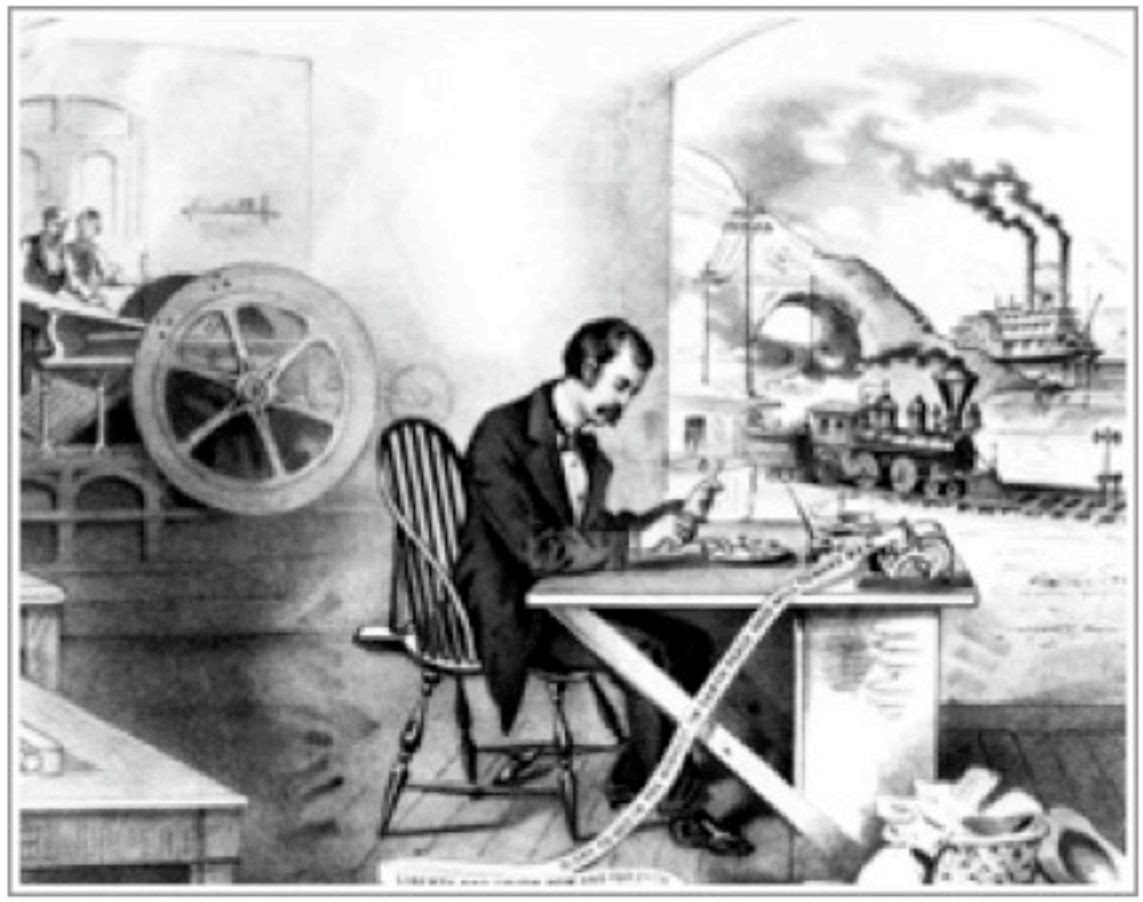
Science in Cyberspace

S. G. Djorgovski
Caltech

China-VO and Astroinformatics Conference, Nov. 2011



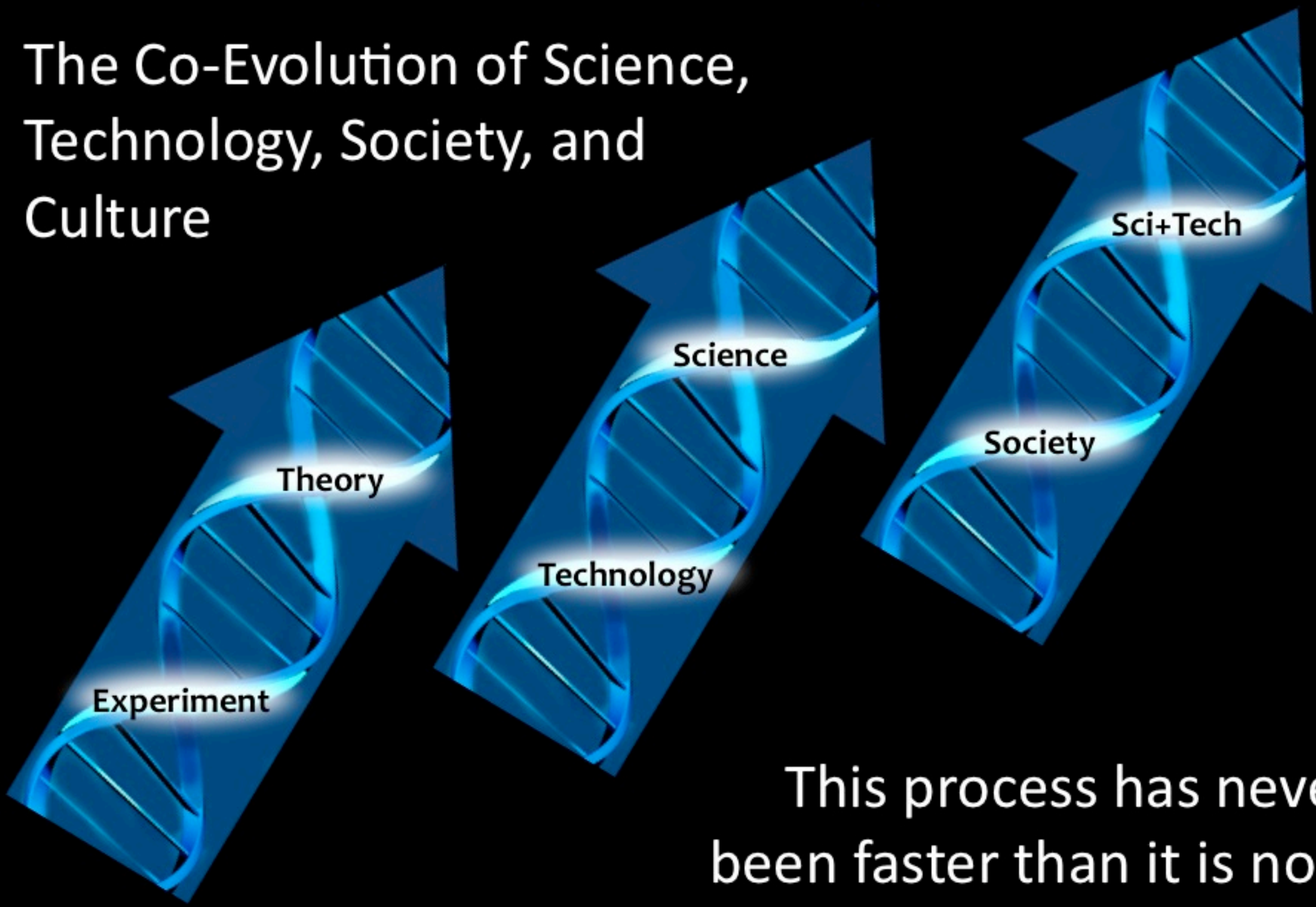
Information technology revolution is historically unprecedented - in its impact it is like the industrial revolution and the invention of printing combined

Science and scholarship are slowly adopting the new tools and technologies and there are great scientific and leadership opportunities in this arena

We are effectively developing a new methodology of science and scholarship for the 21st century

A Virtuous Helix of Progress

The Co-Evolution of Science, Technology, Society, and Culture

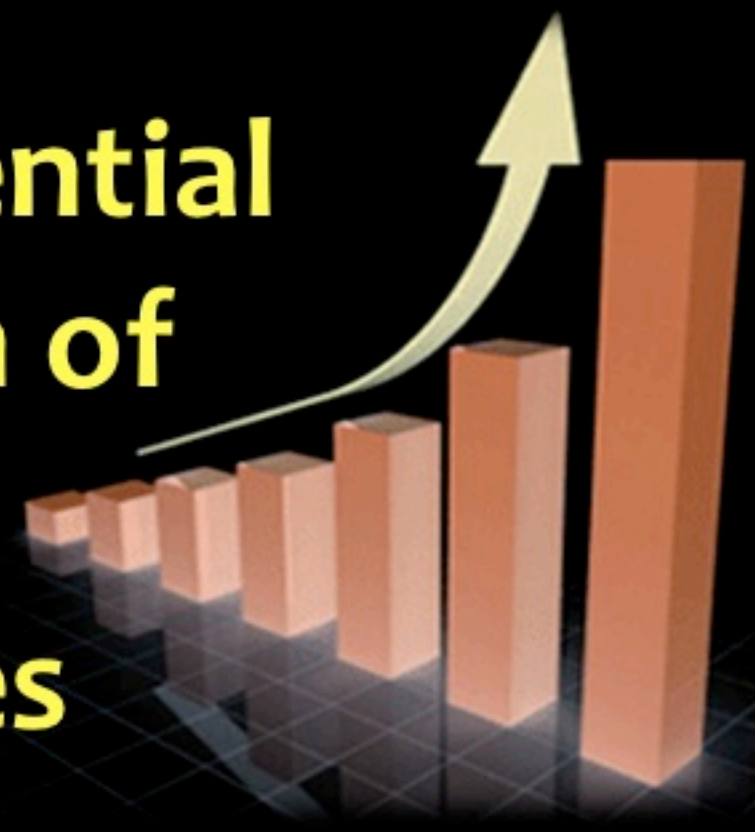


This process has never been faster than it is now

Transformation and Synergy

- ***All science*** in the 21st century is becoming cyber-science (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 technology

Exponential Growth of Data Volumes



… and Complexity

on Moore's law time scales

Understanding of complex phenomena requires complex data!

From data poverty to data glut
From data sets to data streams
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise

Astronomy Has Become Very Data-Rich

- Typical digital sky survey generates ~ 10 - 100 TB, plus a comparable amount of derived data products
  - PB-scale data sets are imminent
- 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

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 **K**nowledge **D**iscovery in **D**atabases):

Pattern or correlation search
Clustering analysis, 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.

