

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 HUI
Beihang University



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

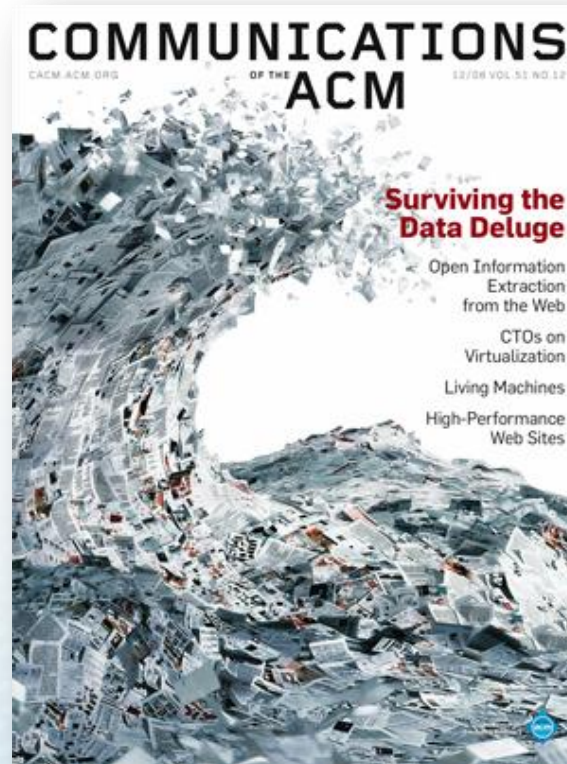


Key-Sun CHOI
Korea Advanced
Institute of Science &
Technology (KAIST)



