Ay/Bi 199b: Methods of Computational Science

(aka "e-Science 101")

S. George Djorgovski,
Mary Kennedy,
and friends

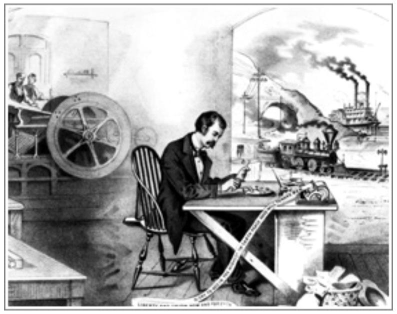
Spring 2011

Class Logistics

- All pertinent info will be on a website:
  http://www.astro.caltech.edu/~george/aybi199/
- Class meets in 100 Powell-Booth, Tue & Thu 1 - 2:30 pm
  - Labs in the cyberspace, on your own time
- Graded P/F only
  - P = If you attend, pay attention, and at least try the labs
- Lectures by subject experts, not professors
- No textbook
  - We will provide useful readings and links
- Open to all members of the Caltech community

Our Motivation and Purpose

- Computation is now an essential component of the scientific method and research practice, and growing in importance and ubiquity
- There are many skills needed by any researcher, going far beyond the usual numerical methods and programming
  - E.g., data-driven computing, software systems, etc., etc.
  - We do not yet teach these skills at Caltech in any formal way
- Our goal is to provide a practical knowledge of these emerging and useful skills and areas
  - See the schedule of lectures for the topics covered
  - In a single term, we can at best provide a quick overview of these subjects, but we hope to start you on your own learning path for further study and exploration



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

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…

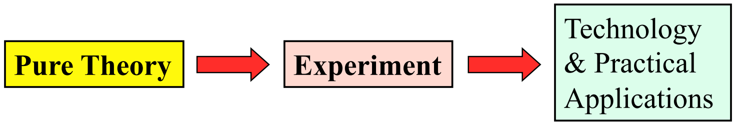
It is a matter of developing a new methodology of science and scholarship for the 21st century

The Evolving Role of Computation

- Computation is no longer just a subsidiary (inferior?) part of the scientific method; it is a ***necessary and increasingly dominant component***
  - Understanding of complex phenomena requires complex data
  - The inevitability of non-analytical theory
- From number crunching to information manipulation
  - The rise of data-driven science
- ***All science*** in the 21st century is becoming e-Science, and with this change comes the need for ***a new scientific methodology***, with common challenges:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data ➔ new knowledge
- There is ***a great emerging synergy*** of the computationally enabled science, and the science-driven IT

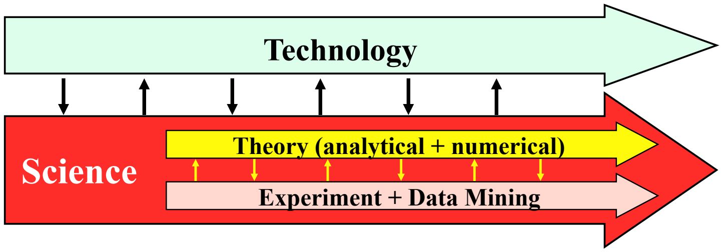
Scientific and Technological Progress

A traditional, "Platonistic" view:



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

A more modern and realistic view:



