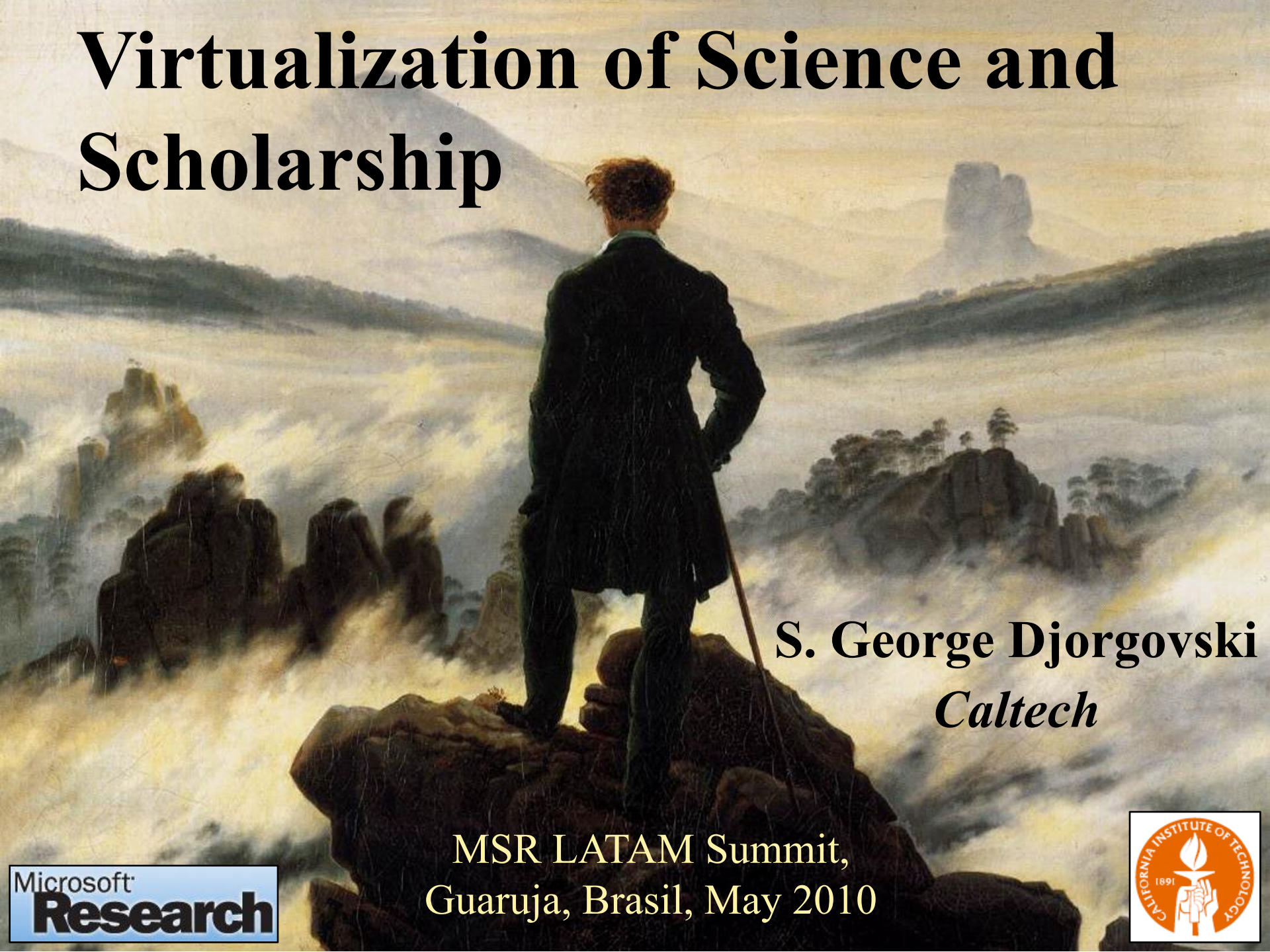
# Virtualization of Science and Scholarship



**S. George Djorgovski**
*Caltech*

MSR LATAM Summit,
Guaruja, Brasil, May 2010

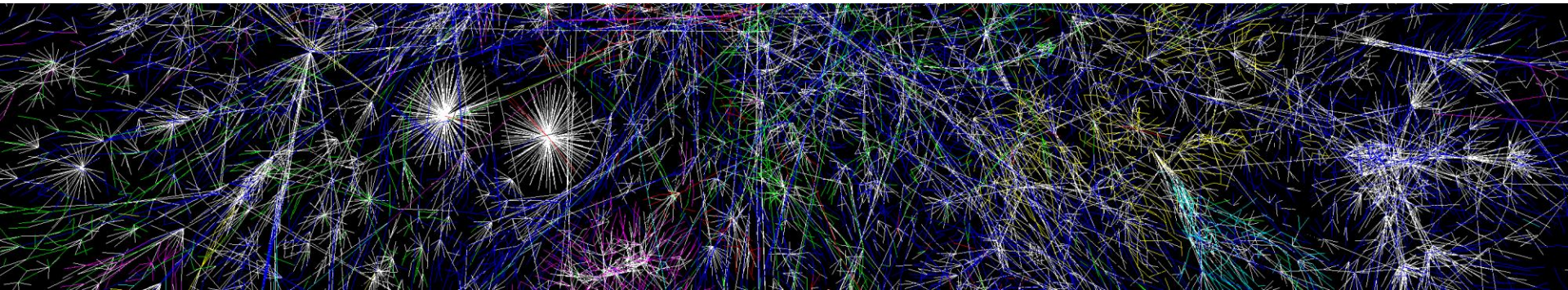
Microsoft Research

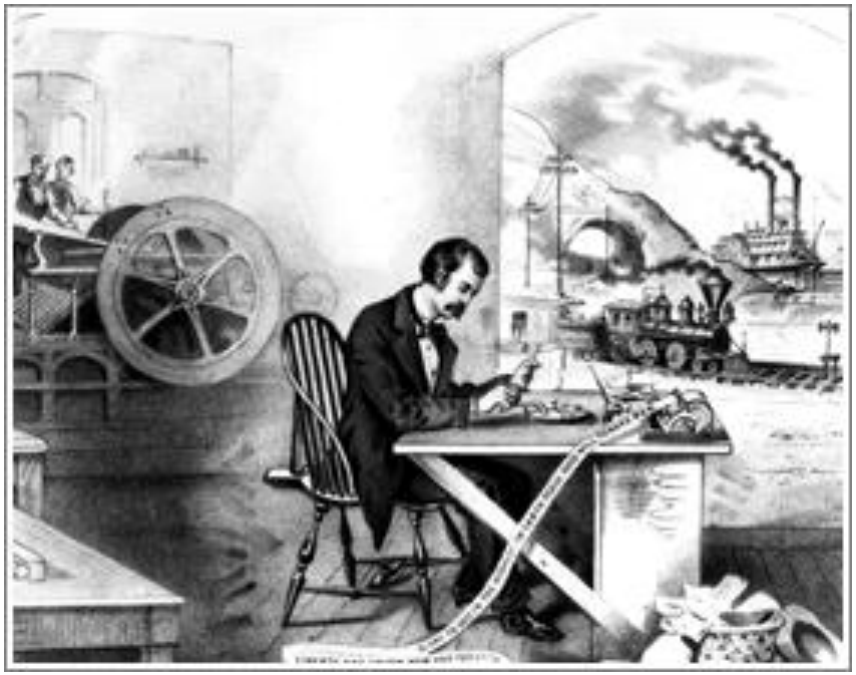
CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

**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





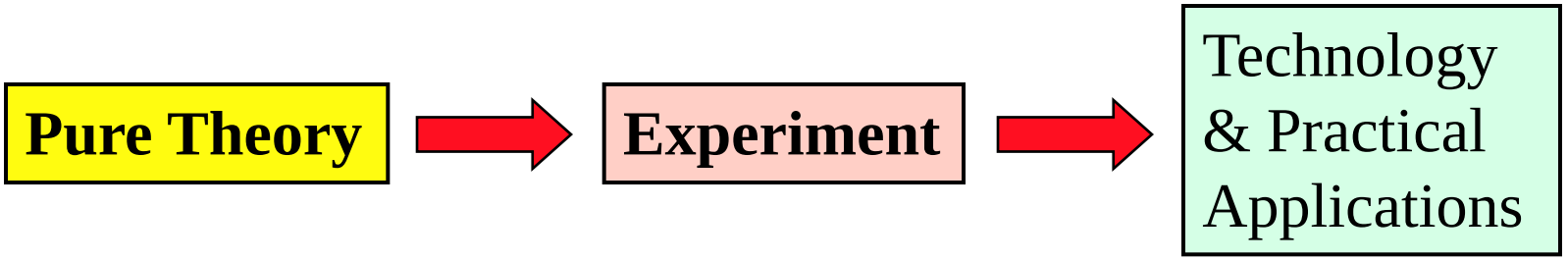
**Information technology revolution is historically unprecedented - in its impact it is like the industrial revolution and the invention of printing combined**


