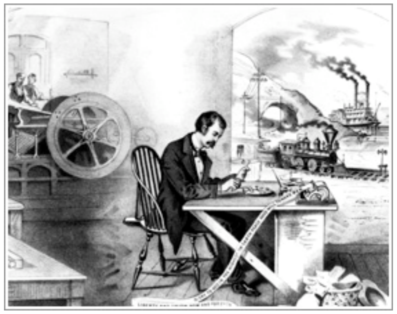
The Big Picture: Information Technology Revolution, and Science in the 21st Century

S. George Djorgovski

Lecture 4

Inaugural BRAVO Lecture Series,
São José dos Campos, July 2007

Roy & George's Excellent Adventure



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…

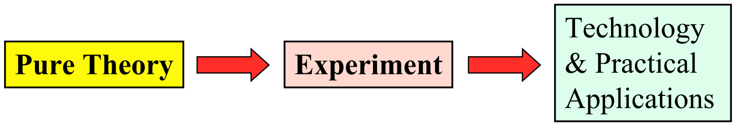
It is a matter of developing a new methodology of science and scholarship for the 21st century

Transformation and Synergy

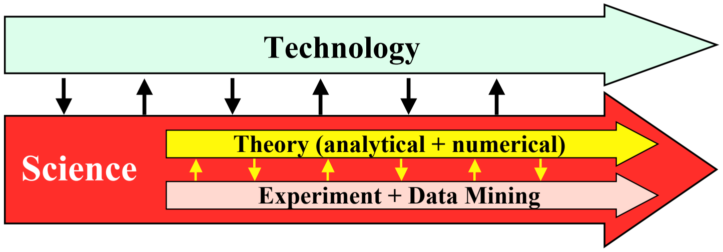
- We are entering the second phase of the IT revolution: the rise of the ***information/data driven computing***
  - The impact is like that of the industrial revolution and the invention of the printing press, combined
- ***All science*** in the 21st century is becoming cyber-science (aka e-science) - 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**
- There is a great emerging synergy of the computationally enabled science, and the science-driven IT

Scientific and Technological Progress

A traditional, “Platonistic” view:

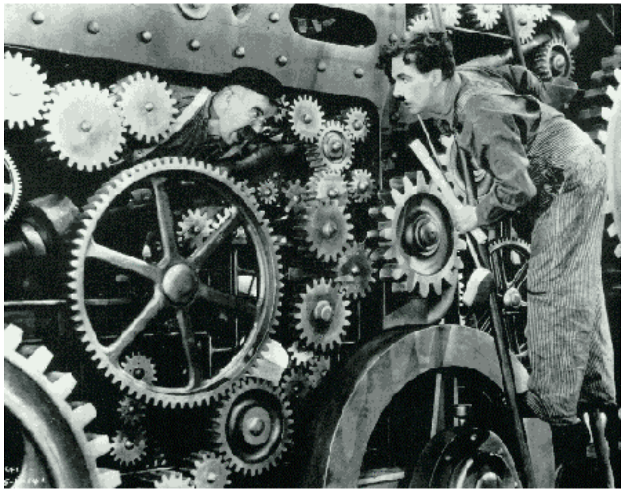


A more modern and realistic view:



This synergy is stronger than ever and growing

Let’s Take a Closer Look at Some Relevant Technological Trends …



The rate of the overall computing power has been amazingly growing for more than one hundred years

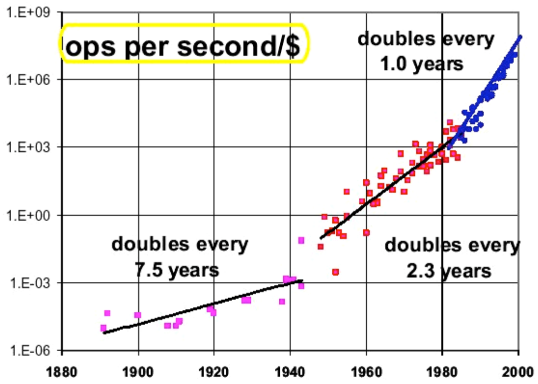
Computing efficiency in ops/s/$ had 3 growth curves:

