Data-Intensive Scientific Discovery: eScience and the Fourth Paradigm

Tony Hey
Corporate Vice President
Microsoft Research

Microsoft Research Connections
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
Example #1: Gene Sequencing

Source: George Church, Harvard Medical School, as reported in IEEE Spectrum, Feb ’10. Figures represented in USD
Genomics and Personalized Medicine

Adapting treatments to a person’s specific genetic make-up:

- Targeting patients who can benefit (e.g. 10% of people cannot respond to codeine), and not develop toxicities (e.g. Abacavir for HIV).
- Appropriate dosage of a drug by using genetic variants to understand drug metabolism (e.g. anti-depressants, beta blockers, opioid analgesics)
- More drug approvals (re-approvals) because can now target the right sub-group based on genetics.
Stop Press!

- “On-the-go DNA sequencing”
- “The USB stick is an absolute game-changer”

17 Feb 2012
Example #2: Sloan Digital Sky Survey

“The Cosmic Genome Project”

• Two surveys in one
  • Photometric survey in 5 bands
  • Spectroscopic redshift survey

• Data is public
  • 2.5 Terapixels of images
  • 40 TB of raw data => 120TB processed
  • 5 TB catalogs => 35TB in the end

• Started in 1992, finished in 2008

• SkyServer Public Database built at JHU (SkyServer) by Alex Szalay and his research group with Jim Gray from Microsoft Research
Open Data: Public Use of the Sloan Data

Posterchild in 21st century data publishing

• Set up SkyServer web service
• Over 400 million web hits in 6 years
• About 1M distinct users vs 10,000 astronomers
• >1600 refereed papers!
• Delivered 50,000 hours of lectures to high schools

➡️ New publishing paradigm: data is published before analysis by astronomers
Citizen Scientists and Data Analysis

Galaxy Zoo activities give a useful indication of the latent appetite for scientific engagement in society. This is a collection of online astronomy projects which invite members of the public to assist in classifying galaxies.

In the first year, **50 million classifications were made by 150,000 individuals in the general public** – it quickly became the world's largest database of galaxy shapes. The original project that it spawned Galaxy Zoo 2 in February 2009 to classify another 250,000 SDSS galaxies.

The project included unique scientific discoveries such as Hanny’s Voorwerp and ‘Green Pea’ galaxies.
Hanny Van Arkel, a Dutch schoolteacher and Galaxy Zoo volunteer, posted an image to the Galaxy Zoo forum and asked "What's the blue stuff below?" No one knew. The object became known as the "Voorwerp", Dutch for "object". 
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
**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

Thanks to Jim Gray
Keynote by Dan Fay, director of E3 at Microsoft Research Connections, on “The Rise of X-Informatics.”
Archaeo-Informatics

- Archaeology is about piecing together the past
- Archaeologists must capture and organize artifacts and data
- Multiple sources, from excavation and hand sifting, to advanced geophysics and aerial surveying
- Context is everything
- Ultimately visualize and synthesize
- Advanced computational tools from data management and processing, to analysis and visualization
- Allow breakthroughs such as AHRC Portus Project, Rome

PI: Graeme Earl
Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**
- Description of natural phenomena

Last few hundred years – **Theoretical Science**
- Newton’s Laws, Maxwell’s Equations...

Last few decades – **Computational Science**
- Simulation of complex phenomena

Today – **Data-Intensive Science**
- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks

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

**eScience** is the set of tools and technologies to support data federation and collaboration
- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

Thanks to Jim Gray
An edited collection of 26 short technical essays, divided into 4 sections
Free PDF Download; Amazon Kindle version; Paperback print-on-demand option

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

Published under a Creative Commons License
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
Machine Learning and eScience

Tackling societal challenges

Identifying genetic and environmental causes of disease

Fighting HIV/AIDS

Increasing energy yield of sugar cane through genome assembly
Fighting HIV with ML and HPC

- PhyloD.Net is a Bayes-net-based tool that deciphers evolution of HIV within a patient
- Developed by eScience research group and published in *Science*, March 2007
- Used by dozens of HIV research groups
- Analysis requires HPC to do tens of thousands of independent computations
- Integrated into .NET Bio open source library
Carbo-Climate Synthesis

• Role of photosynthesis in global warming?
  • Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict
  • Difference is either due to plants or ocean absorption.