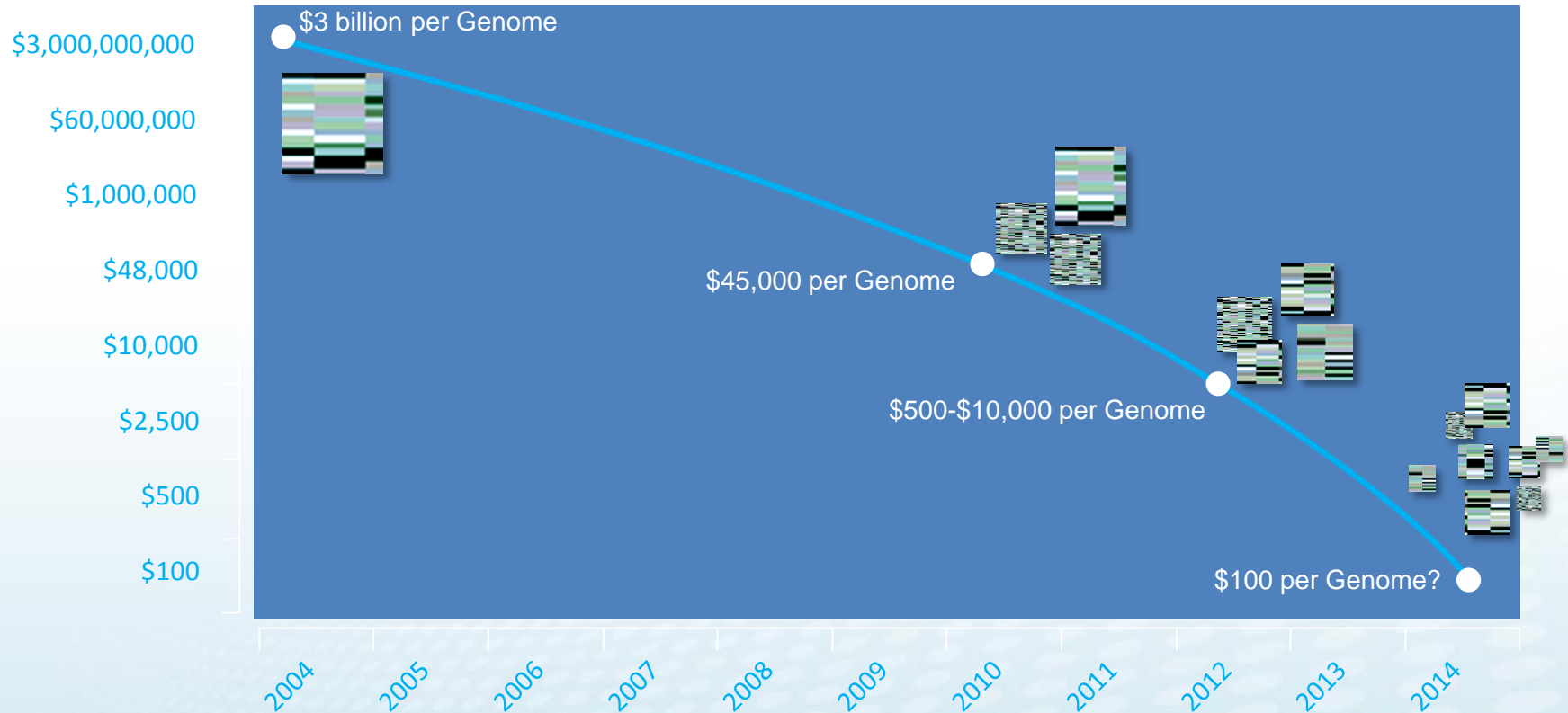
Junichi TSUJII
The University of Tokyo

A Tidal Wave of Scientific Data



This work is licensed under a [Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/).

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

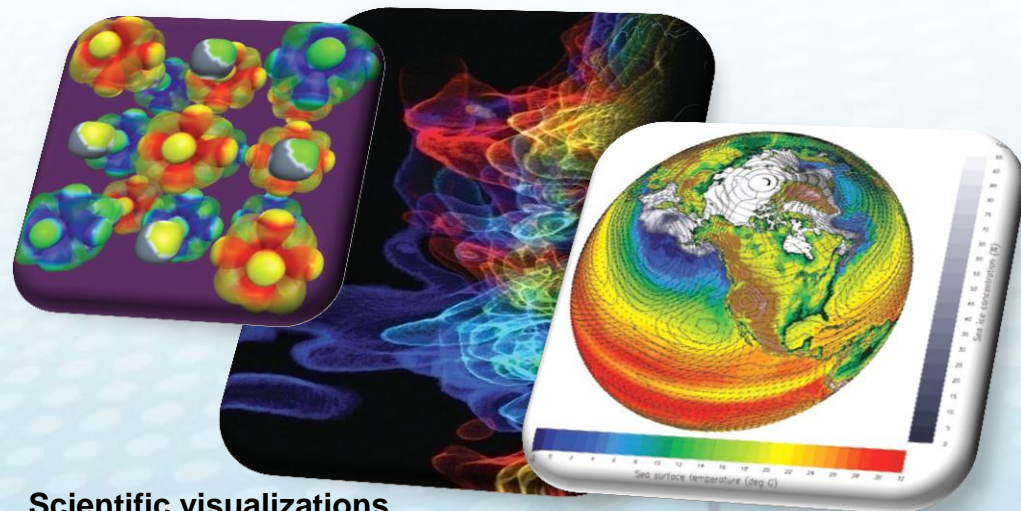
A Digital Data Deluge in Research

- 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.



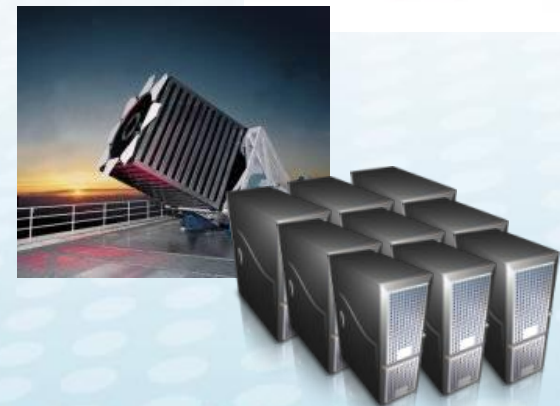
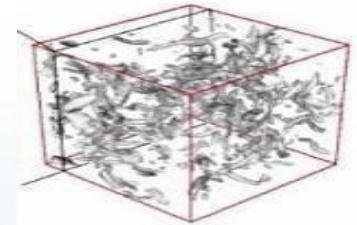
Scientific visualizations

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

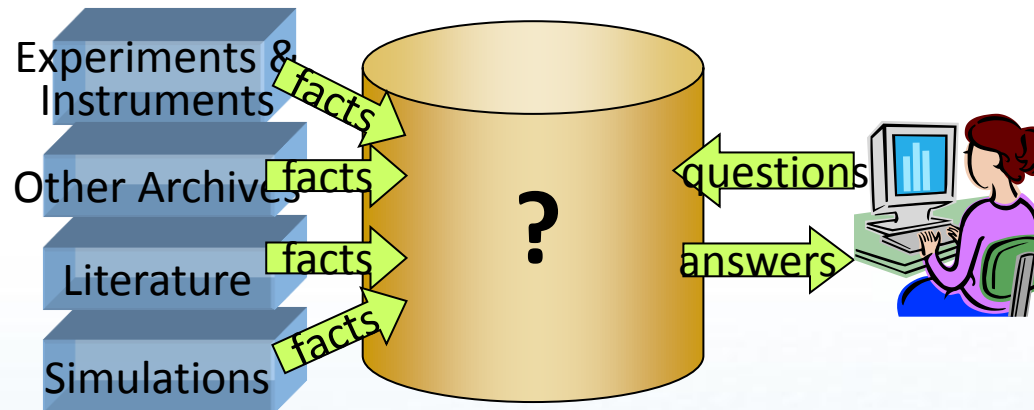
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K \frac{c^2}{a^2}$$



