Cyberinfrastructure and scholarly publishing futures

KAN Min-Yen

(credits due to “The Fourth Paradigm”)

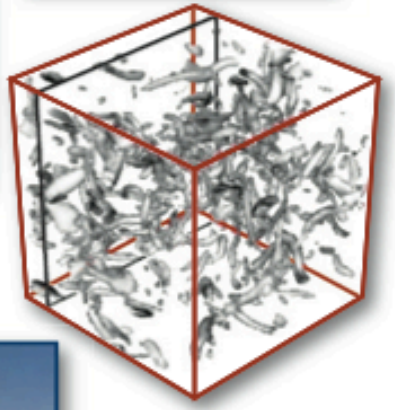
Current Challenges in eSciences

- Data Capture
  - Generating more data than is transferrable: do query and computation at instrument site
- Data Curation
  - Crowdsourcing: putting up raw data for others to discover / tour / curate / teach; citizen science
- (Automated) Data Analysis
  - Need linear scale algorithms; $O(n^2)$ doesn't work for large (PB/EB) datasets
  - Support and make actionable policy; increase information velocity towards Singularity for all spectrum of users

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



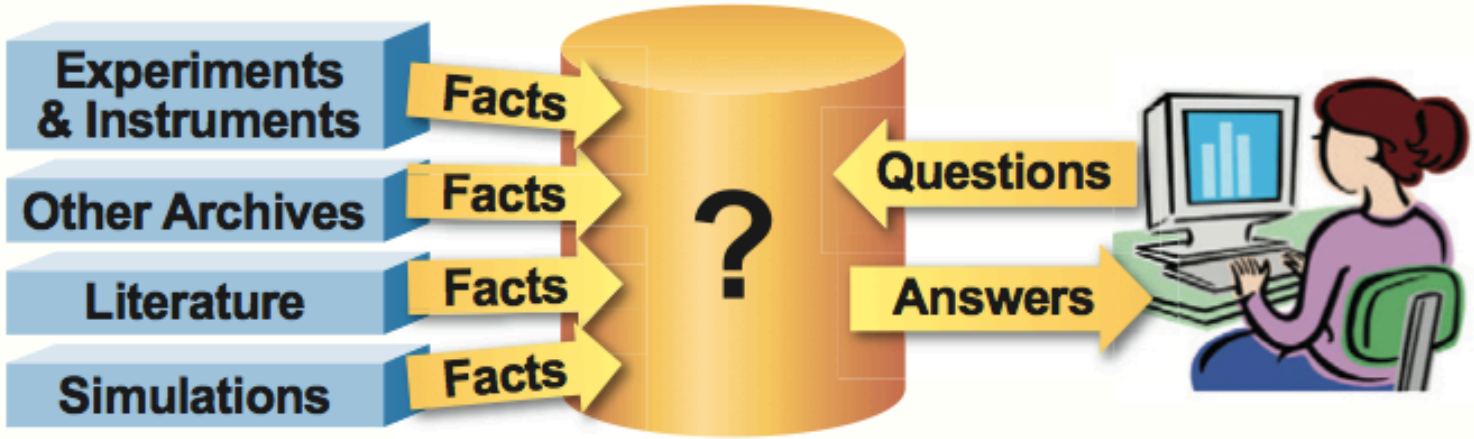
Third Leg: Computational Science

- Experimental Science
- Theoretical Science
- Computational (Simulation) Science

- Don’t look at the object by human eye but through instruments generating lots of data.