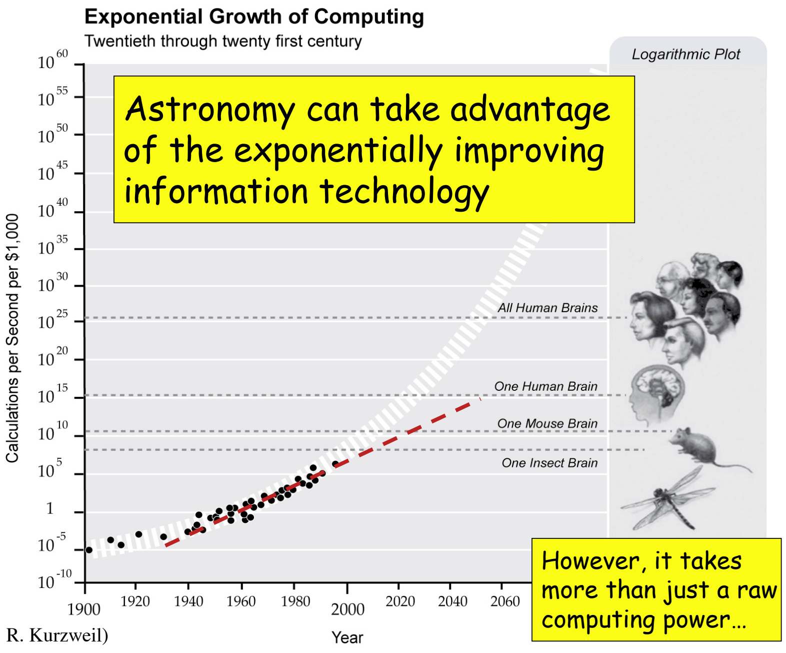
Combination of Hans Moravac + Larry Roberts + Gordon Bell
WordSize*ops/s/sysprice

1890-1945
Mechanical
Relay
7-year doubling

1945-1985
Tube, transistor,..
2.3 year doubling

1985-2000
Microprocessor
1.0 year doubling



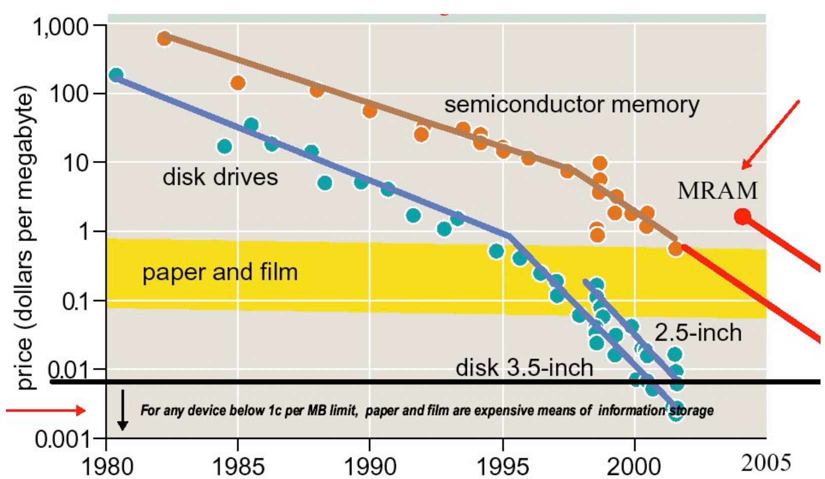


Astronomy can take advantage of the exponentially improving information technology

However, it takes more than just a raw computing power…

(figure: R. Kurzweil)

