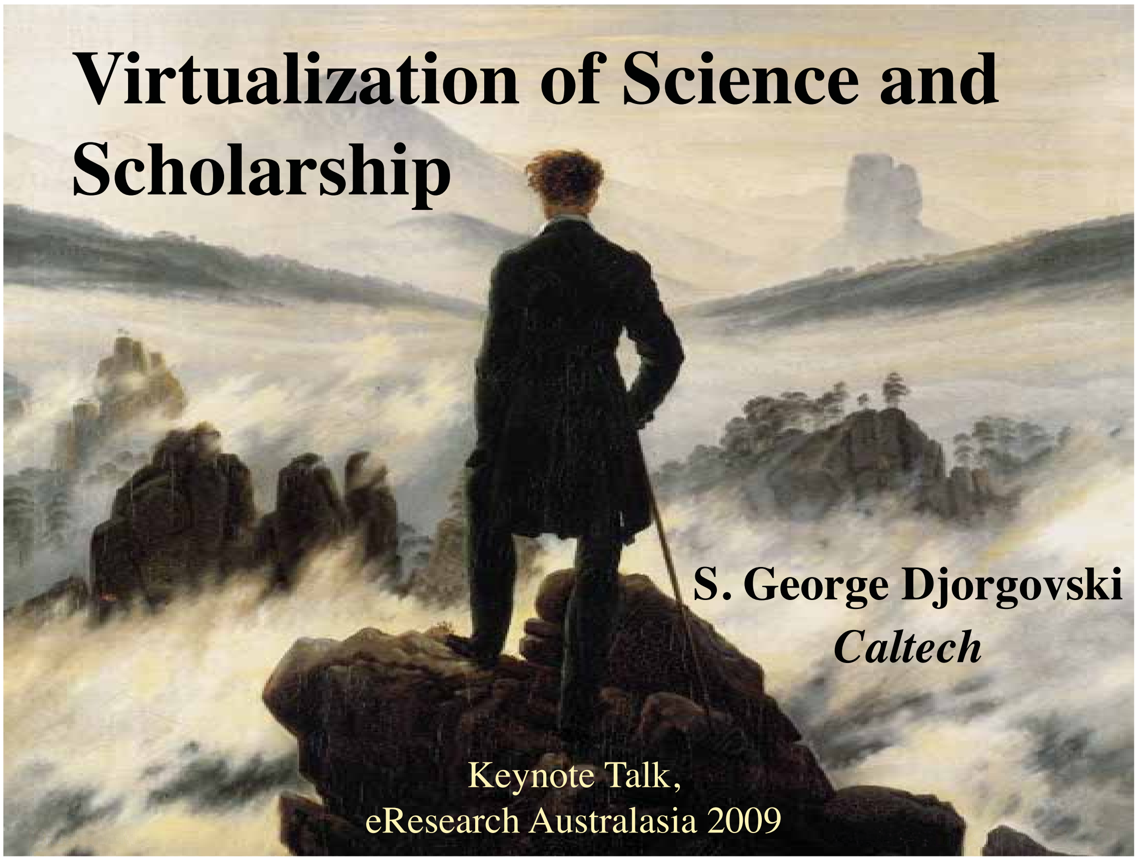


Virtualization of Science and Scholarship

S. George Djorgovski
Caltech

Keynote Talk,
eResearch Australasia 2009

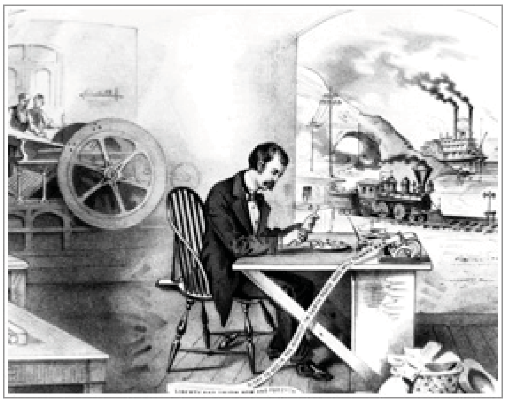
Overture

- The world transformed
- Climbing the S-Curve
  - Science in the exponential world
  - Virtual Observatory: a case study
- The modern scientific process
  - eScience and the new paradigms
  - The evolution of computing
- Scientific communication and collaboration
  - The rise of immersive virtual environments: Web 3.0?
- The growing synergies
  - Exploring and building in cyberspace

Definition: By ***Virtualization***, I mean a migration of the scholarly work, data, tools, methods, etc., to cyber-environments, today effectively the Web

This process is of course not limited to science and scholarship; essentially all aspects of the modern society are undergoing the same transformation

Cyberspace (today the Web, with all information and tools it connects) is increasingly becoming the principal arena where humans interact with each other, with the world of information, where they work, learn, and play



Information technology revolution is historically unprecedented - in its impact it is like the industrial revolution and the invention of printing combined

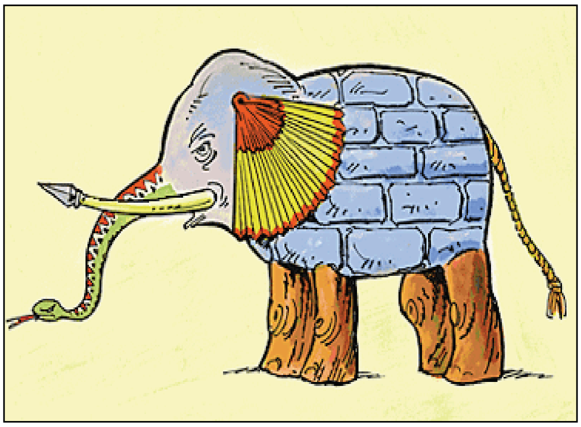
Yet, most fields of science and scholarship have not yet fully adopted the new ways of doing things, and in most cases do not understand them well…

It is a matter of developing a new methodology of science and scholarship for the 21st century

What Is This Beast Called e-Science?

It depends on whom you ask, but some general properties include:

- Computationally enabled
- Data-intensive
- Geographically distributed resources (i.e., Web-based)



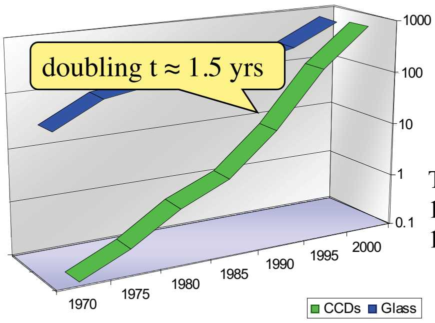
However:

- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) – so this is just a transitional phase
- There is a great emerging synergy of the computationally enabled science, and the science-driven IT

Facing the Data Tsunami

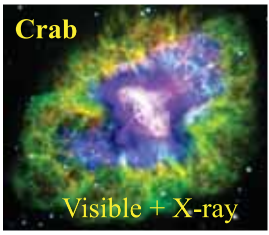
Astronomy, all sciences, and every other modern field of human endeavor (commerce, security, etc.) are facing ***a dramatic increase in the volume and complexity of data***

- We are entering the second phase of the IT revolution: the rise of the ***information/data driven computing***
- The challenges are universal, and growing:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data → **new knowledge**



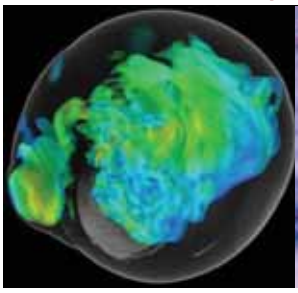
Exponential Growth in Data Volumes and *Complexity*

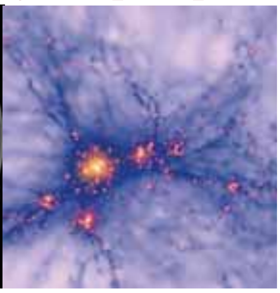
TB's to PB's of data,
10^8 - 10^9 sources,
10^2 - 10^3 param./source



Multi-λ data fusion leads to a more complete, less biased picture (also: multi-scale, multi-epoch, …)

Understanding of complex phenomena requires complex data!





Numerical simulations are also producing many TB's of very complex "data"

Data + Theory = Understanding

Astronomy Has Become Very Data-Rich

- Typical digital sky survey now generates ~ 10 - 100 TB, plus a comparable amount of derived data products
  - PB-scale data sets are on the horizon
- Astronomy today has ~ 1 - 2 PB of archived data, and generates a few TB/day
  - Both data volumes and data rates grow exponentially, with a doubling time ~ 1.5 years
  - Even more important is the growth of *data complexity*
- For comparison:

  Human memory ~ a few hundred MB

  Human Genome < 1 GB

  1 TB ~ 2 million books

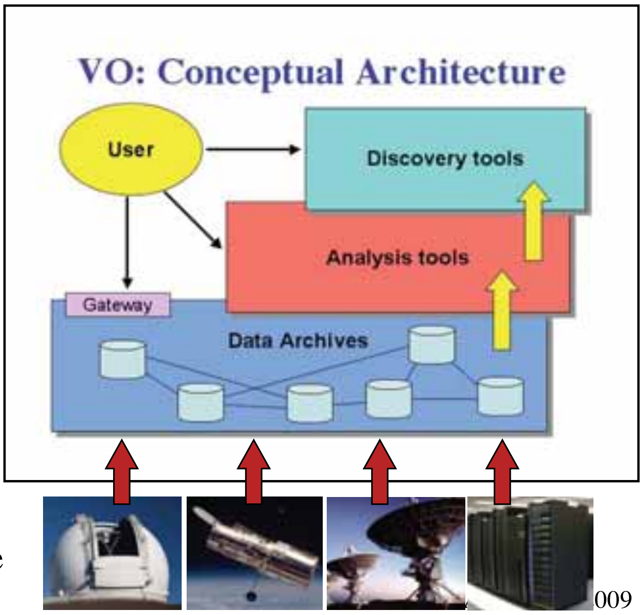
  Library of Congress (print only) ~ 30 TB

