Microsoft Research

Microsoft Research Asia
Faculty Summit 2010

Fourth Paradigm - Exploring Trends and Talents for Data-Intensive Science

Session Chair: Tony Hey
Corporate Vice President – External Research
Microsoft Corporation

Panel Members

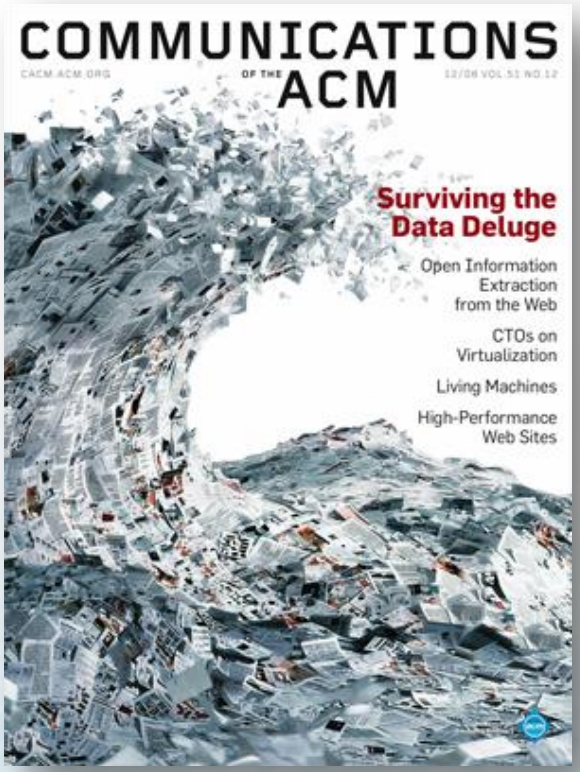
Jinpeng HUAI
Beihang University

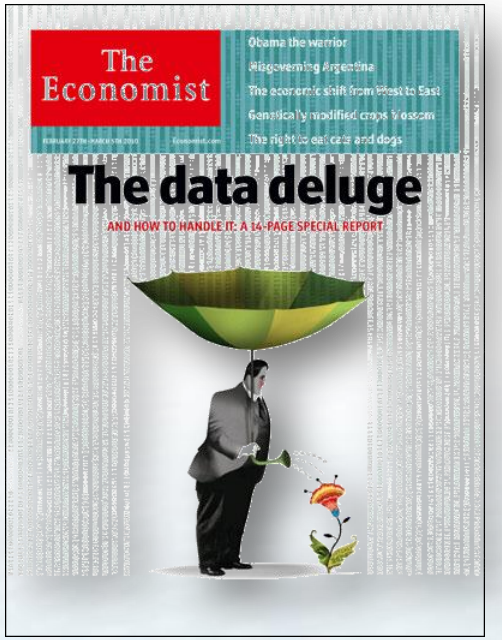
Jimmy LIU
Agency for Science, Technology, and Research, Singapore (A*STAR)

Key-Sun CHOI
Korea Advanced Institute of Science & Technology (KAIST)

