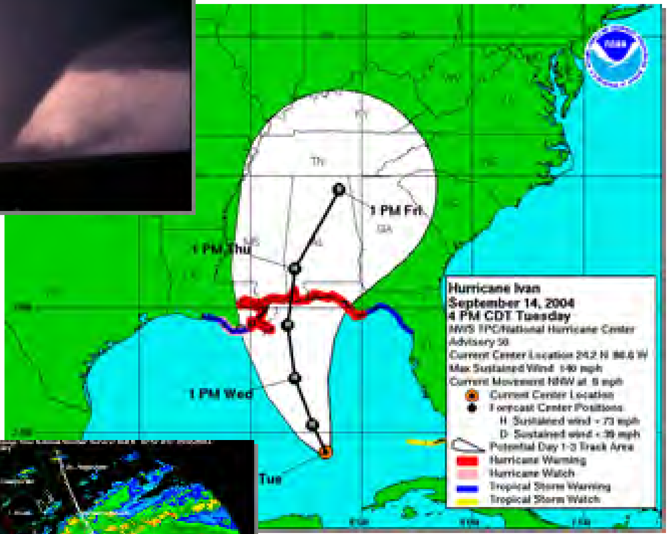


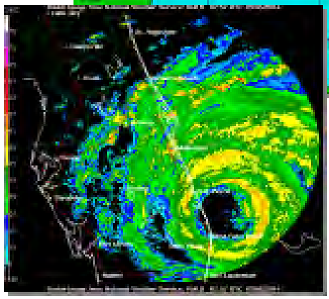
Source: Ray Plante, NCSA

Weather and Economic Loss

- **$10T U.S. economy**
  - 40% is adversely affected by weather and climate
- **$1M in loss to evacuate each mile of coastline**
  - we now over warn by 3X!
  - average over warning
    - 200 miles, or $200M per event
- **Improved forecasts**
  - lives saved and reduced cost
- **LEAD national Grid**
    - Oklahoma, Indiana, UCAR
    - Colorado State, Howard, Alabama
    - Millersville, NCSA, North Carolina

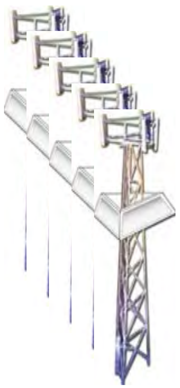






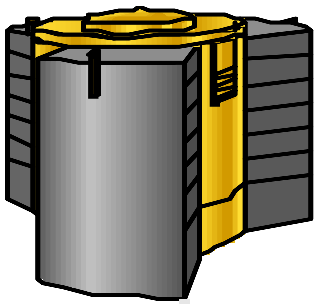
Source: Kelvin Droegemeier, Oklahoma

The LEAD Vision: A Paradigm Shift



Dynamic Observations

Analysis and Assimilation



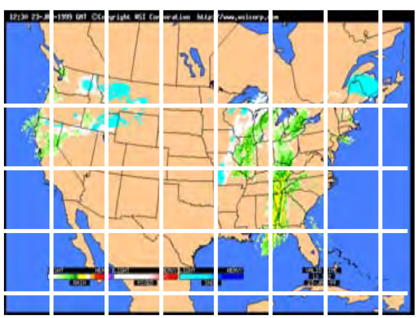
Prediction and Detection

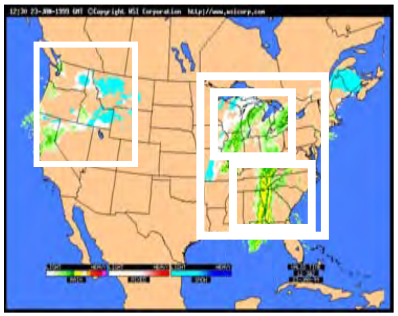
Product Generation and Dissemination

Models/Algorithms Driving Sensors

End Users

The challenge: Build cyberinfrastructure services that provide adaptability, scalability, availability, usability, and real-time response.