Technology

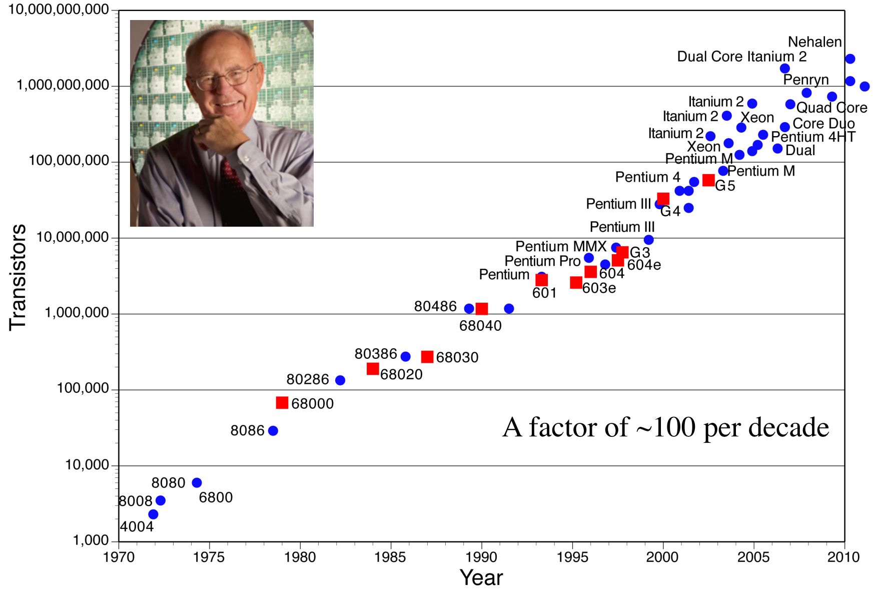
Science

Theory (analytical + numerical)

Experiment + Data Mining

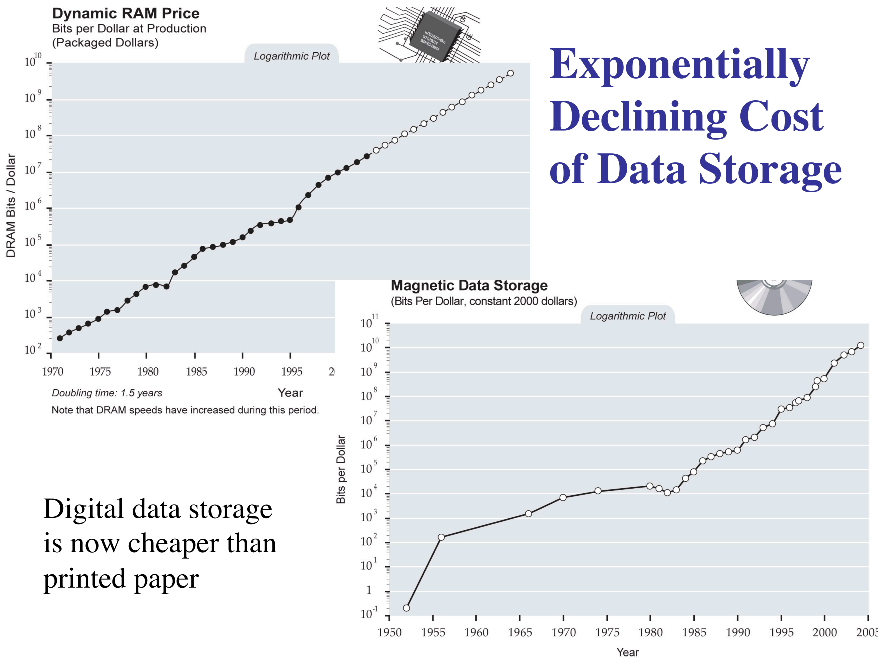
This synergy is stronger than ever and growing; it is greatly enhanced by the IT/computation

Moore's Law



A factor of ~100 per decade

Exponentially Declining Cost of Data Storage


Dynamic RAM Price

Bits per Dollar at Production (Packaged Dollars)

Doubling time: 1.5 years

Note that DRAM speeds have increased during this period.

Magnetic Data Storage

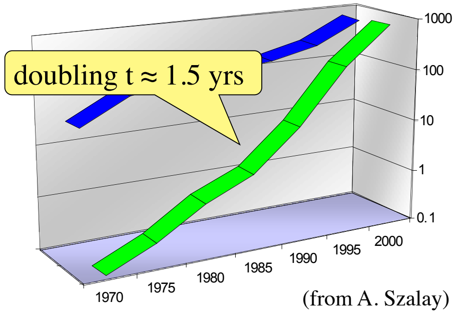
(Bits Per Dollar, constant 2000 dollars)

Digital data storage is now cheaper than printed paper

Facing the Data Tsunami

- All sciences, and other modern fields of human endeavor (commerce, security, etc.) are facing ***a dramatic increase in the volume and complexity of data***
  - For example, in astronomy, large digital sky surveys are the dominant source of data: ~ 10-100 TB/survey (soon PB), ~ 10^6 - 10^9 sources/survey, ~ 10^3 attributes/source, many wavelengths…
- But the real challenge is the growth of ***data complexity***
- The exponential growth of data volume (and quality) driven by the exponential growth in detector and computing technology

Data → Knowledge ?