It is transforming science and scholarship as much as any other field of the modern human endeavor, as they become data-rich, and computationally enabled

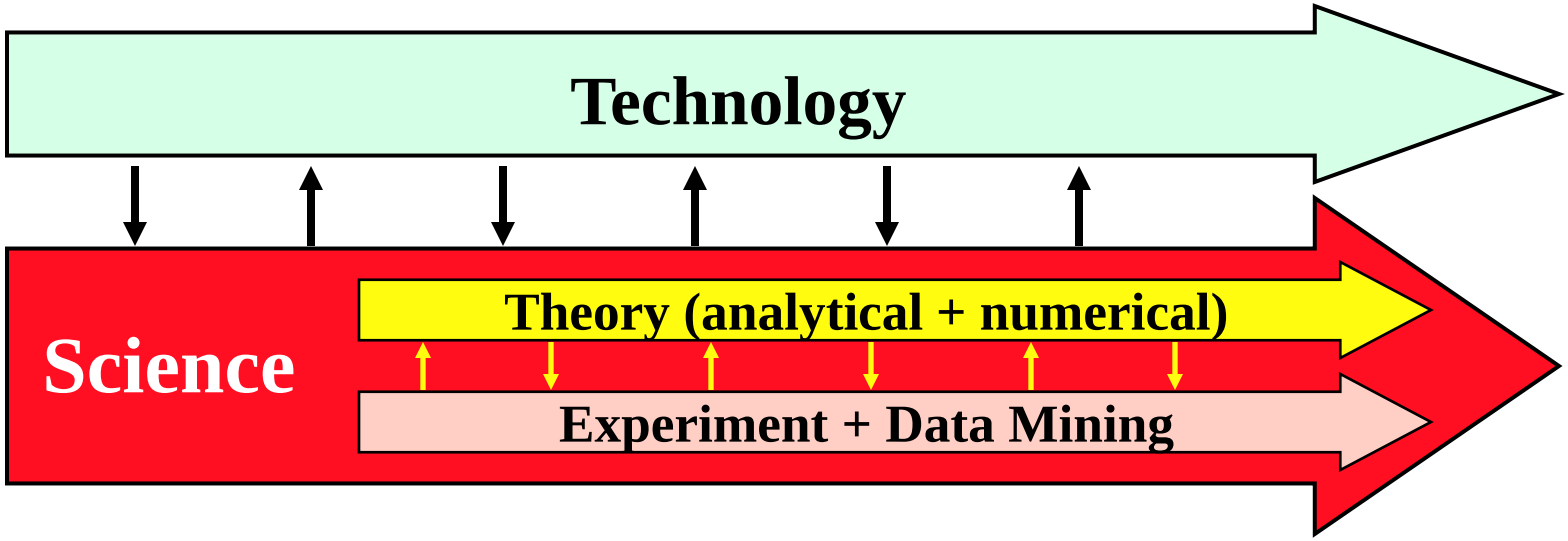
***Through e-Science, we are developing a new scientific methodology for the 21st century***

# Scientific and Technological Progress

A traditional, "Platonistic" view:



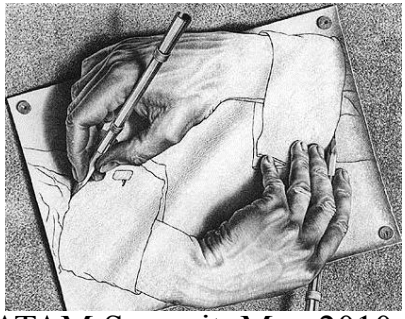

A more modern and realistic view:




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

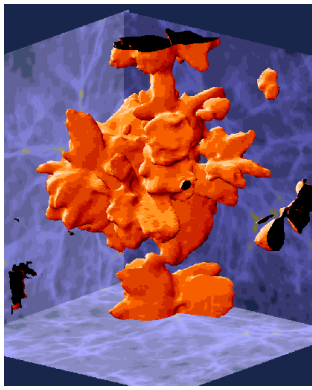
# Transformation and Synergy

- We are now in the second phase of the IT revolution: the rise of the ***information/data driven computing***
  - In addition to the traditional numerically-intensive science
  - IT as a primary publishing and communication technology
- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) - and with this change comes the need for ***a new scientific methodology***
- The challenges we are tackling:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data ⇗ new knowledge
  - **These challenges are universal**
- A great synergy of the computationally enabled science, and the science-driven IT

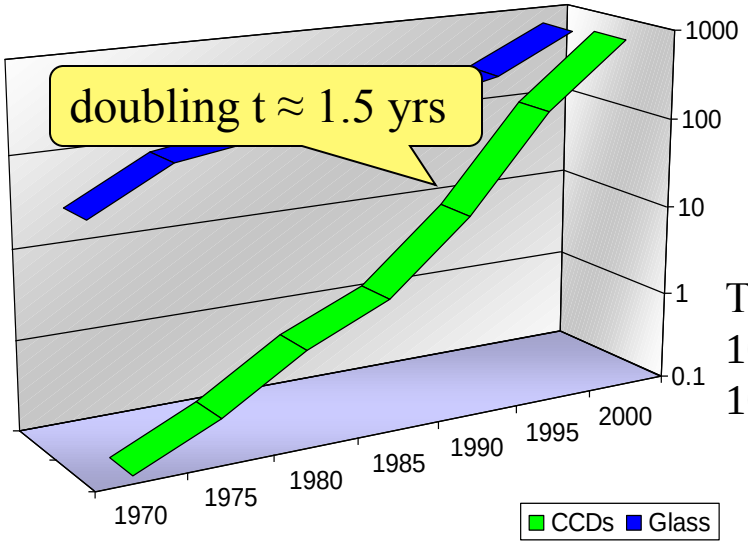


# Some Thoughts About e-Science

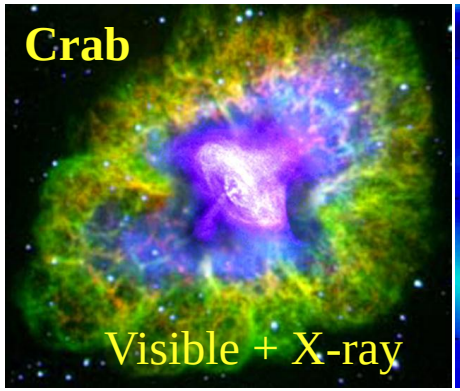
- Comput***ational*** science ≠ Comput***er*** science
- Computational science { Numerical modeling ⬇ Data-driven science
- Data-driven science is *not* about data, it is about ***knowledge extraction*** (the data are incidental to our real mission)
- Information and data are (relatively) cheap, but 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

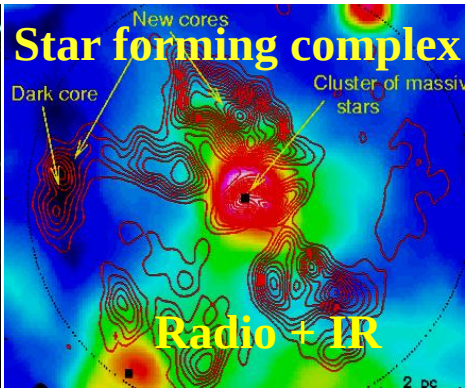


# Exponential Growth in Data Volumes and *Complexity*



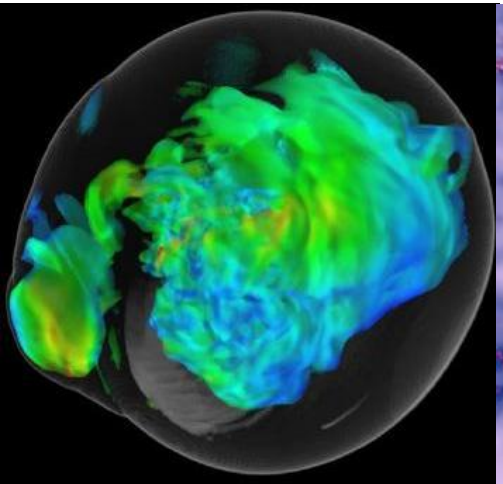

TB's to PB's of data,
$10^8$ - $10^9$ sources,
$10^2$ - $10^3$ param./source

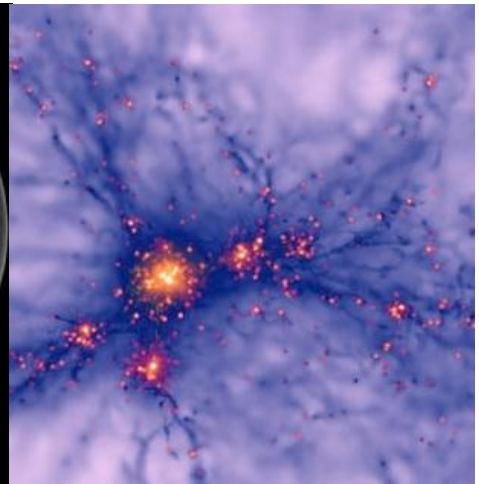



Multi-λ data fusion leads to a more complete, less biased picture
(also: multi-scale, multi-epoch, …)

