

The Fourth Paradigm and X-Informatics

Tony Hey

Corporate Vice President

Microsoft Research



The Data Deluge



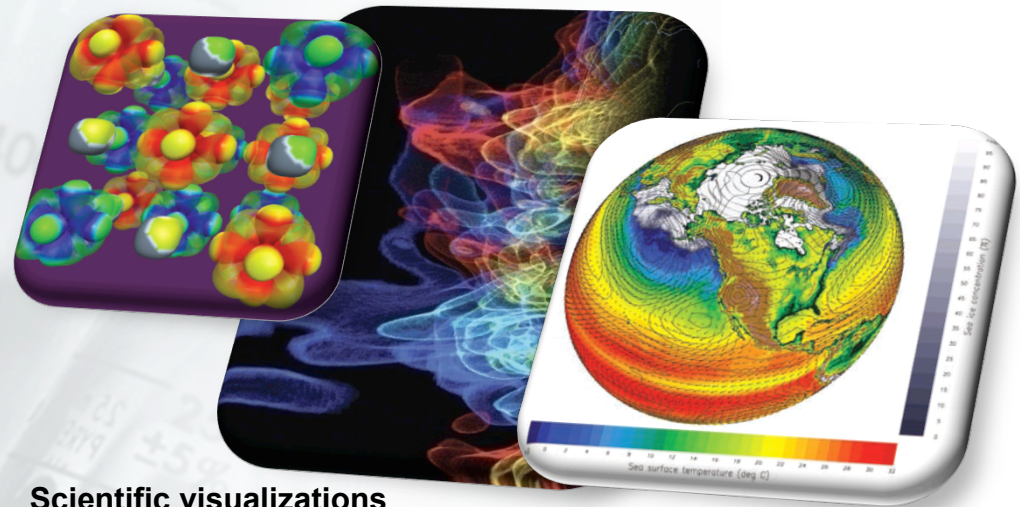
A Digital Data Deluge in Research

- 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.

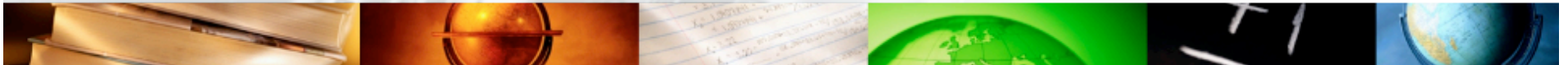


Scientific visualizations

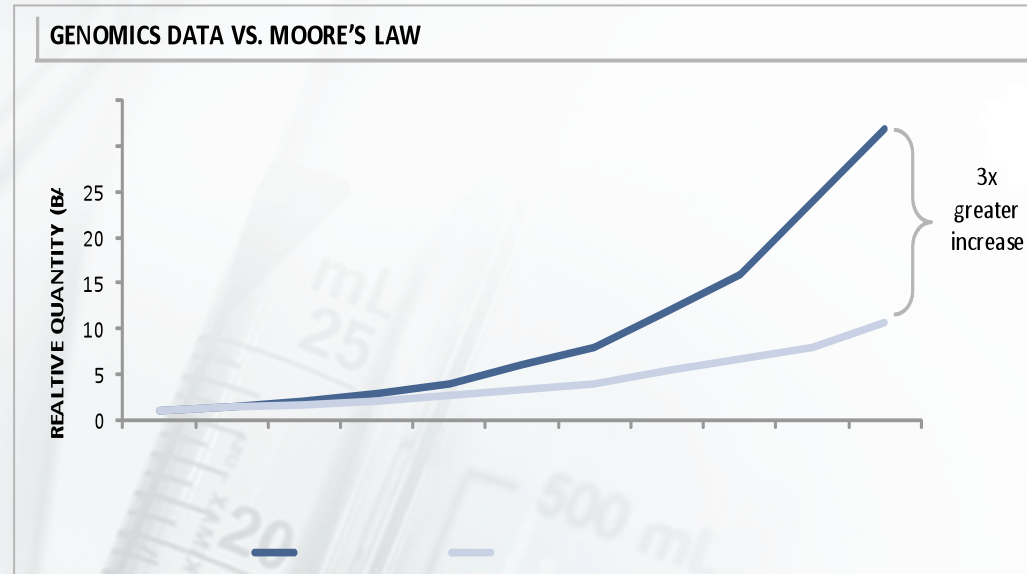
NSF Cyberinfrastructure report, March 2007