(from A. Szalay)

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.

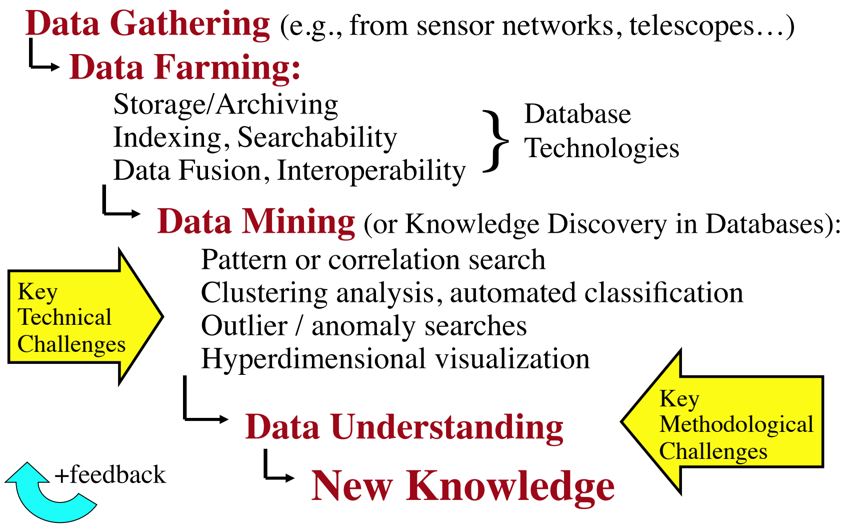
- Information complexity is also increasing greatly

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
- Yet, most scientists are very poorly equipped to do the 21st century, computationally enabled, data-rich science…

A Modern Scientific Discovery Process



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

↳ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability
} Database Technologies

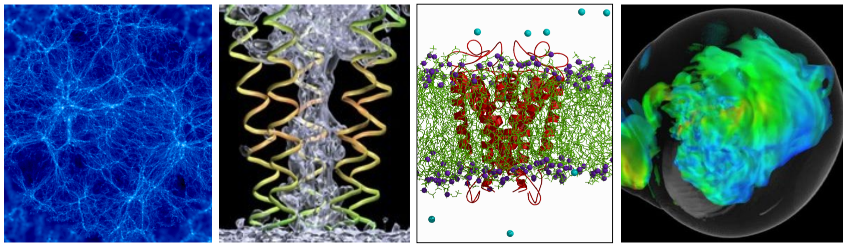
↳ **Data Mining** (or Knowledge Discovery in Databases):

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

↳ **Data Understanding**

↳ **New Knowledge**

Theoretical Simulations Are Also Becoming More Complex and Generate TB's of Data



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

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

Comparing the massive, complex output of such simulations to equally massive and complex data sets is a non-trivial problem!

A Computational Thinking Approach to Understanding of the Universe

- Many (most? all?) complex systems a priori cannot be described analytically, but only computationally
- What does it mean if a theory is not analytical, but expressed as an algorithm, or a computation?
  - It has to be analytical at some "atomic" level (?)
  - Even if we manage to reproduce numerically the behavior of some natural system, does that mean that we understand it?
- Is there a similar qualitative shift in the experimental domain, the empirical basis of science?
  - Is an increase in data volumes by many orders of magnitude a qualitative or just a quantitative change?
  - Or is it a matter of some "critical complexity" of data?