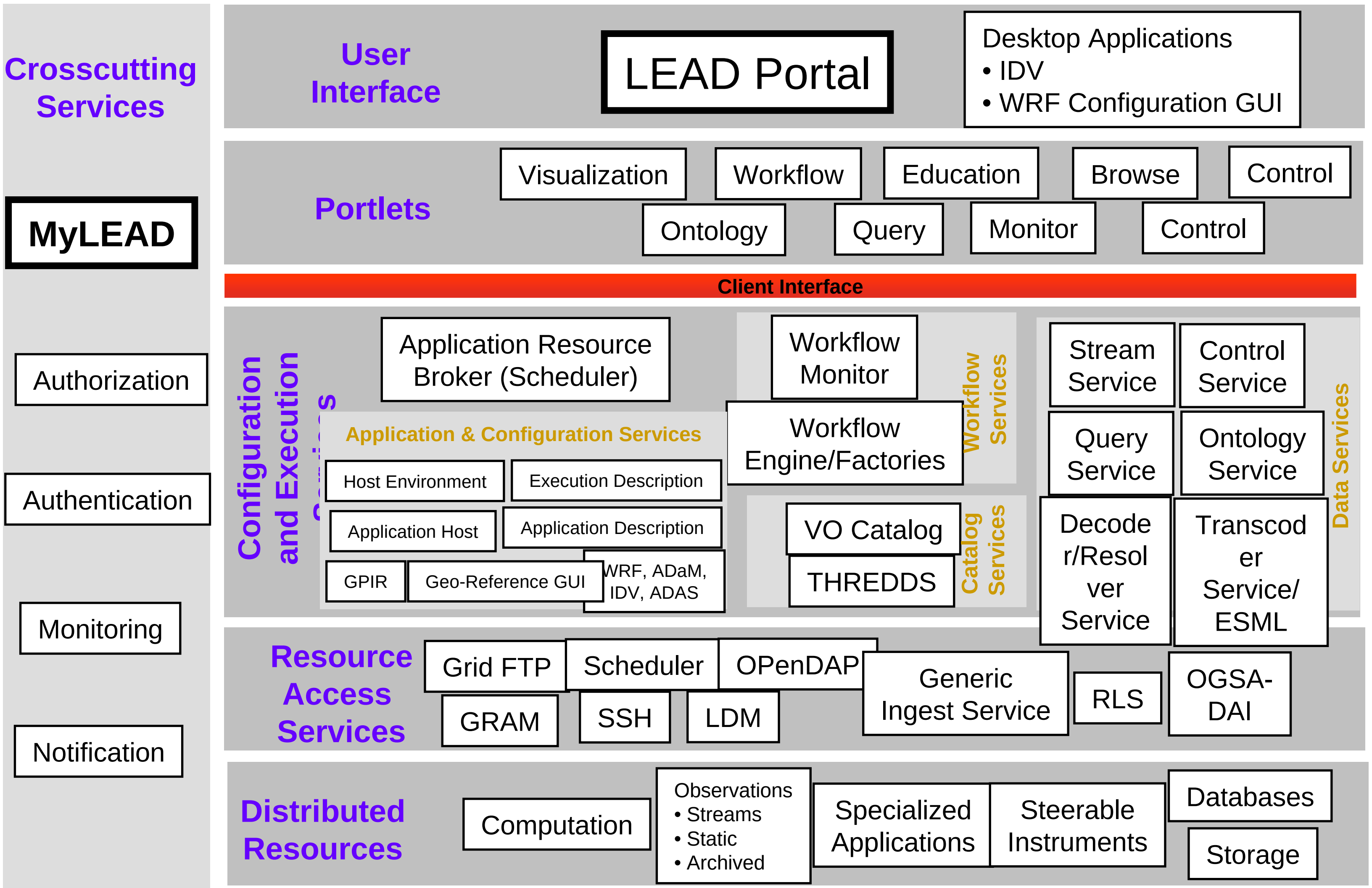


Source: Plale, Indiana

COMPUTING RESEARCH CRA ASSOCIATION

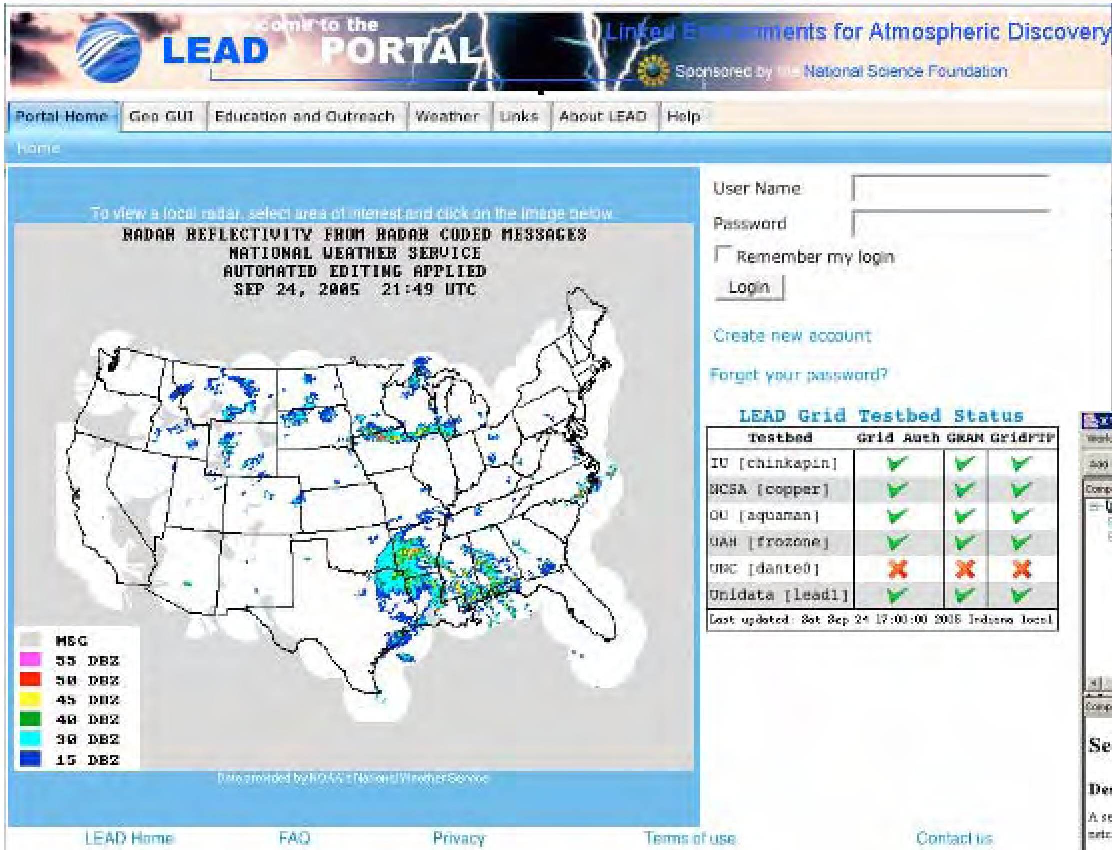
renci renaissance computing institute

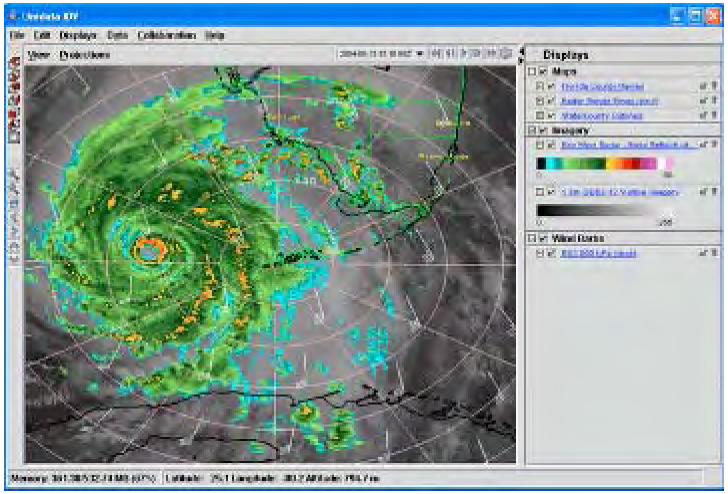