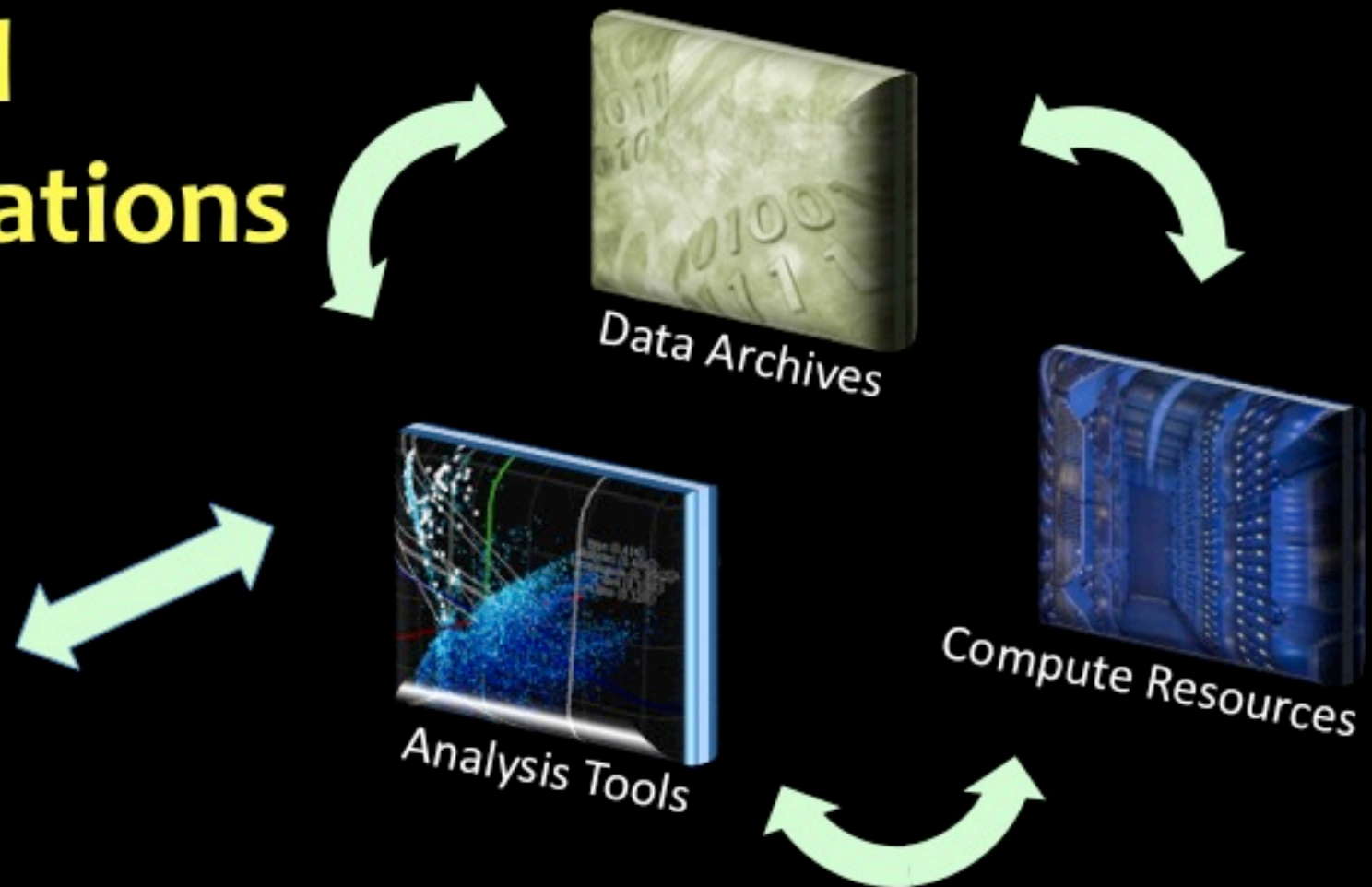
- Information complexity is also increasing greatly

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 Comp.Sci. and Info.Tech.
- Important for practical applications beyond science – knowledge economy, etc.

The Rise of Virtual Scientific Organizations



- A grassroots response of scientific communities to the challenges and opportunities brought by the data glut
- Domain-specific, not institution-based; inherently distributed
  - The human, data, and compute resources are distributed
  - A new type of a scientific organization, needing new management models
- Should VO's have a finite lifetime, as they fulfill their role?

The Virtual Observatory Concept

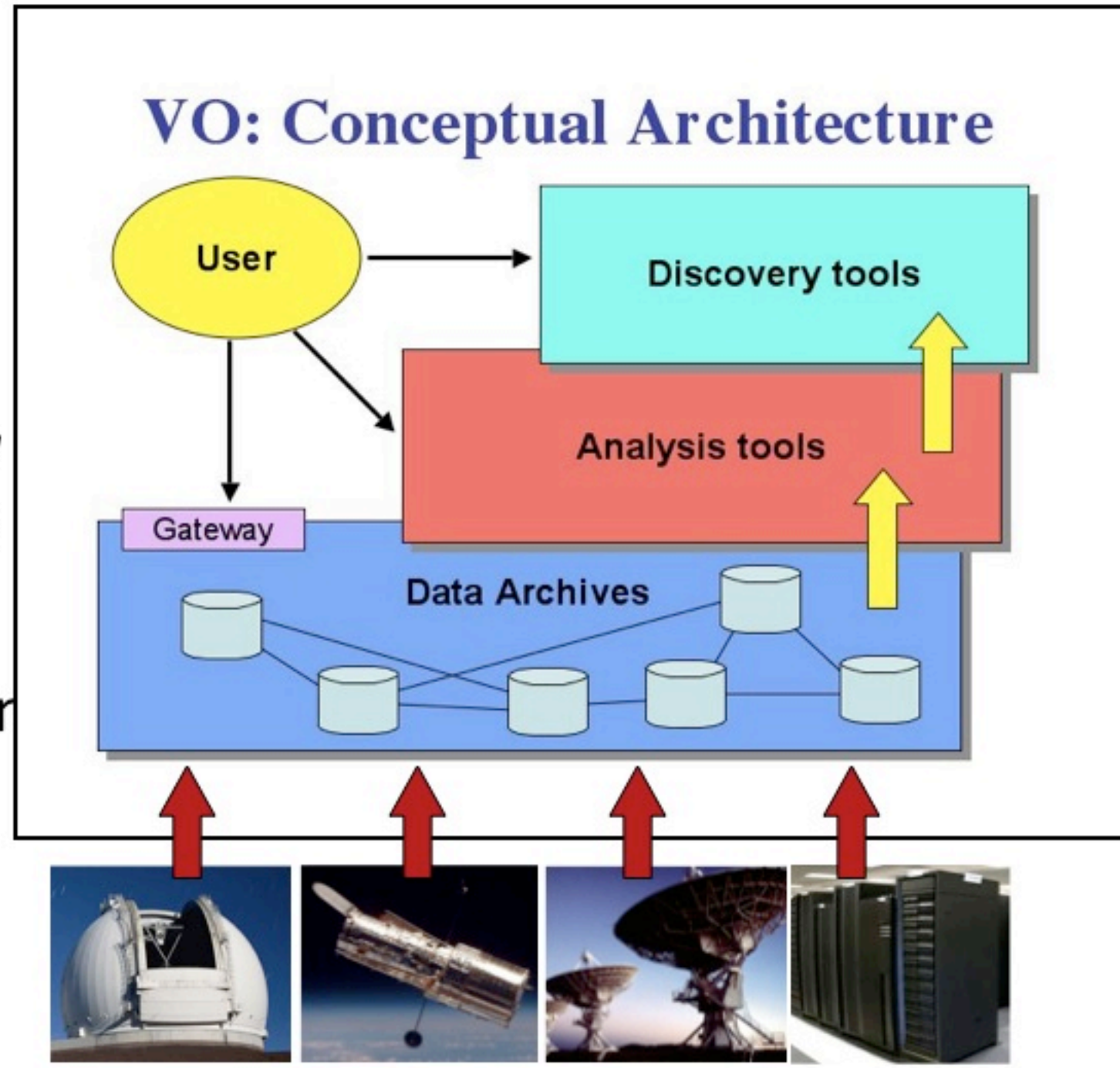
- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***

– Provide and federate content (data, metadata) services, standards, and analysis/compute services

– Develop and provide ***data exploration and knowledge discovery tools***

– Harness the IT revolution in the service of astronomy

– A part of the broader e-Science /Cyber-Infrastructure



Information-Rich Astronomy in the 21st Century

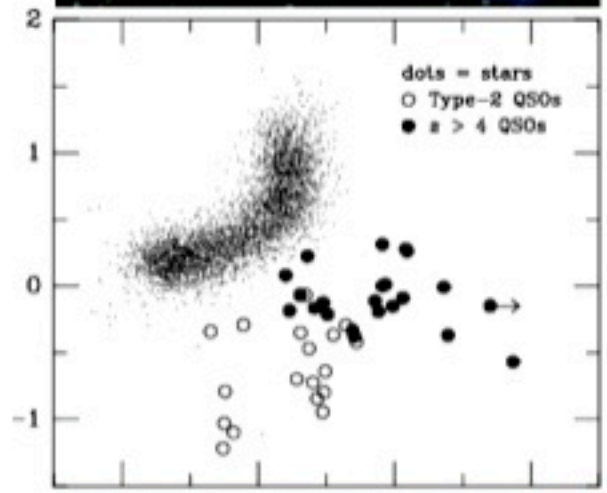
- Technological revolutions as the drivers/enablers of the bursts of scientific growth
  - Detectors, computers + WWW, now data technologies
- Historical examples in astronomy:
  - 1960's: the advent of electronics and access to space
    Quasars, CMBR, x-ray astronomy, pulsars, GRBs, …
  - 1980's - 1990's: computers, digital detectors (CCDs etc.)
    Galaxy formation and evolution, extrasolar planets, CMBR fluctuations, dark matter and dark energy, GRBs, …
  - **2000's and beyond: information technology**

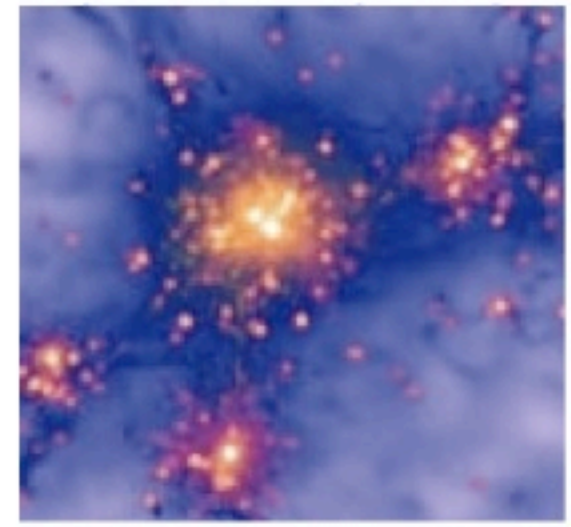
The next golden age of discovery in astronomy?

Virtual Observatory Science Examples

- Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, X-ray, etc., for:
  - Precision large scale structure in the universe
  - Precision structure of our Galaxy
- Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources
  - E.g., extremely distant or unusual quasars, brown dwarfs, new types, etc.
- Probe the evolution of quasars, galaxies, or clusters discovered using different techniques over the cosmic time
- Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations

… etc., etc.