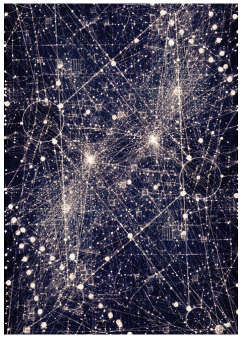
The Key Challenge: Data Complexity

Or: The Curse of Hyper-Dimensionality

1. Data mining algorithms scale very poorly:

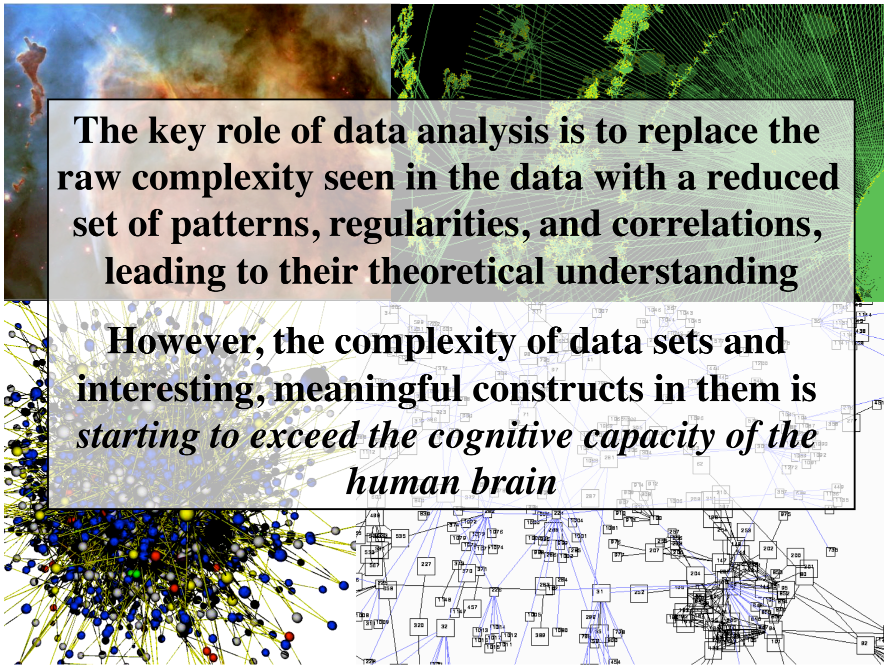
N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

- 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 ($k \geq 1$)
- Better algorithms are the key



2. Visualization in >> 3 dimensions

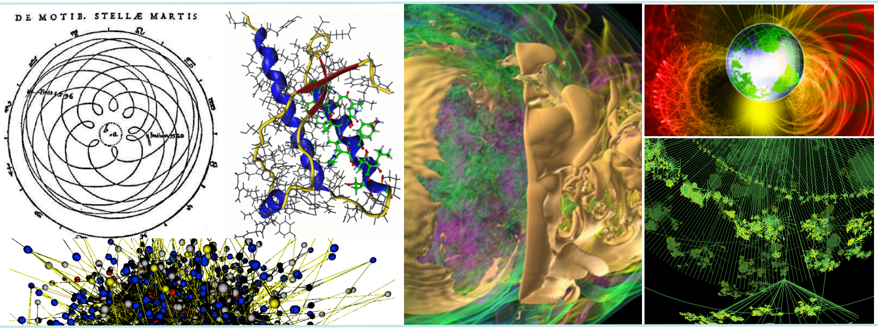
- Data complexity is often representable as a high dimensionality (D >> 3)
  - And yet, we are biologically limited to perceiving D ~ 3 - 10(?)
- Effective visualization has to be a part of the data mining and exploration process

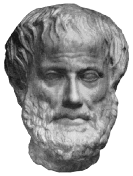


The key role of data analysis is to replace the raw complexity seen in the data with a reduced set of patterns, regularities, and correlations, leading to their theoretical understanding

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

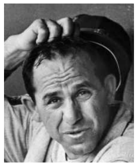
Effective visualization is the bridge between quantitative information and human intuition