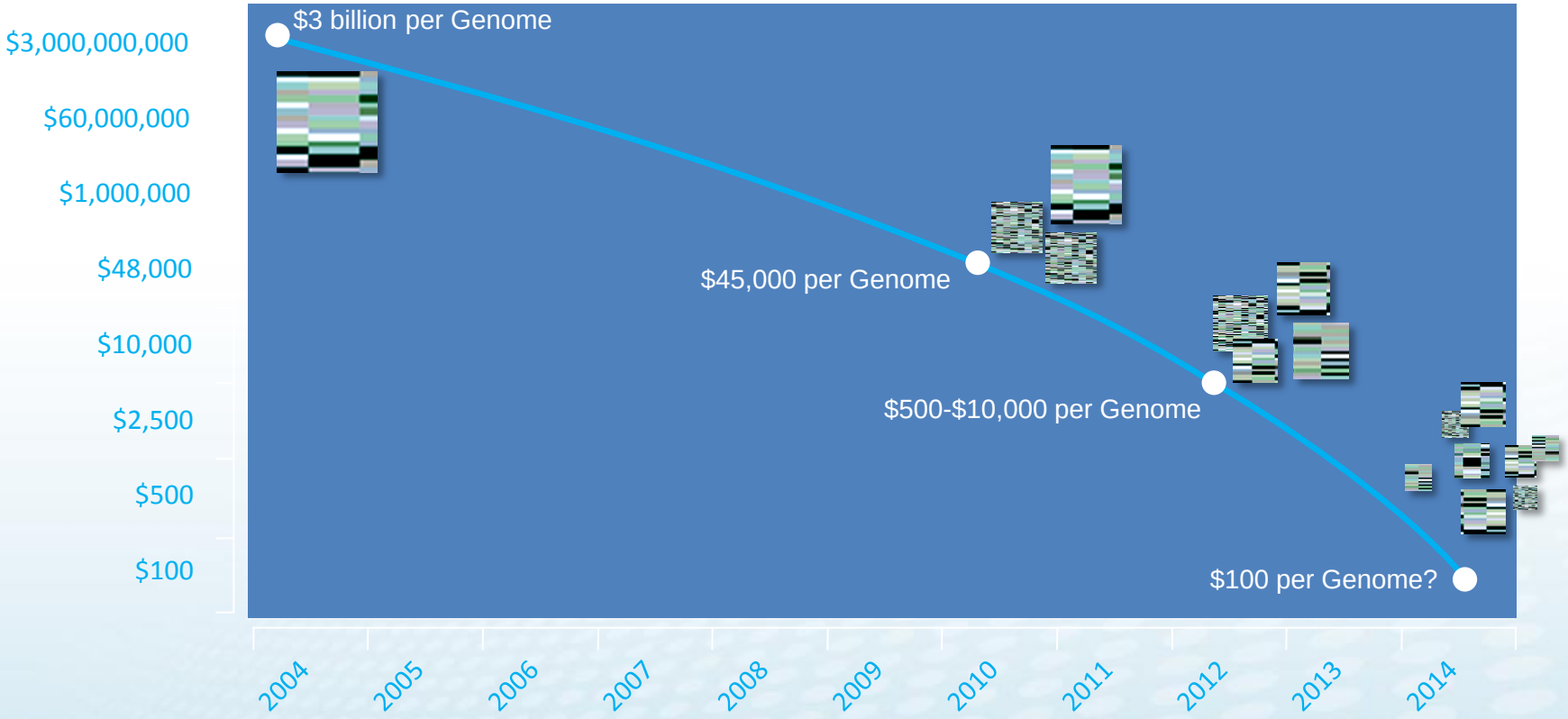
Junichi TSUJII
The University of Tokyo

A Tidal Wave of Scientific Data



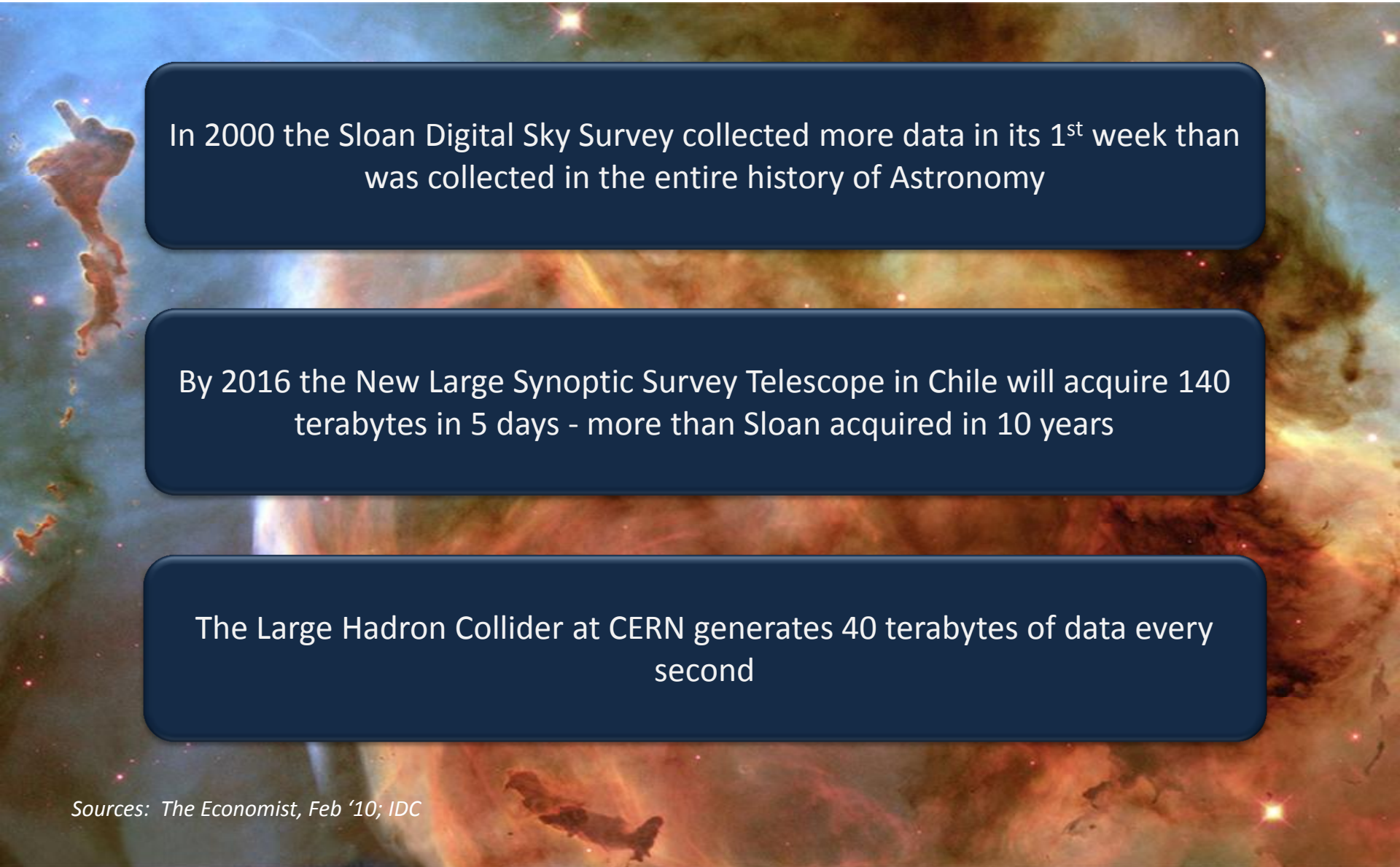


The Gene Sequencing Explosion



Source: George Church, Harvard Medical School, as reported in IEEE Spectrum, Feb '10. Figures represented in USD

Astronomy and Particle Physics



In 2000 the Sloan Digital Sky Survey collected more data in its 1st week than was collected in the entire history of Astronomy

By 2016 the New Large Synoptic Survey Telescope in Chile will acquire 140 terabytes in 5 days - more than Sloan acquired in 10 years