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



The Genomics Revolution

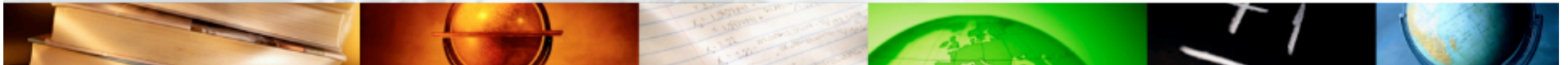


Genomic data is growing exponentially

Statistical methods and machine learning over these large datasets are key technologies to providing insights



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



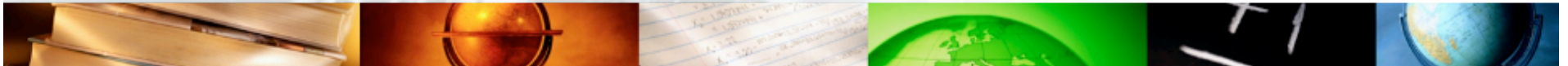
Genomics and Personalized Medicine

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.



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



The 'Cosmic Genome' Project

- The Sloan Digital Sky Survey is the first major astronomical survey project:
 - 5 color images and spectra of $\frac{1}{4}$ of the sky
 - Pictures of over 300 million celestial objects
 - Distances to the closest 1 million galaxies
- Jim Gray from Microsoft Research worked with astronomer Alex Szalay to build the public 'SkyServer' archive for the survey
- New model of scientific publishing
 - Have to publish the data before astronomers publish their analysis



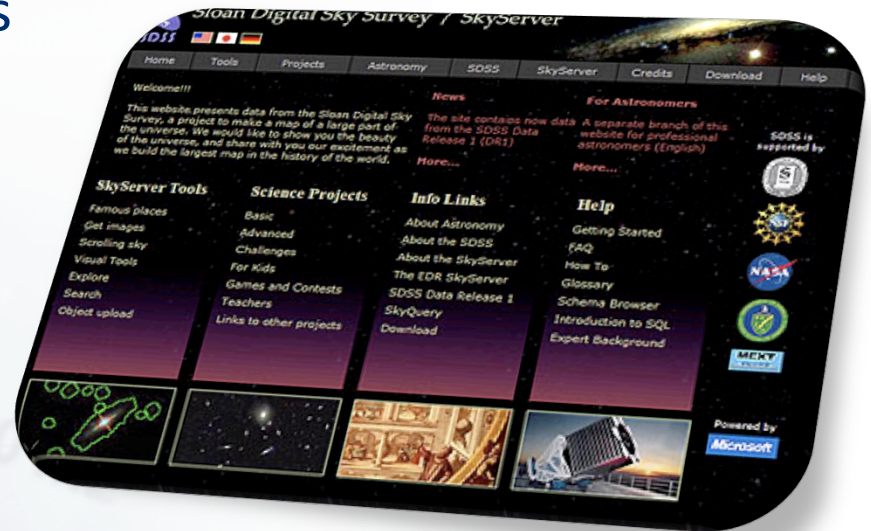
This work is licensed under a
[Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



Public Use of the SkyServer

- **Posterchild in 21st century data publishing**

- 380 million web hits in 6 years
- 930,000 distinct users vs 10,000 astronomers
- 1600 refereed papers!
- Delivered 50,000 hours of lectures to high schools
- Delivered 100B rows of data