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



The Generic Problems

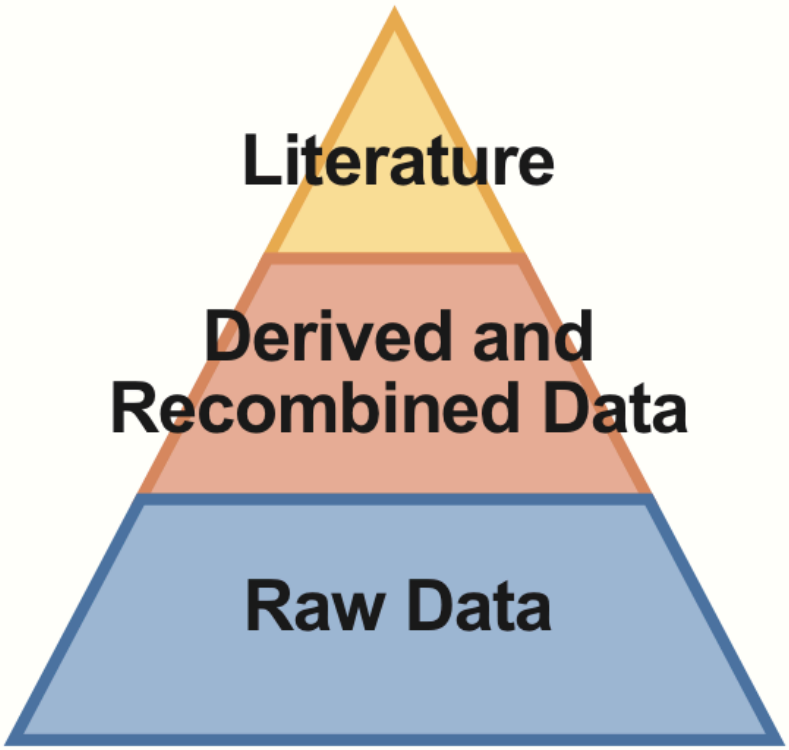
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share it with others
- Query and Vis tools
- Building and executing models
- Integrating data and literature
- Documenting experiments
- Curation and long-term preservation

4th Paradigm: Data-intensive Science

- All projects generate lots of data/files
  - Two goals: outlier detection, clustering
  - Outlier Detection Example: Higgs Boson
- Large Tier 1 projects have software budgets but not many Tier 3 projects?
  - Use COTS, but not really enough

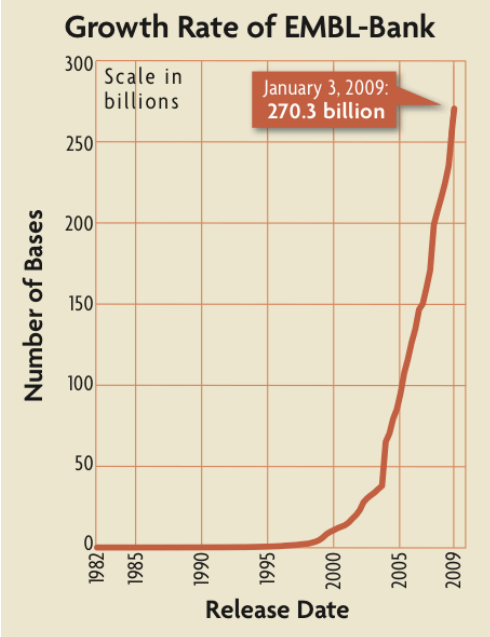
All Scientific Data Online

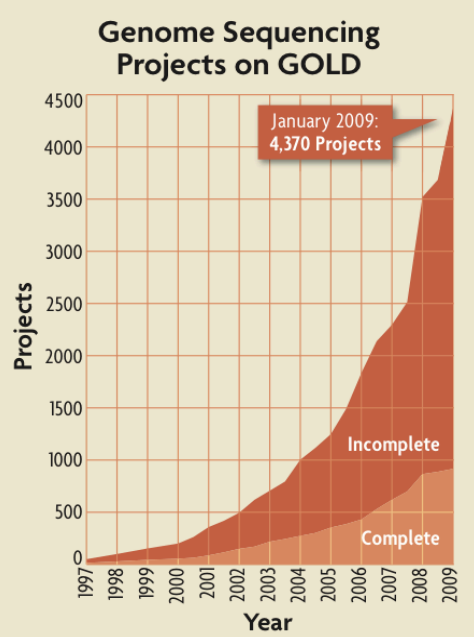
- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere



- Increase Scientific Information Velocity
- Huge increase in Science Productivity