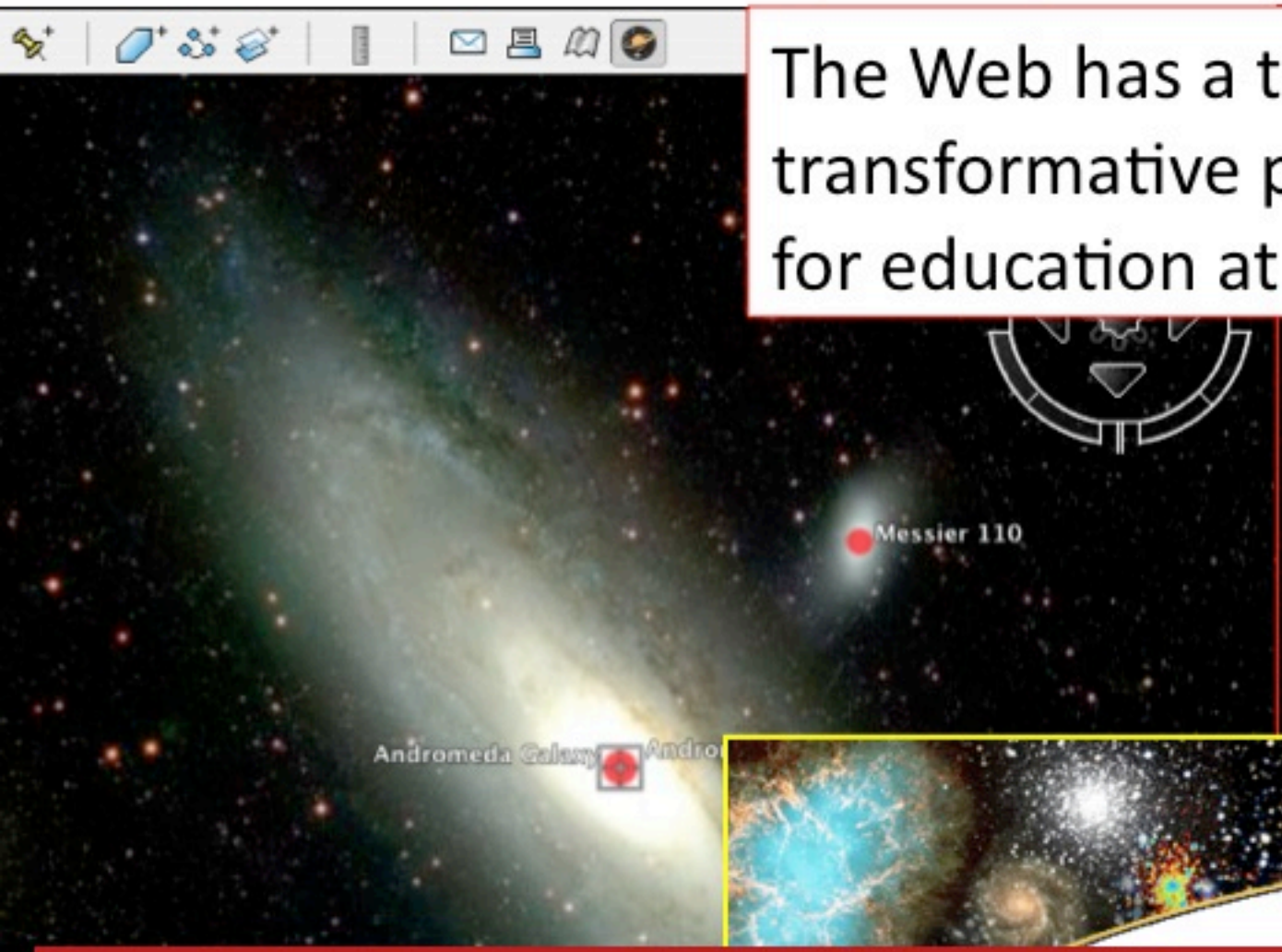


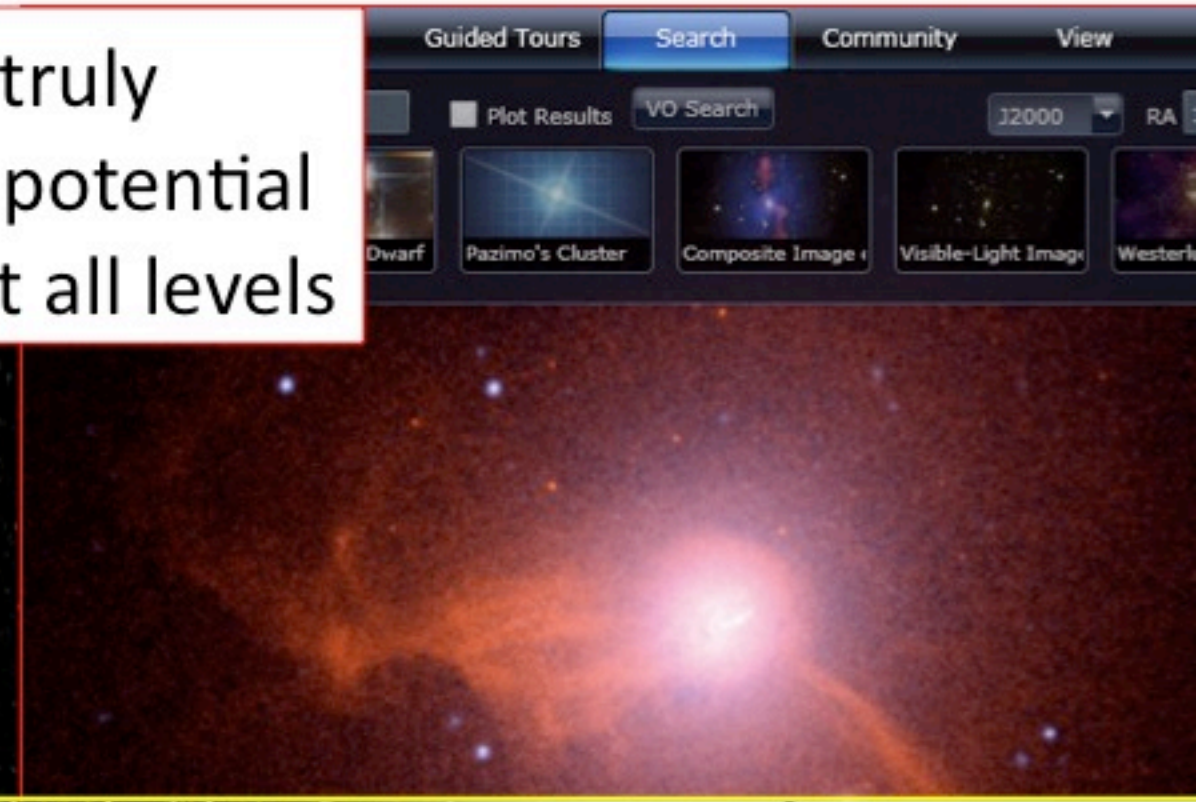


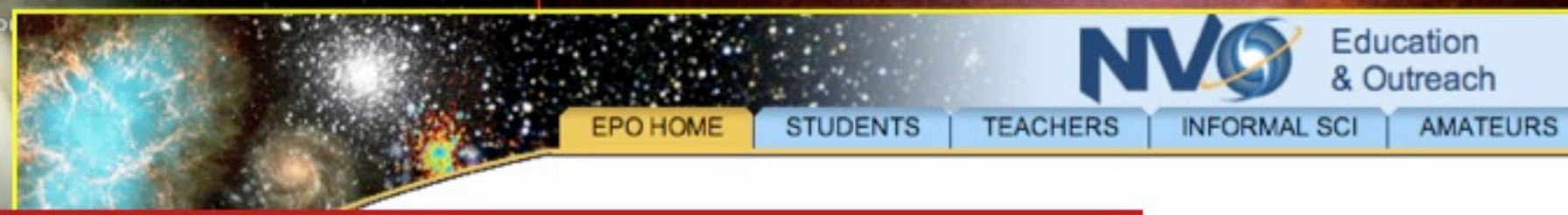
VO Education and Public Outreach

"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 0h42m44.30$treaming

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

The Cyberworld Is Also Flat

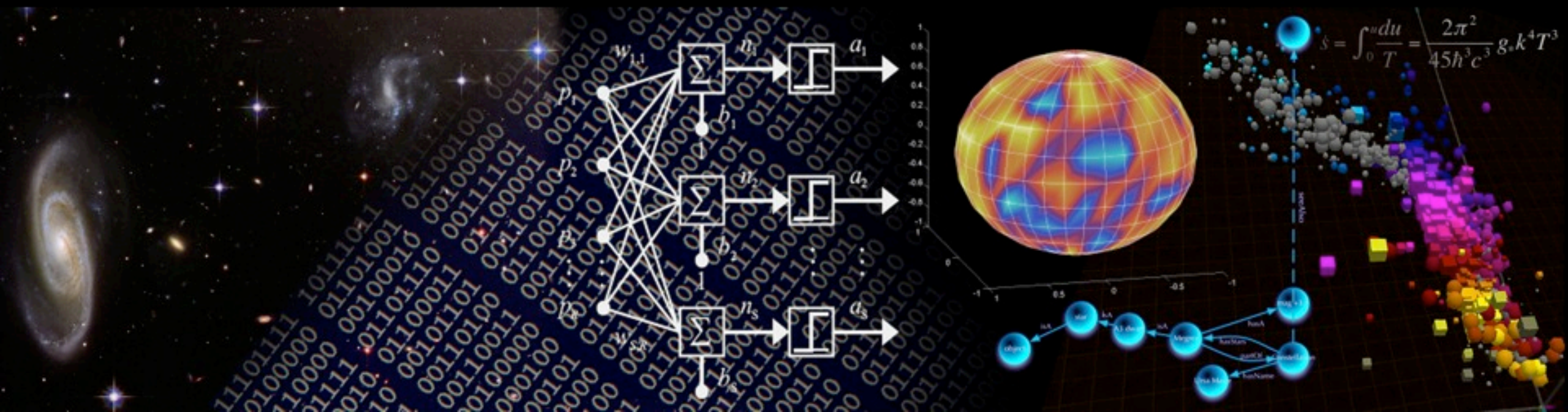
Probably the most important aspect of the IT revolution in science

- **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 democratization of science
  - They can also be substantial contributors, not only consumers of scientific content
- 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 facilities

Beyond Virtual Scientific Organizations:

The Rise of X-Informatics (X = Astro, Bio, Geo, ..)