(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

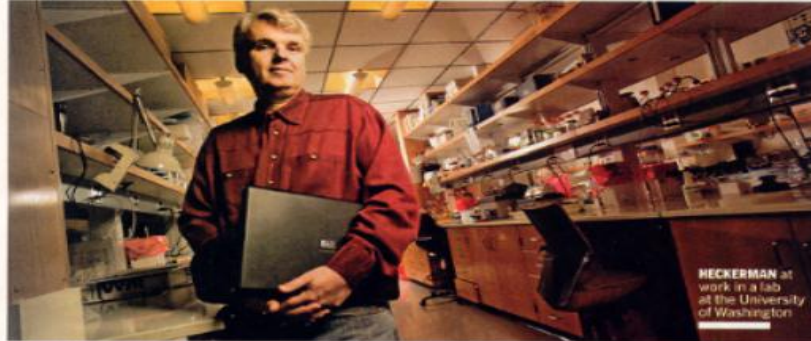
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

With thanks to Jim Gray

Bio-informatics: Machine Learning and HIV



InfoTech | Research



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 Viagra. 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.

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 450 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 congaer? Not exactly. Heckerman's nook

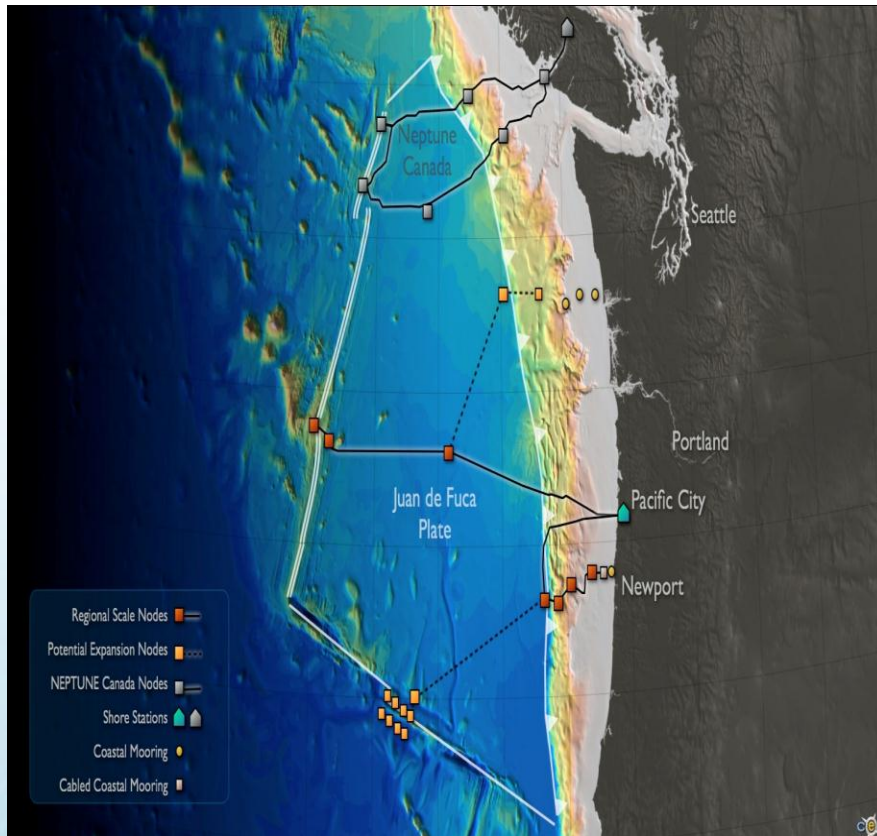
Similar mutations may crop up in computer and medical viruses

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 Koerber, an HIV researcher at Los Alamos National Laboratories. The key query they all face is the virus itself, which is proving wiler 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