- **GalaxyZoo ‘Citizen Science’ site has over 200,000 volunteers**

- Hanny’s ‘Voorwerp’
- Green Pea galaxies

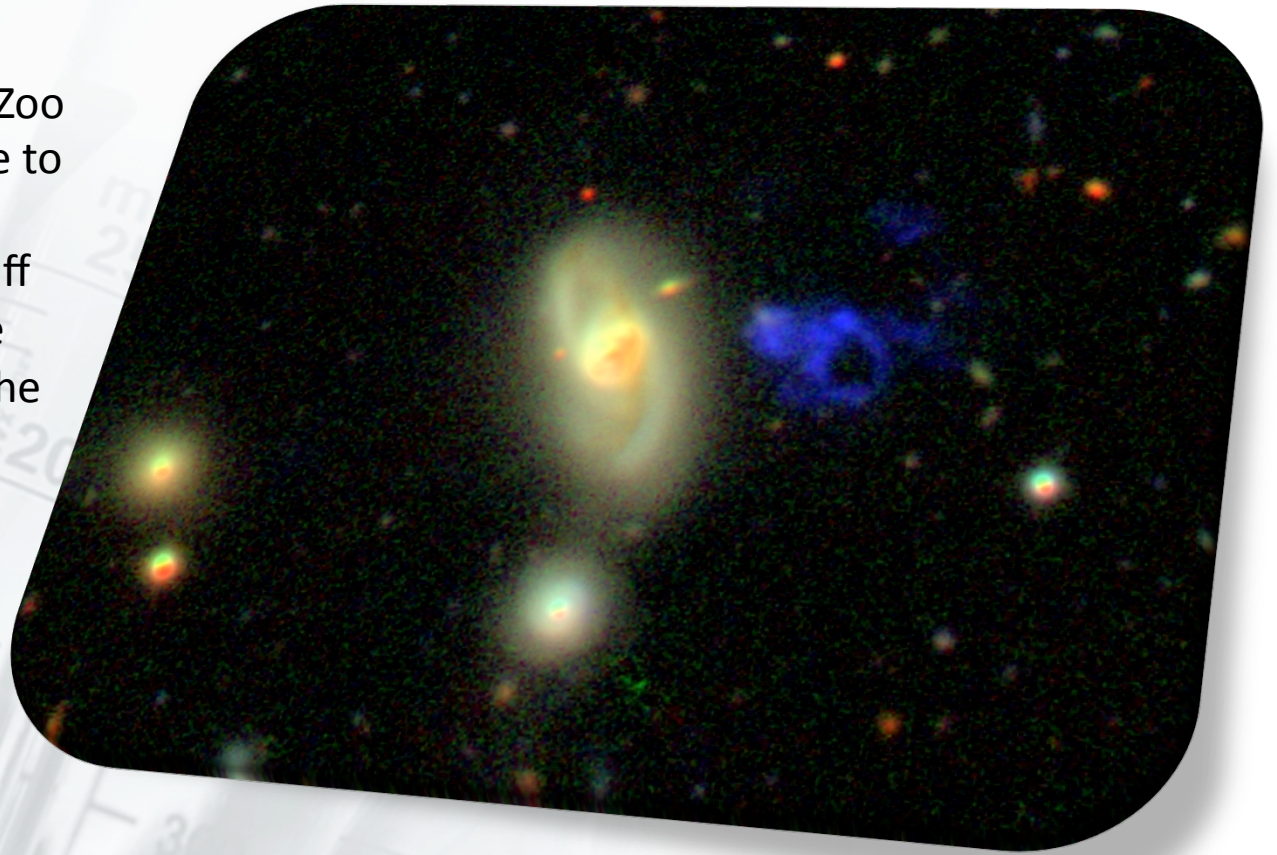


This work is licensed under a [Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).