• Communal field science – each investigator acts independently.
  • Cross site studies and integration with modeling increasingly important

• Sharepoint site: www.fluxdata.org
  • 921 site-years of data from 240 sites around the world; 80+ site-years now being added
  • 60+ paper writing teams
  • American data subset is public and served more widely
  • Summary data products greatly simplify initial data discovery

(Dennis Baldocchi (Berkeley Water Center)
And Catharine van Ingen (Microsoft Research))
Outline

- The Data Deluge
- The Fourth Paradigm and eScience
- Examples of Data-Intensive Science
- Supporting the Data Life Cycle
- Research Libraries and Repositories
- Open Data and Open Science
- Clouds, Search and Semantic Computing
Data-intensive Science

Data

- Acquisition & modelling
- Collaboration and visualisation
- Archiving and preserving
- Dissemination & sharing
- Analysis & data mining

Microsoft Research Connections
Big data requires new types of data visualization tools

Collaborators:
- Alyssa Goodman; Harvard University
- Alex Szalay; Johns Hopkins University
- Curtis Wong, Jonathan Fay; Microsoft Research

- Integration of data sets and one-click contextual access
- Easy access and use

- As of May 2010, over 4M unique users
- Average number of WWT users is over 8K per day

See TED talk by Roy Gould and Curtis Wong
http://www.youtube.com/watch?v=NPU2j3JVmnw&feature=related
Layerscape
Tohoku events (shallow) and subduction slab
ChronoZoom – History in its broadest possible context...

The challenge: exploration of all known time series data with the ability to smoothly transition from billions of years down to individual nanoseconds...

This is what Walter Alvarez, Professor of Earth and Planetary Science at University of Berkeley set out to do.

Our vision is to create an application that allows researchers to browse, overlay, and explore interdisciplinary data sources.

www.chronozoomtimescale.org
‘Big History’
MS Research & Cal-Berkeley: A “zoomable” timescale platform

See the demo live at
www.chronozoomtimescale.org
Data Curation Add-in for Excel

- **Microsoft Research with California Digital Library’s Curation Center**
  - Collaboration with Trisha Cruse & John Kunze
  - Part of the DataONE (an NSF DataNet Project)

- **Some of the proposed functionality under consideration:**
  - Support for versioning,
  - Standardized date/time stamps
  - Ability to export metadata in a standard format
  - Ability to select from a globally shared vocabulary of terms for data descriptions
  - Ability to import term descriptions from the shared vocabulary and annotate them locally
  - Ability to deposit data and metadata directly into a data archive
Accessible Reproducible Research

Jill P. Mesirov

Scientific publications have at least two goals: (i) to announce a result and (ii) to convince readers that the result is correct. Mathematics papers are expected to contain a proof complete enough to allow knowledgeable readers to fill in any details. Papers in experimental science should describe the results and provide a clear enough protocol to allow successful repetition and extension.

Over the past ~35 years, computational science has posed challenges to this traditional paradigm—from the publication of the four-color theorem in mathematics (1), in which the proof was partially performed by a computer program, to results depending on computer simulation in chemistry, materials science, astrophysics, geophysics, and climate modeling. In these settings, the scientists are often sophisticated, skilled, and innovative programmers who develop large...
GenePattern Reproducible Research Add-in

Services: Connects to GenePattern database

Relationships: Inline graphics are synchronized to dataset

Data: Resulting data (and provenance) stored within Word document

Data: Control and execute query pipelines into GenePattern

http://GenepatternWordAddin.codeplex.com
Microsoft Research Connections Contributions

**Project Trident:** Toolset based on Windows Workflow Foundation that provides scientists’ need for a flexible, powerful way to analyze large, diverse datasets.

**Chemistry Add-in for Word:** Chem4Word is an add-in for Microsoft Word that enables semantic authoring of chemical structures.

**ConferenceXP:** Platform for real-time collaboration that seamlessly connects people or groups over a network, providing high-quality, low-latency videoconferencing and a rich set of collaboration capabilities.