LEAD Service Oriented Architecture

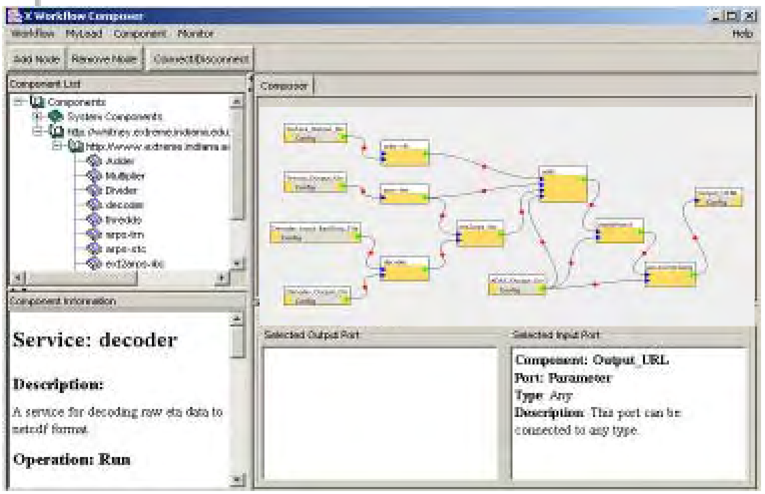


Source: Dennis Gannon, Indiana

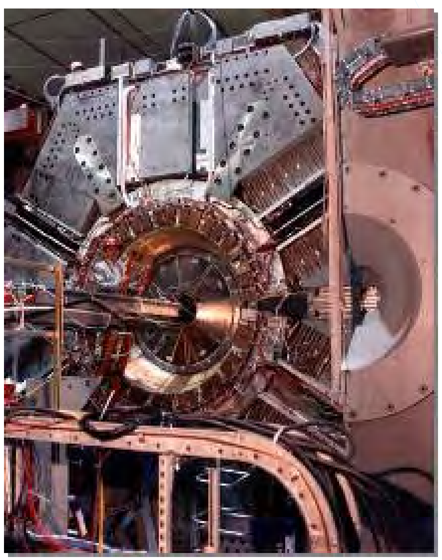
LEAD Mesoscale Meteorology

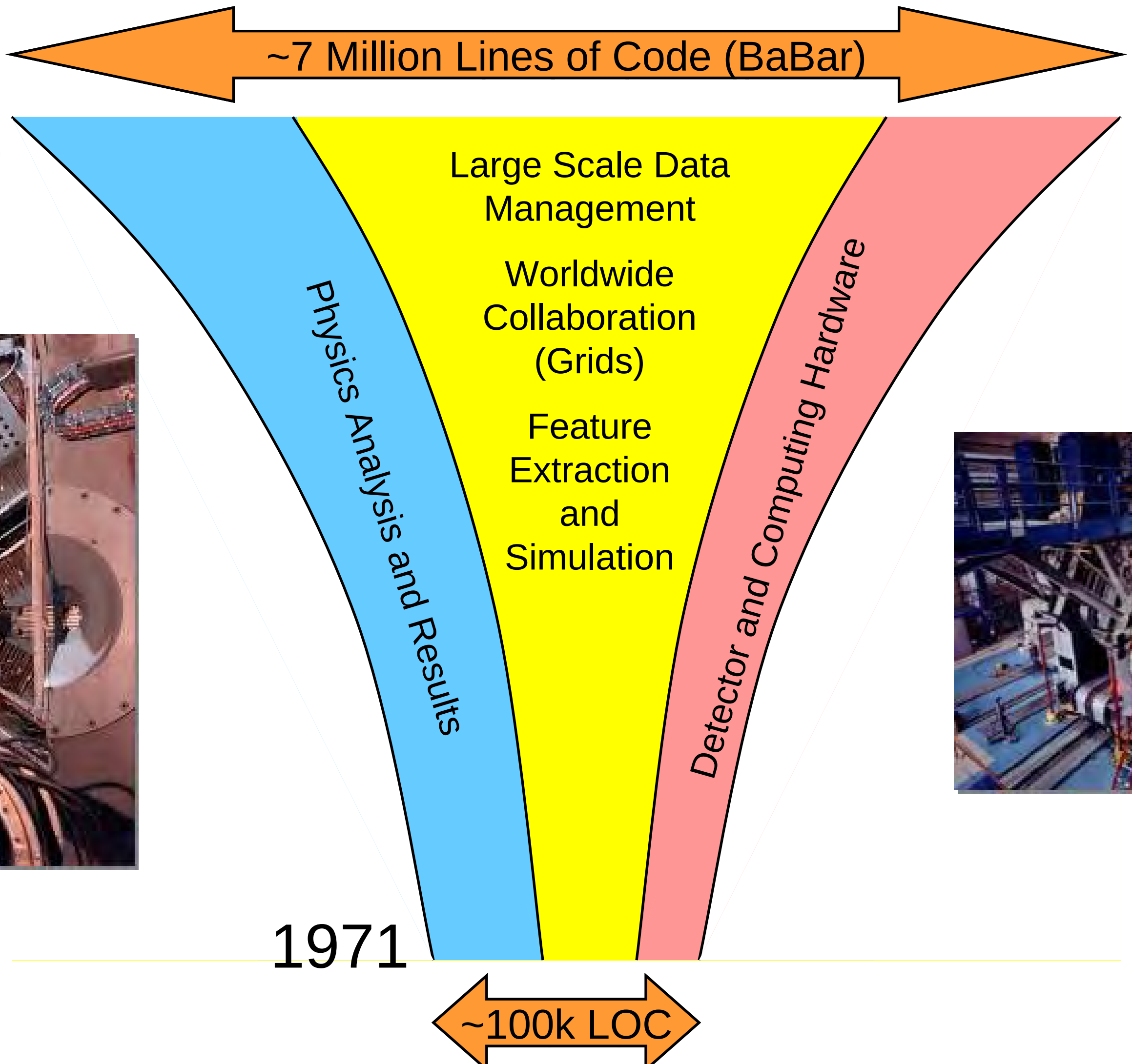