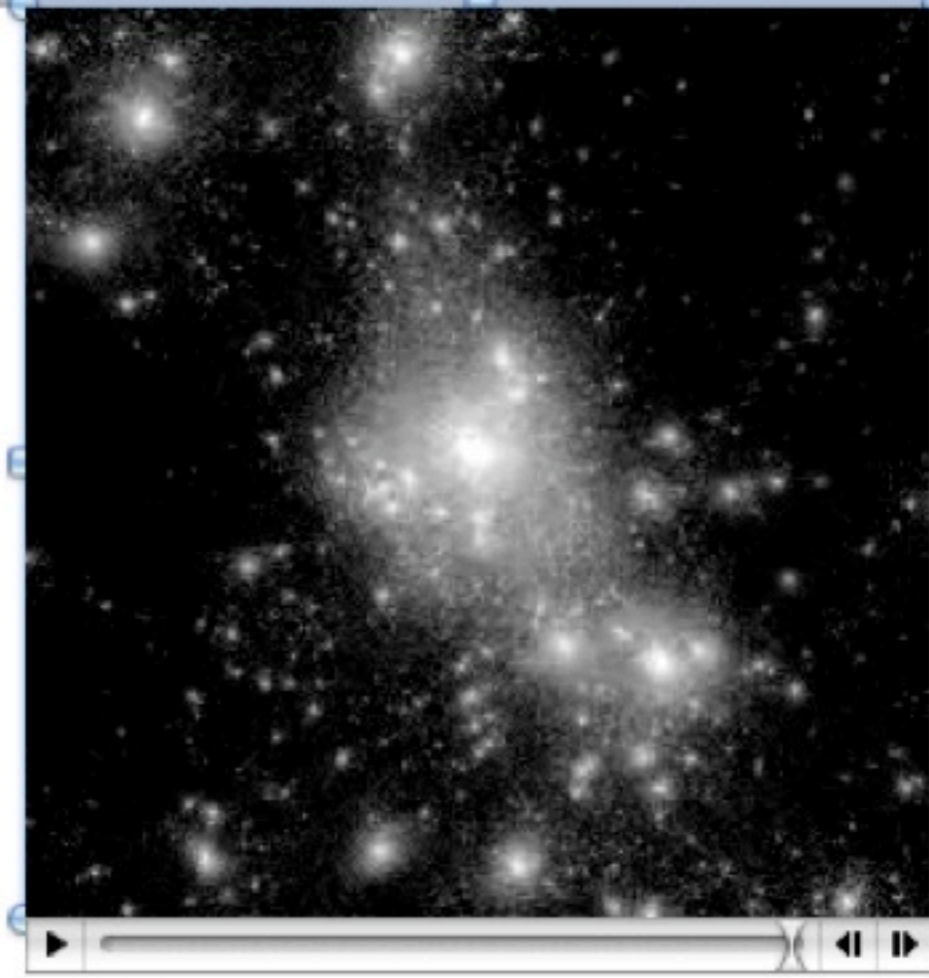
- Domain-specific amalgam fields (science + CS + ICT)
- A mechanism for a broader community inclusion (both as contributors and as consumers)
- A mechanism for interdisciplinary e-Science methodological sharing



Numerical Simulations:

A qualitatively different and necessary way of doing theory, beyond analytical approach

Theory is expressed as ***data*** (an output of a numerical simulation), not as a set of equations



↑ Formation of a cluster of galaxies

← Turbulence

And then, complex theory data must be matched against complex measurements

The Key Challenge: Data Complexity

Or: The Curse of Hyper-Dimensionality

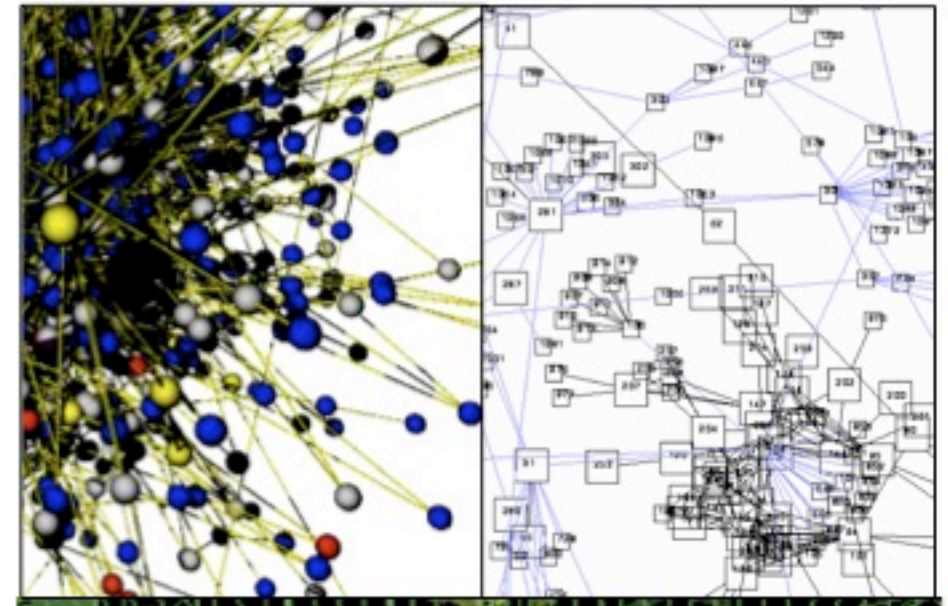
1. **Data mining algorithms scale very poorly:**

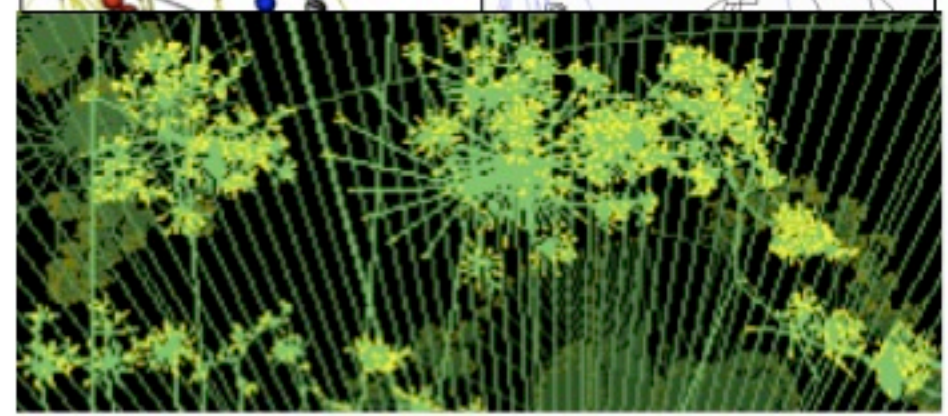
N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

- 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 ($k \geq 1$)

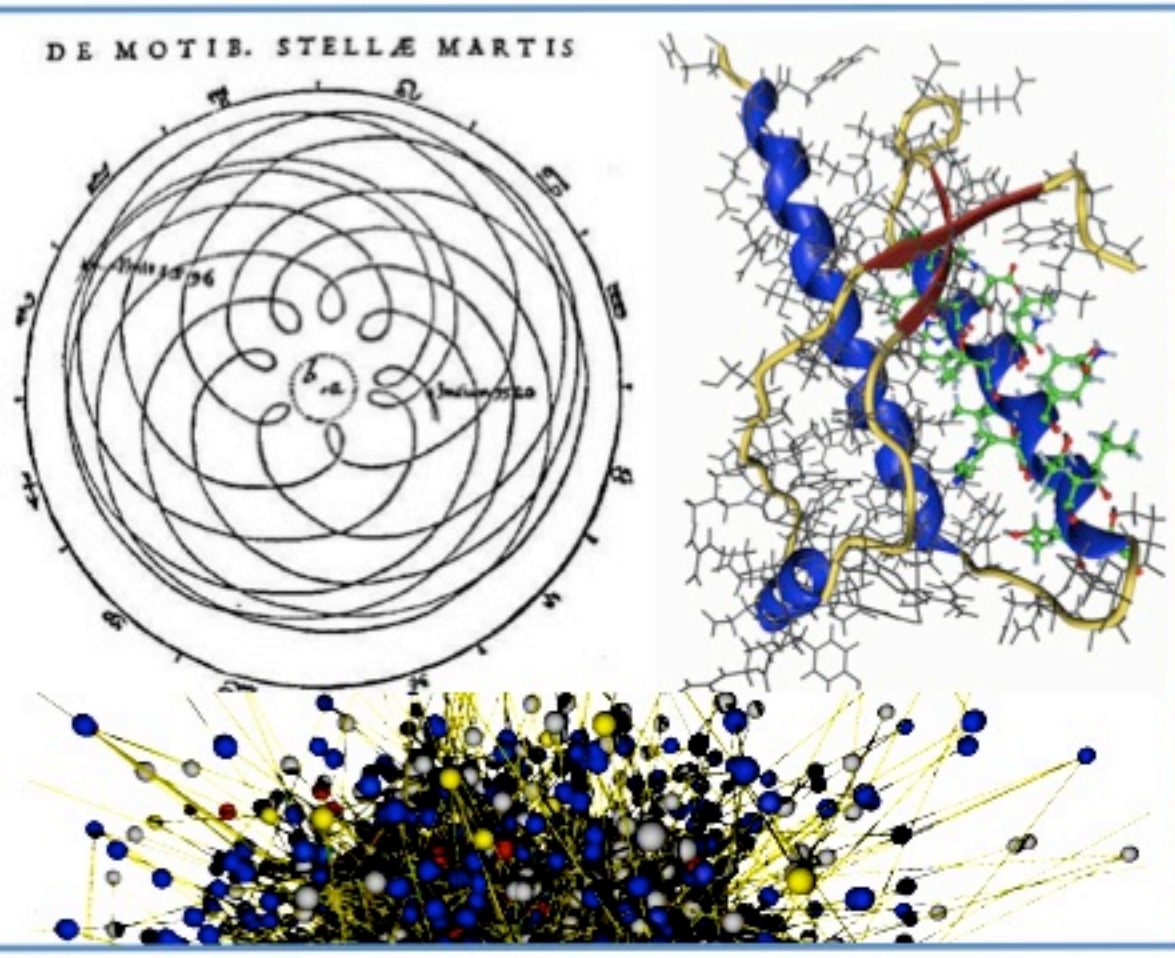
2. **Visualization in >> 3 dimensions**

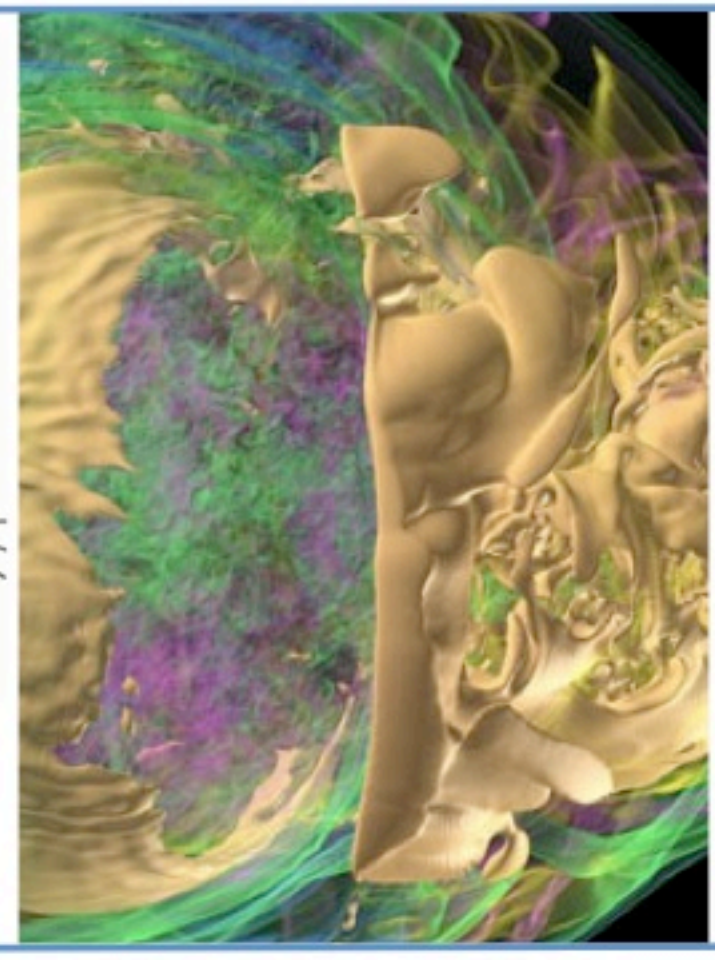
- The complexity of data sets and interesting, meaningful constructs in them is *exceeding the cognitive capacity of the human brain*
- We are biologically limited to perceiving ~ 3 - 12(?) dimensions
- Visualization must be a component of the data mining / exploration process