***Understanding of complex phenomena requires complex data!***





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

**Data + Theory = Understanding**

# The Virtual Observatory Concept

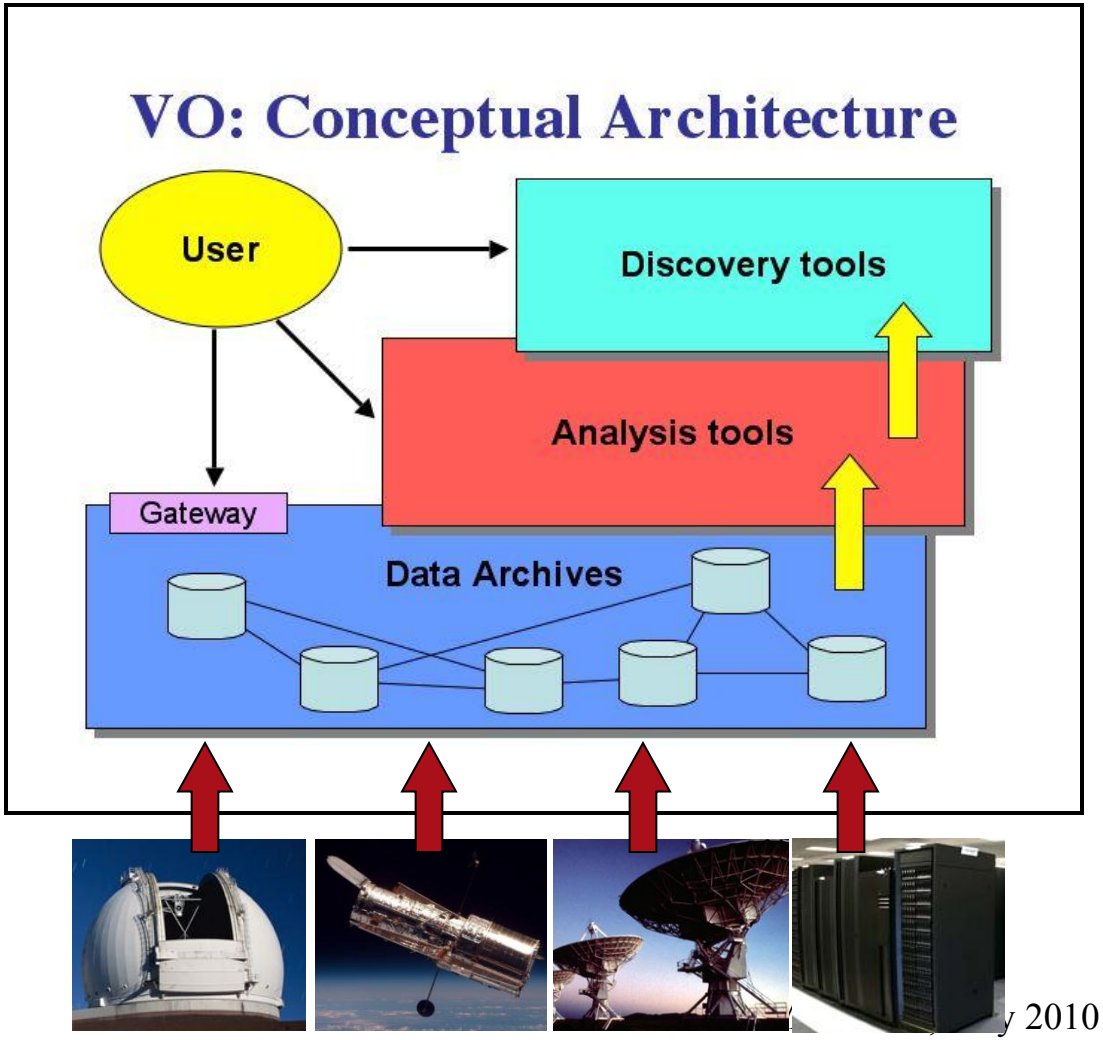
- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

– 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!

NATIONAL VIRTUAL OBSERVATORY
...the Universe at your fingertips

http://us-vo.org

Welcome to the New NVO Home Page! We

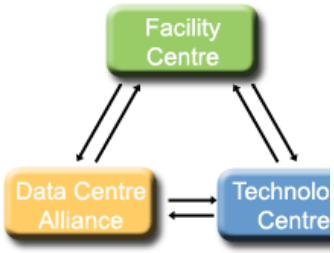
Discover, retrieve, and analyze astronomical data from archives and data centers around the world.

EURO VO

The Euro-VO projects: VOTECH EuroVO-DCA

Science
- Software
- Recipes User Manual
- Scientific Workflows
- Research Initiative
- Science Cases
- Scientific Papers
- Science Advisory Committee
- Acknowledging
- Helpdesk

Technical
- Software




From AVO to EURO-V

The Astrophysical Virtual Observa
of a regional-scale infrastructure b
requirements and technologies. A
was jointly funded by the
(HPRI-CT-2001-50030). The EUF
deployment of an operational VO i

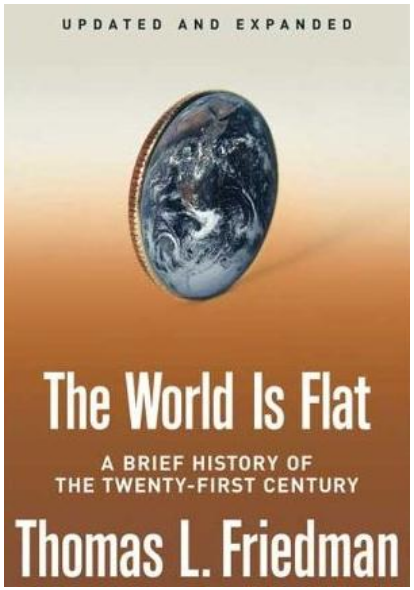
Astro Grid
Aus-VO
VObs.it
vo-i
CVO
SVO
BRAVO
IVOA
ArVO
Armenian Virtual Observatory
China-VO
EURO VO
JVO
HVO
Hungarian Virtual Observatory
GERMAN ASTROPHYSICAL VIRTUAL OBSERVATORY
GAVO
France

NATIONAL VIRTUAL OBSERVATORY

http://www.euro-vo.org

http:// ivoa.net

Djorgovski




# The Sky Is Also Flat

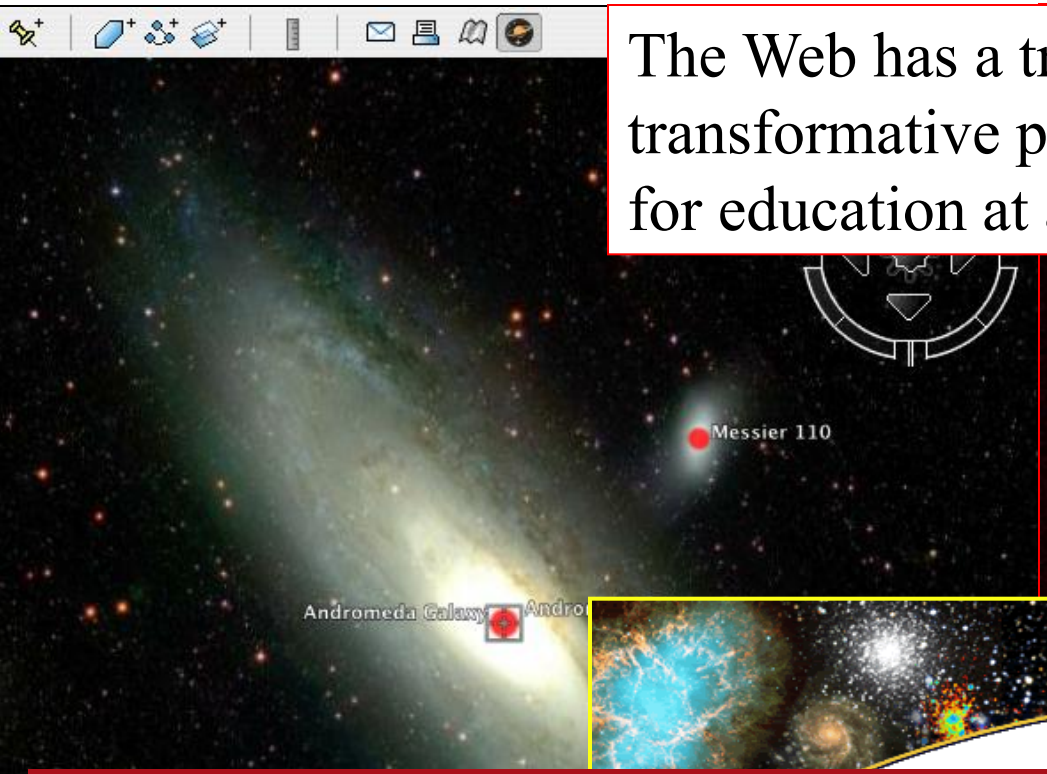
Probably the
most important aspect of the IT revolution in science