**.NET Bio:** This open-source platform features a library of commonly used bioinformatics functions plus applications built upon that framework, and can be extended by using any Microsoft .NET language.

http://www.outercurve.org/
Virtual Research Environment Toolkits

Currently Available Kits
- Cancer Imaging for SharePoint (Oxford University)
- Collaborator Search Kit (British Library)
- Content Archiving Kit (British Library)
- Content Lifecycle Integration Framework (University of Hull)
- Data Management Kit (University of Southampton)
- Document Review Workflow Kit (British Library)
- LabTrove Research Blog (University of Southampton)
- Literature Review Kit (British Library)
- Repository Integration Kit (University of Southampton)
- RSS Feeds Kit (British Library)
- User Administration (FBA) Kit (British Library)
- X-Ray Diffraction Data Converter Kit (University of Delhi)

Templates
- Researcher and Project Templates (British Library)
- La Trobe Templates for VRE – RIC with Active Directory (La Trobe University)

VRE Based on Sharepoint
Archiving and Preservation: A Document Conversion Service

Convert Word Perfect and Word documents to OOXML, ODF and UOF

View documents in various formats

Compare original and converted documents

http://odf-converter.sourceforge.net/
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
The **NIH Public Access Policy** ensures that the public has access to the published results of NIH funded research.

- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive **PubMed Central** upon acceptance for publication.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

**Entrez cross-database search**
Library Budgets and Journal Prices

MIT Libraries Materials Purchases vs. CPI % Increase
1986-2006

- Consumer Price Index % +
- Serial Expenditures % +
- # Serials Purchased % +
- # Books Purchased % +
- Book Expenditures % +
University of Michigan’s libraries are canceling some journal subscriptions because of budget cuts and the increasing costs of the subscriptions.

University Librarian Paul Courant said that about 2,500 were canceled in the 2007 fiscal year.

The University Library budget has gone up by an average of 3.1 percent per year since 2004.

According to Library Journal magazine, the average subscription price of national arts and humanities journals has increased 6.8% per year since 2003; social science journals by 9.2%; and science journals increased by 8.3%.
UCSD Journal Cancellations 2010

- UC San Diego Libraries continue to cancel journal subscriptions because of:
  - **Budget cuts**
  - **Increasing costs of subscriptions**
  - **Shift to e-only**

- **523 titles ($180,000 worth) were canceled in FY 2009/10**, and at least as many will be cancelled in FY 2010/11. We will also be canceling high cost-per-use databases.

- **State funding support for the UCSD Libraries’ materials budget is projected to be almost 20% less in FY 2011/12 than it was in FY 2007/08.**

- According to *Library Journal*, the average subscription price of arts and humanities journals will increase 6.6% nationally in 2012; of social science journals by 7%; and science journals by 7.6%.

- The “big deals” negotiated by CDL have slowed the double-digit journal price inflation that was typical in the scholarly publishing industry over the last decade.

- Thanks to these journal packages, over the last five years, UCSD has only sustained subscription cost increases in the range of 3-5% per annum.
The Inevitability of Open Access Repositories

As Dean of Engineering at Southampton I was ‘responsible’ for monitoring the research output of over 200 Faculty and 500 Post Docs and Grad Students

- The University library could not afford to subscribe to all the journals that my staff published in, not to mention conference proceedings and workshop contributions, so we insisted on keeping a digital copy of all output in a University Repository ...

‘Green Open Access’ or ‘Self-Archiving’ requires authors to make peer-reviewed final drafts of their articles accessible by depositing them in their Institution's OA Repository either on submission or on acceptance for publication

- Note that individual papers need not be set to be immediately visible outside the institution but can be set to ‘delayed open access’ as in NLM’s PubMedCentral. Web copies of non-journal versions are allowed by over 80% of publishers ...
### Some Facts about VT ETDs