The Response of the Scientific Community to the IT Revolution

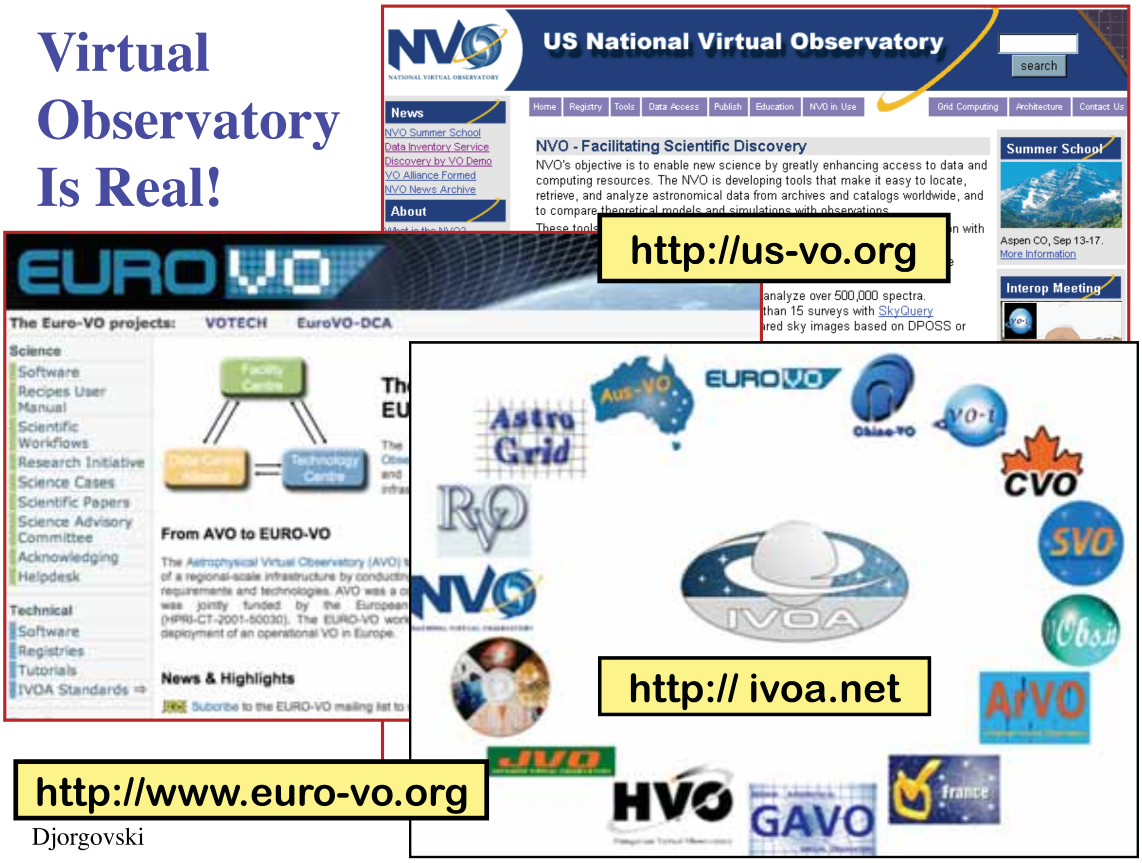
- The rise of **Virtual Scientific Organizations:**
  - Discipline-based, not institution based
  - Inherently distributed, and web-centric
  - Always based on deep collaborations between domain scientists and applied CS/IT scientists and professionals
  - Based on an exponentially growing technology and thus rapidly evolving themselves
  - Do not fit into the traditional organizational structures
  - Great educational and public outreach potential
- However: Little or no coordination and interchange between different scientific disciplines
- Sometimes, entire new fields are created, e.g., bioinformatics, computational biology

The Virtual Observatory Concept

- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***
  - Provide and federate content (data, metadata) services, standards, and analysis/compute services
  - Develop and provide data exploration and discovery tools
  - Harness the IT revolution in the service of astronomy
  - A part of the broader e-Science /Cyber-Infrastructure



Virtual Observatory Is Real!



NVO - Facilitating Scientific Discovery

NVO's objective is to enable new science by greatly enhancing access to data and computing resources. The NVO is developing tools that make it easy to locate, retrieve, and analyze astronomical data from archives and catalogs worldwide, and to compare theoretical models and simulations with observations.

These tools

n with

analyze over 500,000 spectra.

than 15 surveys with SkyQuery

red sky images based on DPOSS or

From AVO to EURO-VO

The Astrophysical Virtual Observatory (AVO) of a regional-scale infrastructure by conductin requirements and technologies. AVO was a c was jointly funded by the European (HPRI-CT-2001-50030). The EURO-VO work deployment of an operational VO in Europe.

News & Highlights

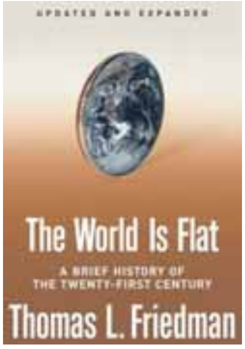
Subcribe to the EURO-VO mailing list to

http://us-vo.org

http:// ivoa.net

http://www.euro-vo.org

Djorgovski



The Sky Is Also Flat

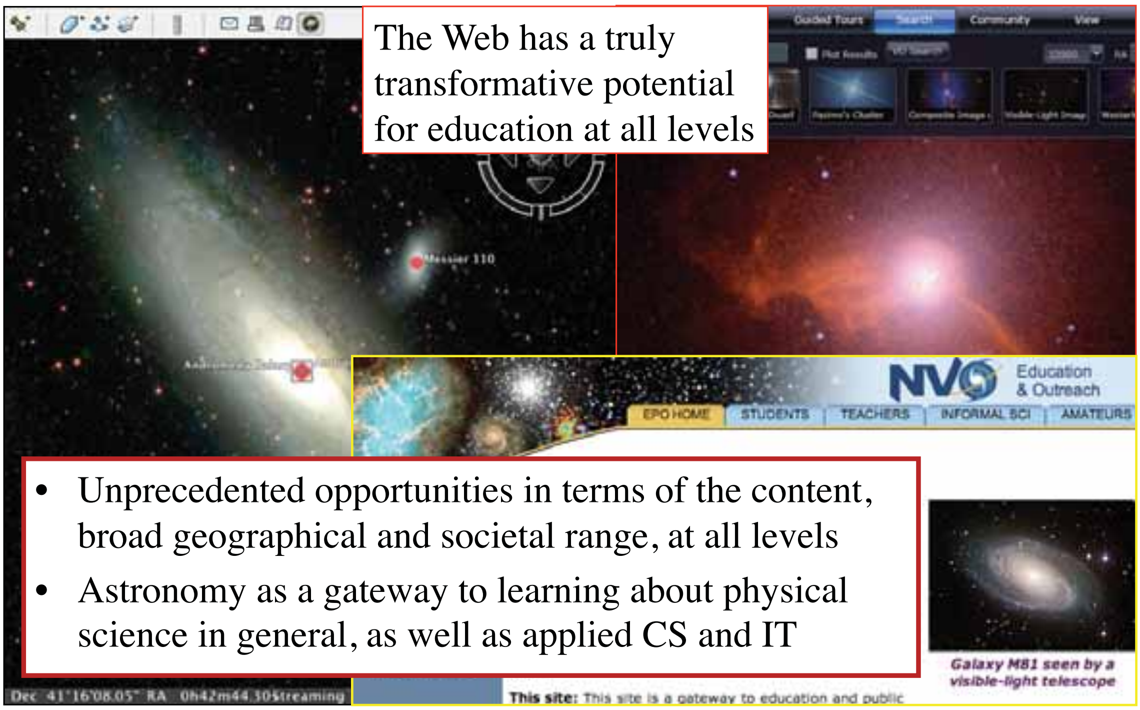
Probably the most important aspect of the IT revolution in science

- **Professional Empowerment:** Scientists and students anywhere with an internet connection should be able to do a first-rate science (access to data *and* tools)
  - A broadening of the talent pool in astronomy, leading to a substantial democratization of the field
- They can also be substantial contributors, not only consumers
  - Riding the exponential growth of the IT is far more cost effective than building expensive hardware facilities, e.g., big telescopes
  - Especially useful for countries without major research facilities

Djorgovski

VO Education and Public Outreach

"Weapons of Mass Instruction"


Guided Tours
Search
Community
View
Messier 110
NVO Education & Outreach
EPO HOME
STUDENTS
TEACHERS
INFORMAL SCI
AMATEURS
Galaxy M81 seen by a visible-light telescope
Dec 41°16'08.05" RA 0h42m44.30Streaming
This site: This site is a gateway to education and public

The Web has a truly transformative potential for education at all levels

- Unprecedented opportunities in terms of the content, broad geographical and societal range, at all levels
- Astronomy as a gateway to learning about physical science in general, as well as applied CS and IT

VO Functionality Today

What we did so far:

- Lots of progress on interoperability, standards, etc.
- An incipient ***data grid of astronomy***
- Some useful web services
- Community training, EPO

What we did not do (yet):

- Significant data exploration and mining tools

That is where the science will come from!