Software Complexity and Collaboration



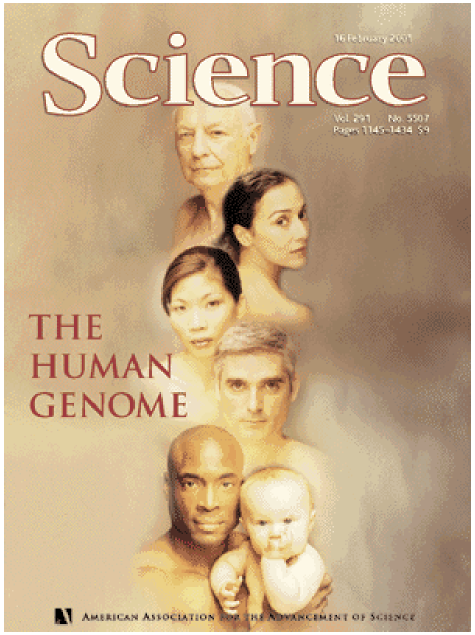


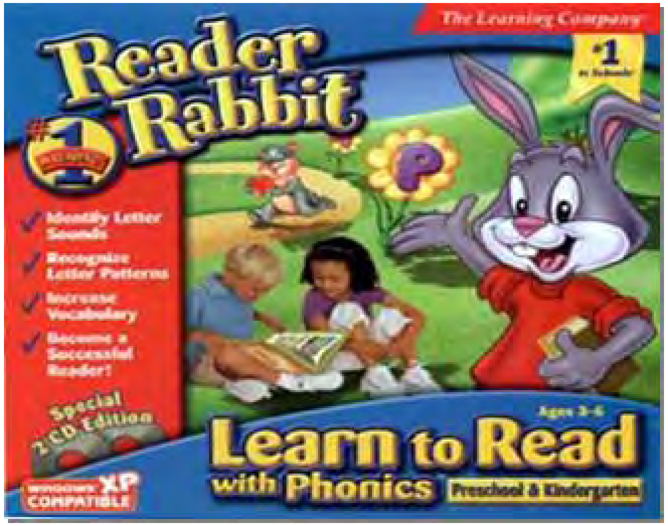
Source: Richard Mount, SLAC

Need: Simple, Easy-To-Use Tools

"Genome. Bought the book. Hard to read."