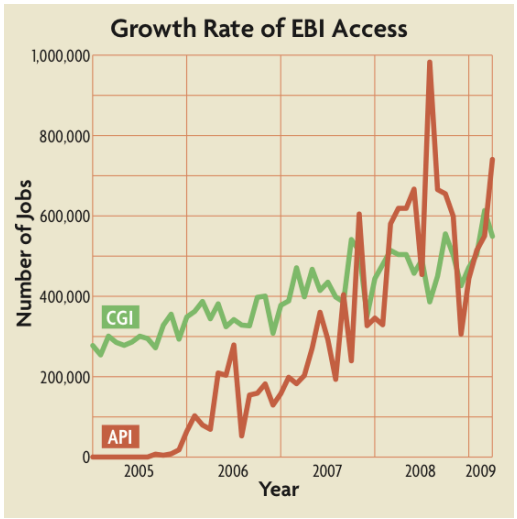
Growth in Genome Sequencing





Data capture growing exponentially
Growth rate partially limited by writing throughput

But data access growing only linearly



Ideas to address data bottleneck

- Look back to previous computing era
  - Shared nothing networks of PCs connected by Ethernet (70's–80's)
- **Now**: Multicore CPUs and general purpose GPU (GPGPU)
  - Some data shared
- **Now**: Cloud computing
  - Data segmented, tasks done map-reduce style
  - Access to mobile phone (computational and network appliance) may be easier than material resources

SCIENTIFIC COMMUNICATION

Full text to be opened, will be the norm

But make the whole chain of data also accessible

- Overlay journals: linking data to the article
- Use a mix of peer review and other means to assess article quality

Objectify knowledge

- To be used to link data in semantic web style
- Schematizing objects