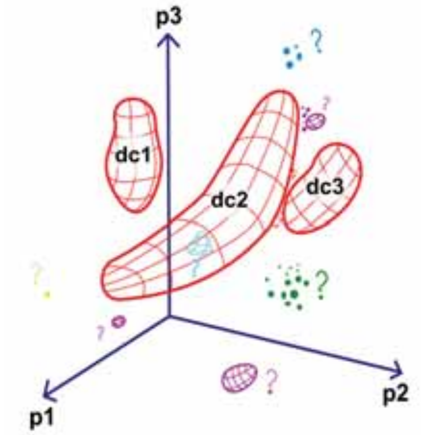
Thus, little VO-enabled science so far

Thus, a slow community buy-in

→ Development of powerful knowledge discovery tools should be a key priority

Donald Rumsfeld's Epistemology

There are known knowns,
There are known unknowns, and
There are unknown unknowns



Or, in other words:

1. Optimized detection algorithms
2. Supervised clustering
3. Unsupervised clustering

The Mixed Blessings of Data Richness

Modern digital sky surveys typically contain ~ 10 - 100 TB, detect $N_{obj} \sim 10^8 - 10^9$ sources, with $D \sim 10^2 - 10^3$ parameters measured for each one -- and multi-PB data sets are on the horizon

Potential for discovery $\left\{ \begin{array}{ll} N_{obj} \text{ or data volume} & \rightarrow \text{Big surveys} \\ N_{surveys}^2 \text{ (connections)} & \rightarrow \text{Data federation} \end{array} \right.$

Great! However … **DM algorithms scale very badly:**

- Clustering ~ $N \log N \rightarrow N^2, \sim D^2$
- Correlations ~ $N \log N \rightarrow N^2, \sim D^k \ (k \geq 1)$
- Likelihood, Bayesian ~ $N^m \ (m \geq 3), \sim D^k \ (k \geq 1)$

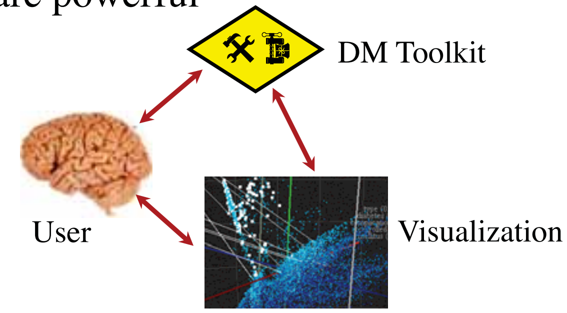
Scalability and ***dimensionality reduction*** (without a significant loss of information) are ***critical needs!***

The Curse of Hyperdimensionality

- **Visualization!**

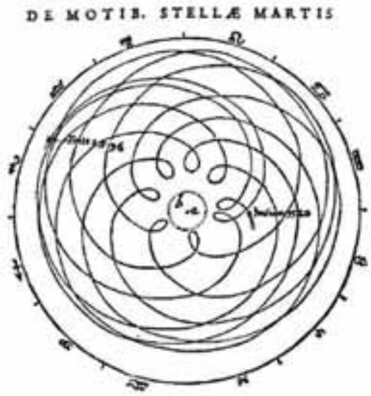
Not a matter of hardware or software, but ***new ideas***

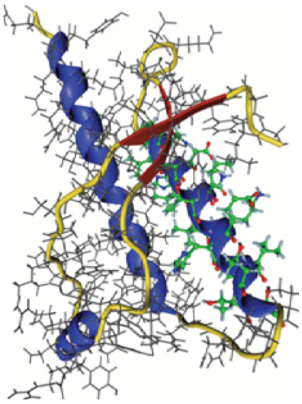
- A fundamental limitation of the human perception: D_{MAX} = 3? 5? 10? (We can understand mathematically much higher dimensionalities, but cannot really visualize them; our own Neural Nets are powerful pattern recognition tools)
- Interactive visualization must be a key part of the data mining process:

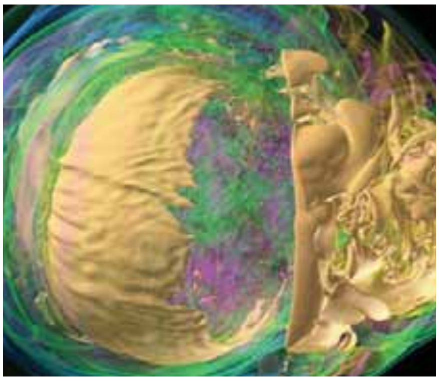


- Dimensionality reduction via machine discovery of patterns / substructures and correlations in the data?

Effective visualization is the bridge between quantitative information, and human intuition

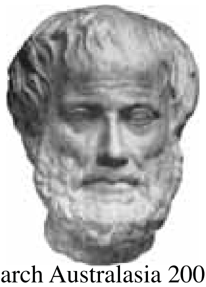


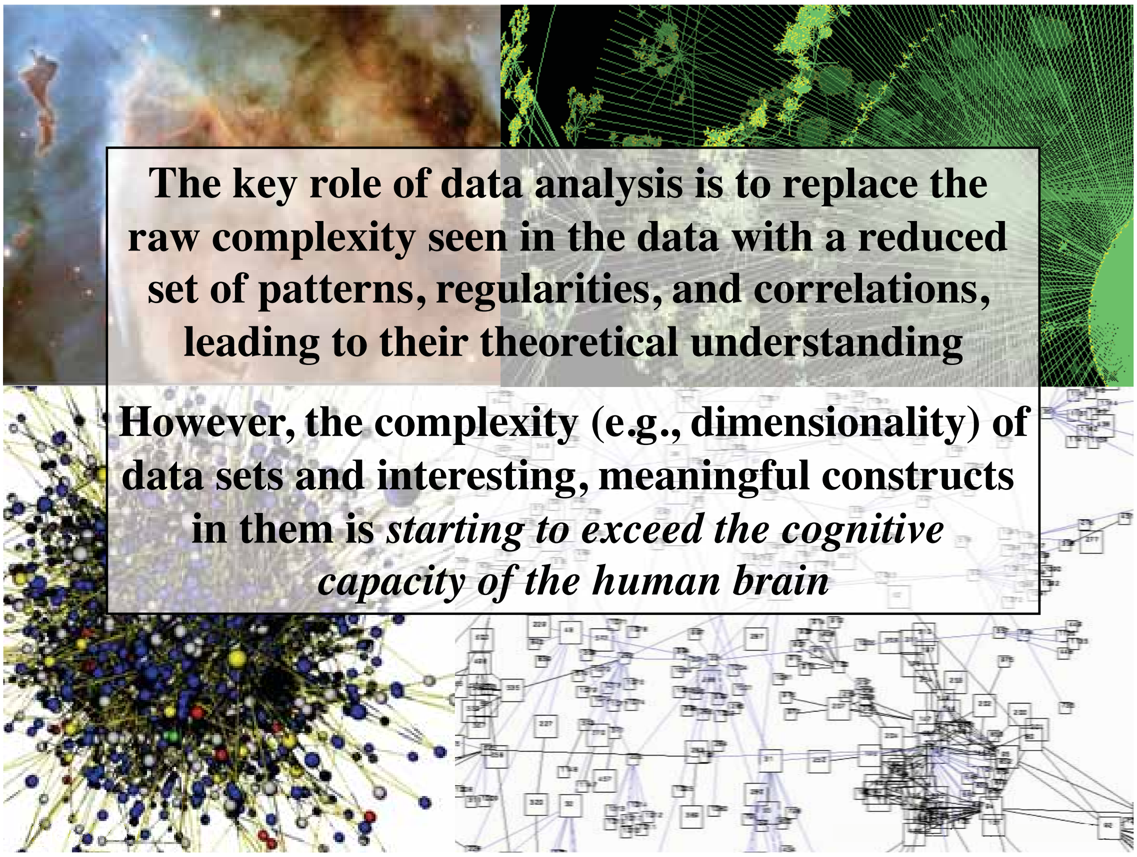




Man cannot understand without images;
the image is a similitude of a corporeal thing,
but understanding is of universals which are
to be abstracted from particulars

Aristotle, *De Memoria et Reminiscentia*



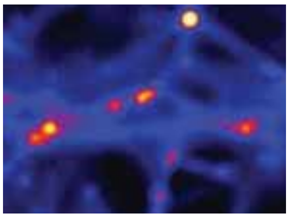


The key role of data analysis is to replace the raw complexity seen in the data with a reduced set of patterns, regularities, and correlations, leading to their theoretical understanding

However, the complexity (e.g., dimensionality) of data sets and interesting, meaningful constructs in them is *starting to exceed the cognitive capacity of the human brain*

This is a Very Serious Problem

- Hyperdimensional structures (clusters, correlations, etc.) are likely present in many complex data sets, whose dimensionality is commonly in the range of D ~ $10^2 – 10^4$, and will surely grow
- It is not only the matter of ***data understanding***, but also of choosing the appropriate data mining algorithms, and interpreting their results
  - Things are seldom Gaussian in reality
  - The clustering topology can be complex

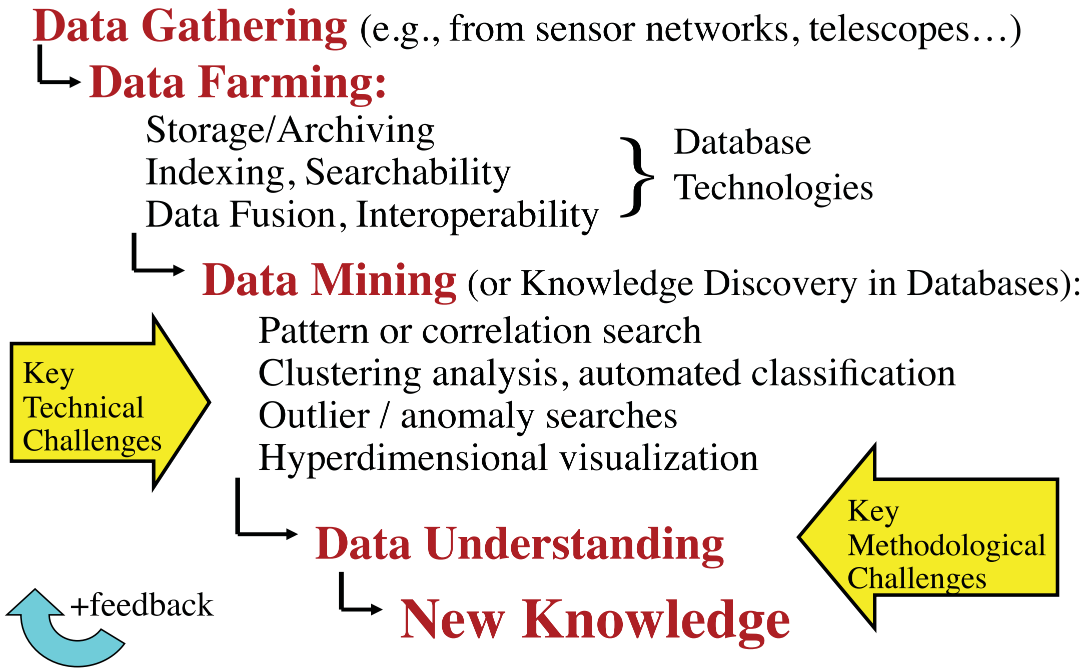


What good are the data if we cannot effectively extract knowledge from them?