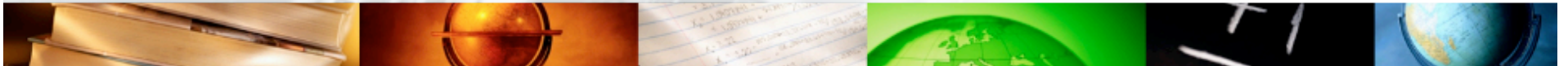


Hanny van Arkel's Voorwerp

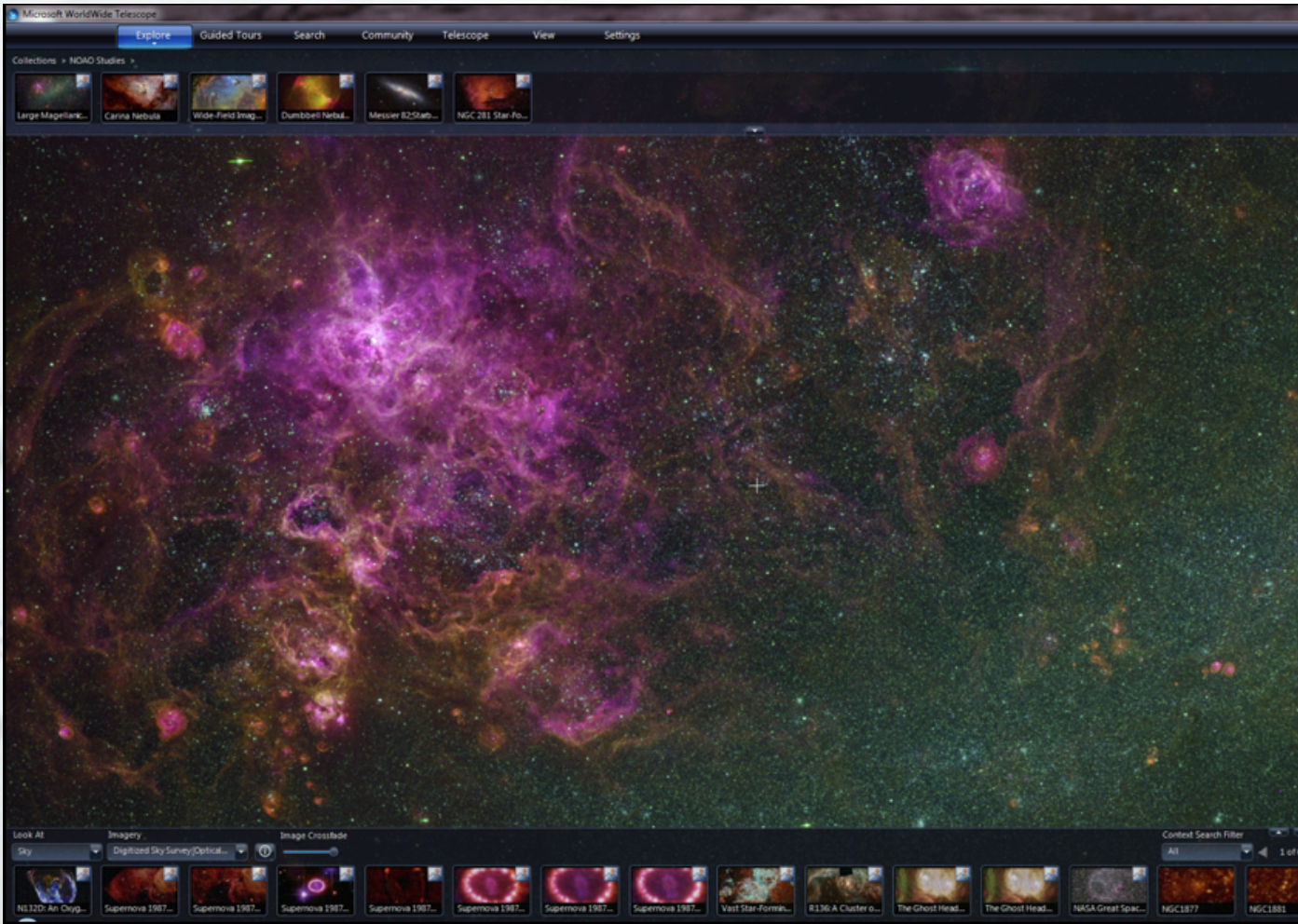
Hanny Van Arkel, a Dutch schoolteacher and Galaxy Zoo volunteer, posted an image to the Galaxy Zoo forum and asked "What's the blue stuff below?" No one knew. The object became known as the "**Voorwerp**" — Dutch for "object".



