Towards a New Scientific Methodology

S. George Djorgovski *(Caltech)*

Panel Discussion on

"What do Scientists Really Need to Facilitate Time to Discovery?"

Microsoft eScience Workshop, Indianapolis, Dec. 2008

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

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):

Key Technical Challenges

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

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

The Key Challenge: Data Complexity

Or: The Curse of Hyper-Dimensionality

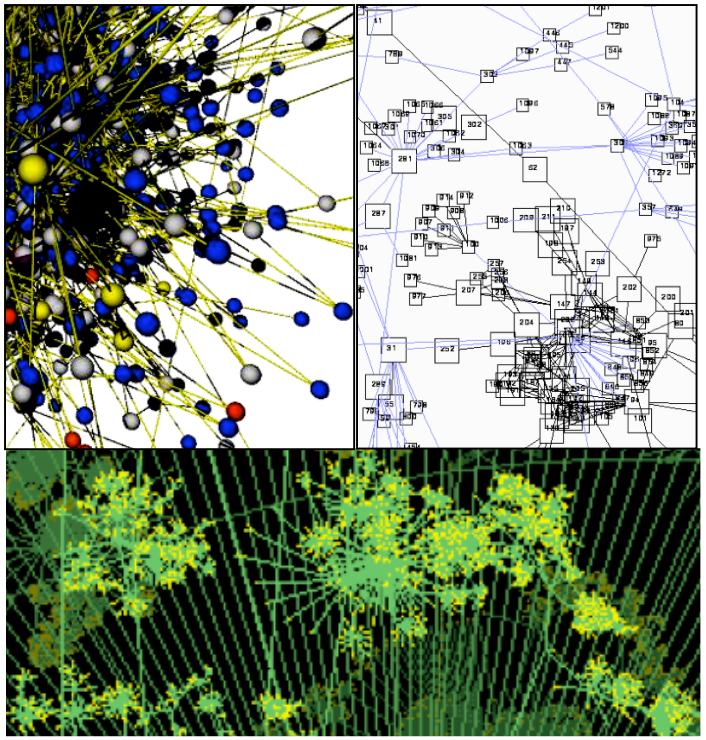
1. **Data mining algorithms scale very poorly:**

   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 ($k \geq 1$)

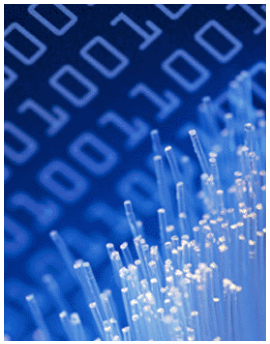
2. **Visualization in >> 3 dimensions**

   - 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

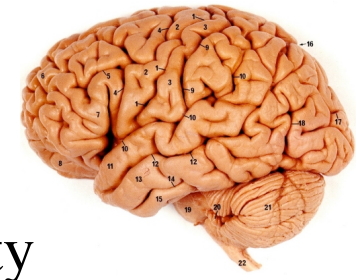


Some Thoughts About e-Science

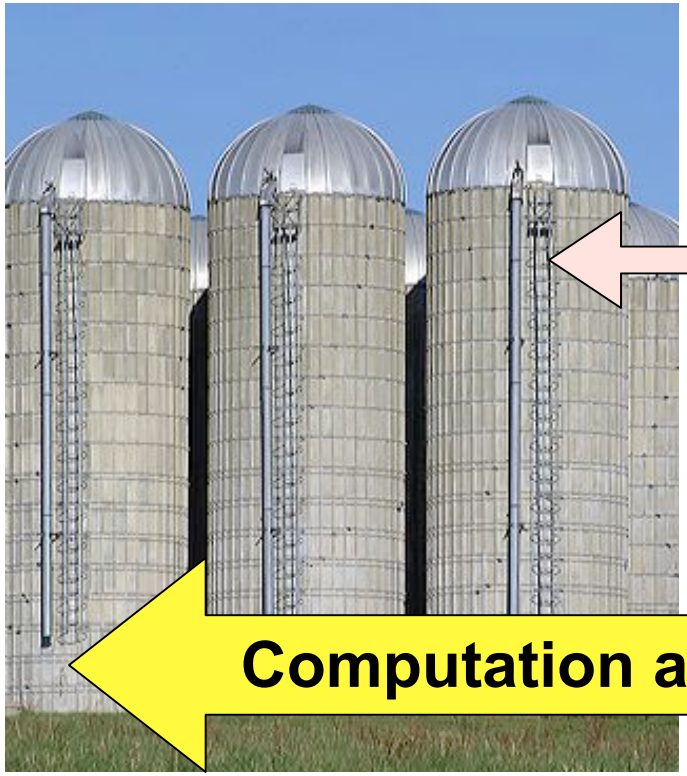
- Comput***ational*** science ≠ Comput***er*** science
- Computational science { Numerical modeling ↓ Data-driven science



- 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

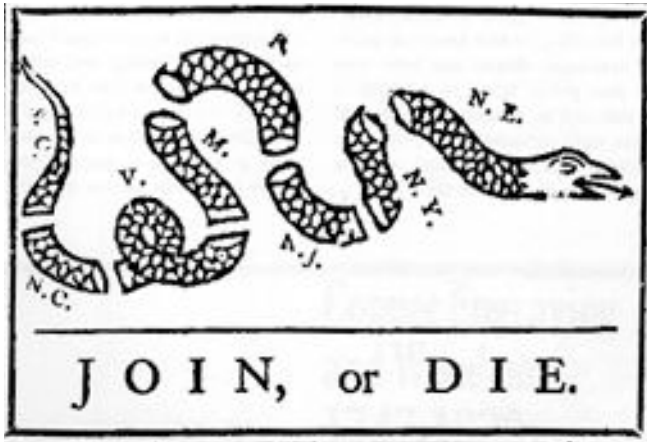


The Structure of Academia / Science



Note:
No windows

Computation and Information Technology



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

-- *Ben Franklin*

e-Science is unified by a common methodology and tools

The Key Computational Science Needs

- Better scalable algorithms for data mining and knowledge discovery in large and complex data sets
  - Including a more extensive use of AI/ML tools
- Hyperdimensional visualization tools and methods
  - A key bridge to human intuition and understanding
- The art and science of scientific software systems
  - Architecture, design, implementation, validation …
- Effective virtual forums and marketplaces of ideas, expertise
- Teaching scientists and their students how to use these tools (and to think computationally)
- These methodologies are:
  - ***Necessary*** - all sciences (and the economy, national security, etc.) are becoming intensely computational and exponentially data-rich and complex
  - ***Shareable*** between all fields of science (and beyond)