- **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 observatories

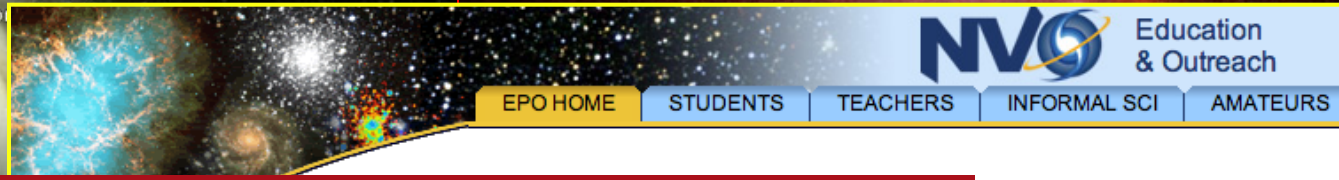
# VO Education and Public Outreach

## *"Weapons of Mass Instruction"*

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

*Galaxy M81 seen by a visible-light telescope*

Dec 41°16'08.05" RA 0h42m44.30s streaming

**This site:** This site is a gateway to education and public

# 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

Key Technical Challenges

↳ **Data Understanding**

↳ **New Knowledge**

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!***

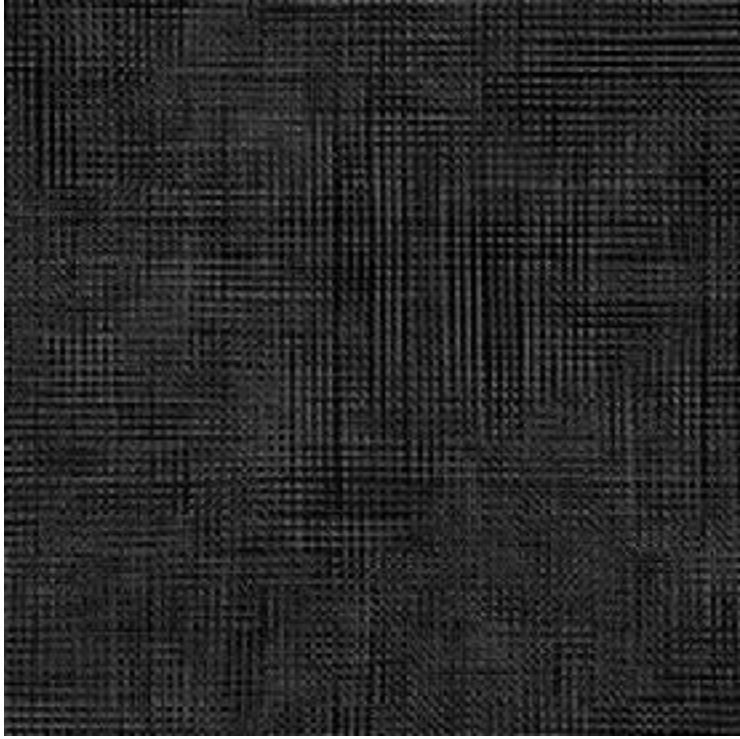
➡ The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery …

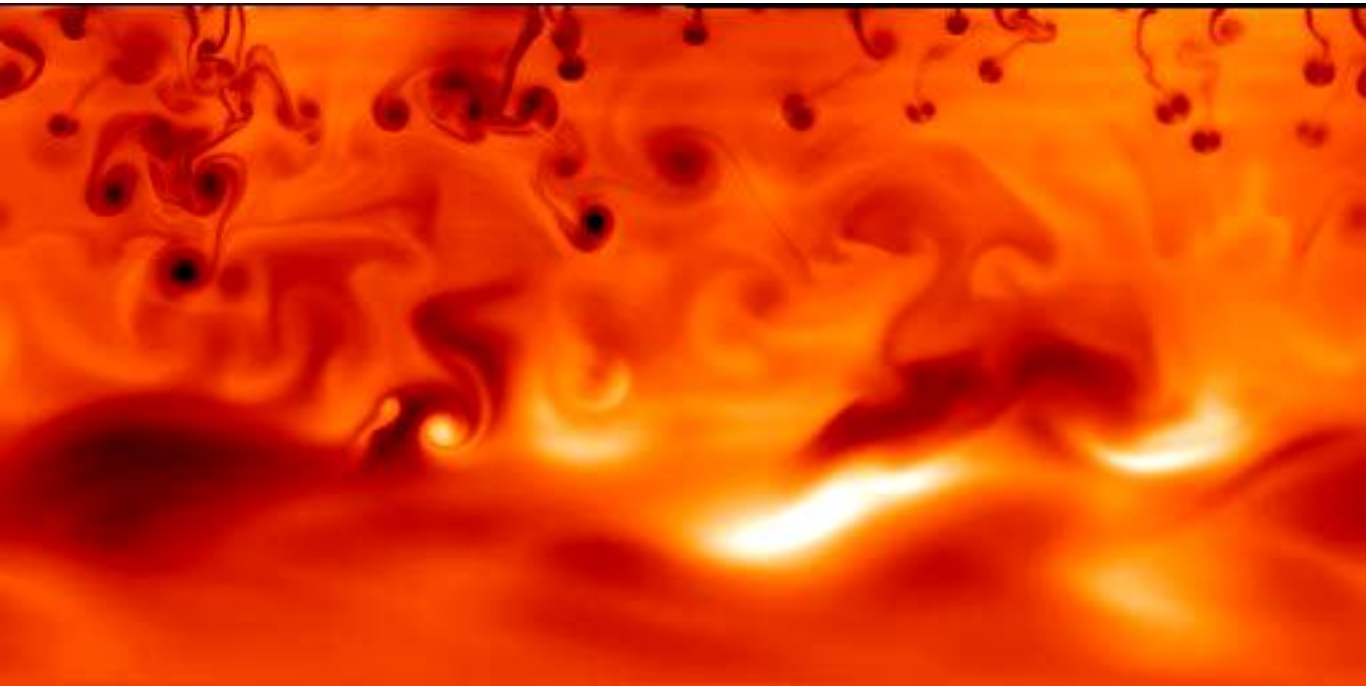
- We need to create ***a new scientific methodology*** on the basis of applied CS and IT
- Important for practical applications beyond science

# 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 Key Challenge: Data Complexity

## Or: The Curse of Hyper-Dimensionality

1. **Data mining algorithms scale very poorly:**

   N = data vectors, ~ $10^8$ - $10^9$, D = dimension, ~ $10^2$ - $10^3$

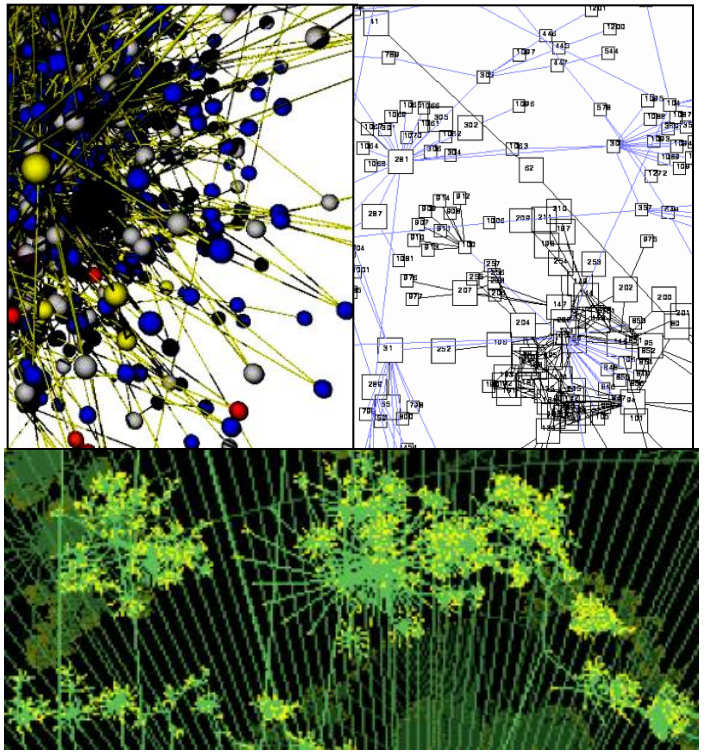
   – Clustering ~ N log N ⌫ $N^2$, ~ $D^2$

   – Correlations ~ N log N ⌫ $N^2$, ~ $D^k$ ($k \geq 1$)