"What better contribution could a scholar make than an article which could … provide a clear, but vivid argument to the [secondary school student] but which, if unraveled, could provide the rigor demanded by the most crusty specialist?" Gregory Crane (of the Perseus DL)

Producer and consumer divide

Scientists are (unwittingly) the bottleneck to the advancement of science

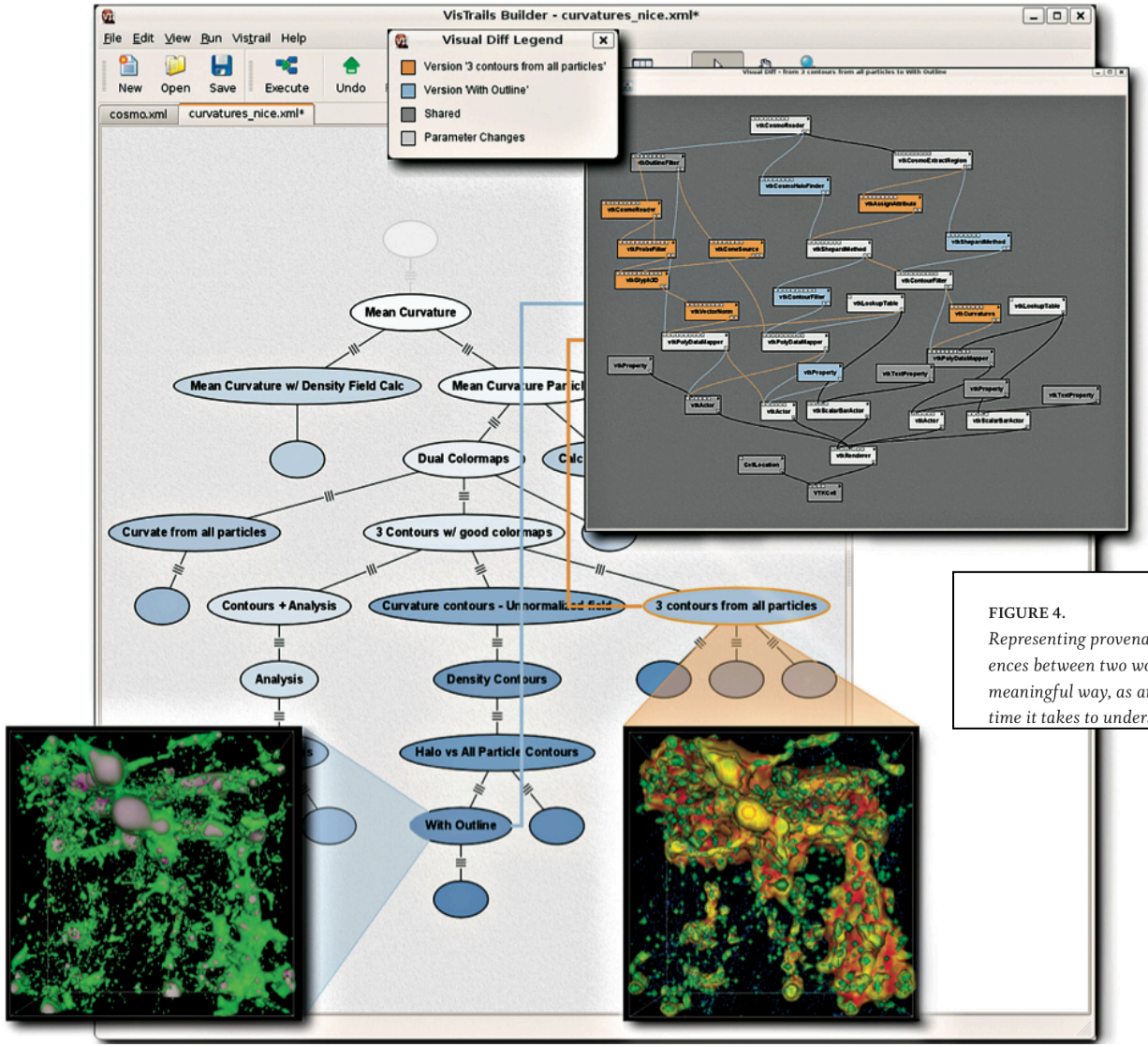
- “How do I use this data that I did not generate?”
- “How to use this data type, new to me, with the data I use every day?”
- “What to do if I need data from another discipline but cannot understand its terms?”
- “Do these data represent the same thing, at the same vertical / geographic position or at the same time, and does that matter?”
- “I found this particular species in an unexpected location. What are the geophysical parameters (e.g.,temperature, humidity) for this area, and how has it changed over the last weeks, months, years?”