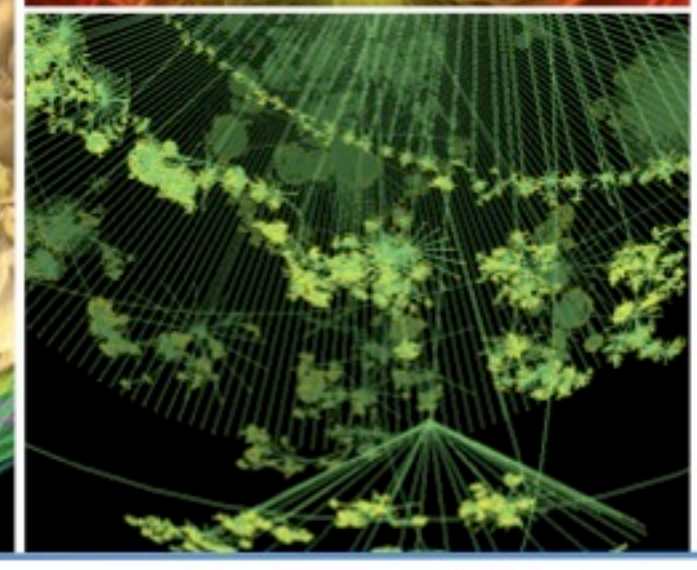


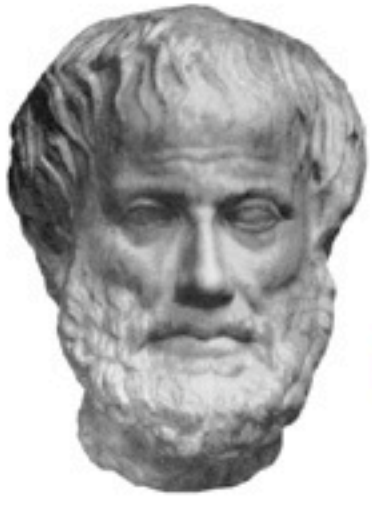


Effective visualization is the bridge between quantitative information and human intuition









Man cannot understand without images

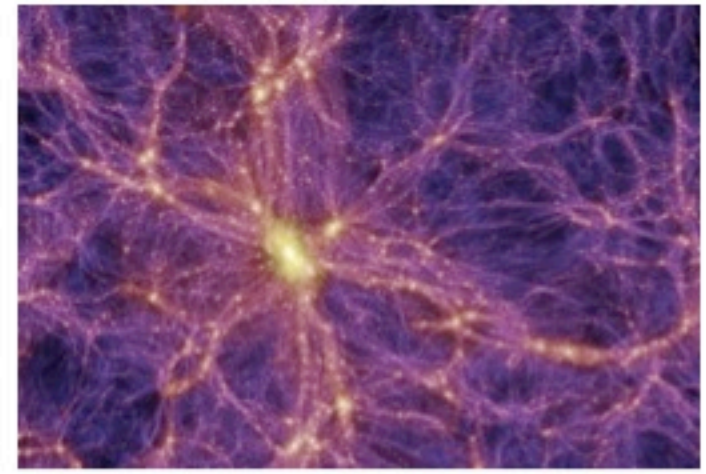
Aristotle, *De Memoria et Reminiscentia*

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 ~ 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 Uses of Machine Intelligence: Science on the Carbon-Silicon Interface

- **Data processing:**
  - Automated object / event classification, pattern recognition
  - Automated data quality control (anomaly/fault detection and repair)

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

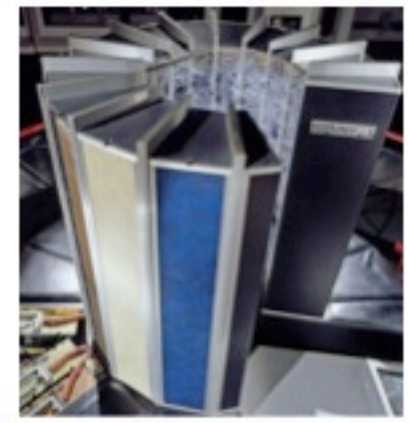
The Evolving Paths to Knowledge

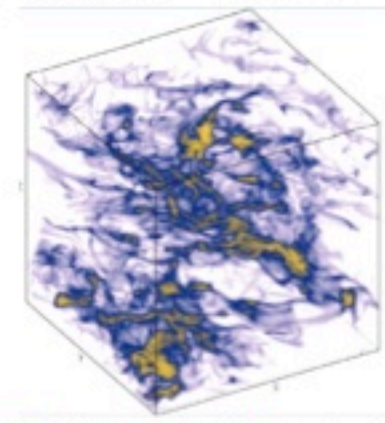
- The First Paradigm:
  Experiment/
  Measurement

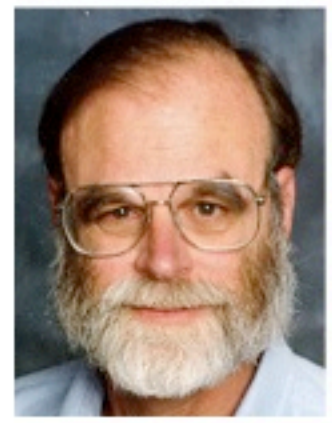
- The Second Paradigm:
  Analytical Theory

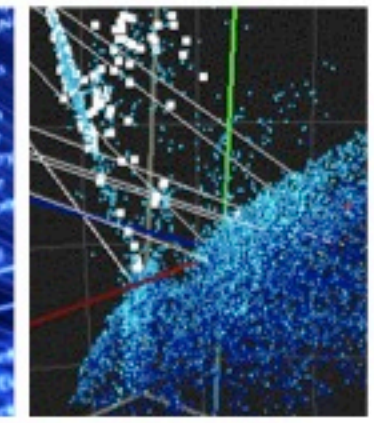
- The Third Paradigm:
  Numerical Simulations

- The Fourth Paradigm:
  Data-Driven Science









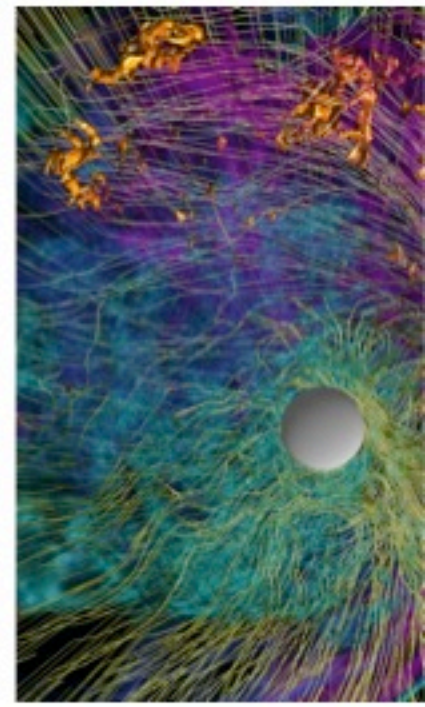
The Fourth Paradigm

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

Some Thoughts About e-Science



- Comput***ational*** science ≠ Compute***r*** 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: an interdisciplinary glue/lubricant
  - Many important problems (e.g., climate change) are inherently inter/multi-disciplinary

The Revolution in Scholarly Publishing

- Increasing complexity and diversity of scientific data and results
  - Data, archives, 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:
  - Curation by domain experts
  - Effective peer review and quality control
  - Persistency and integrity of data and pointers
  - Interoperability and metadata standards
  - ***From the ownership of storage media to the ownership of access to the bits***

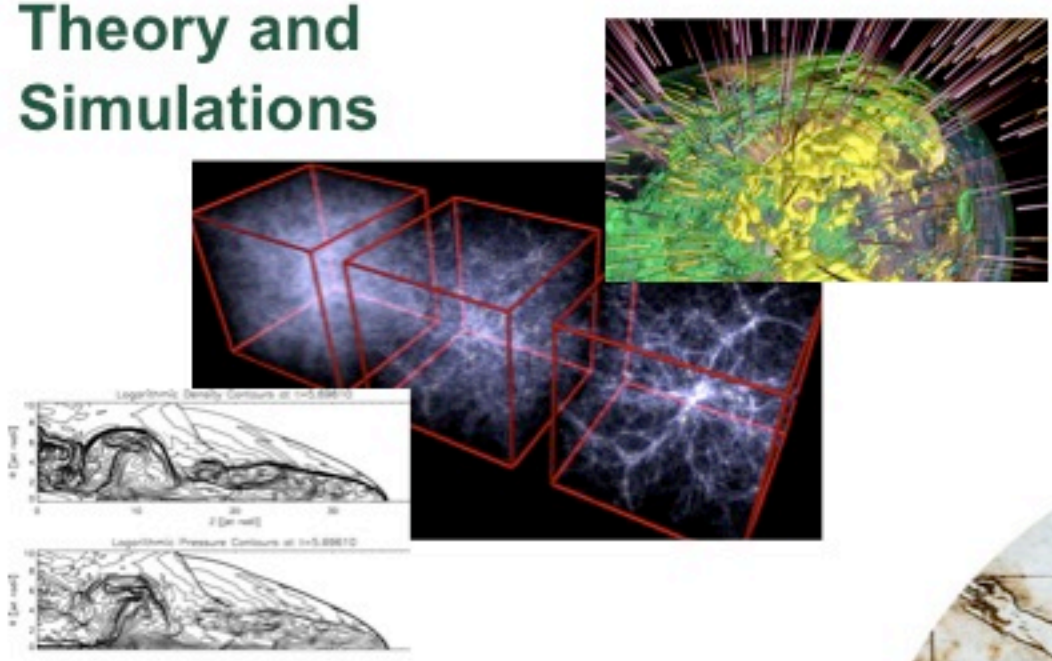
As the science evolves, so does its publishing

Cyberspace (today the Web, with all the 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

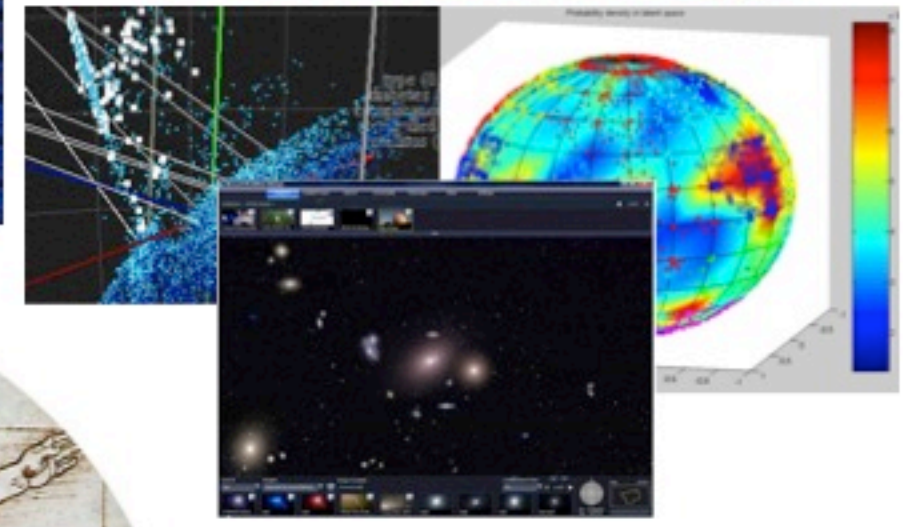
Essentially all aspects of the modern society are migrating to cyberspace, science and scholarship included, with their data, methods, publications, etc.