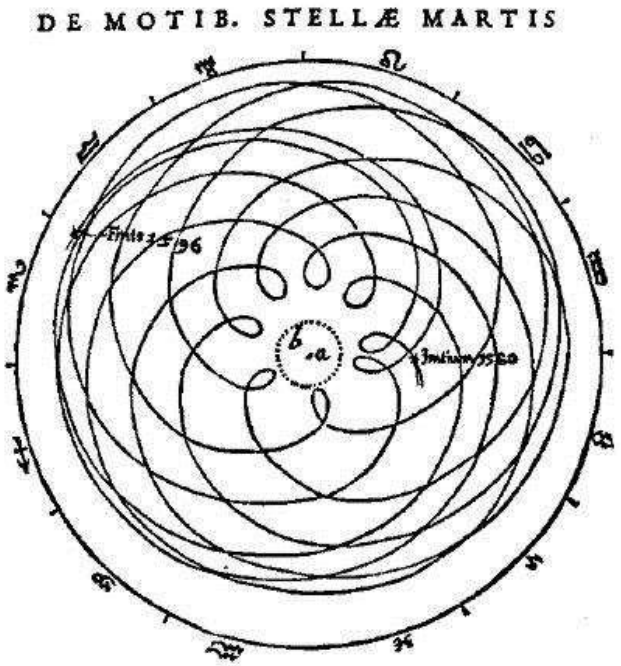
   – Likelihood, Bayesian ~ $N^m$ ($m \geq 3$), ~ $D^k$ ($k \geq 1$)

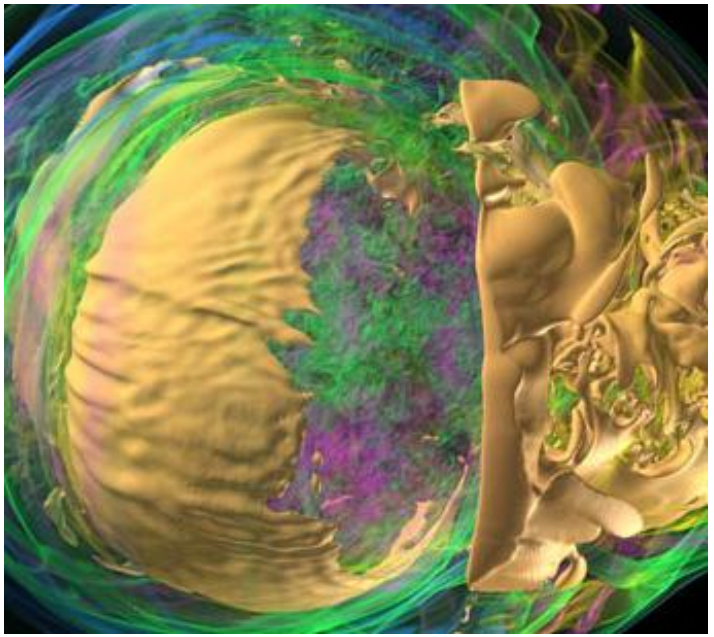
2. **Visualization in >> 3 dimensions**

   - The complexity of data sets and interesting, meaningful constructs in them is *exceeding the cognitive capacity of the human brain*
   - We are biologically limited to perceiving D ~ 3 - 10(?)
   - Visualization is a bridge between data and human intuition/understanding



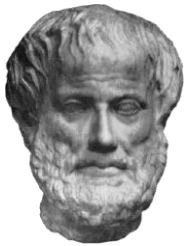
# Effective visualization is the bridge between quantitative information, and human



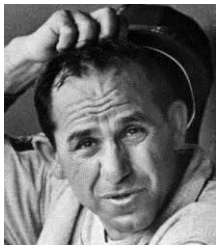
***Man cannot understand without images; the image is a similitude of a corporeal thing, but understanding is of universals which are to be abstracted from particulars***

Aristotle, *De Memoria et Reminiscentia*



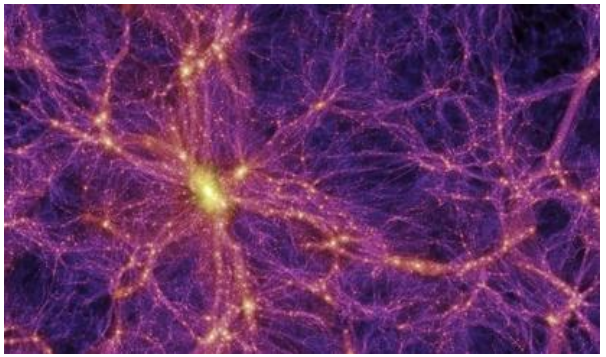
***You can observe a lot just by watching***

Yogi Berra, an American philosopher



# 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 $D \sim 10^2 - 10^4$, and will surely grow
- It is not only the matter of ***data understanding***, but also of choosing the appropriate data mining algorithms, and interpreting the 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, another American philosopher

# The Roles for Machine Learning and Machine Intelligence in CyberScience:

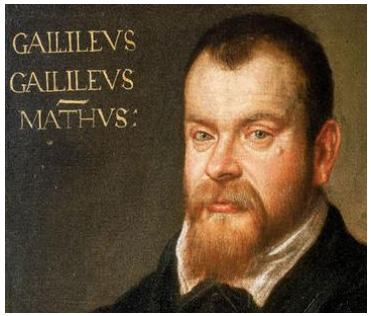
- **Data processing:**
  - Object / event / pattern classification
  - Automated data quality control (glitch/fault detection and repair)

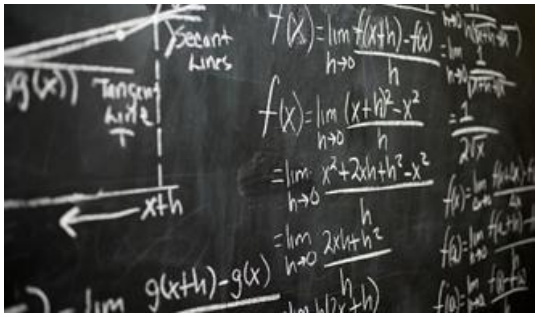
+

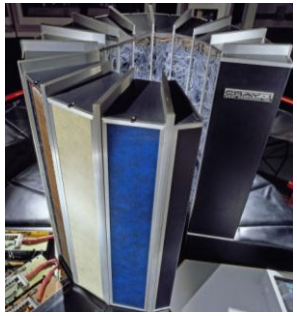
- **Data mining, analysis, and understanding:**
  - Clustering, classification, outlier / anomaly detection
  - Pattern recognition, hidden correlation search
  - Assisted dimensionality reduction for hyperdim. visualisation
  - Workflow control in Grid-based apps
- **Data farming and data discovery:**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

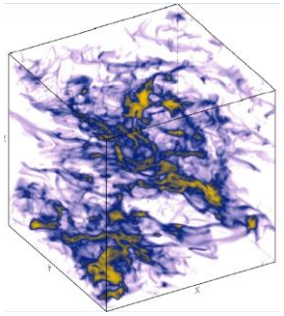
# The Evolving Paths to Knowledge

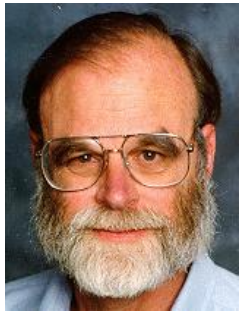
- The First Paradigm: Experiment/Measurement
- The Second Paradigm: Analytical Theory
- The Third Paradigm: Numerical Simulations
- The Fourth Paradigm: Data-Driven Science?

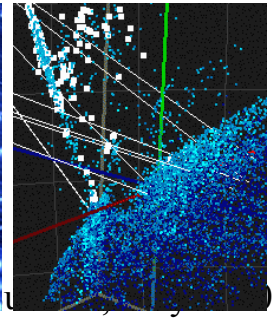






# The Fourth Paradigm

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 Revolution in Scholarly Publishing

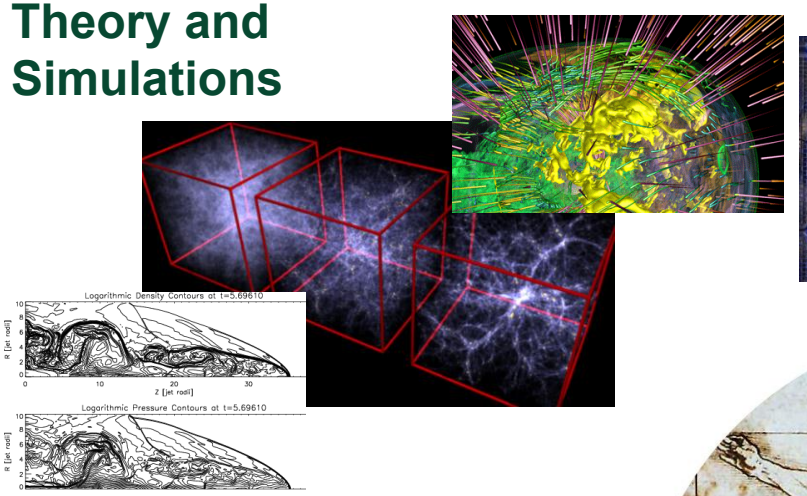
## Information and Knowledge Management Challenges