The Large Hadron Collider at CERN generates 40 terabytes of data every second

Sources: The Economist, Feb '10; IDC

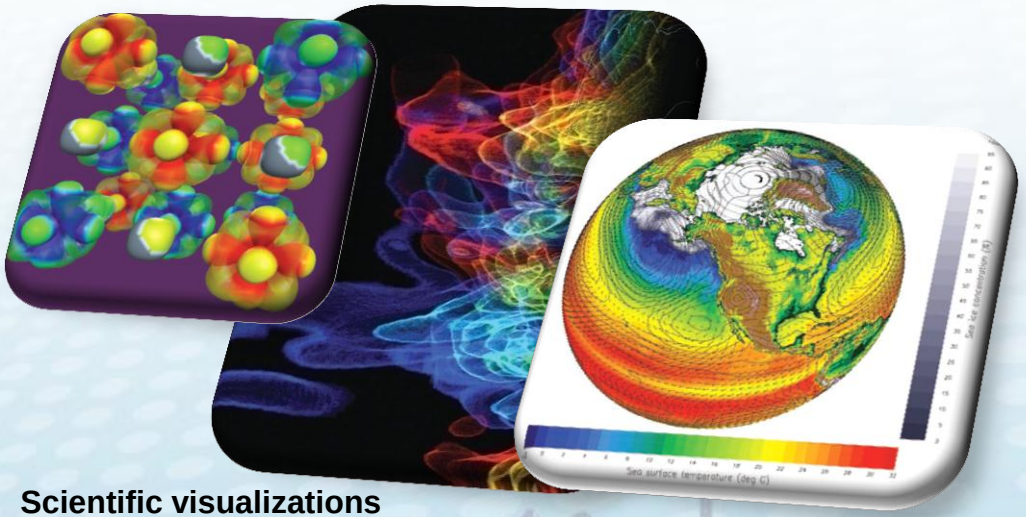
A Digital Data Deluge in Research

- Data collection
  - Sensor networks, satellite surveys, high throughput laboratory instruments, observation devices, supercomputers, LHC …
- Data processing, analysis, visualization
  - Legacy codes, workflows, data mining, indexing, searching, graphics …
- Archiving
  - Digital repositories, libraries, preservation, …

SensorMap

Functionality: Map navigation

Data: sensor-generated temperature, video camera feed, traffic feeds, etc.



Scientific visualizations

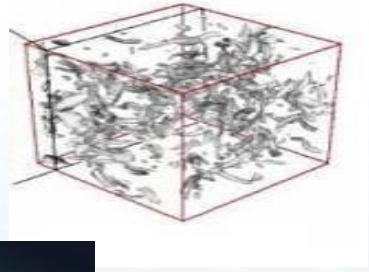
NSF Cyberinfrastructure report, March 2007

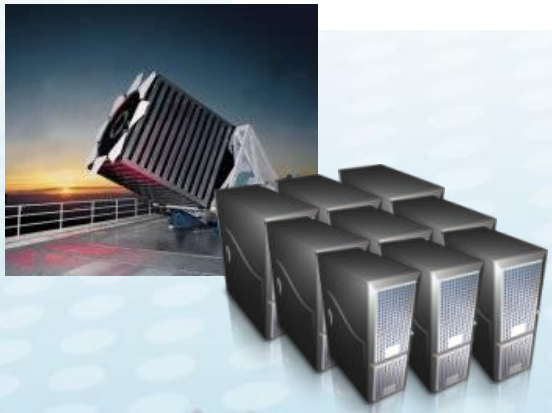
Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks

➢ eScience is the set of tools and technologies to support this data-intensive science
   - For analysis and data mining
   - For data visualization and exploration
   - For scholarly communication and dissemination

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

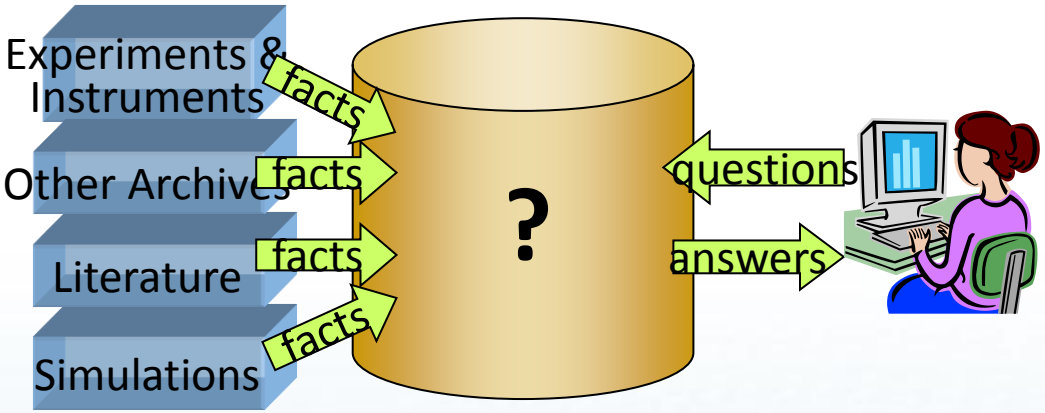




(With thanks to Jim Gray)