etc.

Leads to the needs of semantics in describing data

A current bottleneck in data-intensive science

Visualization



Workflows: need to record provenance so that basic questions can be answered

FIGURE 4.
Representing provenance as a series of actions that modify a pipeline makes visualizing the differences between two workflows possible. The difference between two workflows is represented in a meaningful way, as an aggregation of the two. This is both informative and intuitive, reducing the time it takes to understand how two workflows are functionally different.

Visualization

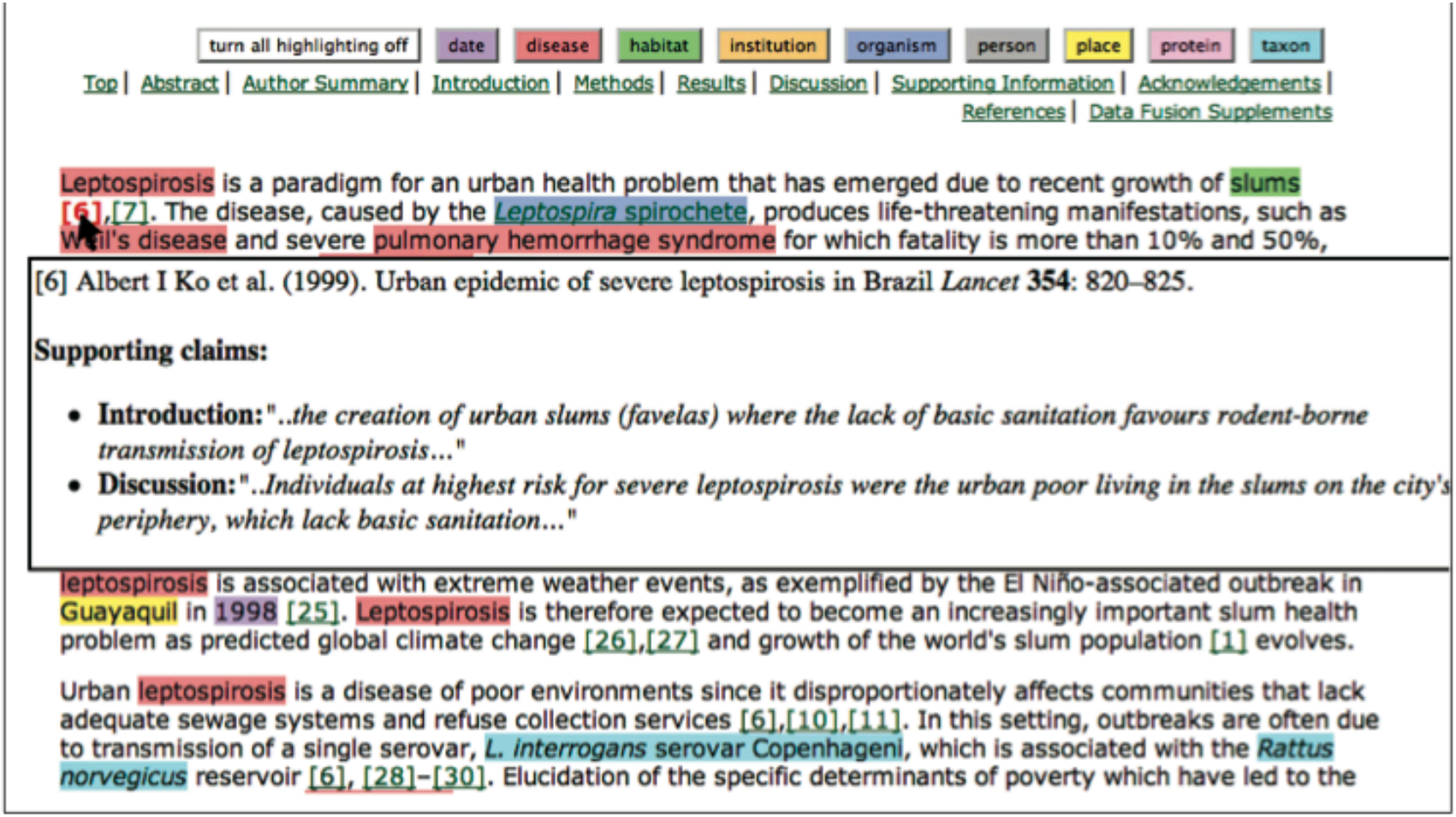
Tours in World Wide Telescope: Allow experts and citizen science to annotate, record interesting objects

To *produce* information suitable for others from either curated or raw data

Scholarly Artifacts

- Grey Literature
  - Used to be: tech reports, unpublished manuscripts (theses)
  - Now: tweets, blogs, news articles and other **ephemera**
- Data and software as first class objects
  - But with massive scale, problems with versioning, fixity