Exponentially Declining Cost of Data Storage



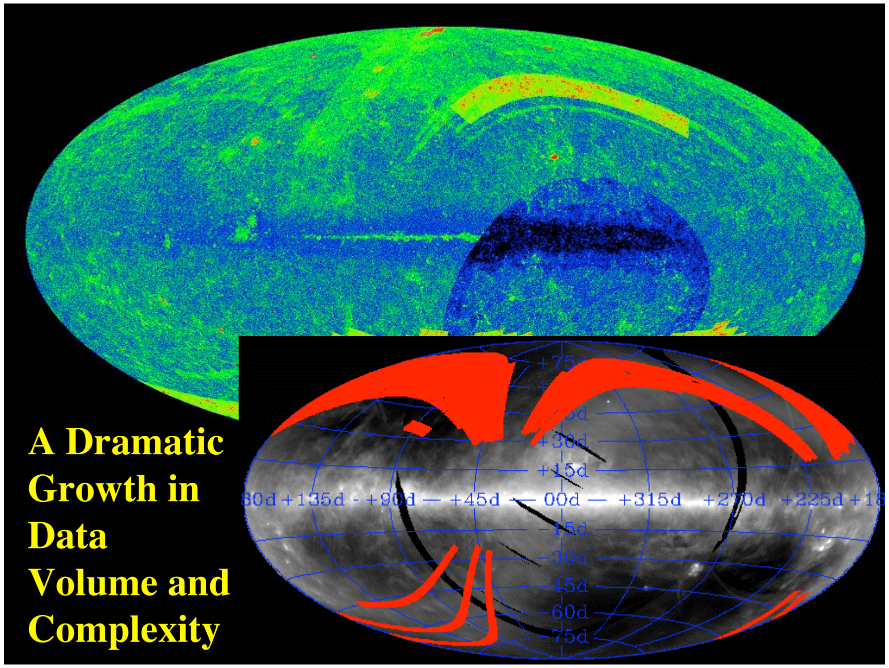
Hayes, Grochowski: American Scientist, 2002

An Early Disk for Information Storage

- Phaistos Disk:
  Minoan, 1700 BC

- No one can read it ☺

(From Jim Gray)



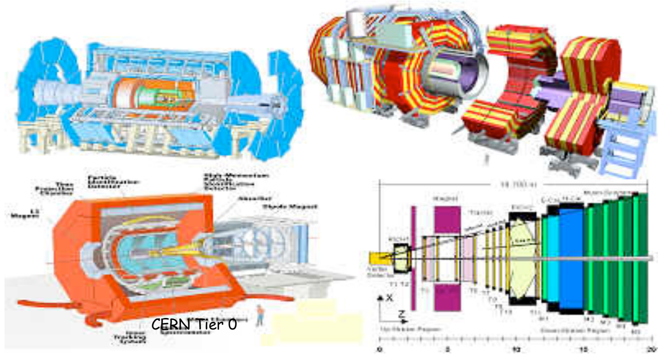
A Dramatic Growth in Data Volume and Complexity

High Energy Physics Instruments (e.g., the LHC): Exabytes to Petabytes per Year

Looking for the Higgs Particle:

- Sensors: 1000 GB/s (1TB/s ~ 30 EB/yr !)
- Events 75 GB/s
- Filtered 5 GB/s
- Reduced 0.1 GB/s

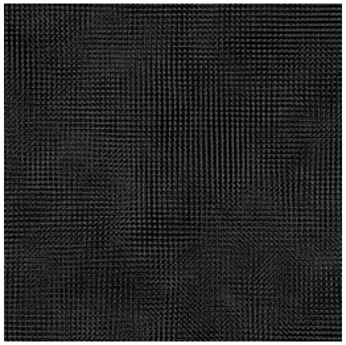
} Thus, *very reduced data* ~ 2 PB/yr !

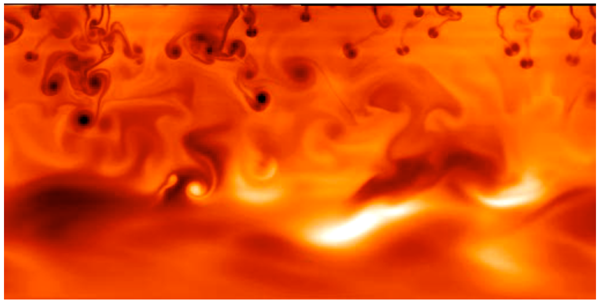


Numerical Simulations:

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

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



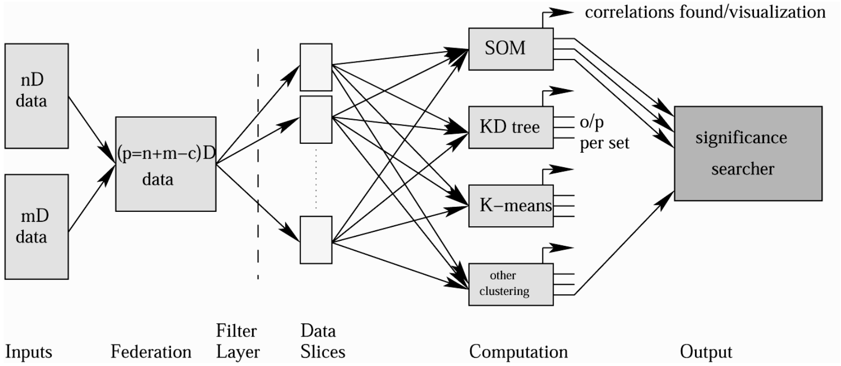


⬆ Formation of a cluster of galaxies

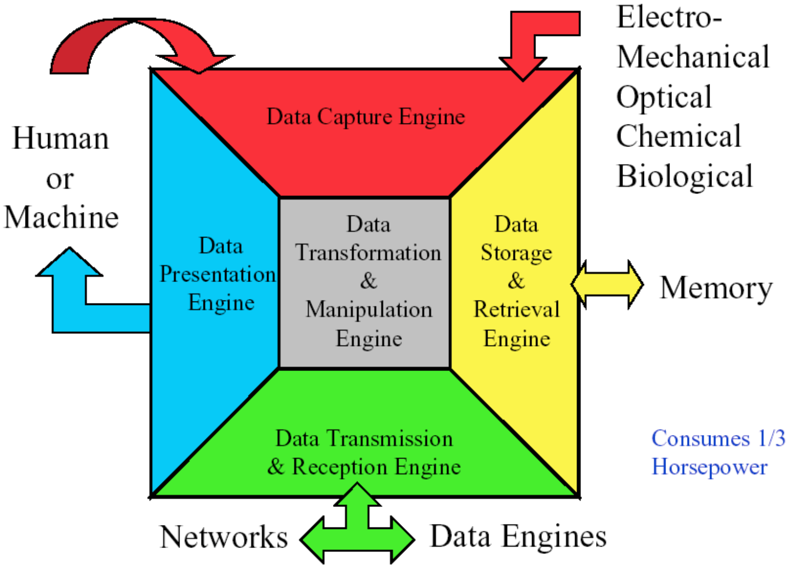
← Turbulence

A New Generation of Scientific Data Analysis Systems

Seamless interactive combinations of data mining, exploration, visualization, and analysis services, operating on standardized format data from any source (astronomy, biology, …)



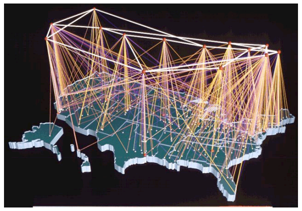
A Modern Data Analysis Engine?



2000 ISSCC Where Will Processor Performance Come From in the Next Ten Years? Frank Fox frank.fox@ieee.org

The Book and the Cathedral …

… and the Web, and the Computer …

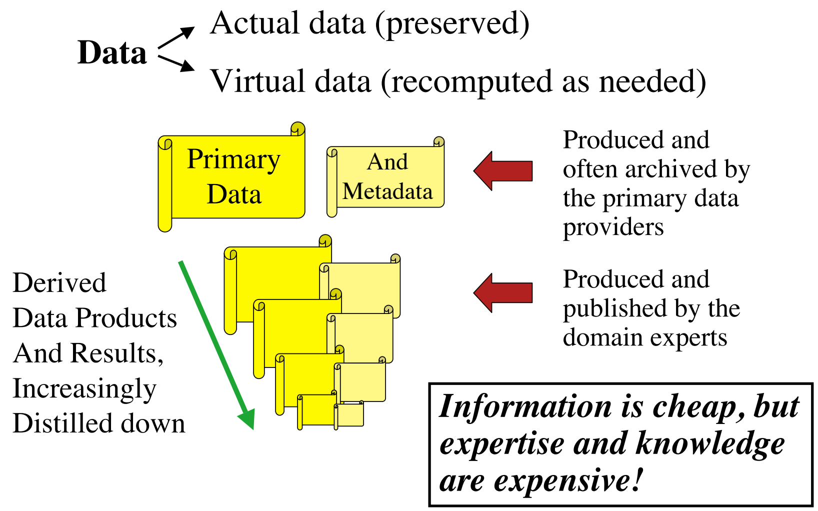


Revolution in Scientific Publishing and Curation

Information and Knowledge Management Challenges

- The concept of scientific data and results is becoming increasingly more complex
  - Data, metadata, virtual data, a hierarchy of products
  - From static to dynamic: revisions and growing data sets
  - ***From print-oriented to web-oriented***
- The changing nature of scientific publishing
  - Massive data sets can be only published as electronic archives, and should be curated by domain experts
  - Peer review / quality control for data and algorithms?
  - The rise of un-refereed archives and a low-cost of web publishing
  - Persistency and integrity of data and pointers
  - Interoperability and metadata standards
- The changing roles of university/research libraries

The Concept of Data (*and* Scientific Results) is Becoming More Complex



Data

Actual data (preserved)

Virtual data (recomputed as needed)

Primary Data

And Metadata

Produced and often archived by the primary data providers

Derived Data Products And Results, Increasingly Distilled down

Produced and published by the domain experts

Information is cheap, but expertise and knowledge are expensive!