X-Info

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

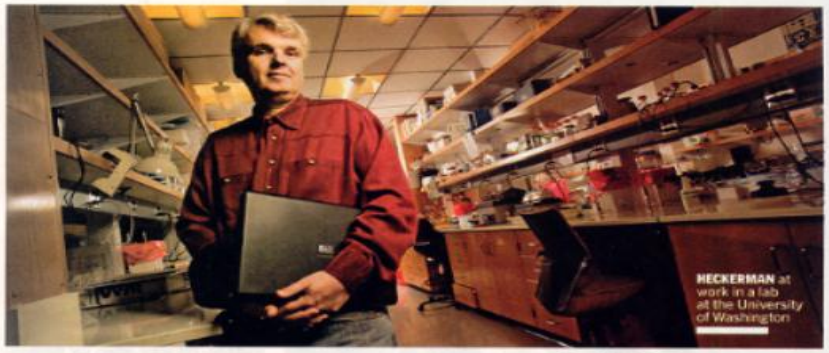


The Generic Problems

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

With thanks to Jim Gray

Bio-informatics: Machine Learning and HIV



HECKERMAN at work in a lab at the University of Washington

Using Spam Blockers To Target HIV, Too

A Microsoft researcher and his team make a surprising new assault on the AIDS epidemic

BY STEPHEN BAKER AND JAY GREENE

CUT-RATE PAINKILLERS! Unclaimed riches in Nigeria!! Most of us quickly identify such e-mail messages as spam. But how would you teach that skill to a machine? David Heckerman needed to know. Early this decade, Heckerman was leading a spam-blocking team at Microsoft Research. To build their tool, team members meticulously mapped out thousands of signals that a message might be junk. An e-mail featuring "Viagra," for example, was a good bet to be spam—but things got complicated in a hurry.

If spammers saw that "Viagra" messages were getting zapped, they switched to Viagra, or Vi agra. It was almost as if spam, like a living thing, were mutating.

This parallel between spam and biology resonated for Heckerman, a physician as well as a PhD in computer science. It didn't take him long to realize that his spam-blocking tool could extend far beyond junk e-mail, into the realm of life science. In 2003, he surprised colleagues in Redmond, Wash., by refocusing the spam-blocking technology on one of the world's deadliest, fastest-mutating conundrums: HIV, the virus that leads to AIDS.

Similar mutations may crop up in computer and medical viruses

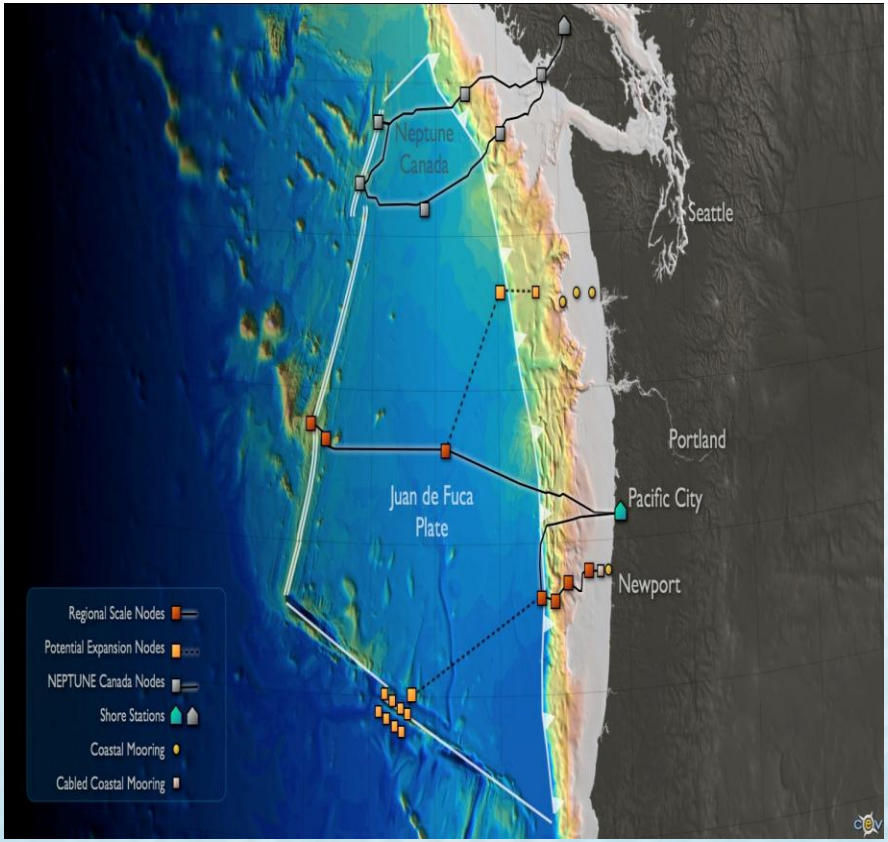
Heckerman was plunging into medicine—and carrying Microsoft with him. When he brought his plan to Bill Gates, the company chairman "got really excited," Heckerman says. Well versed on HIV from his philanthropy work, Gates lined up Heckerman with AIDS researchers at Massachusetts General Hospital, the University of Washington, and elsewhere.