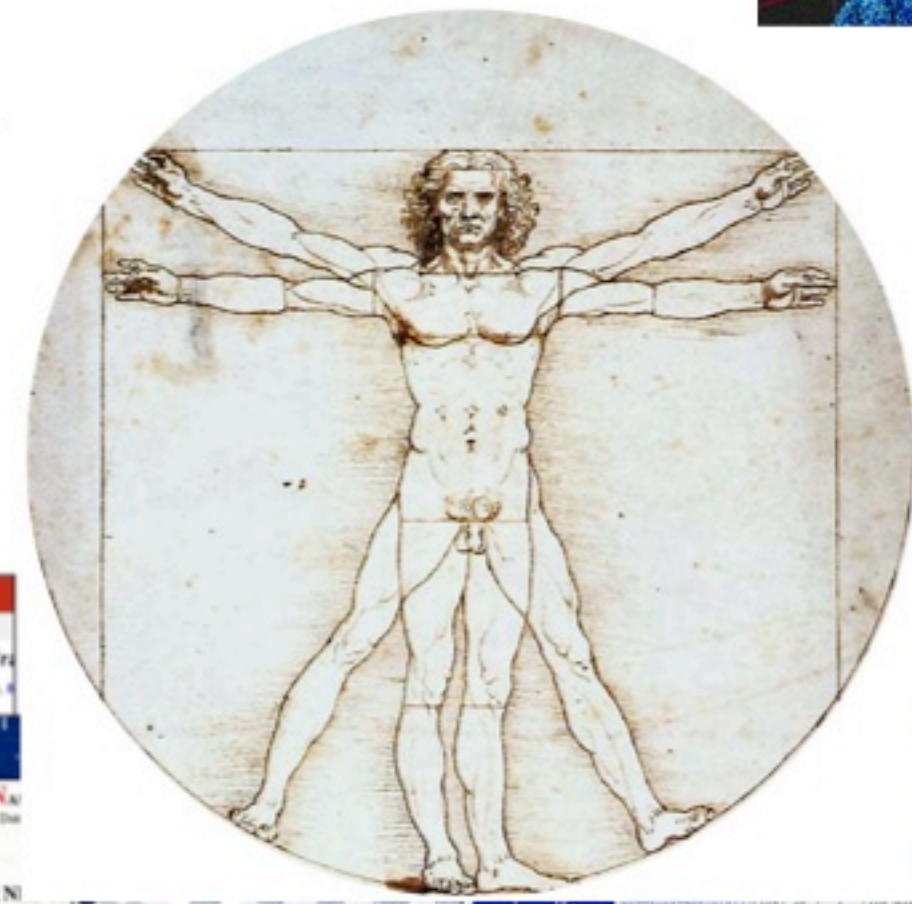
Science in Cyberspace

Theory and Simulations



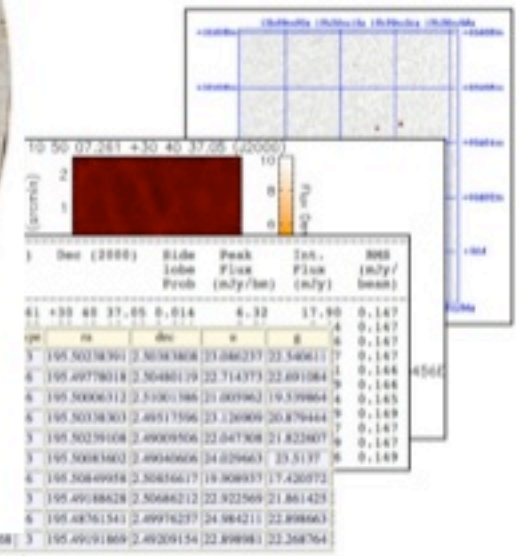
Visual Displays and Linking of Data and Knowledge





Published Literature

Data Archives

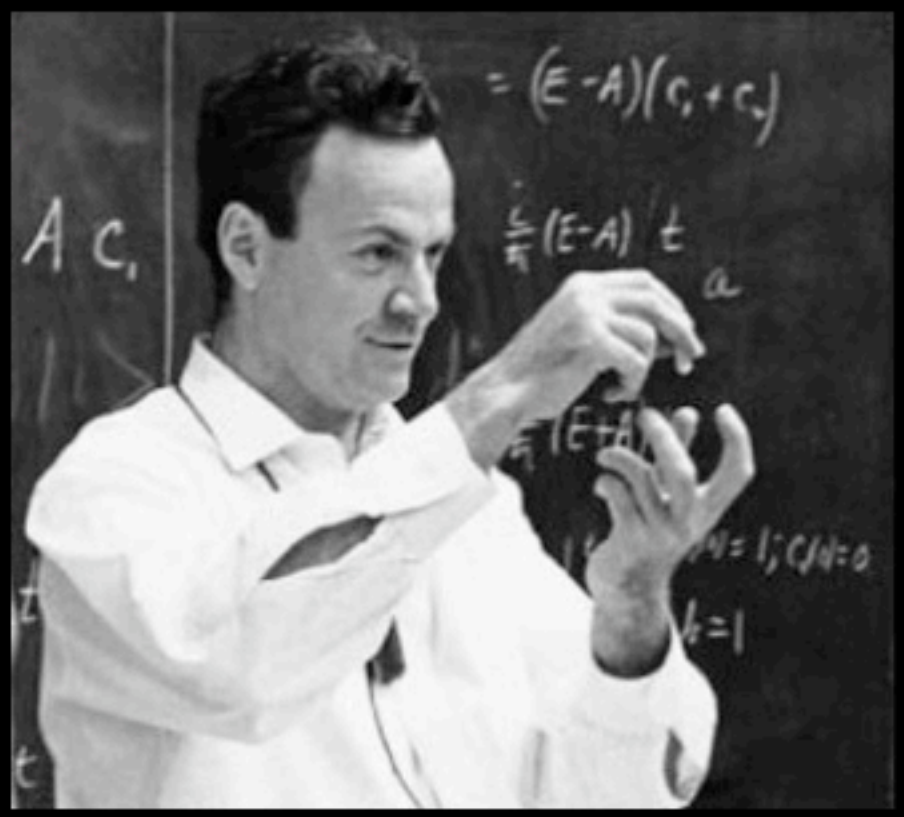


Semantic Web

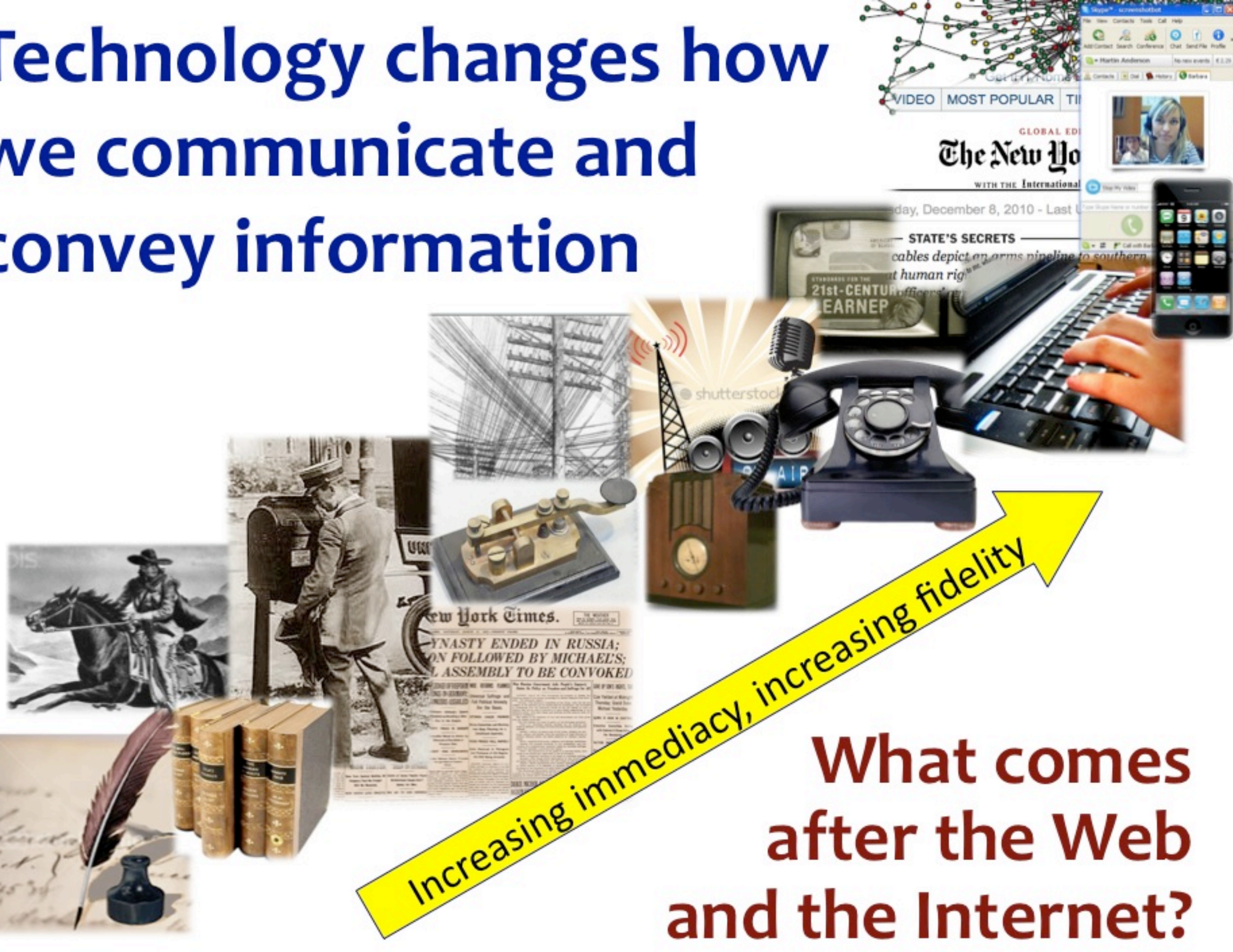
Virtual Observatory

Science Originates on Interfaces

… between human minds, data, and other informational constructs

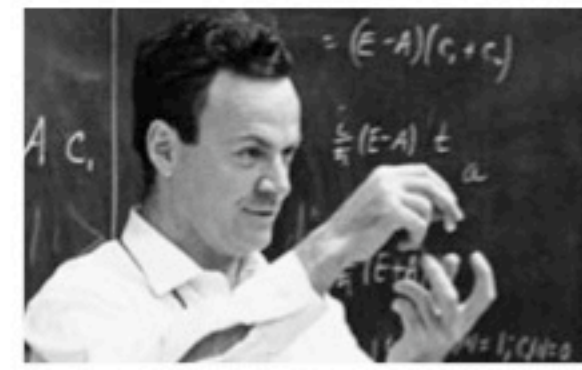


Technology changes how we communicate and convey information



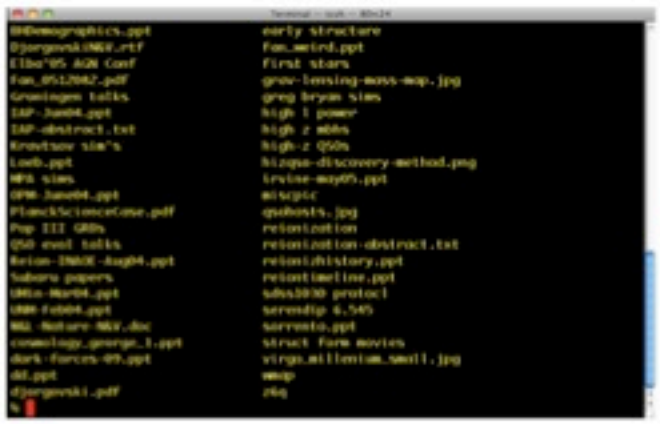
What comes after the Web and the Internet?