Electronic Theses and Dissertations

What the server logs reveal about accesses to VT ETDs. (Fiscal Years)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful requests</td>
<td>441,480</td>
<td>976,587</td>
<td>1,436,279</td>
<td>2,725,773</td>
<td>6,759,779</td>
<td>9,455,258</td>
<td>13,627,721</td>
<td>23,327,315</td>
<td>27,199,853</td>
<td>24,934,678</td>
</tr>
<tr>
<td>Requests for PDF files</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mostly full ETDs)</td>
<td>221,679</td>
<td>481,038</td>
<td>578,152</td>
<td>2,173,420</td>
<td>4,497,199</td>
<td>7,320,818</td>
<td>10,697,468</td>
<td>17,461,678</td>
<td>21,113,555</td>
<td>18,580,199</td>
</tr>
<tr>
<td>Requests for HTML files</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mostly tables of contents and abstracts)</td>
<td>165,710</td>
<td>215,539</td>
<td>260,699</td>
<td>400,149</td>
<td>471,917</td>
<td>367,767</td>
<td>410,988</td>
<td>517,684</td>
<td>555,518</td>
<td>547,237</td>
</tr>
<tr>
<td>Requests for Multimedia</td>
<td>1,714</td>
<td>4,468</td>
<td>12,633</td>
<td>44,237</td>
<td>169,186</td>
<td>121,241</td>
<td>54,584</td>
<td>87,911</td>
<td>88,996</td>
<td>123,648</td>
</tr>
<tr>
<td>Distinct files requested</td>
<td>6,419</td>
<td>21,451</td>
<td>16,409</td>
<td>*</td>
<td>50,982</td>
<td>31,884</td>
<td>43,280</td>
<td>53,606</td>
<td>112,260</td>
<td>135,874</td>
</tr>
<tr>
<td>Distinct hosts served</td>
<td>29,816</td>
<td>57,901</td>
<td>87,804</td>
<td>*</td>
<td>425,475</td>
<td>680,771</td>
<td>985,146</td>
<td>1,594,913</td>
<td>18,92,653</td>
<td>1,530,570</td>
</tr>
<tr>
<td>Average data transferred daily</td>
<td>156,089 Kb</td>
<td>219,132 Kb</td>
<td>382 Mb</td>
<td>945 Mb</td>
<td>2.15 Gb</td>
<td>3.49 Gb</td>
<td>5.641 Gb</td>
<td>27.85 Gb</td>
<td>38.18 Gb</td>
<td>38.46 Gb</td>
</tr>
<tr>
<td>Data transferred</td>
<td>55,637 Mb</td>
<td>78,107 Mb</td>
<td>137 Gb</td>
<td>332 Gb</td>
<td>780 Gb</td>
<td>1.2 Tb</td>
<td>2.06 Tb</td>
<td>9.93 Tb</td>
<td>13.97 Tb</td>
<td>13.71 Tb</td>
</tr>
</tbody>
</table>

* no data available

---

200,000 requests to 20M requests from 1997 to 2007

➢ Graphic demonstration of the power of Open Access
Webometrics overall rankings are obtained by combining four different rankings with the following weights:

- Visibility (external inlinks) 50%
- Size (Web Pages) 20%
- Rich Files (pdf, doc, ps, ppt) 15%
- Scholar (papers and citations) 15%
# Webometrics: Top 25 Universities Worldwide