Eric Lander



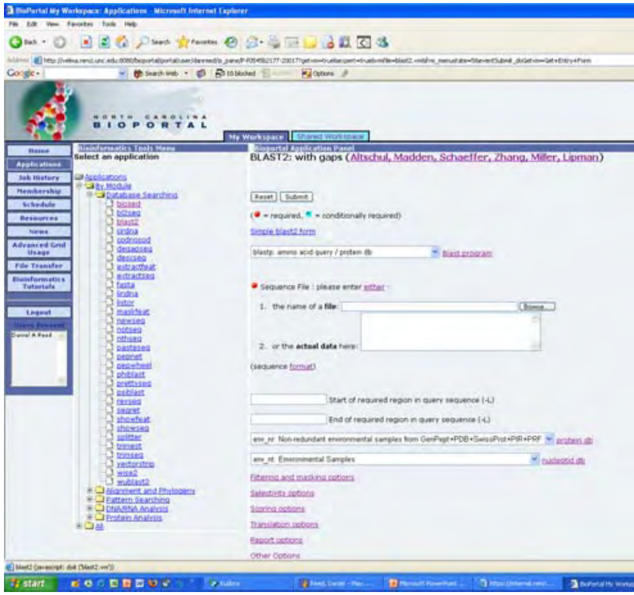


renci
renaissance computing institute

Carolina Bioportal

- **Three overlapping target groups**
  - undergraduate education
  - graduate education and research
  - academic/industrial research
- **Features**
  - access to common bioinformatics tools
  - extensible toolkit and infrastructure
    - OGCE and National Middleware Initiative (NMI)
    - leverages emerging international standards
  - remotely accessible or locally deployable
  - packaged and distributed with documentation
- **National reach and community**
  - NSF TeraGrid deployment
    - science gateway
- **Education and training**
  - hands-on workshops
    - clusters, Grids, portals and bioinformatics