The Changing Nature of Scientific Data and Results:

Static ➡ Dynamic

- Recalibrations: Which versions to save?
- Intrinsically growing data sets: Which versions to save?
- Virtual data:
  - Re-compute on demand, save just the algorithm, but operating on which input version?
  - What about improved algorithms?
- ***Domain expertise is necessary!***
  - Synergy between curation institutions (libraries, archives, museums) and research institutions (and other scholarly content creators) is essential
  - New hybrid types of (virtual) institutions / organizations?

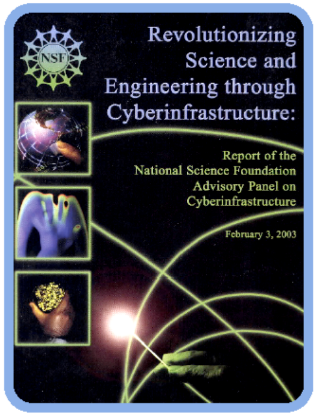
The Response of the Scientific Community to the IT Revolution

- Sometimes, the entire new fields are created
  - e.g., bioinformatics, computational biology
- 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
- However:
  - Little or no coordination and interchange between different scientific disciplines
  - A slow general community buy-in

The Cyber-Infrastructure Movement

"a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today's challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive "cyberinfrastructure" on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy."

(aka "The Atkins Report")



Revolutionizing Science and Engineering through Cyberinfrastructure:

Report of the National Science Foundation Advisory Panel on Cyberinfrastructure

February 3, 2003

The Rise of Virtual Scientific Organizations

- There is an ever growing number of them:
  - NVO = National Virtual Observatory
  - NEESgrid = Network for Earthquake Engineering Simulation
  - CIG = Computational Infrastructure for Geophysics
  - NEON = National Ecological Observatory Network
  - GriPhyN = Grid Physics Network
  - BIRN = Brain Imaging Research Network

  … etc. etc.
- These are the effective responses of various scientific disciplines to the IT/data-related challenges and opportunities
- Note: they are ***discipline-based***, not institution-based!
- And generally global in reach
- The next step: a cross-disciplinary communication, collaboration, and exchange of ideas

OK, So … What is Really New Here?

Why is this not the same old science but with more data and computers?

What is ***qualitatively*** new and different?

How is scientific practice in the 21st century going to be different from the past?

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
- VO is the framework to effect this for astronomy