- Increasing complexity and diversity of scientific data and results
  - Data, metadata, virtual data, simulations, algorithms, blogs, wikis, multimedia…
  - ***From static to dynamic***: evolving and growing data sets
  - ***From print-oriented to web-oriented***
- Institutional, cultural, and technical challenges:
  - Massive data sets can be only published as electronic archives, and should be curated by domain experts
  - Effective peer review and quality control
  - Persistency and integrity of data and pointers
  - Interoperability and metadata standards

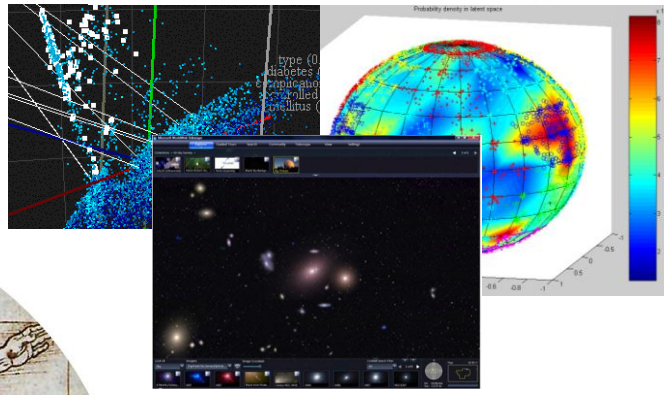
***As the science evolves, so does its publishing***

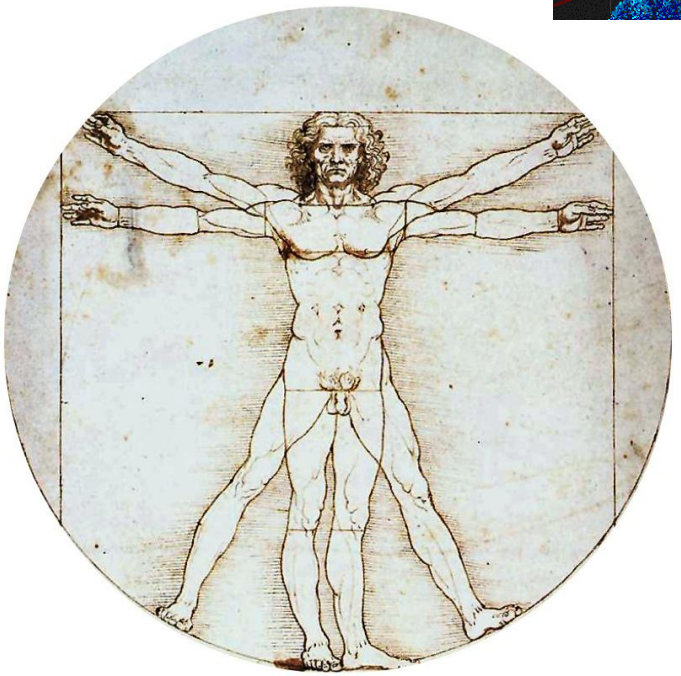
# Science in Cyberspace

## Theory and Simulations

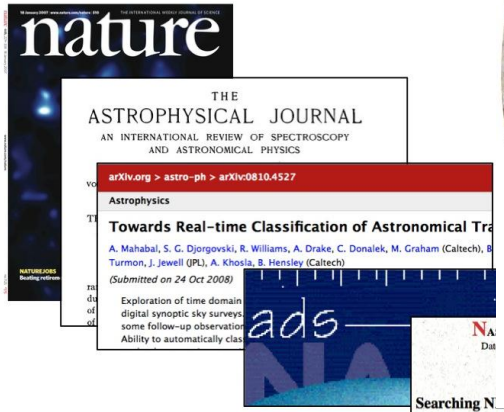



## Visual Displays and Linking of Data and Knowledge

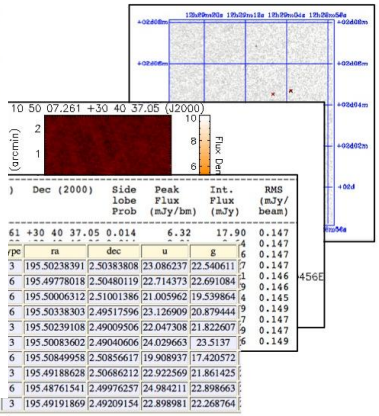




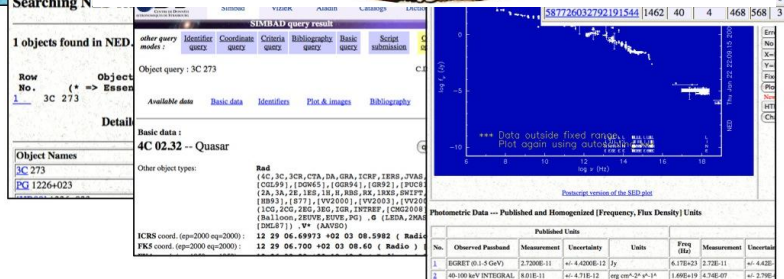
## Published Literature




## Data Archives



## Semantic Web


## Virtual Observatory

# Karl Popper's Three Worlds of Knowledge

Dawkins memes

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

Cognition
Self-System
Consciousness

Production

Development/Recall

Language
Artifacts
Recorded thought
Computer
Memory

World 2
Personal
Knowledge
(emerges
from World
1
processes)

Test
Observe

Regulate/control

Drive/Enable

Describe/Predict

Inferred Logic

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

Energy
Physics
Chemistry

World 1
Existence/Reality

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

# 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

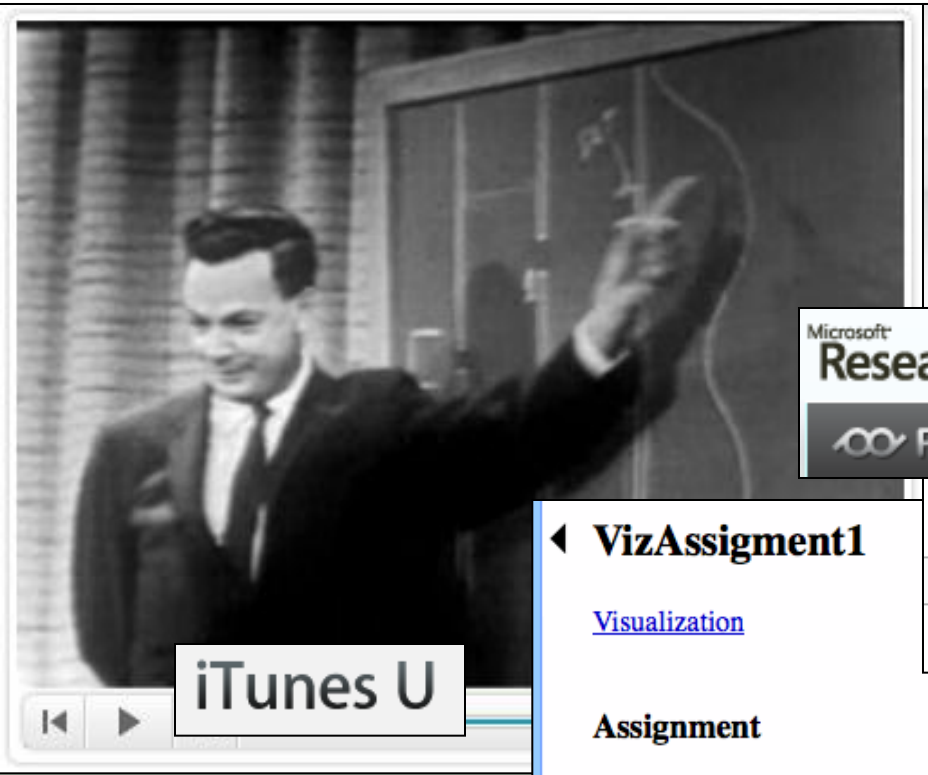
Is the discrepancy of time scales to blame for this slow uptake?