## July 2010

<table>
<thead>
<tr>
<th>DOMAINS</th>
<th>COUNTRY</th>
<th>REGION</th>
<th>SIZE</th>
<th>VISIBILITY</th>
<th>RICH</th>
<th>SCHOLAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>harvard.edu</td>
<td>us</td>
<td>North America</td>
<td>3</td>
<td>3</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>mit.edu</td>
<td>us</td>
<td>North America</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>stanford.edu</td>
<td>us</td>
<td>North America</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>berkeley.edu</td>
<td>us</td>
<td>North America</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>cornell.edu</td>
<td>us</td>
<td>North America</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>umich.edu</td>
<td>us</td>
<td>North America</td>
<td>8</td>
<td>7</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>umn.edu</td>
<td>us</td>
<td>North America</td>
<td>10</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>washington.edu</td>
<td>us</td>
<td>North America</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>54</td>
</tr>
<tr>
<td>wisc.edu</td>
<td>us</td>
<td>North America</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>48</td>
</tr>
<tr>
<td>utexas.edu</td>
<td>us</td>
<td>North America</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>upenn.edu</td>
<td>us</td>
<td>North America</td>
<td>19</td>
<td>9</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>psu.edu</td>
<td>us</td>
<td>North America</td>
<td>1</td>
<td>45</td>
<td>5</td>
<td>116</td>
</tr>
<tr>
<td>columbia.edu</td>
<td>us</td>
<td>North America</td>
<td>14</td>
<td>14</td>
<td>19</td>
<td>89</td>
</tr>
<tr>
<td>cmu.edu</td>
<td>us</td>
<td>North America</td>
<td>11</td>
<td>27</td>
<td>3</td>
<td>74</td>
</tr>
<tr>
<td>uiuc.edu</td>
<td>us</td>
<td>North America</td>
<td>42</td>
<td>12</td>
<td>20</td>
<td>75</td>
</tr>
<tr>
<td>ucla.edu</td>
<td>us</td>
<td>North America</td>
<td>18</td>
<td>15</td>
<td>21</td>
<td>78</td>
</tr>
<tr>
<td>tamu.edu</td>
<td>us</td>
<td>North America</td>
<td>24</td>
<td>31</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>umd.edu</td>
<td>us</td>
<td>North America</td>
<td>25</td>
<td>28</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>purdue.edu</td>
<td>us</td>
<td>North America</td>
<td>33</td>
<td>32</td>
<td>14</td>
<td>39</td>
</tr>
<tr>
<td>unc.edu</td>
<td>us</td>
<td>North America</td>
<td>22</td>
<td>19</td>
<td>22</td>
<td>148</td>
</tr>
<tr>
<td>msu.edu</td>
<td>us</td>
<td>North America</td>
<td>21</td>
<td>24</td>
<td>26</td>
<td>100</td>
</tr>
<tr>
<td>cam.ac.uk</td>
<td>uk</td>
<td>Europe</td>
<td>15</td>
<td>18</td>
<td>54</td>
<td>97</td>
</tr>
<tr>
<td>ufl.edu</td>
<td>us</td>
<td>North America</td>
<td>23</td>
<td>37</td>
<td>12</td>
<td>94</td>
</tr>
<tr>
<td>rutgers.edu</td>
<td>us</td>
<td>North America</td>
<td>17</td>
<td>38</td>
<td>13</td>
<td>127</td>
</tr>
</tbody>
</table>
Metrics and Research Rankings

• Webometrics ‘Scholar’ ranking
  • “Scholar (Sc). The data is a combination of items published between 2006 and 2010 included in Google Scholar and the global output (2004.- 2008) obtained from Scimago SIR.”
  • ‘Scholar’ is clearly not a ‘perfect’ metric - but equally clearly, this must measure something of relevance for the research reputation of a university ...

➢ Institutional Research Repositories must be part of the university’s ‘Reputation Management’ strategy
<table>
<thead>
<tr>
<th>Rank</th>
<th>Institution</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Harvard</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MIT</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>UNAM</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Minnesota</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>UC Madrid</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Munich</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Stanford</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>U Queensland</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Kyoto</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Masaryk</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Toronto</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Michigan</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>UPC Barcelona</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Texas A&amp;M</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ETH Zurich</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Nebraska</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Groningen</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Vienna</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>CUHK</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Georgia Tech</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Southampton</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>Cornell</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Pennsylvania</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Tokyo</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Murcia</td>
<td></td>
</tr>
</tbody>
</table>
# Webometrics: European Universities Scholar Ranking (January 2012)