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

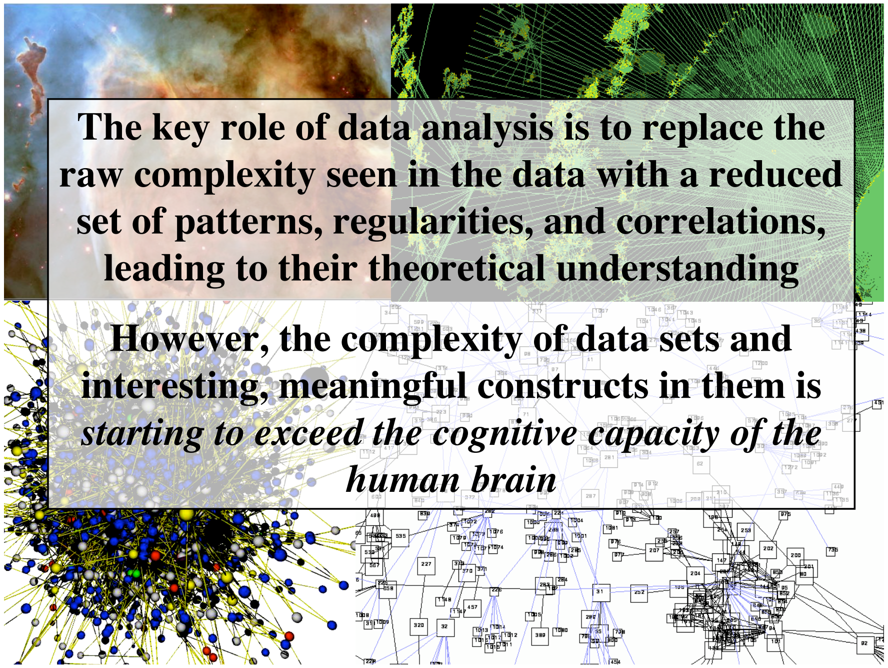
Key Technical Challenges

↳ **Data Understanding**

↳ **New Knowledge**

Key Methodological Challenges

+feedback



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 of data sets and interesting, meaningful constructs in them is *starting to exceed the cognitive capacity of the human brain*

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?

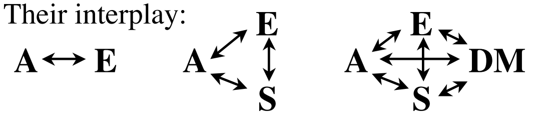
The Evolution of Science

1600 1700 1800 1900 1950 2000 → *t*

Empirical/Descriptive
Analytical+Experimental
A+E+Simulations
A+E+S+**DM/DE/KDD**

Technology

Their interplay:



Computational science rises with the advent of computers
Data-intensive science is a more recent phenomenon

The Evolving Role of Computing:

Number crunching ➔ Data intensive (data farming, data mining)

Some Thoughts on CyberScience

- Enables a broad spectrum of users and contributors
  - From large teams, to small teams, to individuals
  - Data volume ~ team size, but scientific returns ≠ *f* (team size)
  - Human talent is distributed very broadly geographically

    Open, distributed, web-based nature of new science is a key feature
- Transition from data-poor to data-rich science
  - Chaotic ➔ Organized … regulation vs. creative freedom
  - Can we learn to ask ***a new kind of questions?***
- Information is cheap, but 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: a formal, universal framework
  - Scientific discovery is fundamentally a pattern recognition process

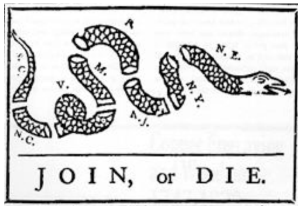
Universal Challenges:
The New Scientific Methodology

- **Data farming and harvesting**
  - Semantic webs, computational and data grids, universal or trans-disciplinary standards and ontologies …
  - Digital scholarly publishing and curation (libraries)

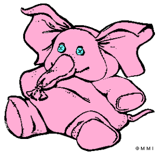
    … data, metadata, virtual data, hierarchical data products; legacy vs. dynamical; open vs. proprietary; data, knowledge, and codes; persistency; peer review; web samizdat vs. officially blessed and supported; mandates; etc., etc.
- **Data mining and understanding, knowledge extraction**
  - Scalable DM algorithms
  - Hyperdimensional visualization
  - Empirical validation of numerical models
  - Computer science as the “new mathematics”
- **The art and science of scientific software systems**
  - Architecture, design, implementation, validation …

Universal Challenges: Sociology

- Breaking the stovepipes
  - From domain-specific cases to a general computational science approach
  - A more efficient use of resources (human and technological)
  - New science on, and across the traditional disciplinary boundaries



- New modes of scientific organization and collaboration
  - Domain-specific virtual organizations, collaboratories
    - Do we know how to run them optimally?
  - Inter-agency, public-private, national-global partnerships …
    - Enlightened self-interest is the key
- Fostering a computational thinking, education
- Community buy-in!



The Participation Challenge:
What if you gave a revolution, and no one came?

- A small subset of the scientific community is understanding the scope of the ongoing fundamental transformation; most are not buying, for 2.5 reasons:
  - “I don’t do it / don’t understand it, therefore it cannot be important”

    … And these upstarts may distract from my glory and divert resources
  - Where are the new IT/CS-enabled discoveries?
- How to go about it?
  - ***Make some discoveries!***

    ➔ Must go beyond data handling, into data mining etc
  - Enable an easy participation
  - Make it glamorous and attractive by political and funding support
- What are the relevant currencies? ($, data, cycles, expertise…)

"There will be opened a gateway and a road to a large and excellent science, into which minds more piercing than mine shall penetrate to recesses still deeper."