NORTH CAROLINA BIOPORTAL



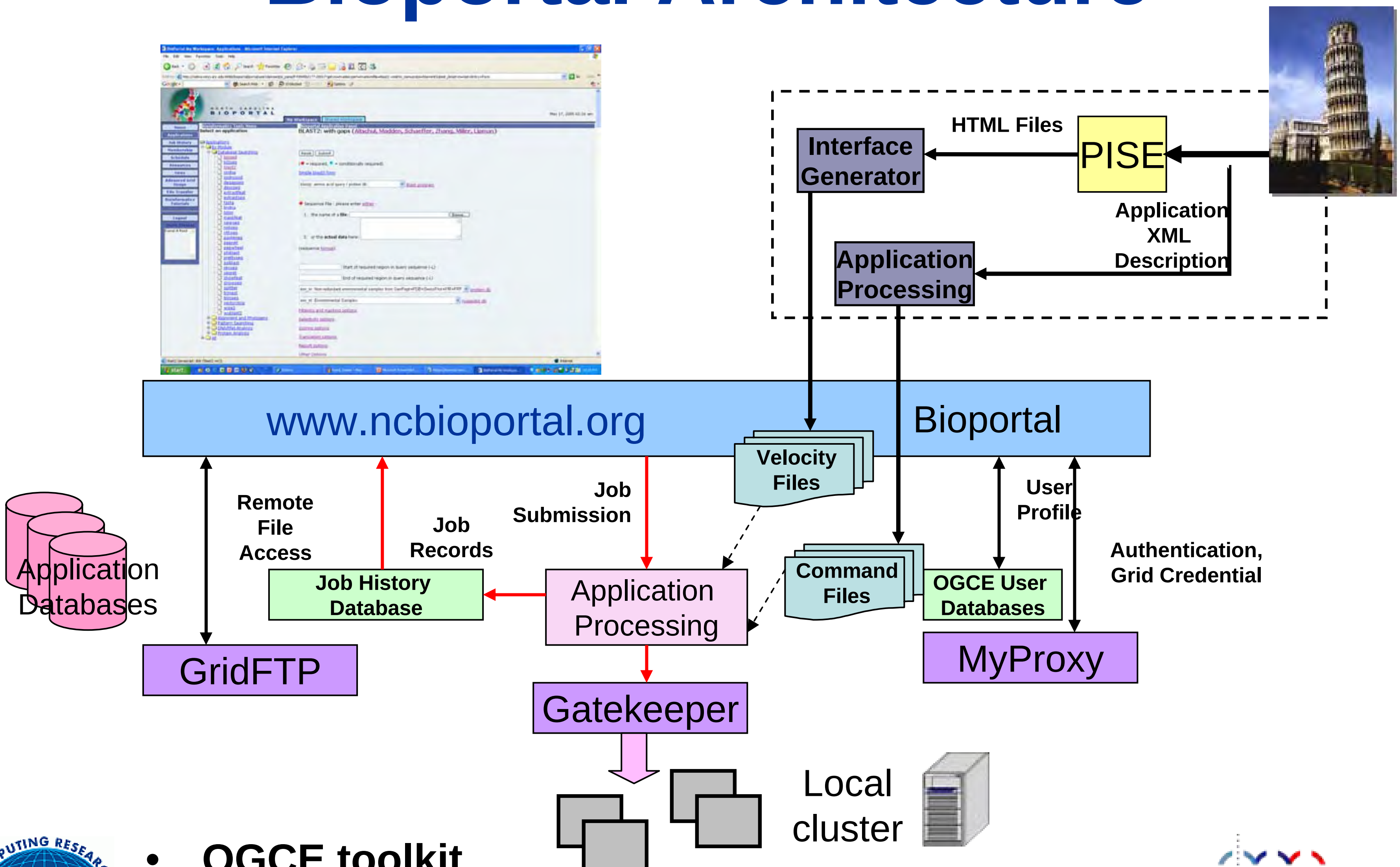
TeraGrid

North Carolina

COMPUTING RESEARCH ASSOCIATION CRA

renci renaissance computing institute

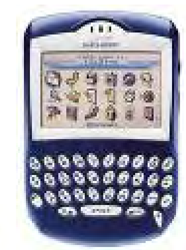
Bioportal Architecture



- **OGCE toolkit**
  - used by cyberinfrastructure projects
    - LEAD, NEES, PACI, DOE, TeraGrid …

Information and Social Processes

- **Google**
  - it's a search engine, it's a verb, …
- **Blogs**
  - published self-expression
- **Instant Messenger**
  - social networks
- **Wireless messaging**
  - semi-synchronous
- **Internet commerce**
  - the dot.com boom/bust
  - EBay, Amazon
- **Spam, phishing, …**
  - anti-social behavior



The Six Computing Eras

- **Big Iron (post WW II)**
  - vacuum tubes and campy science fiction movies
- **Mainframe (‘60s/’70s)**
  - spinning tapes and bad science fiction movies
- **Workstations (‘70s/’80s)**
  - spinning disks and Star Trek™
- **PCs (‘80s/’90s)**
  - spinning CDs and Jurassic Park™
- **Internet (‘90s)**
  - spinning DVDs and Internet pet food companies ☹
- ***Implicit computing (21st century)***
  - IPods™ and The Matrix™
  - embedded intelligence in everyday objects
    - number of processors/person → infinity

COMPUTING RESEARCH ASSOCIATION CRA

renci
renaissance computing institute

Human-Computer Symbiosis

It seems reasonable to envision, for a time 10 or 15 years hence, a 'thinking center' that will incorporate the functions of present-day libraries together with anticipated advances in information storage and retrieval.

The picture readily enlarges itself into a network of such centers, connected to one another by wide-band communication lines and to individual users by leased-wire services. In such a system, the speed of the computers would be balanced, and the cost of the gigantic memories and the sophisticated programs would be divided by the number of users. **J.C.R. Licklider, 1960**

COMPUTING RESEARCH CRA ASSOCIATION

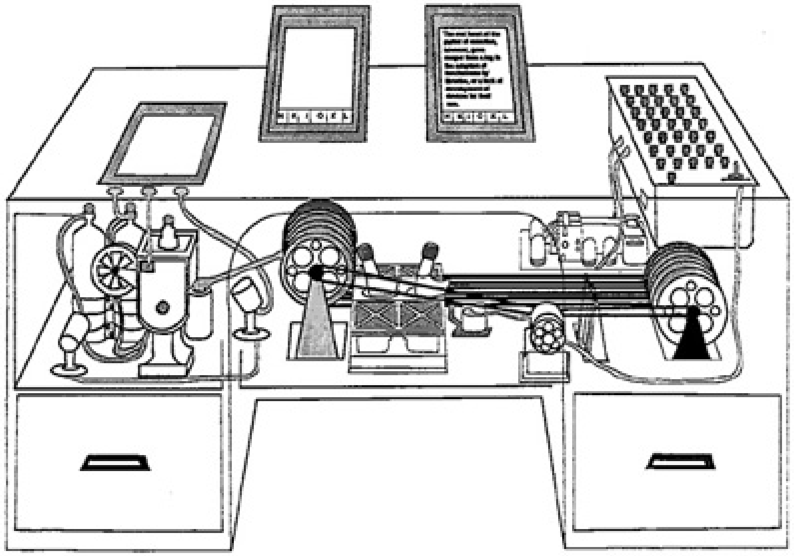
renci renaissance computing institute

Memex: Still Prescient

"Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Vannevar Bush

"As We May Think," 1945



COMPUTING RESEARCH CRA ASSOCIATION

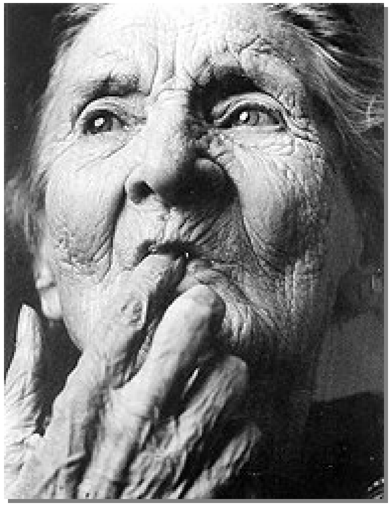
renci renaissance computing institute

Exemplar 21st Century Challenges

- **Population growth**
  - severe weather sensitivity
    - statewide impact
  - geobiology and environment
  - economics and finance
  - sociology and policy