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



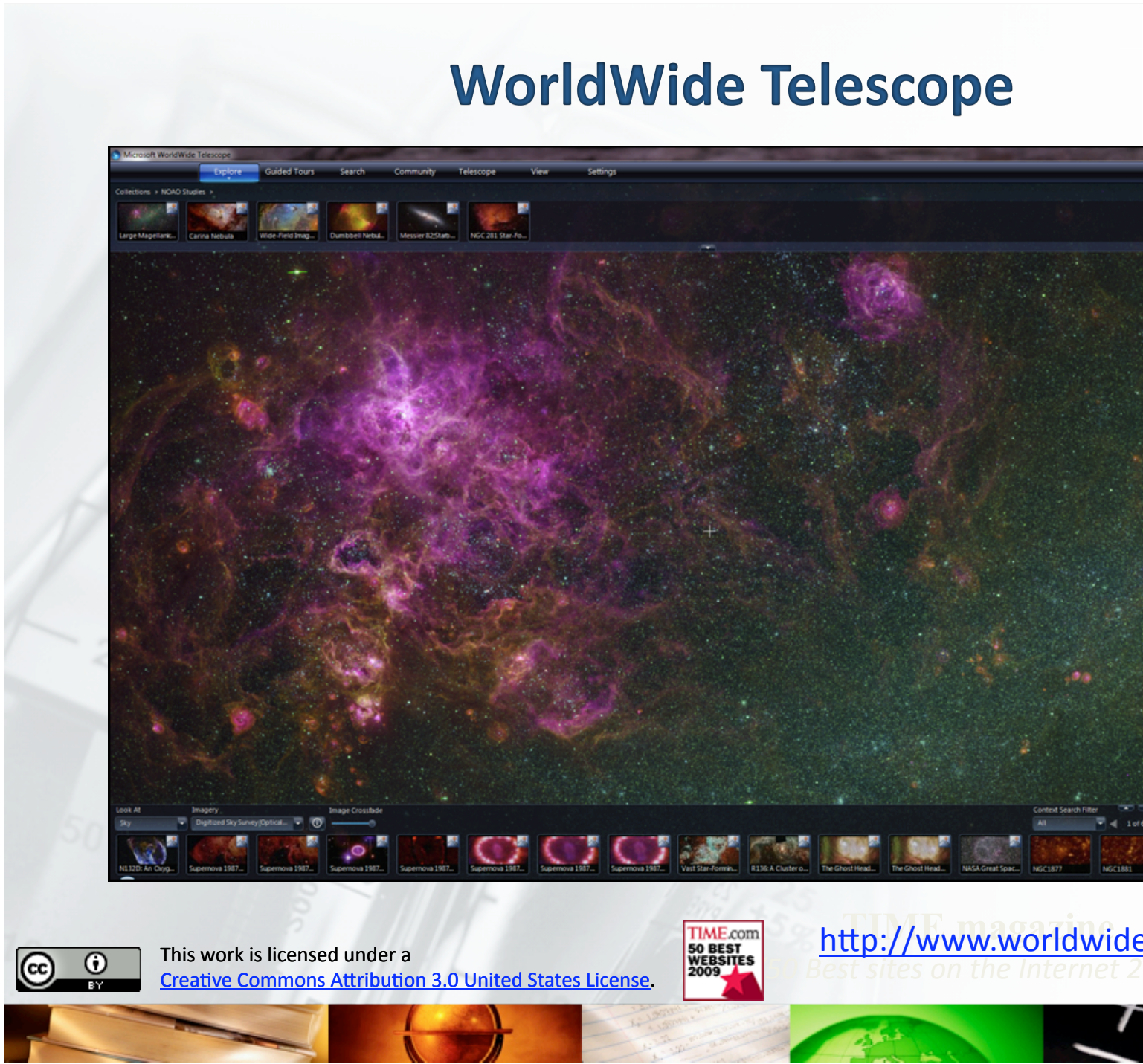