Galileo (1564–1642)

[on the "experimental mathematical analysis of nature"]

Take a look at the presentations at

http://escience.caltech.edu/workshop/

Some Questions to Discuss:

DIALOGO di GALILEO GALILEI LINCEO
AL SER.mo FERD. II. GRAN. DVCA DI TOSCANA

- What is beyond Cyber-Infrastructure, when *all* science is e-Science?

  "We need some new clichés" - S. Goldwyn
- What is the optimal R&D program for the new scientific methodology?

  And the right hardware/software balance?
- How do we get some genuine inter-disciplinary collaborations going?
- How do we optimize the public-private, academia-gov't-industry partnerships?
- How do we accelerate the community participation?
- How do we learn and teach computational thinking?

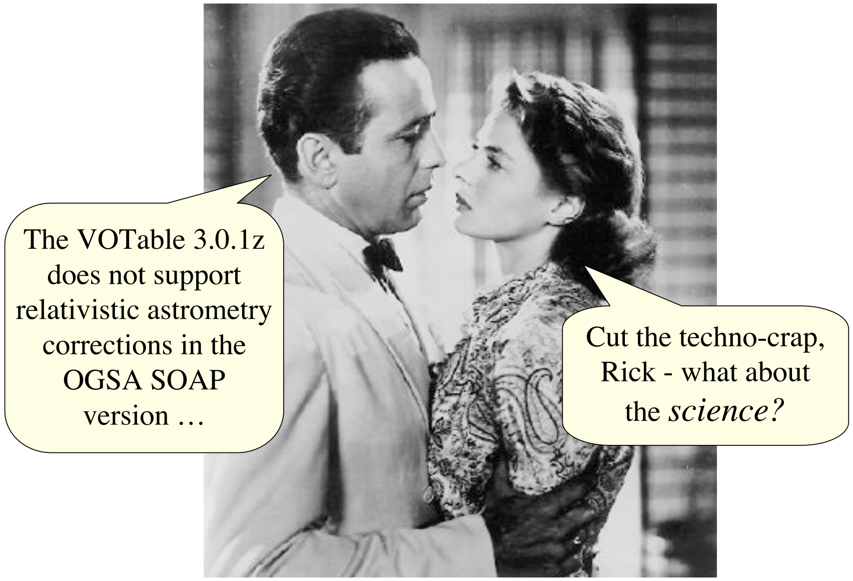
Some Musings on Virtual Observatory and e-Science in General

By Rick and friends

This is going to be rough, kiddo.
Oh, Rick … I love data mining!

And besides, We'll always have Pasadena…

There Has Been Much Progress … Of Sorts …


The VOTable 3.0.1z does not support relativistic astrometry corrections in the OGSA SOAP version …
Cut the techno-crap, Rick - what about the *science?*

So, How Are We Doing?

Uh, Not So Good …

So, Rick - How come the VO is not doing so well?
I blame the management
But darling, you *are* the management!
Ah, well, I …
…Dames …

Good Thing We Are Getting Organized

International VO cooperation is very important!
IVOA
Yes, Roy, I love to collaborate
We'll let that Hungarian fellow do the work …

Things Are Looking Good, Yes?

Virtual Astronomy is just one example of e-Science … driven by the progress in IT, which is getting exponentially cheaper and more powerful - and that is a great trend to ride!
Yes, computers are cheap, but people and software are expensive!

So, Should we be Outsourcing e-Science?

I need some data reduced fast… … lots of data
I have a high bandwidth connection to India, Efendi… … and a chap named Ajit
But you cannot outsource your core competence!
Those Indians are smart folks, Efendi…

Then Why Is There So Little VO Science?

Because the problems really are hard! And because we have no smarts, no imagination, no good tools, no …
Oh, Rick, this is so wonderful! So much to do!

Some Questions to Discuss:

DIALOGO di GALILEO GALILEI LINCEO AL SER.mo FERD. II. GRAN DVCA DI TOSCANA

- What is beyond Cyber-Infrastructure, when *all* science is e-Science?

  "We need some new clichés" - *S. Goldwyn*
- What is the optimal R&D program for the new scientific methodology?

  And the right hardware/software balance?
- How do we get some genuine inter-disciplinary collaborations going?
- How do we optimize the public-private, academia-gov't-industry partnerships?
- How do we accelerate the community participation?
- How do we learn and teach computational thinking?