Where Minds Meet



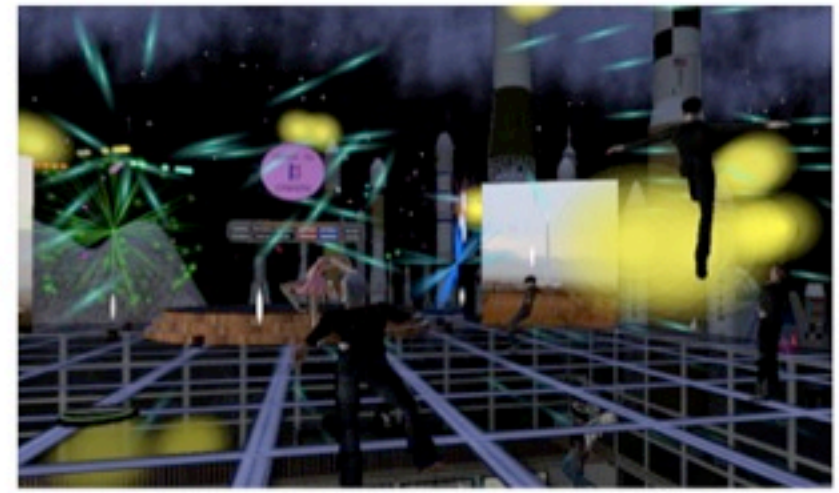
- Exchange, recombination, and cross-pollination of ideas and information on human/data/textual interfaces
- Any technology which facilitates these interactions is enabling science, scholarship, and education

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 …

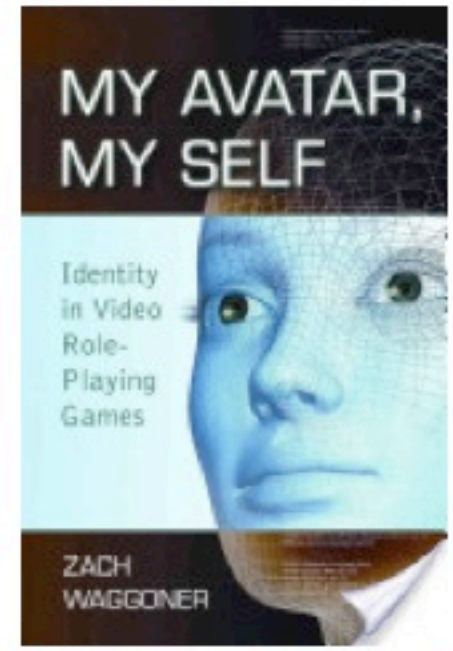
… to Web browsers and hypertext …



… and now immersive virtual environments

Immersive VR and the Emerging 3D Web

- **Immersive VR is a fundamentally transformative technology, on par with the Web browser (maybe more)**
  - The quality of the subjective sense of presence is remarkable, even with the current interfaces
  - The next generation Web will likely be accessed via immersive and augmentative VR interfaces
  - This will change dramatically how we interact with each other and with informational constructs



- **These are still the very early days**
  - 3D display technology is exploding, currently driven by the entertainment industry ($$)
  - Improved interfaces, from cartoony graphics to photorealistic
  - Haptic interfaces, Wii, Kinect, Sixth Sense, …

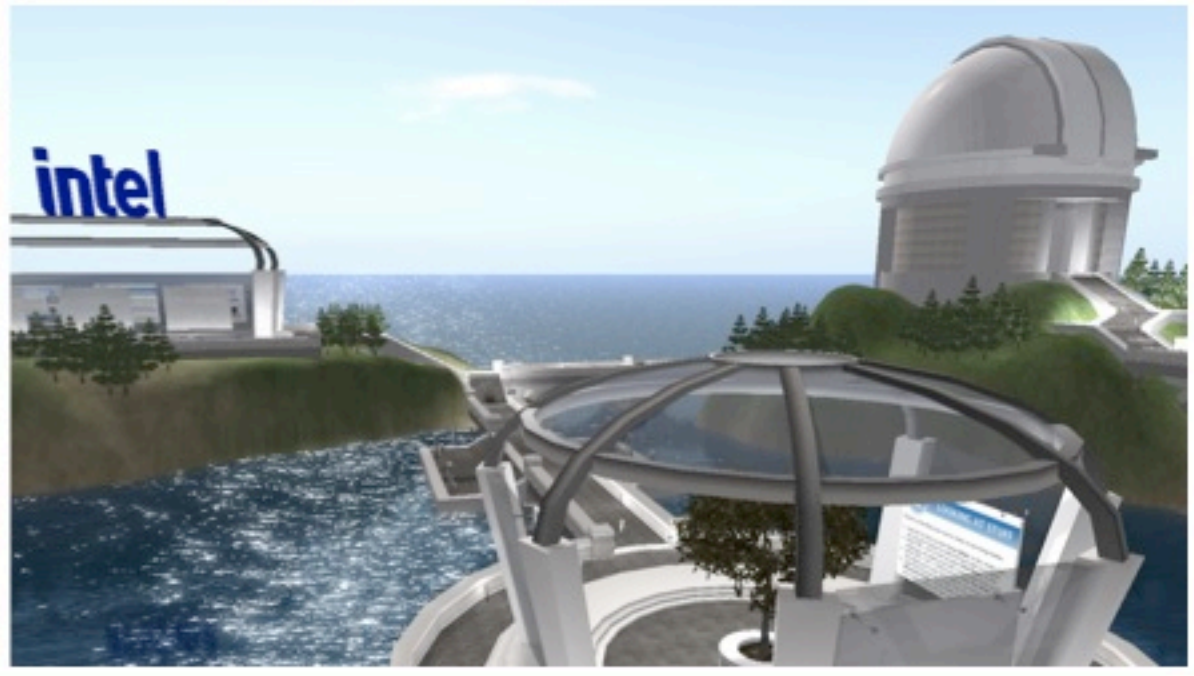
Meta-Institute for Computational Astrophysics

MICA

Meta Institute for Computational Astrophysics

Exploring Astrophysics in Virtual Worlds

http://mica-vw.org/



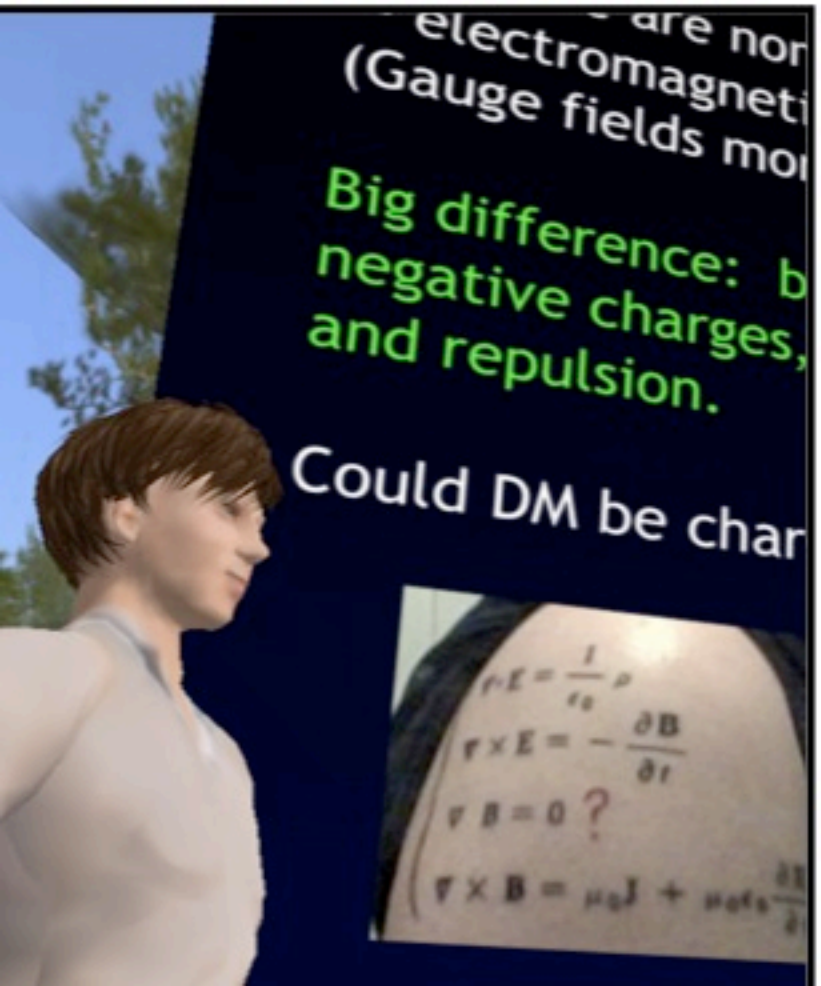
MICA is an experiment in the scholarly use of VWs technologies