Man cannot understand without images

Aristotle, *De Memoria et Reminiscentia*



You can observe a lot just by watching

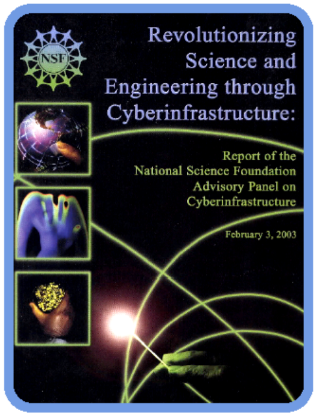
Yogi Berra, an American philosopher

The Cyber-Infrastructure Movement

(or e-Science)

"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 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
- Examples: 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

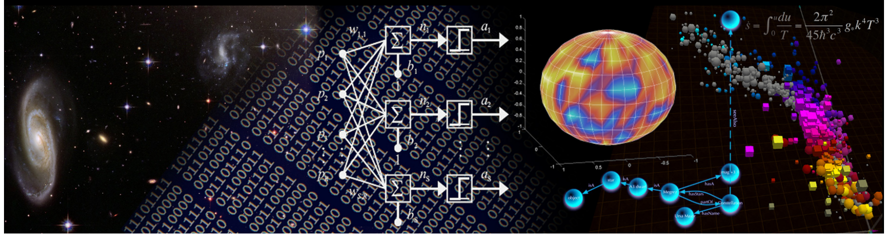
etc., etc.

Beyond Virtual Scientific Organizations:

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

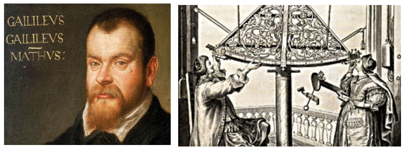
- Domain-specific amalgam fields (science + CS + ICT)
- A mechanism for a broader community inclusion (both as contributors and as consumers)
- A mechanism for interdisciplinary e-Science methodological sharing

➡ **Astroinformatics**

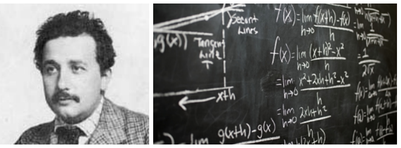


The Evolving Paths to Knowledge

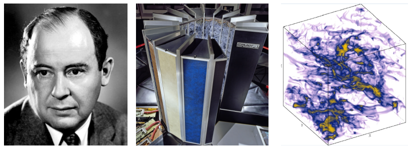
- The First Paradigm: Experiment/Measurement



- 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

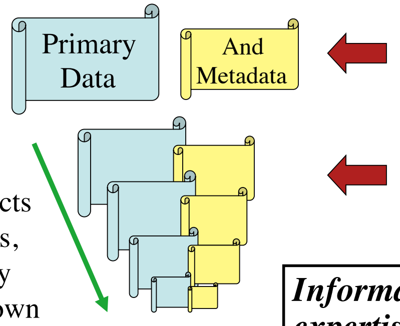
- Data fusion reveals new knowledge which was implicitly present, but not recognizable in the individual data sets

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

Data < Actual data (preserved) / Virtual data (recomputed as needed)