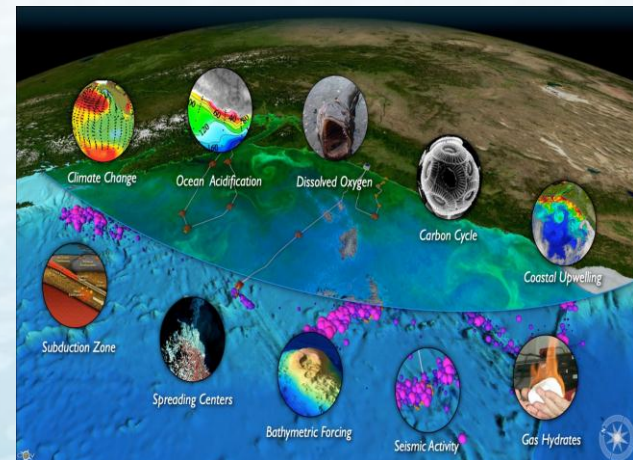
- The NSF Ocean Observatory Initiative
 - Hundreds of cabled sensors and robots exploring the sea floor
 - Data to be collected, curated, mined
 - OOI Architecture plan of record, store this data in the cloud



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**.



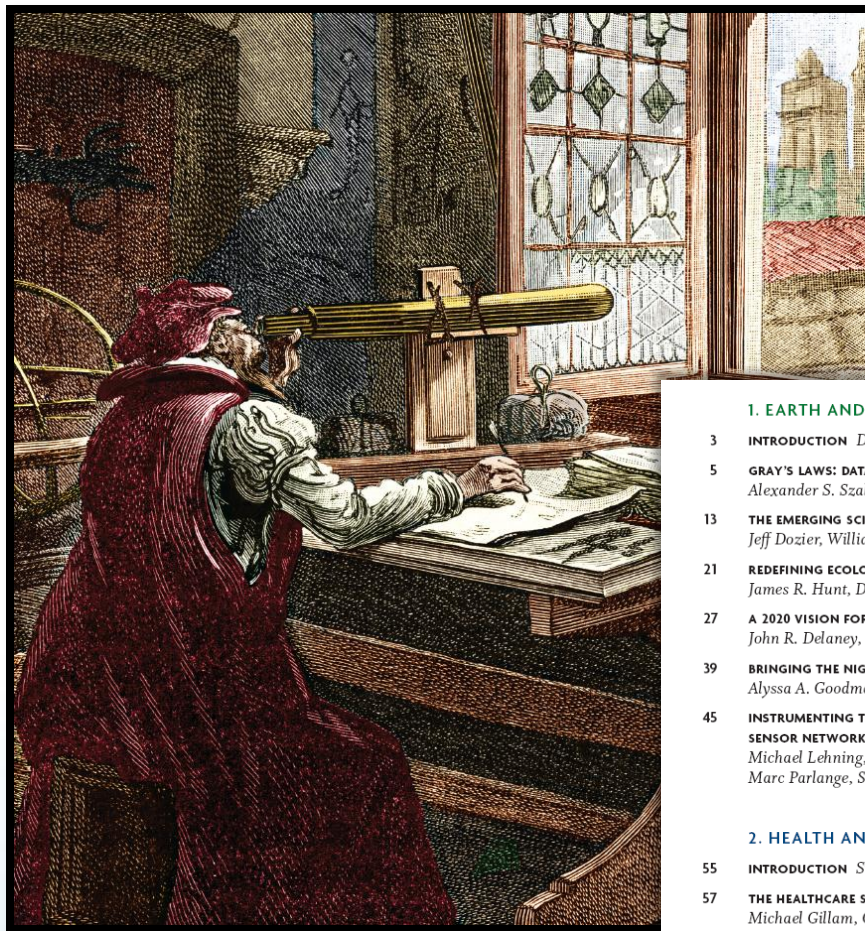


The
F O U R T H
P A R A D I G M

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE





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

1. EARTH AND ENVIRONMENT

- 3 INTRODUCTION *Dan Fay*
- 5 GRAY'S LAWS: DATABASE-CENTRIC COMPUTING IN SCIENCE
Alexander S. Szalay, José A. Blakeley
- 13 THE EMERGING SCIENCE OF ENVIRONMENTAL APPLICATIONS
Jeff Dozier, William B. Gail
- 21 REDEFINING ECOLOGICAL SCIENCE USING DATA
James R. Hunt, Dennis D. Baldocchi, Catharine van Ingen
- 27 A 2020 VISION FOR OCEAN SCIENCE
John R. Delaney, Roger S. Barga
- 39 BRINGING THE NIGHT SKY CLOSER: DISCOVERIES IN THE DATA DELUGE
Alyssa A. Goodman, Curtis G. Wong
- 45 INSTRUMENTING THE EARTH: NEXT-GENERATION
SENSOR NETWORKS AND ENVIRONMENTAL SCIENCE
*Michael Lehning, Nicholas Dawes, Mathias Bavay,
Marc Parlange, Suman Nath, Feng Zhao*
- ### 2. HEALTH AND WELLBEING
- 55 INTRODUCTION *Simon Mercer*
- 57 THE HEALTHCARE SINGULARITY AND THE AGE OF SEMANTIC MEDICINE
*Michael Gillam, Craig Feied, Jonathan Handler, Eliza Moody,
Ben Shneiderman, Catherine Plaisant, Mark Smith, John Dickason*
- 65 HEALTHCARE DELIVERY IN DEVELOPING COUNTRIES:
CHALLENGES AND POTENTIAL SOLUTIONS
Joel Robertson, Del DeHart, Kristin Tolle, David Heckerman
- 75 DISCOVERING THE WIRING DIAGRAM OF THE BRAIN
Jeff W. Lichtman, R. Clay Reid, Hanspeter Pfister, Michael F. Cohen
- 83 TOWARD A COMPUTATIONAL MICROSCOPE FOR NEUROBIOLOGY
Eric Horvitz, William Kristan
- 91 A UNIFIED MODELING APPROACH TO DATA-INTENSIVE HEALTHCARE
Iain Buchan, John Winn, Chris Bishop
- 99 VISUALIZATION IN PROCESS ALGEBRA MODELS OF BIOLOGICAL SYSTEMS
Luca Cardelli, Corrado Priami

3. SCIENTIFIC INFRASTRUCTURE

- 109 INTRODUCTION *Daron Green*
- 111 A NEW PATH FOR SCIENCE? *Mark R. Abbott*
- 117 BEYOND THE TSUNAMI: DEVELOPING THE INFRASTRUCTURE
TO DEAL WITH LIFE SCIENCES DATA *Christopher Southan, Graham Cameron*
- 125 MULTICORE COMPUTING AND SCIENTIFIC DISCOVERY