Consilience

- **Economics and healthcare**
  - longitudinal public health data
    - environmental interactions
  - genetic susceptibility
    - heart disease, cancer, Alzheimer's
  - privacy and insurance
  - public policy and coordination

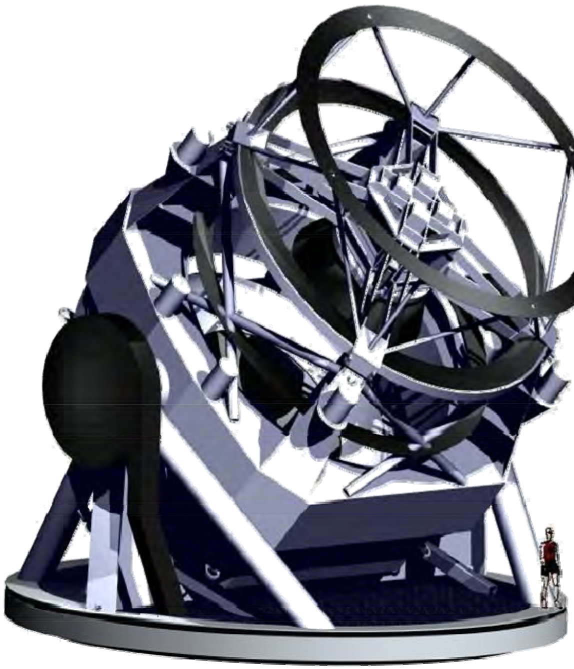


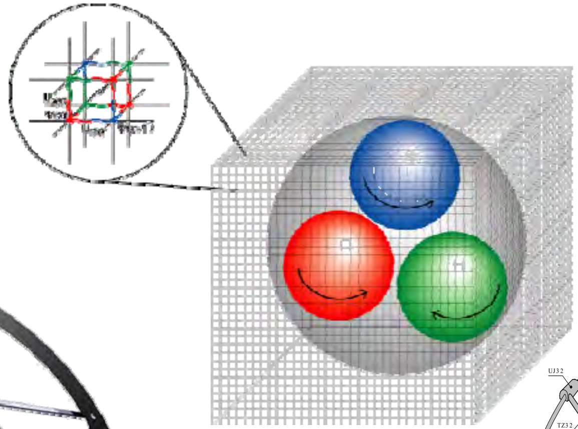
COMPUTING RESEARCH ASSOCIATION CRA

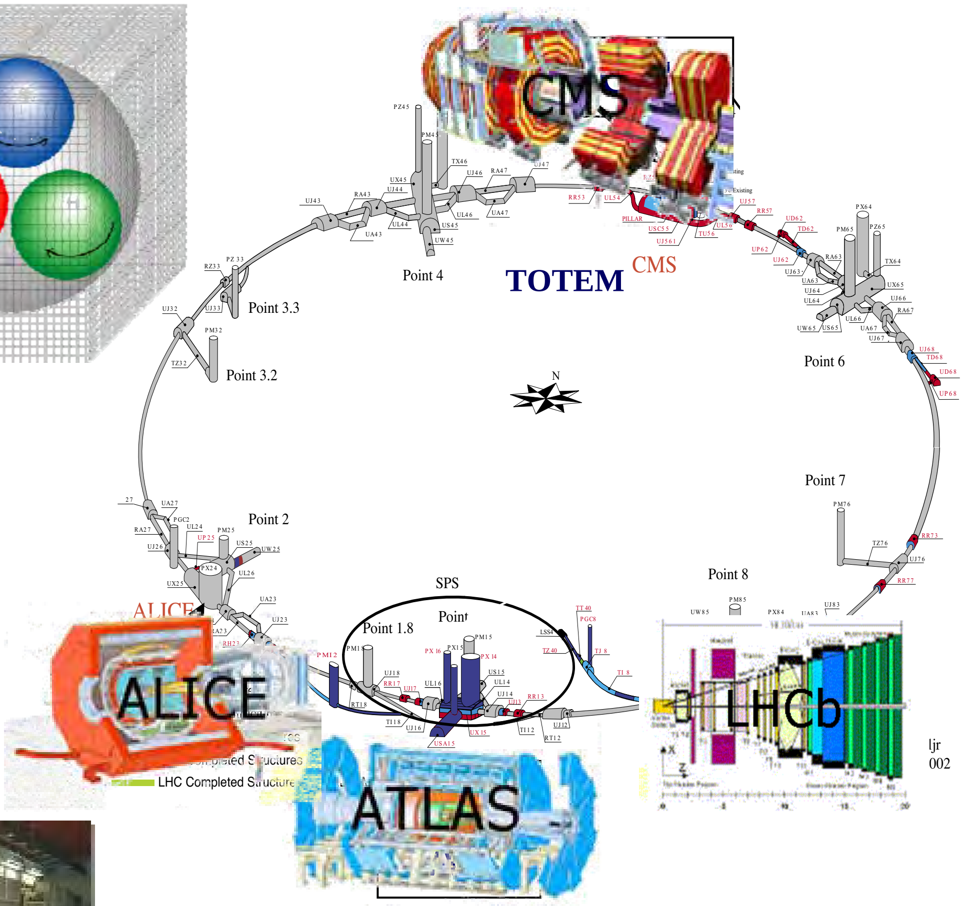
DNAdirect Your Genes In Context

renci renaissance computing institute

LSST, LHC and QCD







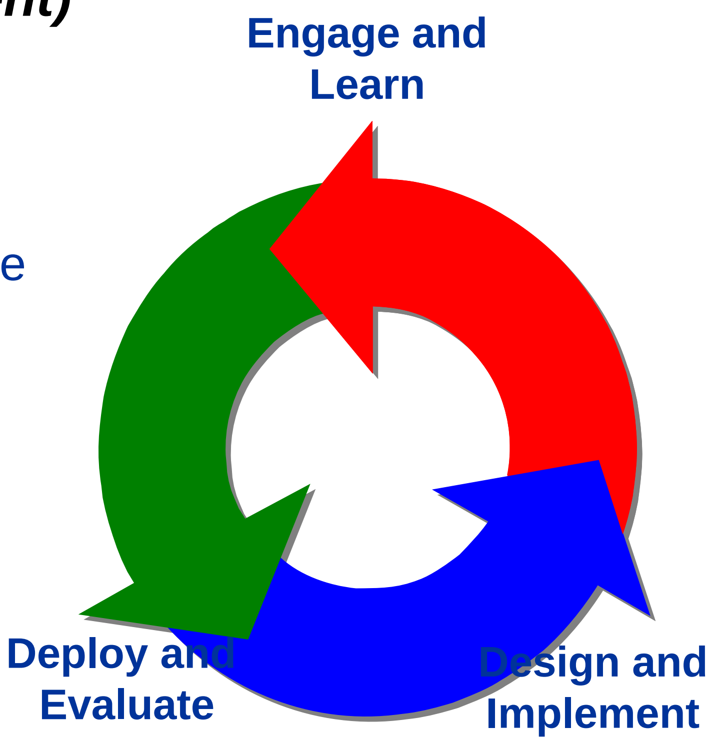
$$D\psi(x)=\frac{1}{2a}\sum_{\mu}\left[U(x)\psi(x+\hat{\mu})-U^{\dagger}(x-\hat{\mu})\psi(x-\hat{\mu})\right]$$

PITAC Report

- **Computational Science: Ensuring America's Competitiveness**
  1. *A Wake-up Call: The Challenges to U.S. Preeminence and Competitiveness*
  2. *Medieval or Modern? Research and Education Structures for the 21st Century*
  3. *Multi-decade Roadmap for Computational Science*
  4. *Sustained Infrastructure for Discovery and Competitiveness*
  5. *Research and Development Challenges*
- **Two key appendices**
  – *Examples of Computational Science at Work*
  – *Computational Science Warnings – A Message Rarely Heeded*
- **Available at www.nitrd.gov**

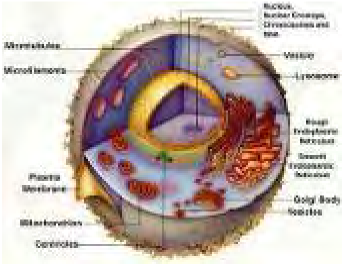
The Virtuous Cycle

- **Live in the future *(research and development)***
  - track evolving infrastructure trends
  - prototype advanced infrastructure
- **Ride the exponentials**
  - see qualitative change from qualitative change
  - recognize the inflection points
    - e.g., personal petabytes are in sight