<table>
<thead>
<tr>
<th>Rank</th>
<th>University</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Utrecht</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>ETH Zurich</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>L.Max.Univ. Munich</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Computense Madrid</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Leuven</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>UCL</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>Ghent</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>Milano</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>Amsterdam</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>Groningen</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>UPC Barcelona</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>Oxford</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>Cambridge</td>
<td>41</td>
</tr>
<tr>
<td>14</td>
<td>Napoli</td>
<td>43</td>
</tr>
<tr>
<td>15</td>
<td>Padova</td>
<td>45</td>
</tr>
<tr>
<td>16</td>
<td>Wien</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>Helsinki</td>
<td>52</td>
</tr>
<tr>
<td>18</td>
<td>Sapienza</td>
<td>53</td>
</tr>
<tr>
<td>19</td>
<td>Pisa</td>
<td>55</td>
</tr>
<tr>
<td>20</td>
<td>Southampton</td>
<td>58</td>
</tr>
<tr>
<td>Rank</td>
<td>University</td>
<td>Score</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>L.Max.Univ. Munich</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Freiburg</td>
<td>61</td>
</tr>
<tr>
<td>3</td>
<td>Heidelberg</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>Frei Univ. Berlin</td>
<td>99</td>
</tr>
<tr>
<td>5</td>
<td>Aachen</td>
<td>115</td>
</tr>
<tr>
<td>6</td>
<td>Frankfurt</td>
<td>128</td>
</tr>
<tr>
<td>7</td>
<td>Bonn</td>
<td>135</td>
</tr>
<tr>
<td>8</td>
<td>T.U.Munich</td>
<td>146</td>
</tr>
<tr>
<td>9</td>
<td>Tubingen</td>
<td>154</td>
</tr>
<tr>
<td>10</td>
<td>Karlsruhe</td>
<td>155</td>
</tr>
<tr>
<td>11</td>
<td>Humboldt Berlin</td>
<td>172</td>
</tr>
<tr>
<td>12</td>
<td>Erlangen</td>
<td>173</td>
</tr>
<tr>
<td>13</td>
<td>T.U.Darmstadt</td>
<td>178</td>
</tr>
<tr>
<td>14</td>
<td>Hamburg</td>
<td>186</td>
</tr>
<tr>
<td>15</td>
<td>Stuttgart</td>
<td>189</td>
</tr>
<tr>
<td>16</td>
<td>Munster</td>
<td>209</td>
</tr>
<tr>
<td>17</td>
<td>Koln</td>
<td>231</td>
</tr>
<tr>
<td>18</td>
<td>T.U.Berlin</td>
<td>240</td>
</tr>
<tr>
<td>19</td>
<td>T.U.Dresden</td>
<td>242</td>
</tr>
<tr>
<td>20</td>
<td>Bochum</td>
<td>255</td>
</tr>
</tbody>
</table>
Future of Research Repositories?

• Repositories will contain full text versions of research papers and also ‘grey’ literature such as workshop papers, presentations, as well as technical reports and theses
  • In the future repositories will also contain data, images and software
• Centralized versus Distributed?
  • NIH National Library of Medicine
  • arXiv
  • COAR – Coalition of Open Access Repositories
➤ Need for good distributed search tools
About COAR

COAR, the Confederation of Open Access Repositories, is a young association of repository initiatives launched in October 2009, uniting 59 institutions in 23 countries from throughout Europe, Latin America, Asia, and North America. Its mission is to enhance greater visibility and application of research outputs through global networks of Open Access digital repositories.

COAR promotes infrastructure interoperability and a joint global data store of Open Access repositories to enable and support the re-use of data by service and portal providers. Currently, COAR has three working groups, each with its own set of responsibilities, objectives, and related activities.

COAR Interoperability Project
We welcome you to get involved with the COAR Interoperability Project and to join the ongoing discussion.
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
The Berlin Declaration 2003

• ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’

• Defines open access contributions as including:
  
  • ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’
The Government, in line with our overarching commitment to transparency and open data, is committed to ensuring that publicly-funded research should be accessible free of charge.

Free and open access to taxpayer-funded research offers significant social and economic benefits by spreading knowledge, raising the prestige of UK research and encouraging technology transfer.
NSF-OCI Task Force on Data and Visualization

Advisory Committee on Cyberinfrastructure

March 2011

Tony Hey, Co-Chair
Microsoft Corporation

Dan Atkins, Co-Chair
University of Michigan

Margaret Hedstrom
University of Michigan

• **Roles and Responsibilities**

• Recognize that responsibility for data stewardship is shared among:
  
  • Principal Investigators
  • Research centers
  • University research libraries
  • Discipline-based libraries and archives
  • National scientific agencies
  • Commercial service providers.
Infrastructure Delivery

• Acknowledge that data infrastructure and services are essential research assets fundamental to today’s science and worthy of long-term investments.

Culture and Sociological Change

• Introduce new funding models that reinforce expectations and institute specific conditions for data sharing.
Economic Value and Sustainability
• Develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services.

Data Management Guidelines
• Identify and share best practices for critical areas of data management.