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

**Are universities structurally obsolete?**

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

# Virtualizing Education




WIKIPEDIA
The Free Encyclopedia

article | discussion | edit this pag

Quasar

From Wikipedia, the free encyclopedia that anyo

*This article is about the astronomical object.*

A **quasi-stellar radio source** (**quasar**) is a very
nucleus. They are the most luminous objects in t

Microsoft Research
PROJECT TUVA

**VizAssigment1**

Visualization

**Assignment**

**create your own a**
+ have at least 100
+ have at least 4 da
*if you have some as*
**+ get mondrian, m**

**create your own v**
+ have at least 100
+ if you have some
*you can used awk/s*

for a sample on the

Course | Programming Methodology
Lecture 4 |
Lecture 5 |
Lecture 2 |
by StanfordUniversity | 28 videos

YouTube

**Methods of Computat**

This is a blog for the Ay/Bi 199ab cl

THURSDAY, MAY 7, 2009

**Special class time set**

The make-up lecture on scientific in
be at 3 pm on Monday, May 18. Thar

Posted by Djorgovski at 4:47 PM

0 comments

MITOPENCOURSEWARE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Home | Courses | Donate | About OCW

Get Started with OCW
VIEW ALL 1900 COURSES
Most Visited

Home > Courses > Most Visited Courses

Most Visited Courses

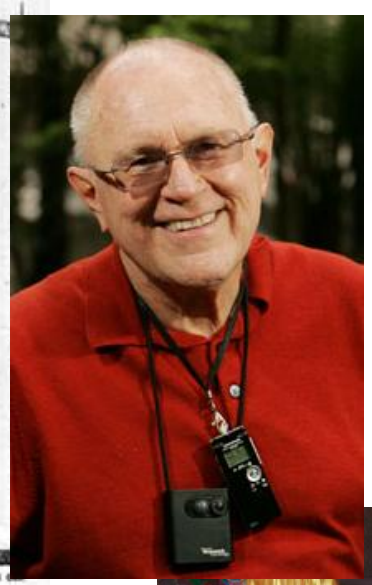
Below is a selection of our most visited courses

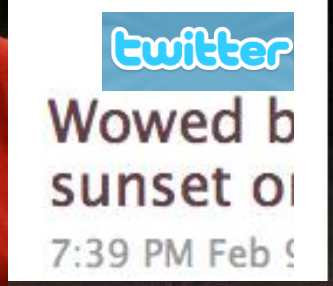
# Personalization of Cyberspace

AS WE MAY THINK

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

by VANNEVAR BUSH



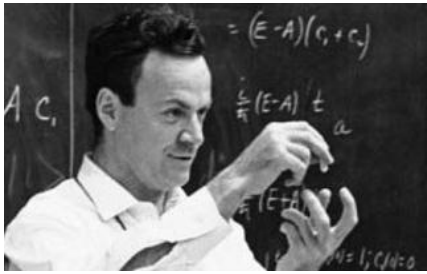




*From MEMEX to Web 2.0*

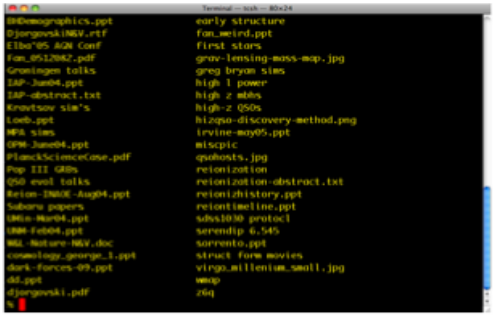
We inhabit the Cyberspace *as individuals*
– and not just for work, but in very personal ways, to express ourselves, and to connect with others ("As we may feel"?)

# Human Interactions



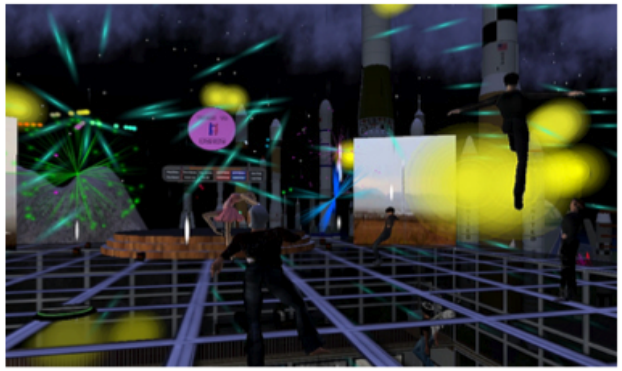
- Science originates on the interface between human minds, and humans and data (measurements, simulations, literature, etc.)
- Any technology which facilitates these interactions is enabling science, scholarship, and education

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 …

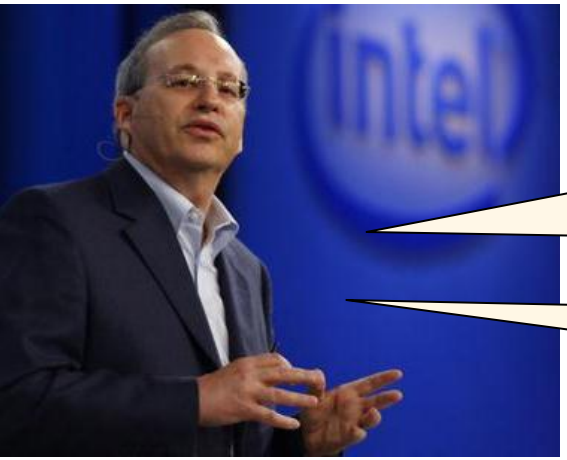


… and now immersive virtual environments

# Immersive VR and the Emerging 3D Web




## … and the future of the Web:




Justin Rattner, Intel CTO, in a keynote talk at the SC'09:

"… There is nothing more important to the long-term health of the HPC industry than the 3D Web…"

"… the 3D Web will be the technology driver that revitalizes the HPC business model …"

***What should the academic community be doing about these emerging technologies? How can we use them?***


**http://mica-vw.org/**

Go
Search

Log in / create account

**Navigation**

- Main Page
- Charter
- Events
- Research
- Organization
- People
- How to Join
- MICA blog
- Publications
- Links

## 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.

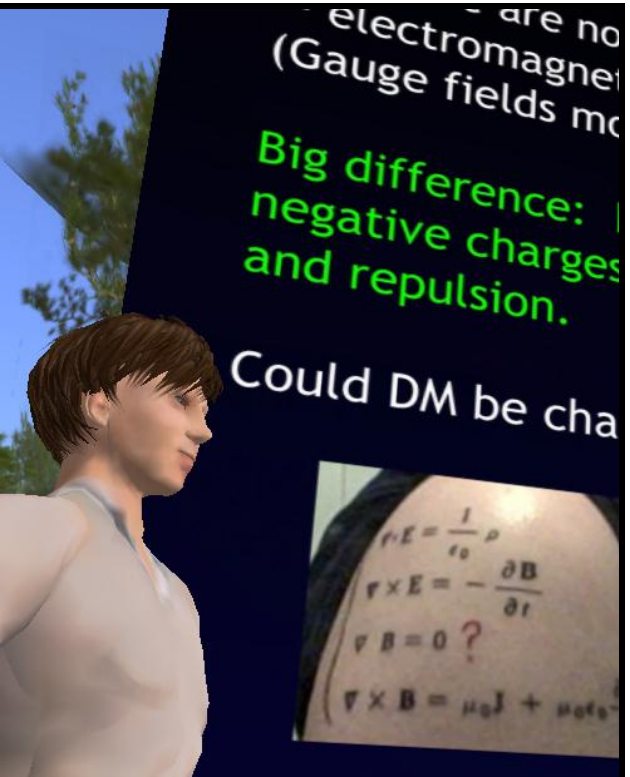
***MICA is an experiment in the scholarly use of VWs technologies***

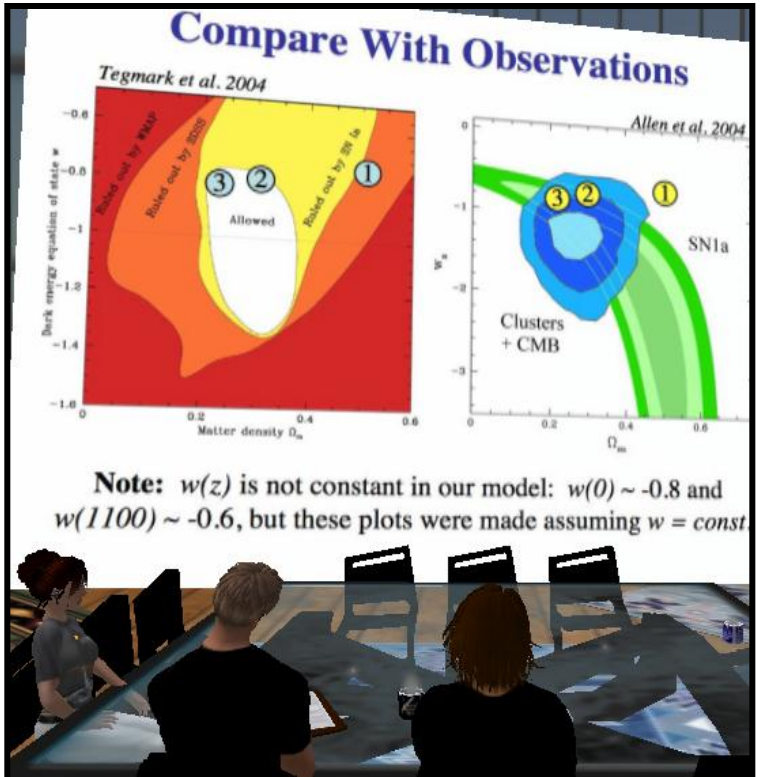
- Currently ~ 50 professional members and > 100 affiliates
- Regular schedule of events: seminars, workshops, public lectures, etc.