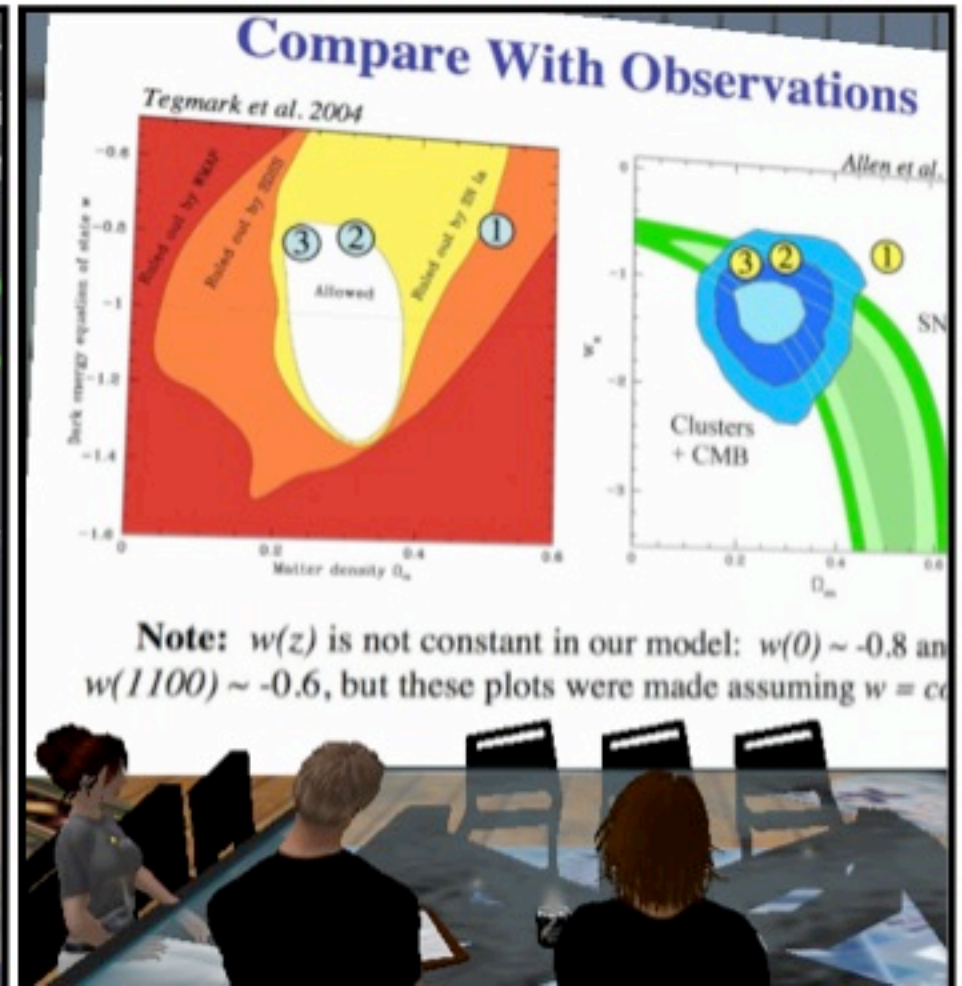
- Virtual worlds are already used extensively for research in the humanities: sociology, psychology, economics…
- Also: business, medicine, situational training, etc.

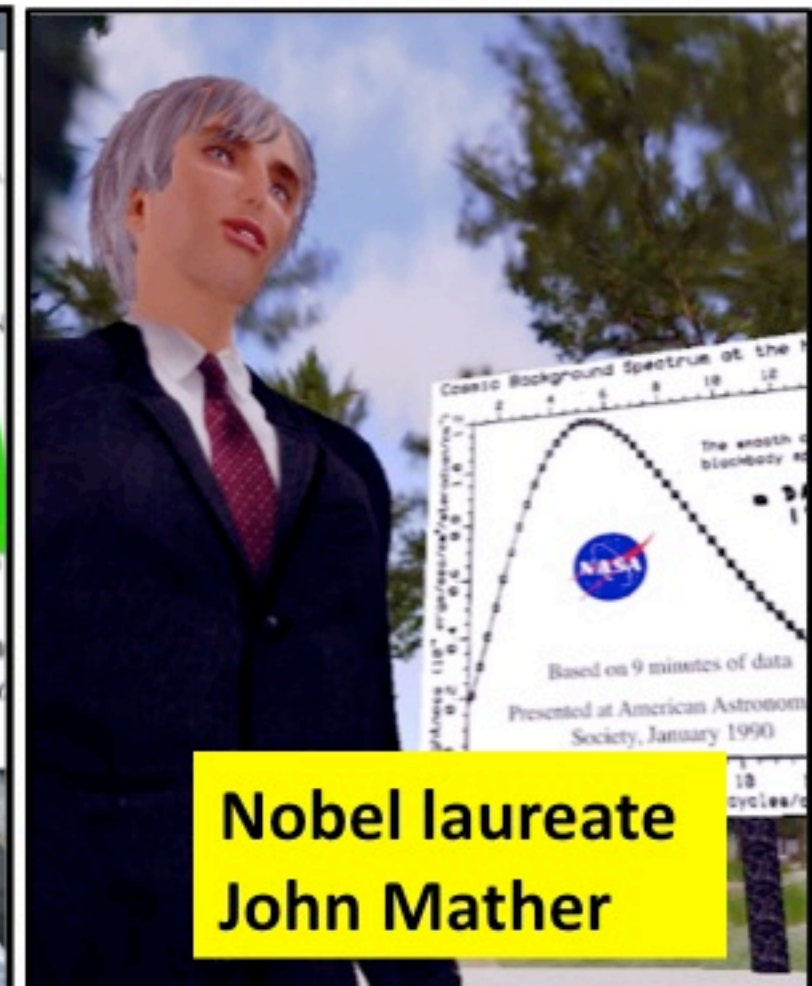
MICA: Scientific Communication and Collaboration in VR Environments

- 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

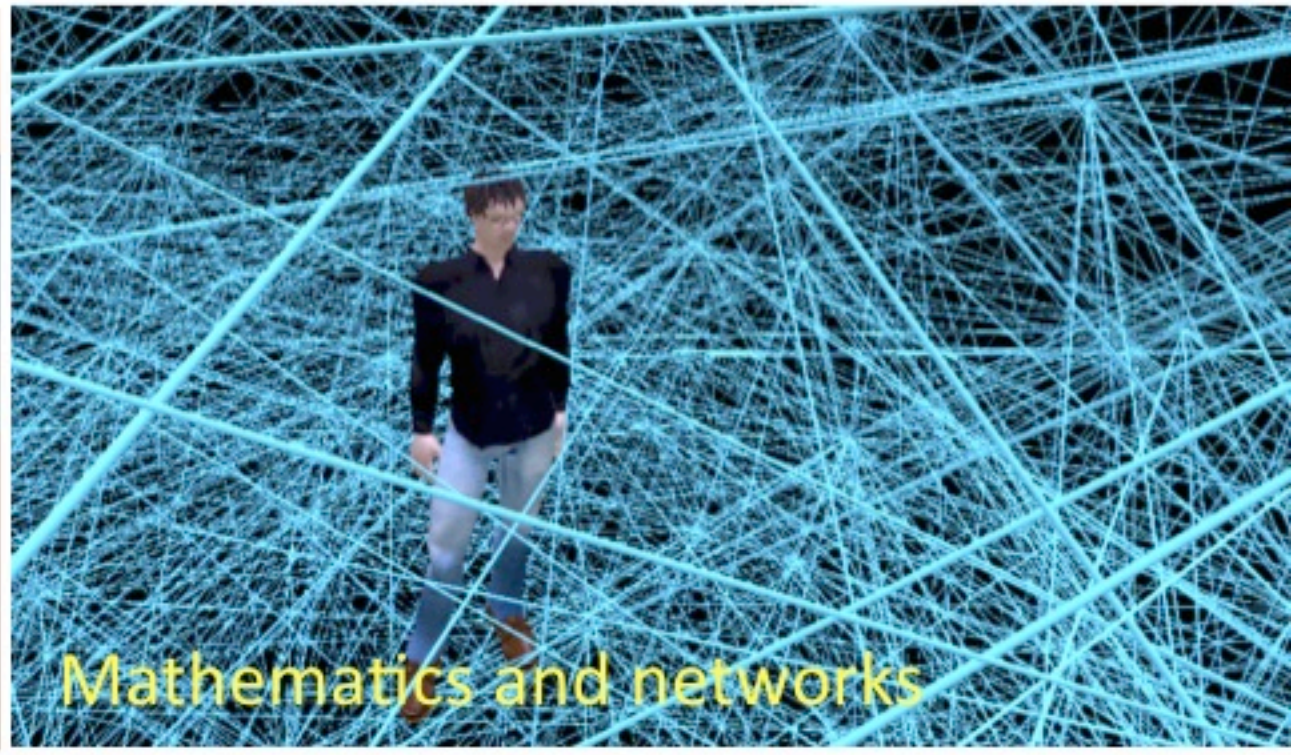
Professional seminars **Collaboration meetings** **Public outreach**

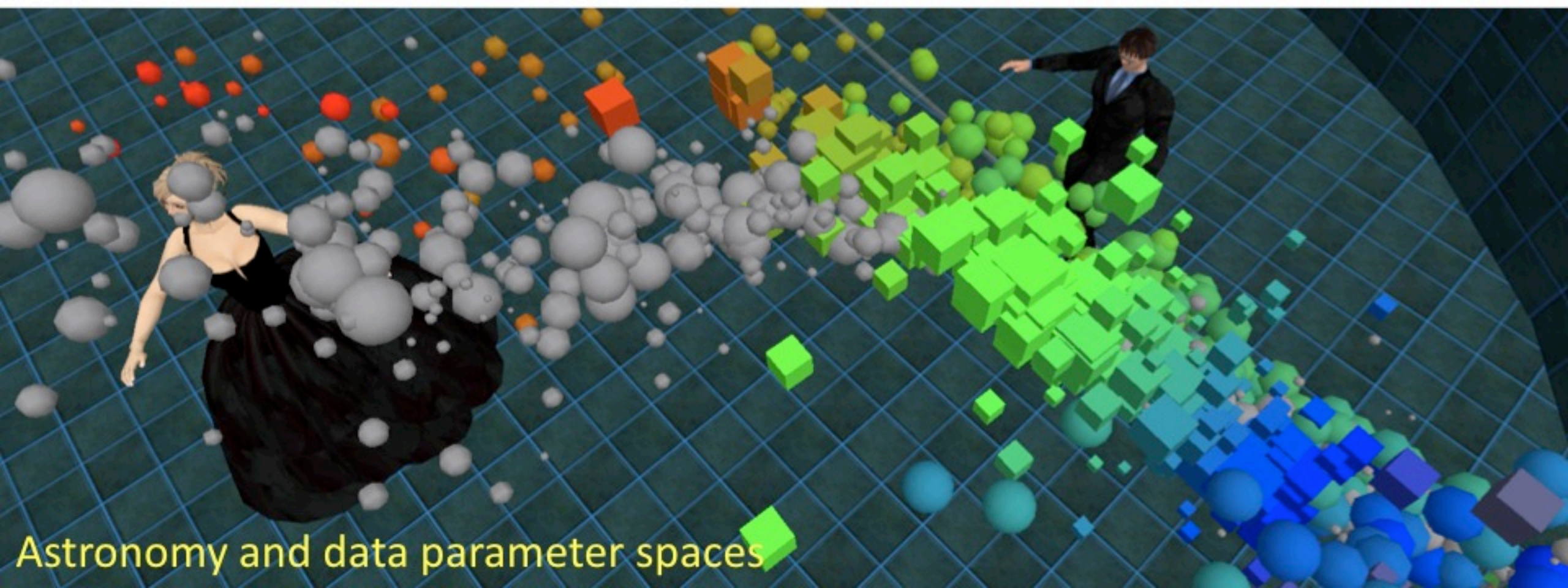






Immersive Data Visualization





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 evidently a very 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?

Summary

- Essentially all of the humanity's activities, science and scholarship included, are **migrating into Cyberspace**, where humans interact with each other, and with the world of information
- Science in the 21st century is increasingly data-rich and computationally enabled, driven by the evolution of technology; thus, **the scientific method evolves**
  - Many specific challenges in the area of data exploration and knowledge discovery: ML, scalability, visualization, …
  - Important well beyond science
- ICT are the fastest developing technologies, paid by the commercial world, and thus strategically well leveraged
- We are witnessing a ***co-evolution*** of humanity, science, and technology