- **Bring insights to science *(infrastructure)***
  - translate prototypes into use
  - glean insights from science applications
  - expand the user community
- **Learn from experience (good and bad)**
  - enhance the good
  - fix the bad and explore alternatives
- **Recognize that the cycles continue *ad infinitum***
  - commit to continual investment



Some Grand Challenges

1. **Ubiquitous invisibility**
   - successful technologies become “invisible”
   - composable, interoperable systems
2. **Intelligence amplification (Memex)**
   - the right information at the right time
   - seamless modality transduction, situated and mobile
3. **Predictive *in silico* biological models**
   - the “other” artificial life
   - multidisciplinary modeling and integration



4. **The Universe in a Box**
   - origins and alternatives
   - the theory of everything (TOE)
5. **The Cultural Encyclopedia**
   - cultural history, context and the digital village
6. **Grand AI, our long-term CS fascination**
   - deep questions about thinking

COMPUTING RESEARCH ASSOCIATION CRA

renci renaissance computing institute

The Cambrian Explosion

- **Most phyla appear**
  - sponges, archaeocyathids, brachiopods
  - trilobites, primitive mollusks, echinoderms
- **Indeed, most appeared quickly!**
  - Tommotian and Atdbanian
  - as little as five million years
- **Lessons for computing and science**
  - it doesn’t take long when conditions are right
    - raw materials and environment
  - leave fossil records if you want to be remembered!