Enrichment of the text itself



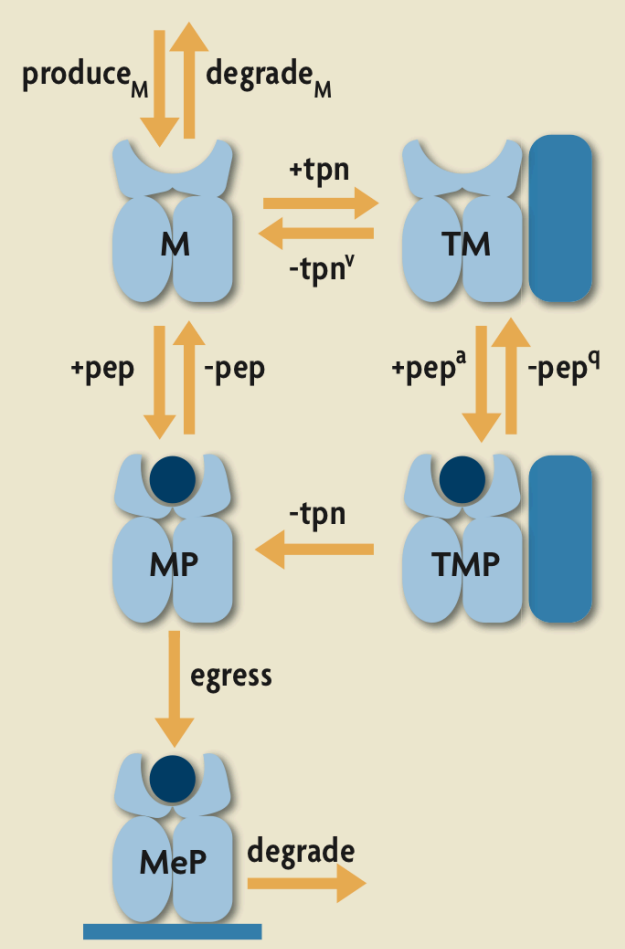
Objectified Knowledge

- Graph (for Bayesian) or schema (for structured records) based objects in future
- Knowledge Representation: "Formalized" into a narrative form for human reading; but in a form suitable for parsing, leading to **actionable data**

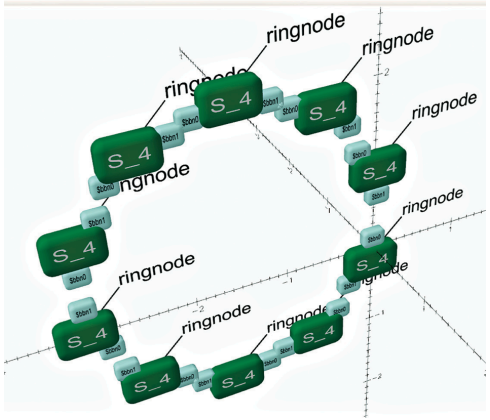
Leads to the (automated) treatment of scholarly communication ***in the large***

- Standards exist today, but not widely adopted
- Barriers to widespread adoption is political and application-specific
  - Jim Gray's "May all of your problems be technical"
- Need to incorporate credit of such sources and annotations in a citation-like manner (needs to count).

An example from biology



This diagram is in 1:1 correspondence with formal stochastic pi-calculus models. One can edit either the diagrams or the models. The nodes represent molecular states, and the labeled arcs represent interactions with environment.



A visualization of a model for use in computer simulation or execution

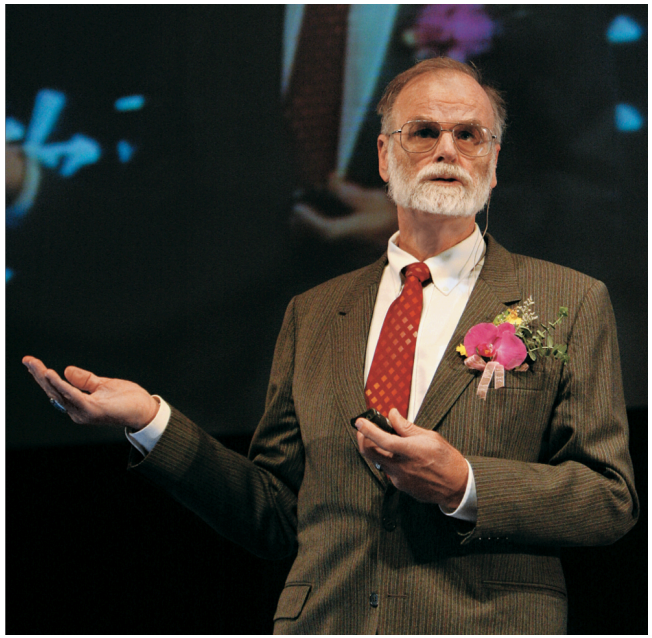
Team Science

Data intensive science requires teams, not individuals

- Rare to find individuals discovering
- "More is different": data-driven changes the nature of science and its advancement
- Teams of collaborators, working as equals

What is the way forward?

Jim's Thoughts



- Fund both the development of software tools and support
- Invest in tools at all levels of the funding pyramid.
- Develop generic Laboratory Information Management Systems (LIMS).
- More research on scientific data management, data analysis, data visualization, and new algorithms and tools.
- Create digital libraries that support other sciences.
- Develop new document authoring tools and publication models.
- Develop digital data libraries that contain scientific data (not just the metadata) and support integration with published literature.