"A man has got to know his limitations"
Dirty Harry, an American philosopher

Djorgovski

A Modern Scientific Discovery Process



Data Gathering (e.g., from sensor networks, telescopes…)

Data Farming:

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability

} Database Technologies

Data Mining (or Knowledge Discovery in Databases):

Pattern or correlation search
Clustering analysis, automated classification
Outlier / anomaly searches
Hyperdimensional visualization

Data Understanding

New Knowledge

Key Technical Challenges

Key Methodological Challenges

+feedback

Information Technology → New Science

- The information volume grows exponentially

Most data will never be seen by humans!

➡ The need for data storage, network, database-related technologies, standards, etc.

- Information complexity is also increasing greatly

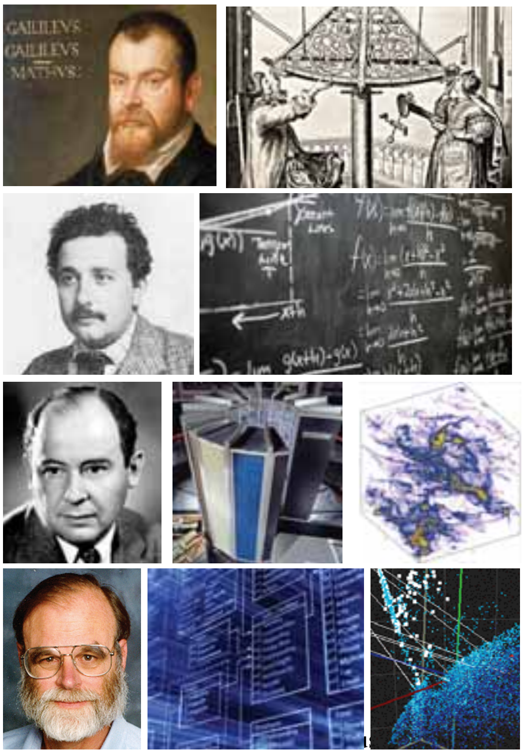
Most data (and data constructs) cannot be comprehended by humans directly!

➡ The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/ Machine-assisted discovery …

- We need to create *a new scientific methodology* on the basis of applied CS and IT
- Important for practical applications beyond science

The Evolving Paths to Knowledge

- The First Paradigm: Experiment/Measurement
- The Second Paradigm: Analytical Theory
- The Third Paradigm: Numerical Simulations
- The Fourth Paradigm: Data-Driven Science?

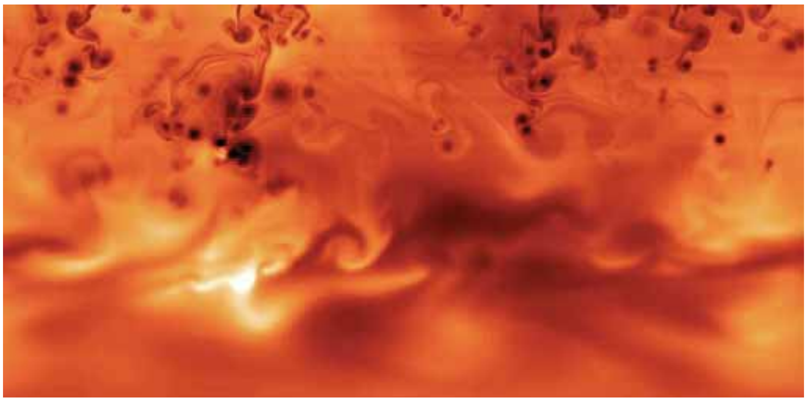


Numerical Simulations:

A qualitatively new (and necessary) way of doing theory - beyond analytical approach

Simulation output - a data set - is the theoretical statement, not an equation

↑ Formation of a cluster of galaxies



← Turbulence

The Fourth Paradigm?

Is this really something ***qualitatively new***, rather than the same old data analysis, but with more data?

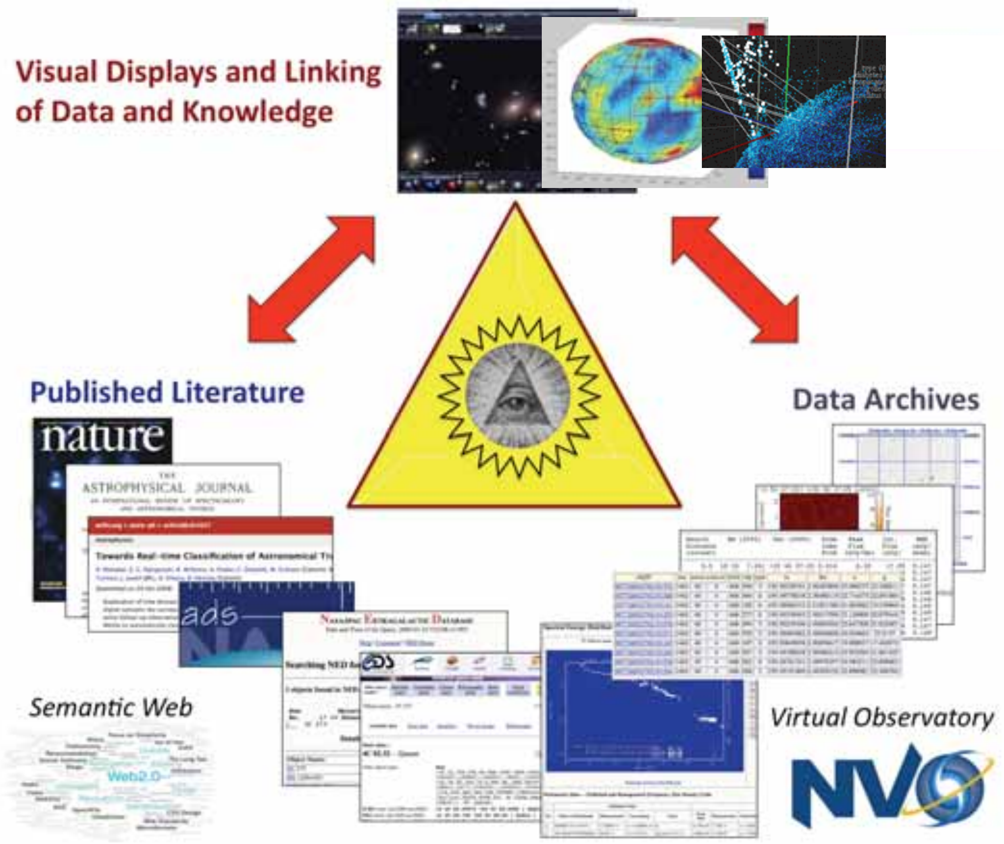
- The information content of modern data sets is so high as to enable discoveries which were not envisioned by the data originators
- Data fusion reveals new knowledge which was implicitly present, but not recognizable in the individual data sets
- Complexity threshold for a human comprehension of complex data constructs? Need new methods to make the data understanding possible

Data Fusion + Data Mining + Machine Learning = The Fourth Paradigm

The Roles for Machine Learning and Machine Intelligence in CyberScience:

+

- **Data processing:**
  - Object / event / pattern classification
  - Automated data quality control (glitch/fault detection and repair)
- **Data mining, analysis, and understanding:**
  - Clustering, classification, outlier / anomaly detection
  - Pattern recognition, hidden correlation search
  - Assisted dimensionality reduction for hyperdim. visualisation
  - Workflow control in Grid-based apps
- **Data farming and data discovery:** semantic web, and beyond
- **Code design and implementation:** from art to science?



Visual Displays and Linking of Data and Knowledge

Published Literature

Data Archives

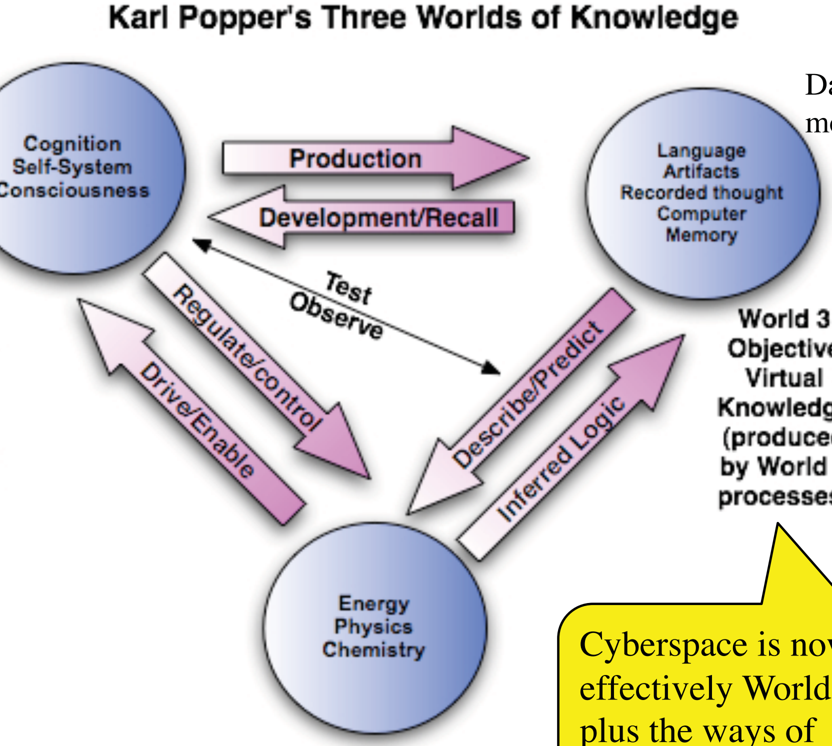
Semantic Web

Virtual Observatory

The Book and the Cathedral …

… and the Web, and the Computer

Technologies for information storage and access are **evolving**, and so does scholarly publishing