Ethics, Privacy and Intellectual Property
• Invest in the research and training of the research community in privacy-preserving data-access so that PIs can embrace privacy by design.
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

*(From Jim Gray's last talk)*
Outline

• The Data Deluge
• The Fourth Paradigm and eScience
• Examples of Data-Intensive Science
• Supporting the Data Life Cycle
• Research Libraries and Repositories
• Open Data and Open Science
• Clouds, Search and Semantic Computing
The Cloud - Options

- Public Cloud
- Secure Cloud Federation
- Private Cloud
- Internal IT
- Enterprise
- Dedicated Cloud
MODIS Azure: Computing Evapotranspiration (ET) in the Cloud

A pipeline for download, processing, and reduction of diverse NASA MODIS satellite imagery

Contributors: Catharine van Ingen (Microsoft Research), Youngryel Ryu (UC Berkeley), and Jie Li (Univ. of Virginia)
MODIS Azure Service

Source Imagery Download Sites

Data Collection Stage

Download Queue

Source Metadata

Request Queue

MODIS Azure Service Web Role Portal

Scientists

Science results

Scientific Results Download

Derivation Reduction Stage

Analysis Reduction Stage

Reduction #1 Queue

Reduction #2 Queue

Catharine van Ingen (Microsoft Research), Jie Li, Marty Humphrey (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL), Keith Jackson (BL), Jay Borenstein (Stanford), Team SICT: Vlad Andrei, Klaus Ganser, Samir Selman, Nandita Prabhu (Stanford), Team Nimbus: David Li, Sudarshan Rangarajan, Shantanu Kurhekar, Riddhi Mittal (Stanford)
Computing ET for 1 US Year

All storage costs assume 3 month project duration

Source Imagery Download Sites

Data Collection Stage

- 400-500 GB
- 60K files
- 10 MB/sec
- 11 hours
- <10 workers

$50 upload
$225 storage

Reprojection Queue

Reprojection Stage

- 400 GB
- 45K files
- 3500 hours

$420 cpu
$60 download

Request Queue

Derivation Reduction Stage

- 5-7 GB
- 5.5K files
- 1800 hours
- 100 workers

$216 cpu

Total Cost $1200

Analysis Reduction Stage

- <10 GB
- ~15K files
- 1800 hours

$216 cpu

Total Cost $1200

$50 upload
$225 storage

$420 cpu
$60 download

$216 cpu
$216 cpu
$420 cpu
$60 download

Scientific Results Download
EU VENUS-C Project: Virtual multidisciplinary Environments Using Cloud infrastructures

Part of the XCG Cloud Initiative for Research led by Dennis Gannon
Seven Pilot Scenarios

T5.1
Structural Analysis for Civil Engineering

T5.2
Building Information Management

T5.3
Data for Science - AquaMaps

T5.4
Civil Protection and Emergencies

T5.5
Bioinformatics

T5.6
System Biology

T5.7
Drug Discovery
Microsoft Research – Academic Research

http://academic.research.microsoft.com

- Experimental search tool
  - Not yet fully developed
- From our MSR Asia Lab (Beijing)
  - For researchers, by researchers
- Historically focused on Computer Science
  - Now targeting all academic fields/domains
- Desired functionality includes
  - Find top papers in a domain
  - Easily search the top papers, authors, conferences, and journals for a topic
  - See details about a specific paper, author, conference or journal
  - Quickly find relationships between authors (with Visual Explorer)
  - Get a related Bing Answer
MSR Academic Search data comes from open access repositories, publishers, and web crawls

- Currently 44M papers across 20+ domains
  - With 100M papers in the queue
- More improvements to come...
# Top 10 Computer Science Organizations

<table>
<thead>
<tr>
<th>Organization</th>
<th>Publications</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft (H-Index: 285)</td>
<td>9846</td>
<td>37983</td>
</tr>
<tr>
<td>Stanford University (H-Index: 365)</td>
<td>6371</td>
<td>26084</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology (H-Index: 362)</td>
<td>6977</td>
<td>23939</td>
</tr>
<tr>
<td>Carnegie Mellon University (H-Index: 279)</td>
<td>8379</td>
<td>23145</td>
</tr>
<tr>
<td>University of California Berkeley (H-Index: 349)</td>
<td>5804</td>
<td>21467</td>
</tr>
<tr>
<td>IBM (H-Index: 244)</td>
<td>7326</td>
<td>17166</td>
</tr>
<tr>
<td>University of Illinois Urbana Champaign (H-Index: 221)</td>
<td>6684</td>
<td>16700</td>
</tr>
<tr>
<td>Georgia Institute of Technology (H-Index: 176)</td>
<td>5685</td>
<td>12749</td>
</tr>
<tr>
<td>The french National Institute for Research in Computer science and Control (H-Index: 134)</td>
<td>4794</td>
<td>12358</td>
</tr>
<tr>
<td>University of Maryland (H-Index: 210)</td>
<td>4435</td>
<td>11647</td>
</tr>
</tbody>
</table>
Public API for Academic Search