Produced and often archived by the primary data providers

Derived Data Products And Results, Increasingly Distilled down

Produced and published by the domain experts

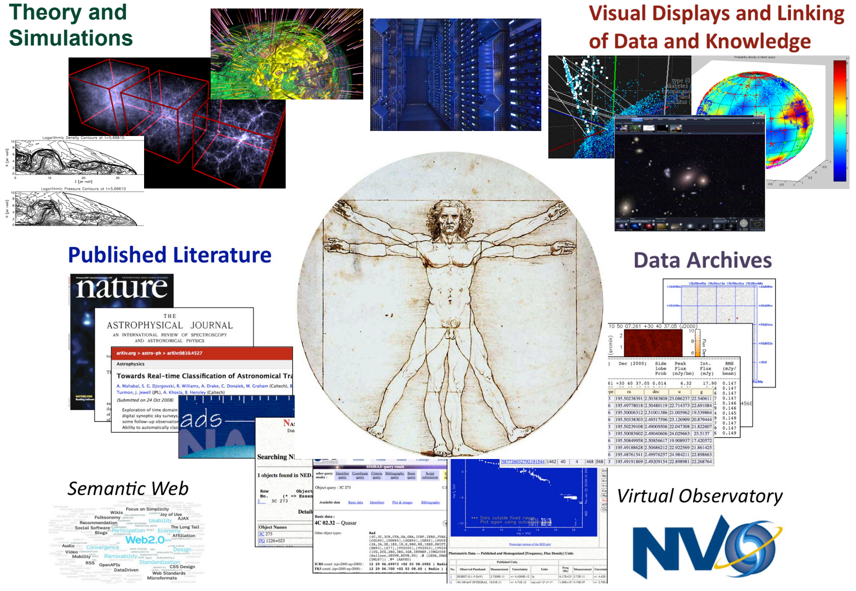
Information is cheap, but expertise and knowledge are expensive!

The Changing Nature of Scientific Publishing

Information and Knowledge Management Challenges

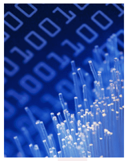
- Increasing complexity and diversity of scientific data and results
  - Data, metadata, virtual data, blogs, wikis, multimedia…
  - From static to dynamic: evolving and growing data sets
  - ***From print-oriented to web-oriented***
- Institutional, cultural, and technical challenges:
  - Massive data sets can be only published as electronic archives, and should be curated by domain experts
  - Peer review/quality control for data and algorithms?
  - A low-cost of web publishing (*samizdat*)
  - Persistency and integrity of data and pointers
  - Interoperability and metadata standards
- The changing roles of scholarly libraries

Science in Cyberspace

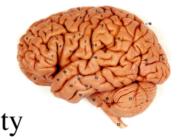


Theory and Simulations
Visual Displays and Linking of Data and Knowledge
Published Literature
nature
THE ASTROPHYSICAL JOURNAL
Towards Real-time Classification of Astronomical Transients
ads
Data Archives
Semantic Web
Web2.0
Virtual Observatory
NVO

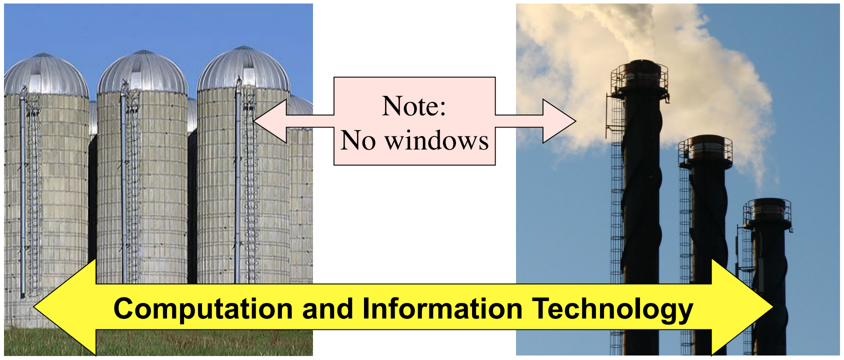
Some Thoughts About e-Science



- Computat*ional* 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

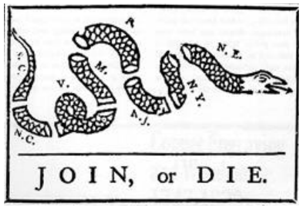


The Structure of Academia / Science



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

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; persistence; 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 …

Many New Things To Learn:

- As science evolves, so does its methodology
  – IT revolution drives a scientific revolution, through an exponential growth of data volumes and complexity, but also the computational power needed to tackle them
- The empirical basis of science
  – Databases, data grids, data mining…
- The role of simulation, inherently non-analytical theory
  – Numerical tools, software systems, code validation, …
- Discovery process
  – Data mining, visualization, data/theory matching…
- Scientific communication, collaboration, publishing
- The hardware tools
  – Massively parallel systems, grids, networks, …