Dawkins memes

World 2
Personal
Knowledge
(emerges
from World
1
processes)

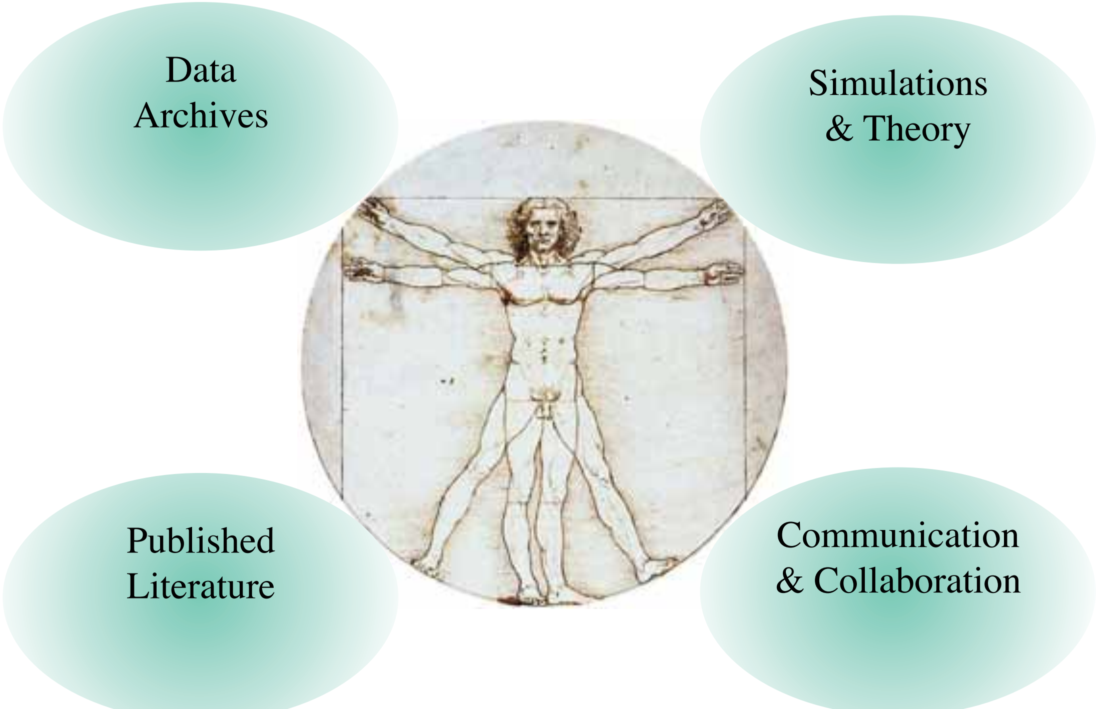
World 3
Objective
Virtual
Knowledge
(produced
by World 2
processes)

Cyberspace is now effectively World 3, plus the ways of interacting with it

K. Popper, *Objective Knowledge: An Evolutionary Approach*, 1972

World 1
Existence/Reality

Science Commons, or Discovery Space



A Lot of Science Originates in Discussions and Constructive Interactions

This creative process can be enabled and enhanced using virtual interactive spaces, including the Web2.0 tools

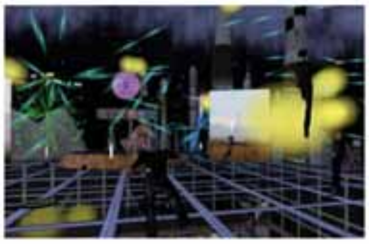
Computing as a Communication Tool

With the advent of the Web, most of the computing usage is not in a number crunching, but in a search, manipulation, and display of data and information, and increasingly also for ***human interactions*** (e.g., much of Web 2.0)

The way in which we interact with computers, and with each other, and with the world of information using computers, is evolving

From ASCI text terminals …

… to Web browsers and hypertext …

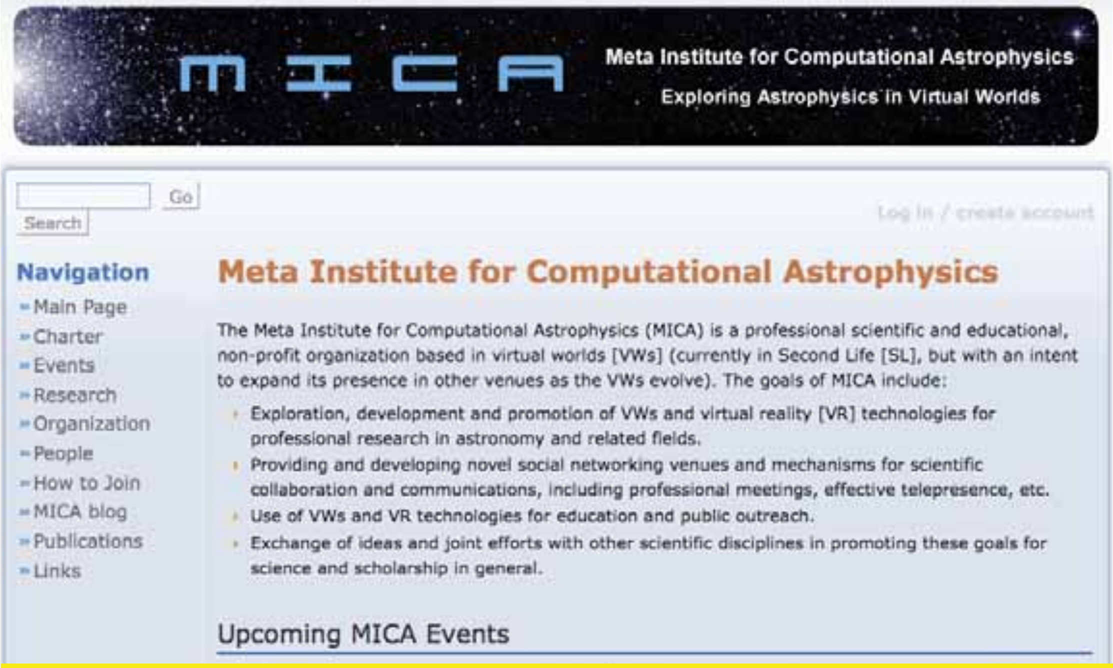


… and now immersive virtual environments

Information Technology as a Communication Medium: Social Networking and Beyond

- Science originates on the interface between human minds, and the human minds and data (measurements, structured information, output of simulations)
- Thus, any technology which facilitates these interactions is an enabling technology for science, scholarship, and intellectual progress more generally
- ***Virtual Worlds*** (or immersive VR) are one such technology, and will likely revolutionize the ways in which we interact with each other, and with the world of information we create
- Thus, we started the *Meta-Institute for Computational Astrophysics (MICA)*, the first professional scientific organization based entirely in VWs

MICA Website: http://mica-vw.org/



Meta Institute for Computational Astrophysics

The Meta Institute for Computational Astrophysics (MICA) is a professional scientific and educational, non-profit organization based in virtual worlds [VWs] (currently in Second Life [SL], but with an intent to expand its presence in other venues as the VWs evolve). The goals of MICA include:

- Exploration, development and promotion of VWs and virtual reality [VR] technologies for professional research in astronomy and related fields.
- Providing and developing novel social networking venues and mechanisms for scientific collaboration and communications, including professional meetings, effective telepresence, etc.
- Use of VWs and VR technologies for education and public outreach.
- Exchange of ideas and joint efforts with other scientific disciplines in promoting these goals for science and scholarship in general.

Upcoming MICA Events

MICA is an experiment in the scholarly use of VWs technologies

StellaNova Sim, MICA's home in SL:

A pleasant virtual environment for scientific communication, collaboration, and experiments

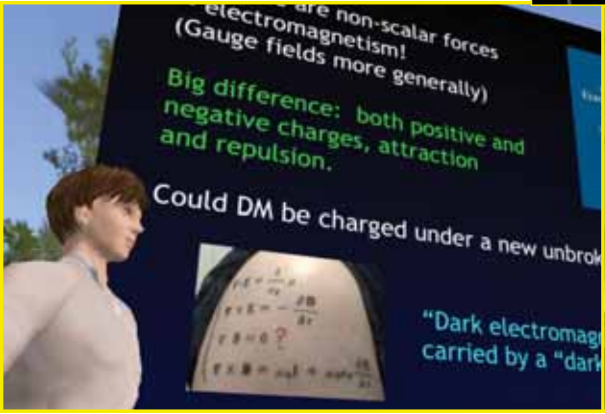
http://slurl.com/secondlife/StellaNova/127/129/32

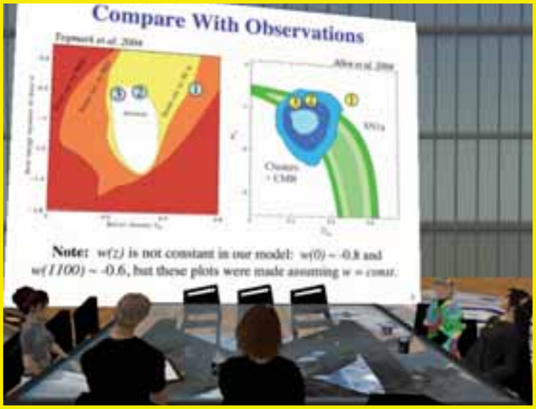
A part of the SciLands virtual continent: **http://www.scilands.org/**

What do we do?

Professional Seminars, Workshops, Group Collaboration Meetings, etc.