Since then, the 50-year-old Heckerman and two colleagues have created their own biology niche at Microsoft, where they build HIV-detecting software. These are research tools to spot infected cells and correlate the viral mutations with the individual's genetic profile. Heckerman's team runs mountains of data through enormous clusters of 320 computers, operating in parallel. Thanks to smarter algorithms and more powerful machines, they're sifting through the data 480 times faster than a year ago. In June, the team released its first batch of tools for free on the Internet.

A new industry for the behemoth to conquer? Not exactly. Heckerman's nook in Redmond represents just one small node in a global AIDS research effort marked largely by cooperation. "The Microsoft group has a different perspective and a good statistical background," says Bette Korber, an HIV researcher at Los Alamos National Laboratories. The key quarry they all face is the virus itself, which is proving wilier than any of Microsoft's corporate foes. While Heckerman has high hopes that his tools will lead to vaccines that can be tested on humans within three years, his research

Environmental Informatics: Smart Sensors and Data Fusion

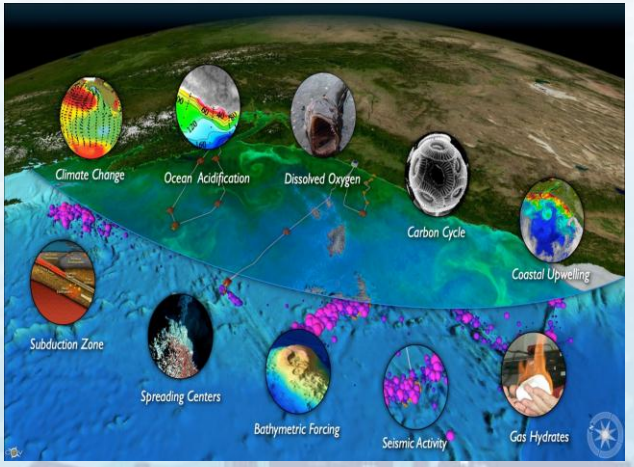
- The NSF Ocean Observatory Initiative
  - Hundreds of cabled sensors and robots exploring the sea floor
  - Data to be collected, curated, mined
  - OOI Architecture plan of record, store this data in the cloud