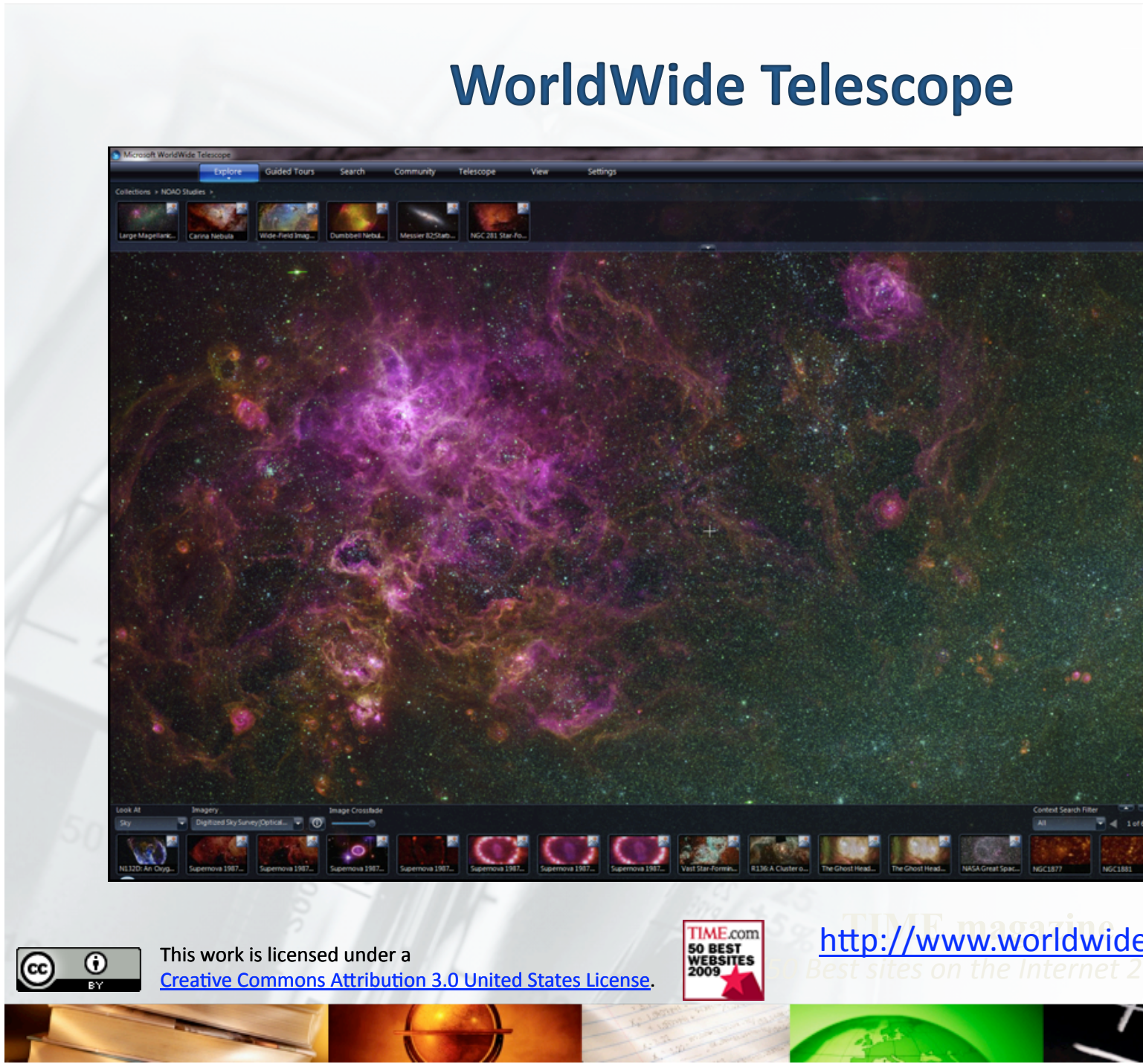
WorldWide Telescope