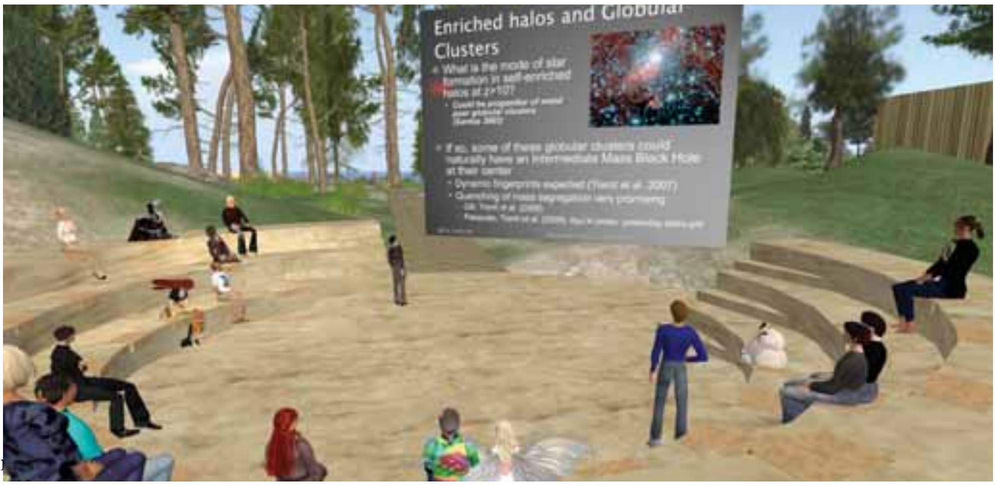
Computation occurs in distributed
environments
Intelligent agent systems managing d
discovery and analysis according to a
beliefs and intents
Programmed with declarative language
Contextualisation data structures


are non-scalar forces
electromagnetism!
(Gauge fields more generally)
Big difference: both positive and
negative charges, attraction
and repulsion.
Could DM be charged under a new unbrok
"Dark electromag
carried by a "dark


Compare With Observations
Note: w(z) is not constant in our model: w(0) ~ -0.8 and
w(1100) ~ -0.6, but these plots were made assuming w = const.

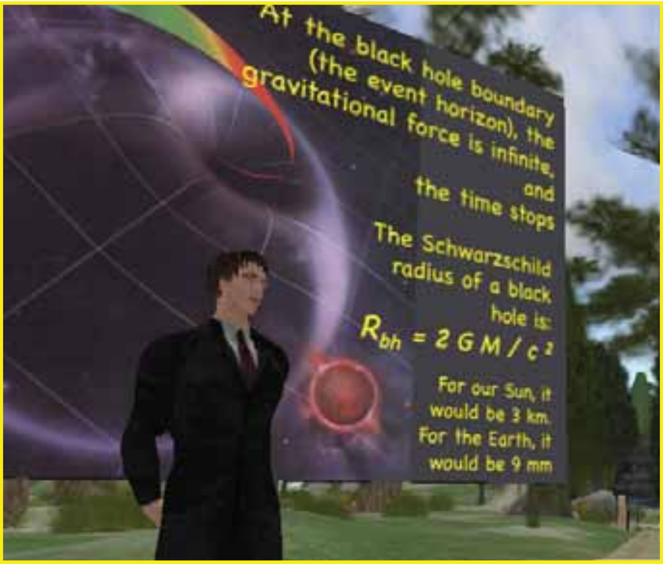
Scientific Communication and Collaboration in VR Environments

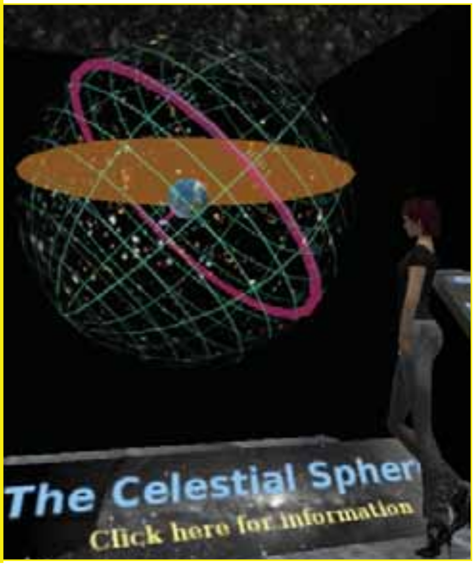
- Subjective experience quality much higher than traditional videoconferencing (and it can only get better as VR improves)
- Effective worldwide telecommuting, at ~ zero cost
- Professional conferences easily organized, at ~ zero cost


Enriched halos and Globular Clusters
What is the mode of star formation in self-enriched halos at z>10?
If so, some of these globular clusters could naturally have an Intermediate Mass Black Hole at their center
Dynamic fingerprints expected (Trenti et al. 2007)
Quenching of mass segregation very promising

Education and Public Outreach

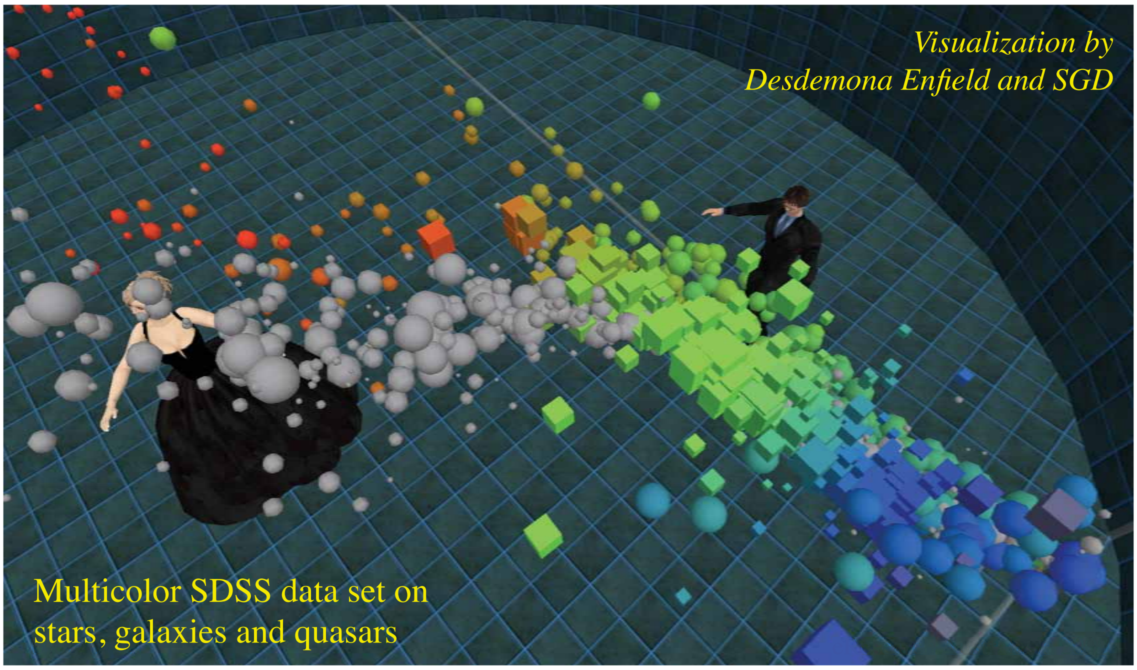
- Already a very powerful platform: public lectures, etc.
- Many virtual science museums already exist
- Classroom instruction being explored


At the black hole boundary (the event horizon), the gravitational force is infinite, and the time stops
The Schwarzschild radius of a black hole is:
$R_{bh} = 2GM/c^2$
For our Sun, it would be 3 km.
For the Earth, it would be 9 mm



Immersive Data Visualization

- Encode up to a dozen dimensions for a parameter space representation
- Interactive data exploration in a pseudo-3D environment

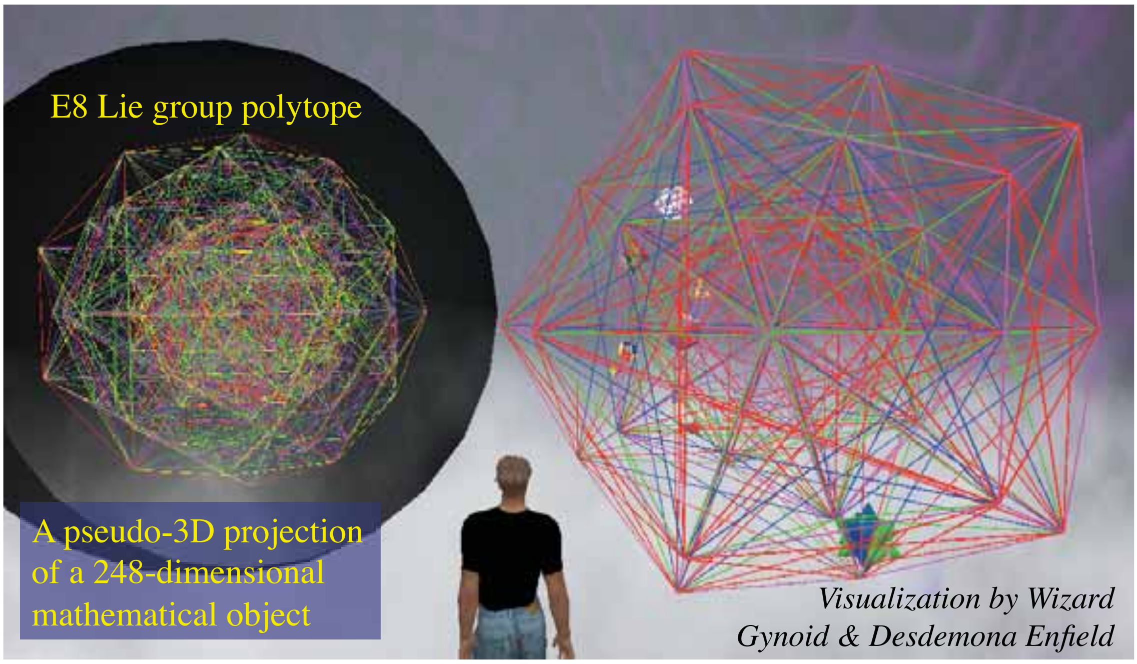


Visualization by Desdemona Enfield and SGD

Multicolor SDSS data set on stars, galaxies and quasars

Immersive Mathematical Visualization

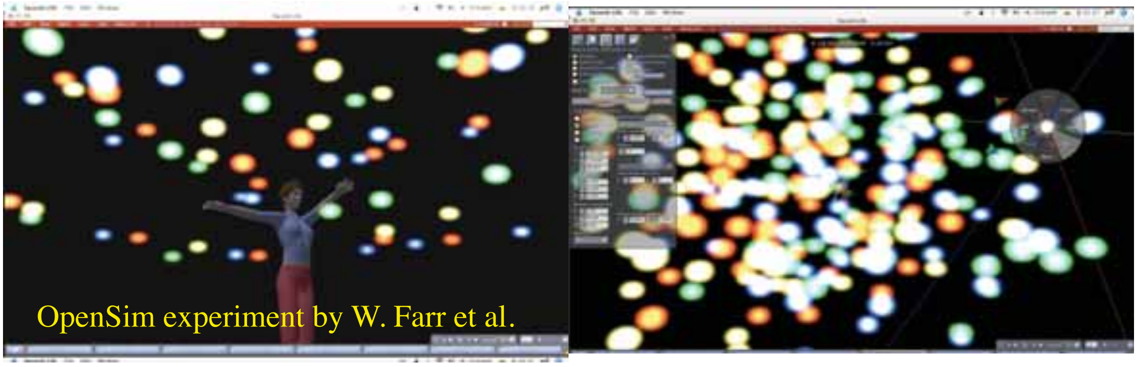
- Pseudo-3D representation of highly-dimensional mathematical objects
- Potential research and educational uses: geometry, topology, etc.