- **Application Programming Interface**
  - Supports queries against all academic entities and their basic information

- **With the API, you can**
  - Work with others to share information
  - Help users to build useful clients

- **Openly available to everyone**
  - Targeting the academic community
  - API is available for non-commercial use only

API details at: [http://academic.research.microsoft.com/About/Help.htm#5](http://academic.research.microsoft.com/About/Help.htm#5)
By uncovering the hierarchical structure of scholarly citation, we can identify key papers pertaining to any search query. For a reader new to the field we can find the classic and foundational papers; for an expert we can find the latest innovations.

From patterns of scholarly citation, we use Rosvall and Bergstrom’s map equation to chart the topography of science and the relations among fields and subfields. [journal map] [paper map]

By integrating a hierarchical clustering of citation networks with semantic analysis, we develop a scalable map of scientific fields and the key research terms and topics therein.

Scientific influence is often quantified using simple citation counts, but the structure of a citation network provides far more information than can be revealed by these simple counts. This is principle behind the Eigenfactor metrics; we can better rank the importance of scientific journals or papers by viewing them in the context of the full citation network.
Semantic Computing

Computers are great tools for storing, computing, managing, and indexing huge amounts of data. We would like computers to also help with the automatic acquisition, discovery, aggregation, organization, correlation, analysis, interpretation, and inference of the world's information and knowledge.
Semantic Chemistry Add-in for Word

- Authoring and rendering of semantic-rich chemical information (CML)
- In partnership with the University of Cambridge
- Support for Office 2007 and Office 2010
- Available under Apache 2.0
- Over 360K downloads since March 22nd, 2010
InnerEye: Semantic Understanding of Medical Images

- InnerEye focuses on the analysis of patient scans using machine learning techniques for automatic detection and segmentation of healthy anatomy as well as anomalies.

- In this image, we see that InnerEye can separate a carotid artery visually from adjacent parts of a human body.
Zentity: Semantically-enabled repository software

Default web UI with CSS support and custom ASP.Net controls

Flexible data model enables many scenarios and can be easily extended over time

A semantic computing platform to store and expose relationships between digital assets

http://research.microsoft.com/zentity/
Introducing Schema.org: Bing, Google and Yahoo Unite to Build the Web of Objects

The Bing Team  6/2/2011 10:01 AM  Comments (3)

We’ve been talking for a while about the need to rethink the search experience to better reflect both the changing web and advancing user habits.

One of the biggest challenges and opportunities we see is to literally create a high-definition proxy of the physical world inside of Bing. In other words, we want to be able to model the world in which we all live to the level that search can actually help you make decisions and get things done in real life by understanding all the options the world presents.

We’ve made great progress on the technical front to begin to model the real world from the messy bits of data scattered across the web. Things like movies have benefitted from this work. We’re now able to understand “Casablanca” is a movie and literally mine the web to re-assemble information about that movie from millions of sites.

But we think we can do better. We want to enable publishers to give us hints about what things they are describing on their sites. Rather than rely solely on machine learning and other AI techniques, we asked “what if we could enable publishers to have a single schema they could use to describe their sites that all search engines could understand?”
Clouds, Search and Semantic Computing

Future Research Infrastructure will use semantic knowledge services on Client + Cloud
Some Resources

- Microsoft Research
  - [http://research.microsoft.com](http://research.microsoft.com)
  - Microsoft Research downloads: [http://research.microsoft.com/research/downloads](http://research.microsoft.com/research/downloads)

- Microsoft Research Connections

- Science at Microsoft
  - [http://www.microsoft.com/science](http://www.microsoft.com/science)

- Scholarly Communications
  - [http://www.microsoft.com/scholarlycomm](http://www.microsoft.com/scholarlycomm)

- CodePlex
  - [http://www.codeplex.com](http://www.codeplex.com)

- Outercurve Foundation
  - [http://www.outercurve.org/](http://www.outercurve.org/)

- Tony Hey on eScience
  - [http://tonyhey.net/](http://tonyhey.net/)