James Larus, Dennis Gannon
- 131 PARALLELISM AND THE CLOUD *Dennis Gannon, Dan Reed*
- 137 THE IMPACT OF WORKFLOW TOOLS ON DATA-CENTRIC RESEARCH
Carole Goble, David De Roure
- 147 SEMANTIC eSCIENCE: ENCODING MEANING IN NEXT-GENERATION
DIGITALLY ENHANCED SCIENCE *Peter Fox, James Hendler*
- 153 VISUALIZATION FOR DATA-INTENSIVE SCIENCE
Charles Hansen, Chris R. Johnson, Valerio Pascucci, Claudio T. Silva
- 165 A PLATFORM FOR ALL THAT WE KNOW: CREATING A KNOWLEDGE-DRIVEN
RESEARCH INFRASTRUCTURE *Savas Parastatidis*

4. SCHOLARLY COMMUNICATION


- 175 INTRODUCTION *Lee Dirks*
- 177 JIM GRAY'S FOURTH PARADIGM AND THE CONSTRUCTION
OF THE SCIENTIFIC RECORD *Clifford Lynch*
- 185 TEXT IN A DATA-CENTRIC WORLD *Paul Ginsparg*
- 193 ALL ABOARD: TOWARD A MACHINE-FRIENDLY SCHOLARLY
COMMUNICATION SYSTEM *Herbert Van de Sompel, Carl Lagoze*
- 201 THE FUTURE OF DATA POLICY
Anne Fitzgerald, Brian Fitzgerald, Kylie Pappalardo
- 209 I HAVE SEEN THE PARADIGM SHIFT, AND IT IS US *John Wilbanks*
- 215 FROM WEB 2.0 TO THE GLOBAL DATABASE *Timo Hannay*

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



The screenshot shows the Microsoft Research website page for "The Fourth Paradigm: Data-Intensive Scientific Discovery". The page features a search bar, navigation tabs (Home, Our Research, Collaboration, Careers), and a main heading "The Fourth Paradigm: Data-Intensive Scientific Discovery". Below the heading, it states "Presenting the first broad look at the rapidly emerging field of data-intensive science". A central image shows a book cover for "The Fourth Paradigm: Data-Intensive Scientific Discovery". To the right, there is a section titled "The Fourth Paradigm Now Available in Paperback and On Demand" with text explaining that the book is available as a free PDF download, but also offers paperback and Kindle versions for purchase on Amazon.com. Links are provided to "Order the paperback from Amazon.com" and "Order the Kindle version from Amazon.com". On the far right, there are sections for "In the News" (listing "A Deluge of Data Shapes a New Era in Computing") and "Related Resources" (listing "Microsoft Research collaborative projects" and "eScience Workshop 2009").

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?