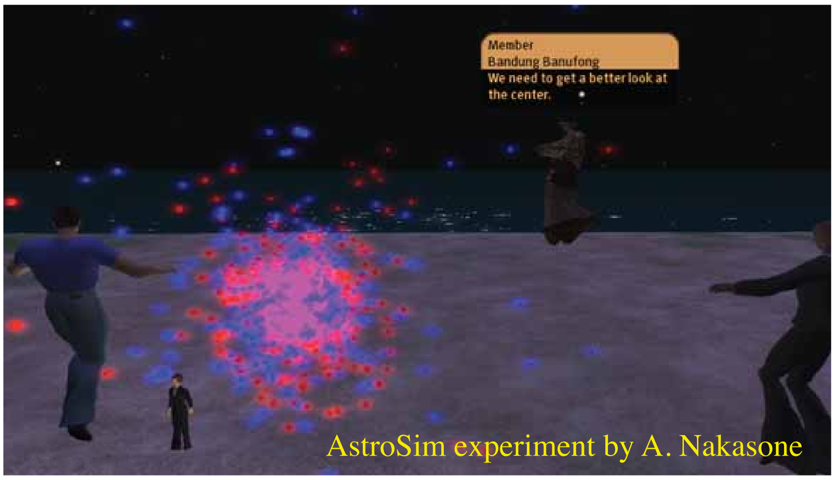
E8 Lie group polytope

A pseudo-3D projection of a 248-dimensional mathematical object

Visualization by Wizard Gynoid & Desdemona Enfield



OpenSim experiment by W. Farr et al.



AstroSim experiment by A. Nakasone

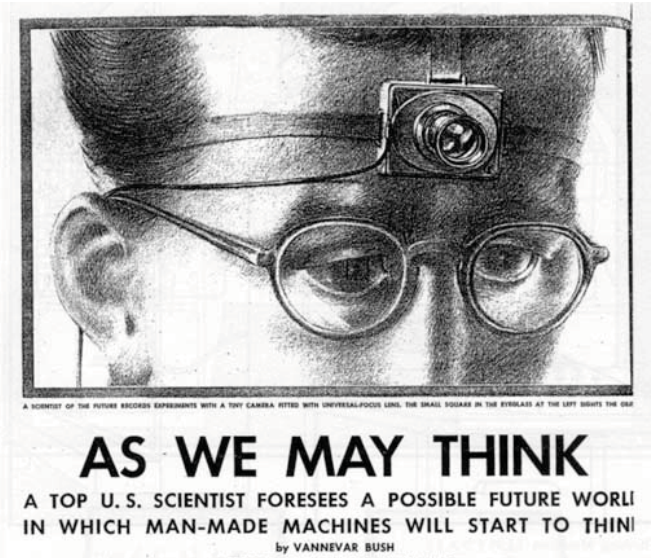
Scientists immersed in, and interacting with, numerical simulations of star clusters

Towards the Immersive Web

- Humanity’s information holdings are largely, and will be, on the Web
- The challenges of information discovery, representation, and understanding, can only get sharper
- Immersive VR is obviously a powerful approach, well suited to a human intuition
- The future is in the synergy of the Web and the VWs technologies

How do we architect effective displays of structured information (e.g., databases, data grids, semantic web constructs, etc.) in immersive, pseudo-3D environments?

Personalization of Cyberspace



A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OB

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORL
IN WHICH MAN-MADE MACHINES WILL START TO THIN

by VANNEVAR BUSH

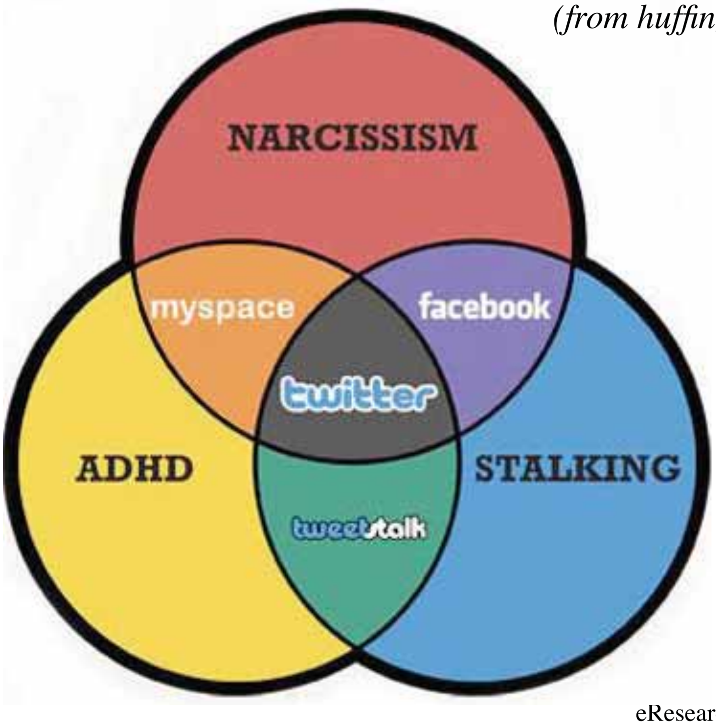
From MEMEX to MyLifeBits

… and of course Facebook,
and other personal-public
Web archives

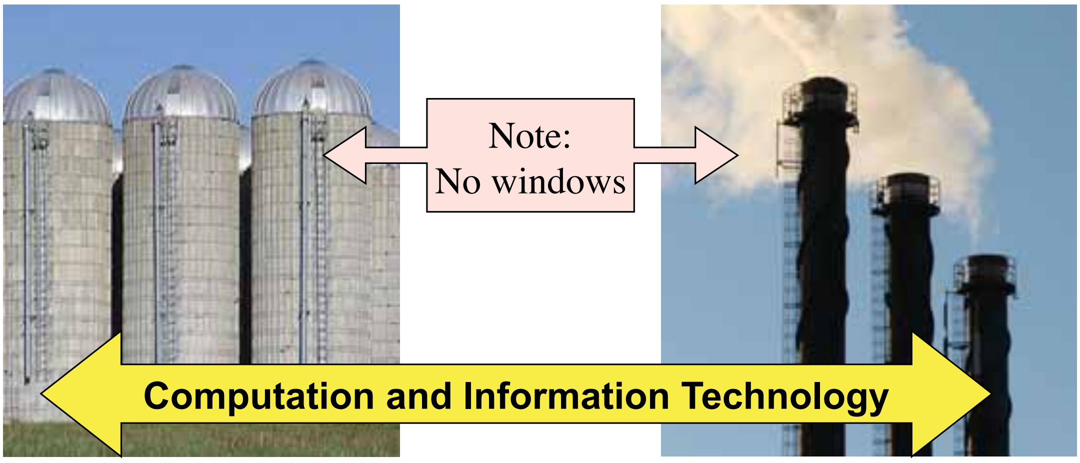
We inhabit the Cyberspace *as individuals*
– and not just for work, but in very personal ways, to express ourselves, and to connect with others (“As we may feel”?)

The Truth About Social Networking

(from huffingtonpost.com)

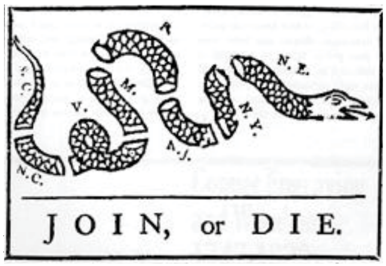


The Structure of Academia / Science



Note:
No windows

Computation and Information Technology



"We must all hang together, or assuredly we will all hang separately"

-- *Ben Franklin*

e-Science is unified by a common methodology and tools

The Core Functions of Academia

- To discover, preserve, and disseminate knowledge
- To serve as a source of scientific and technological innovation
- To educate the new generations, in terms of the knowledge, skills, and tools

But when it comes to the adoption of computational tools and methods, innovation, and teaching them to our students, we are doing very poorly – and yet, the science and the economy of the 21st century depend critically on these issues

Is the discrepancy of time scales to blame for this slow uptake?

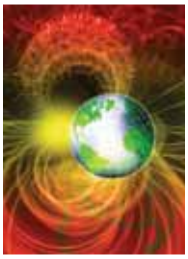
- IT ~ 2 years
- Education ~ 20 years
- Career ~ 50 years
- Universities ~ 200 years

(Are universities obsolete?)