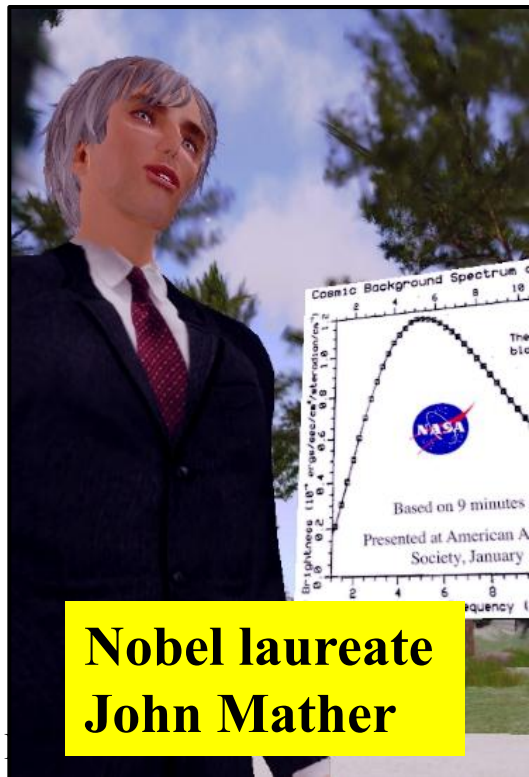
# MICA: Scientific Communication and Collaboration in VR Environments

- 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

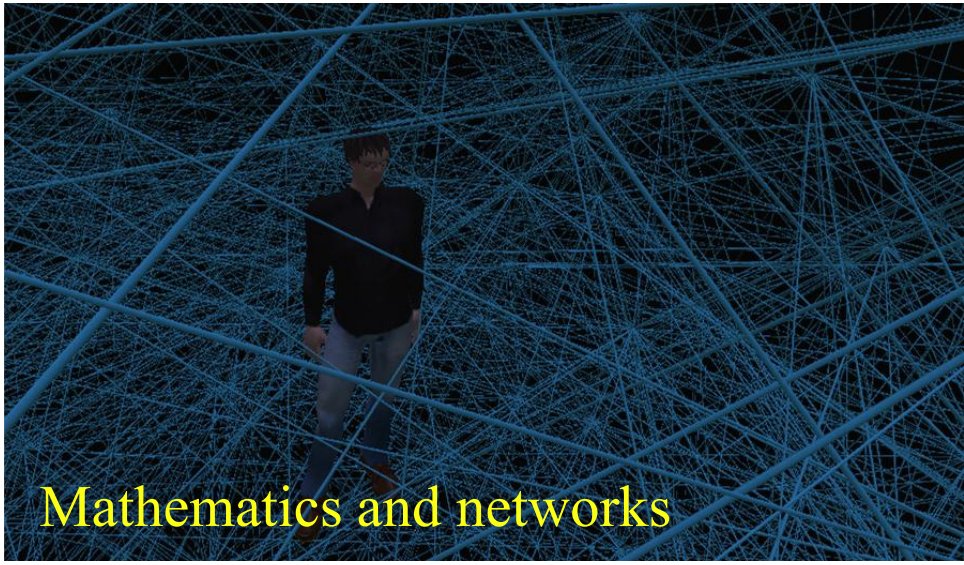
**Professional seminars** **Collaboration meetings** **Public outreach**

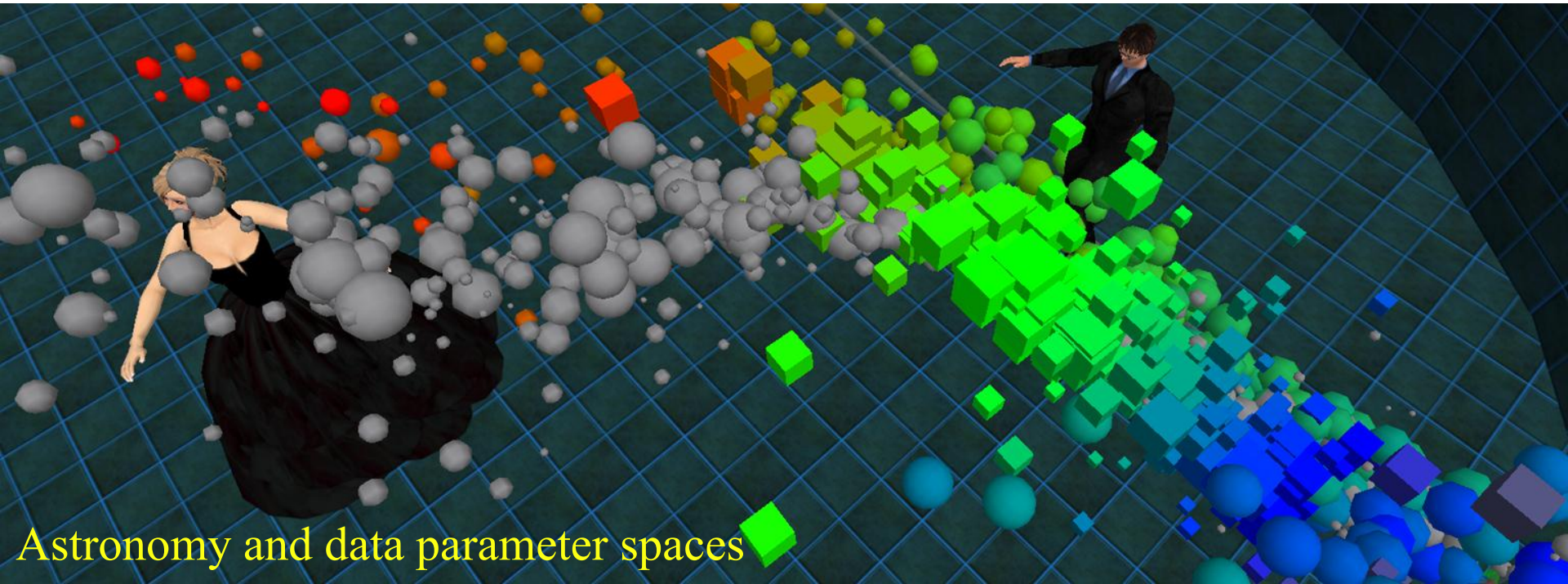


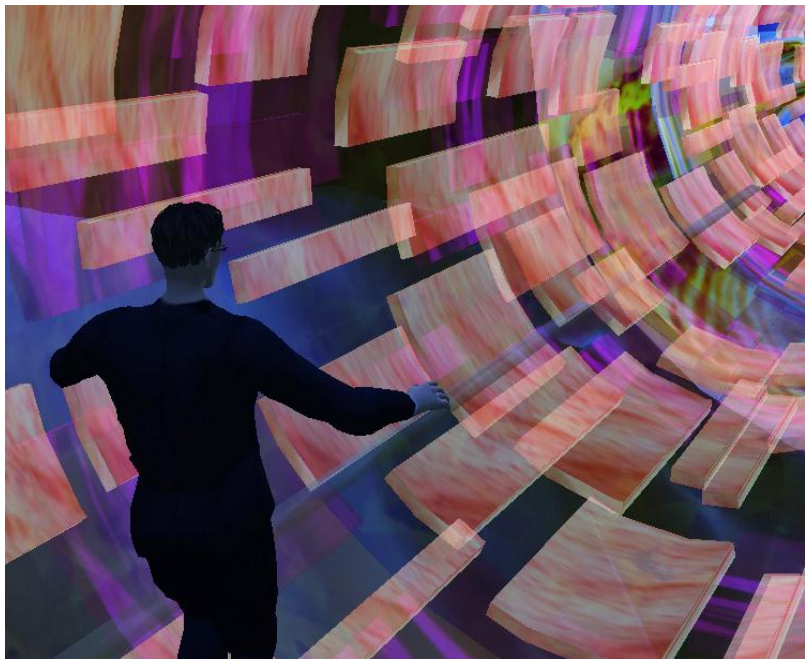
# Immersive Data Visualization

# 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 3-D VR is obviously a powerful approach, well suited to a human intuition

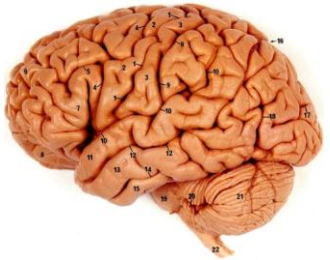


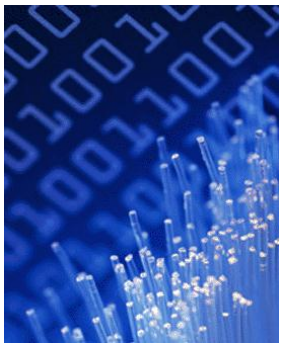
- The future is in the synergy of the Web and the immersive VR technologies as the next generation interface

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

# 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; IT and VR are such technologies
- 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





# Summary

- 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

# 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**