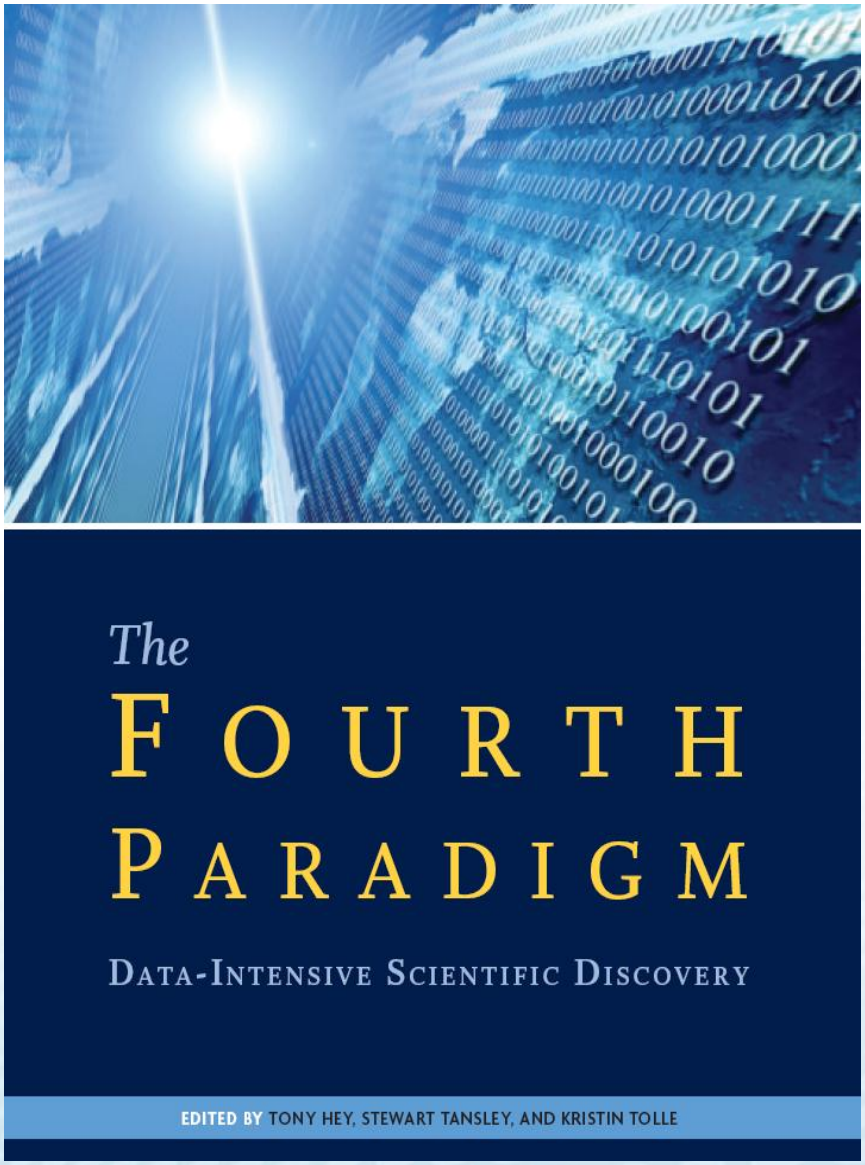


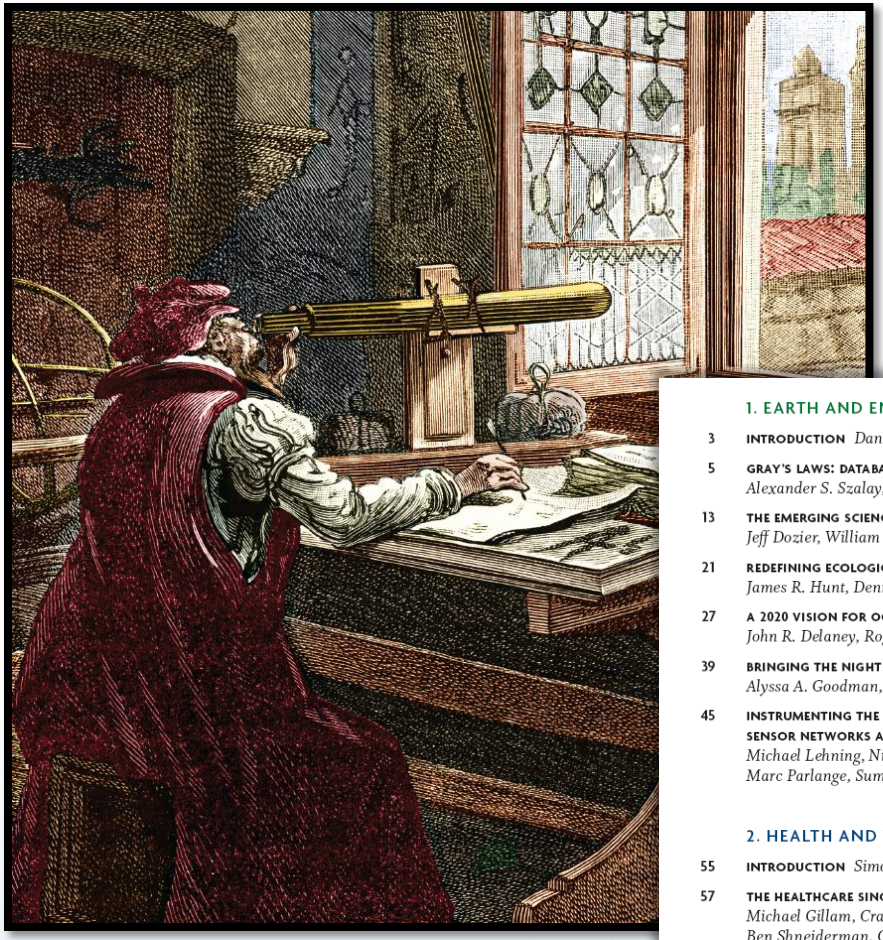
Data collected from:

- Ocean floor sensors, AUV tracks, ship-side cruises, computational models

Data moves from **ocean** to shore side **data center** to the **Azure cloud** to your **computer**.




The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



Microsoft Research Asia
Faculty Summit 2010

An edited collection of 26 short technical essays, divided into 4 sections

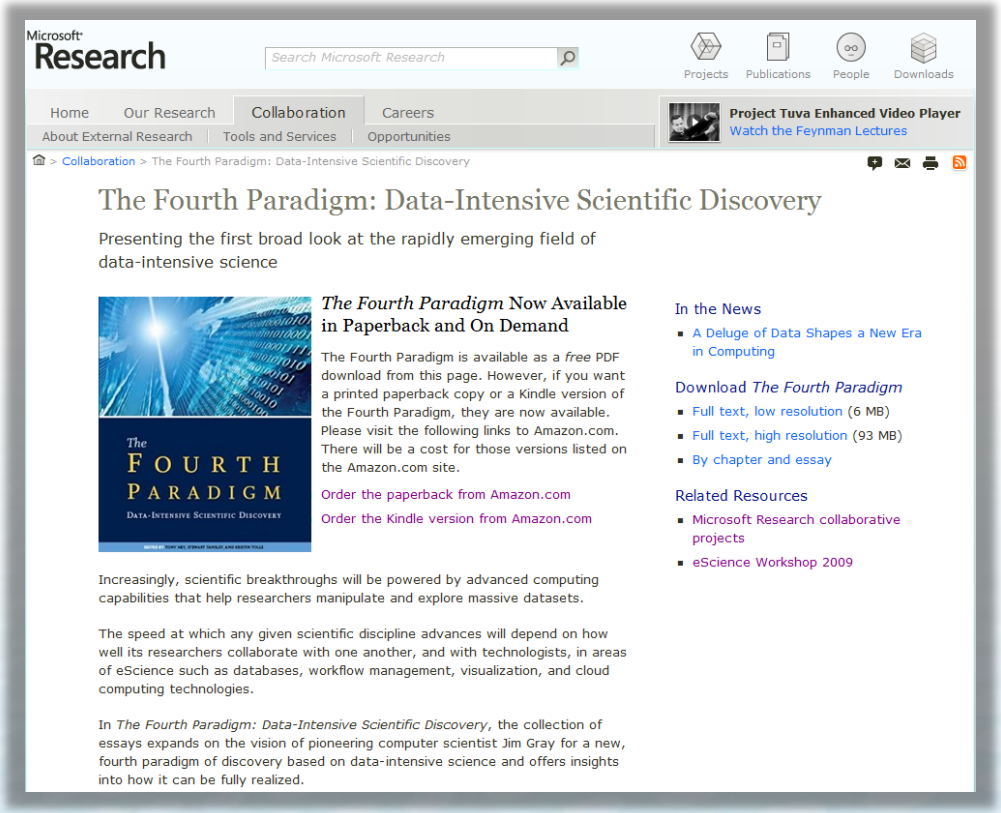
Free PDF Download
Amazon Kindle version; Paperback print on demand

http://research.microsoft.com/fourthparadigm/

- "The impact of Jim Gray's thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."
- — **Bill Gates,** Chairman, Microsoft Corporation

- "One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working."
- — **Douglas Kell**, University of Manchester

- "The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives."
- — **Gordon Bell**, Microsoft Research



Questions for Discussion

- What are the potentially important technologies for applications in data-intensive science?
- What semantic technologies will deliver new tools for the global research community?
- How can the cloud be leveraged for data-intensive science?
- How do we educate a new generation of students and research scientists to have both discipline based skills and knowledge of data technologies?