***"Science progresses through funerals"* –** Max Planck

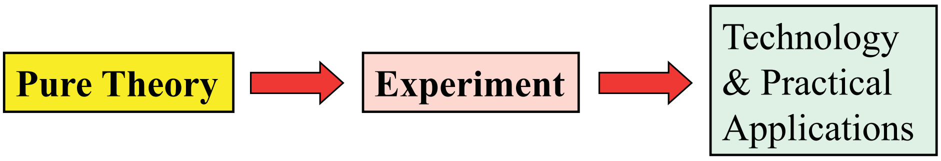
Some Thoughts About e-Science

- Computa*tional* science ≠ Compute*r* science
- Computational science { Numerical modeling ↓ Data-driven science
- Data-driven science is *not* about data, it is about ***knowledge extraction*** (the data are incidental to our real mission)
- Information and data are (relatively) cheap, but the expertise is expensive
  - Just like the hardware/software situation
- Computer science as the "new mathematics"
  - It plays the role in relation to other sciences which mathematics did in ~ 17th - 20th century
  - Computation as a glue / lubricant of interdisciplinarity



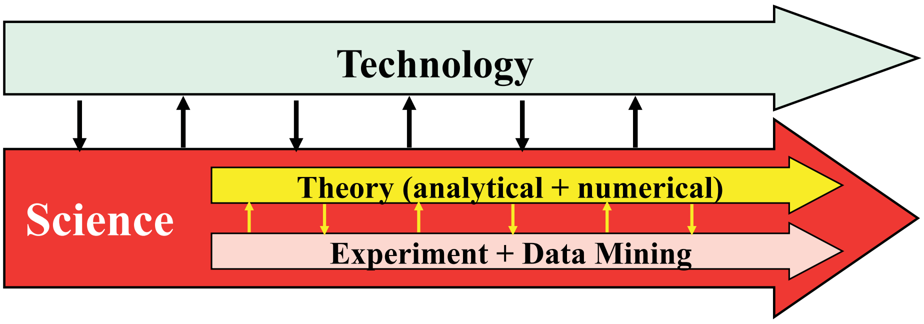
Scientific and Technological Progress

A traditional, "Platonistic" view:



Pure Theory → **Experiment** → Technology & Practical Applications

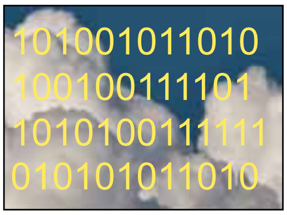
A more modern and realistic view:

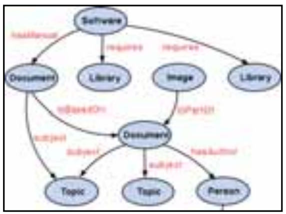


This synergy is stronger than ever and growing; it is greatly enhanced by the IT/computation

Some Transformative Technologies To Watch

- Cloud (mobile, ubiquitous) computing
  - Distributed data and services
  - Also mobile / ubiquitous computing
- Semantic Web
  - Knowledge encoding and discovery infrastructure for the next generation Web
- Immersive & Augmentative Virtual Reality
  - The human interface for the next generation Web, beyond the Web 2.0 social networking
- Machine Intelligence redux
  - Intelligent agents as your assistants / proxies
  - Human-machine intelligence interaction

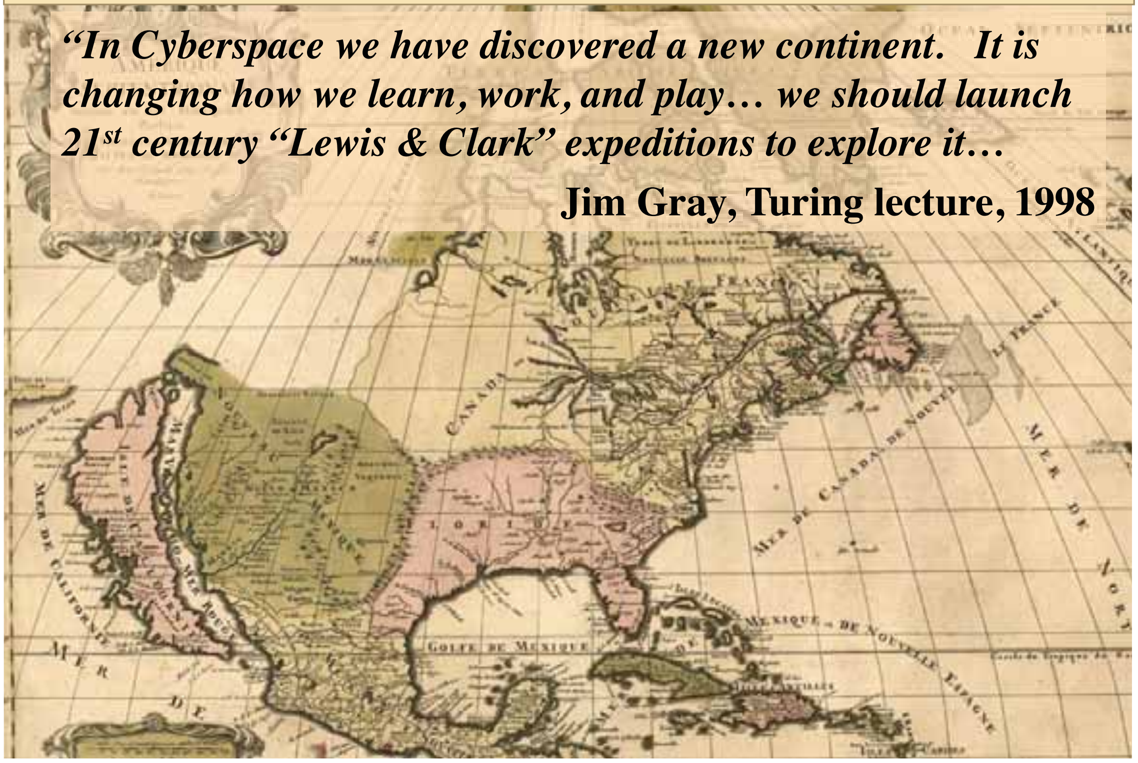




Cyberspace, The Endless Frontier

"In Cyberspace we have discovered a new continent. It is changing how we learn, work, and play… we should launch 21st century "Lewis & Clark" expeditions to explore it…

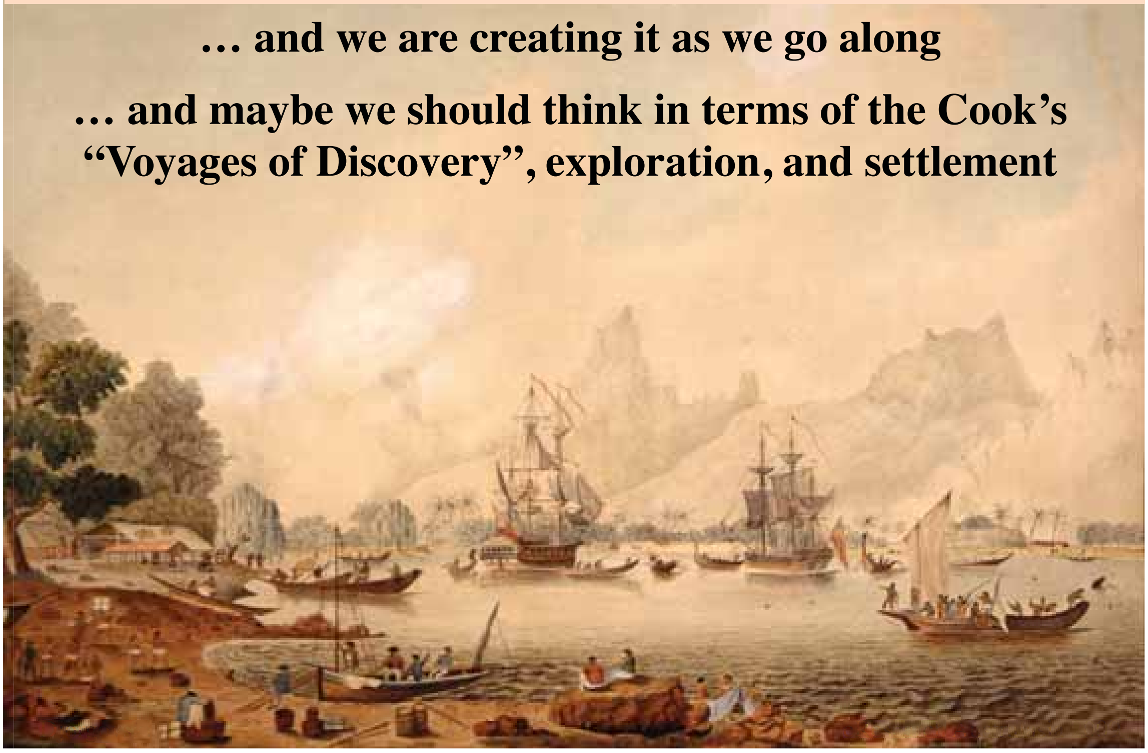
Jim Gray, Turing lecture, 1998



Actually, It’s A Whole New World

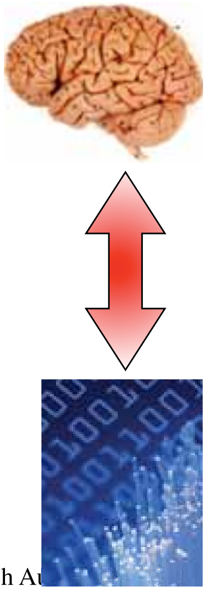
… and we are creating it as we go along

… and maybe we should think in terms of the Cook’s “Voyages of Discovery”, exploration, and settlement



Some Speculations

- We create technology, and it changes us – starting with the grasping of sticks and rocks as primitive tools, and continuing ever since
- When the technology touches our minds, that process can have profound evolutionary impact in the long term; WVs are one such technology
- Development of AI seems inevitable, and its uses in assisting us with the information management and knowledge discovery are already starting
- In the long run, immersive VR may facilitate the co-evolution of human and machine intelligence



Coda

- e-Science is a transitional phenomenon, and will become an overall research environment of the data-rich, computationally enabled science of the 21st century
- Essentially all of the humanity's activities are being virtualized in some way, science and scholarship included
- We see growing synergies and co-evolution between science, technology, society, and individuals, with an increasing fusion of the real and the virtual
- Cyberspace, now embodied though the Web and its participants, is the arena in which these processes unfold
- VR technologies may revolutionize the ways in which humans interact with each other, and with the world of information
- A synthesis of the semantic Web, immersive and augmentative VR, and machine intelligence may shape our world profoundly