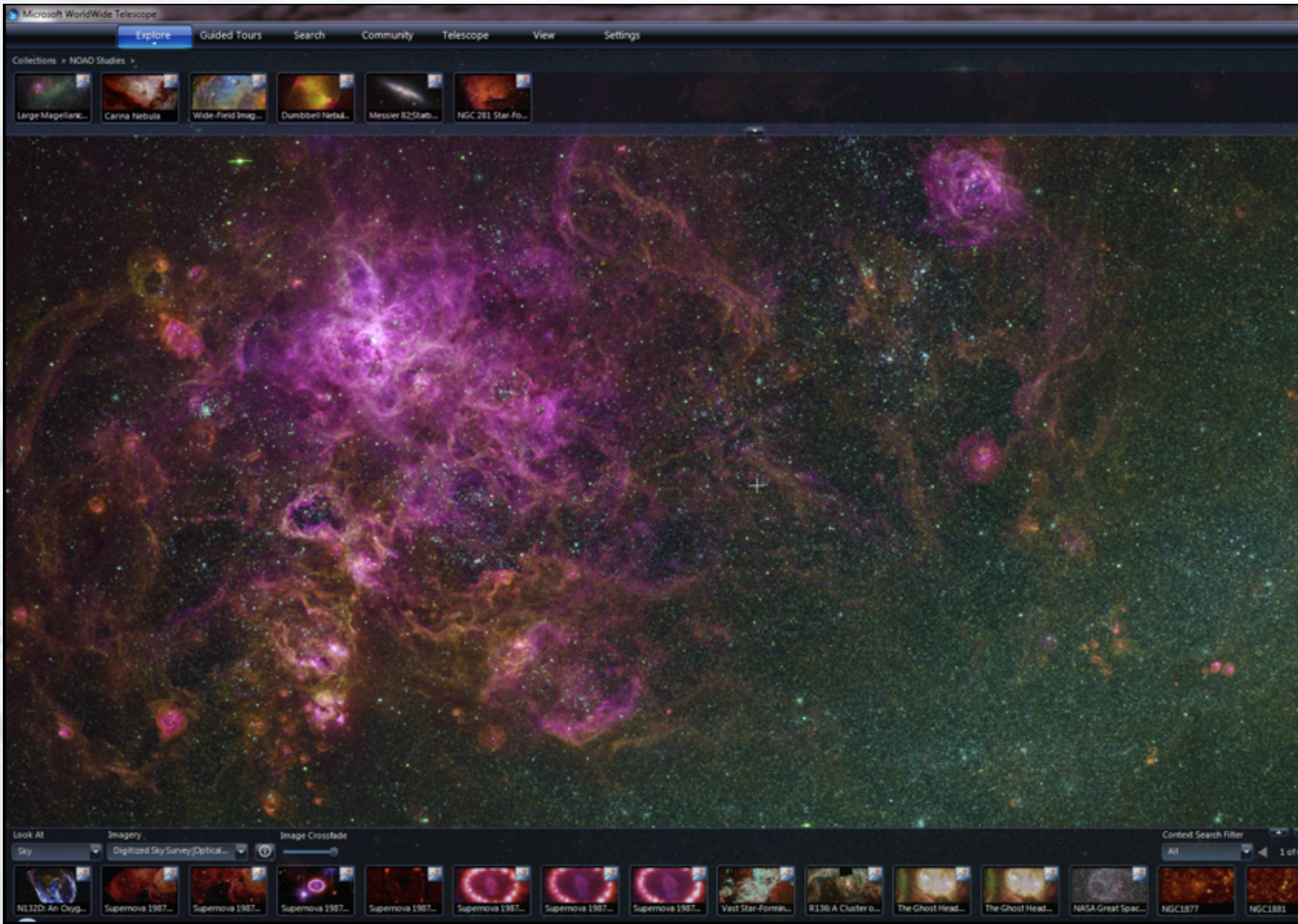
This work is licensed under a [Creative Commons Attribution 3.0 United States License](http://creativecommons.org/licenses/by/3.0/).

 <http://www.worldwidetelescope.org>

TIME magazine
50 Best sites on the Internet 2009



WorldWide Telescope

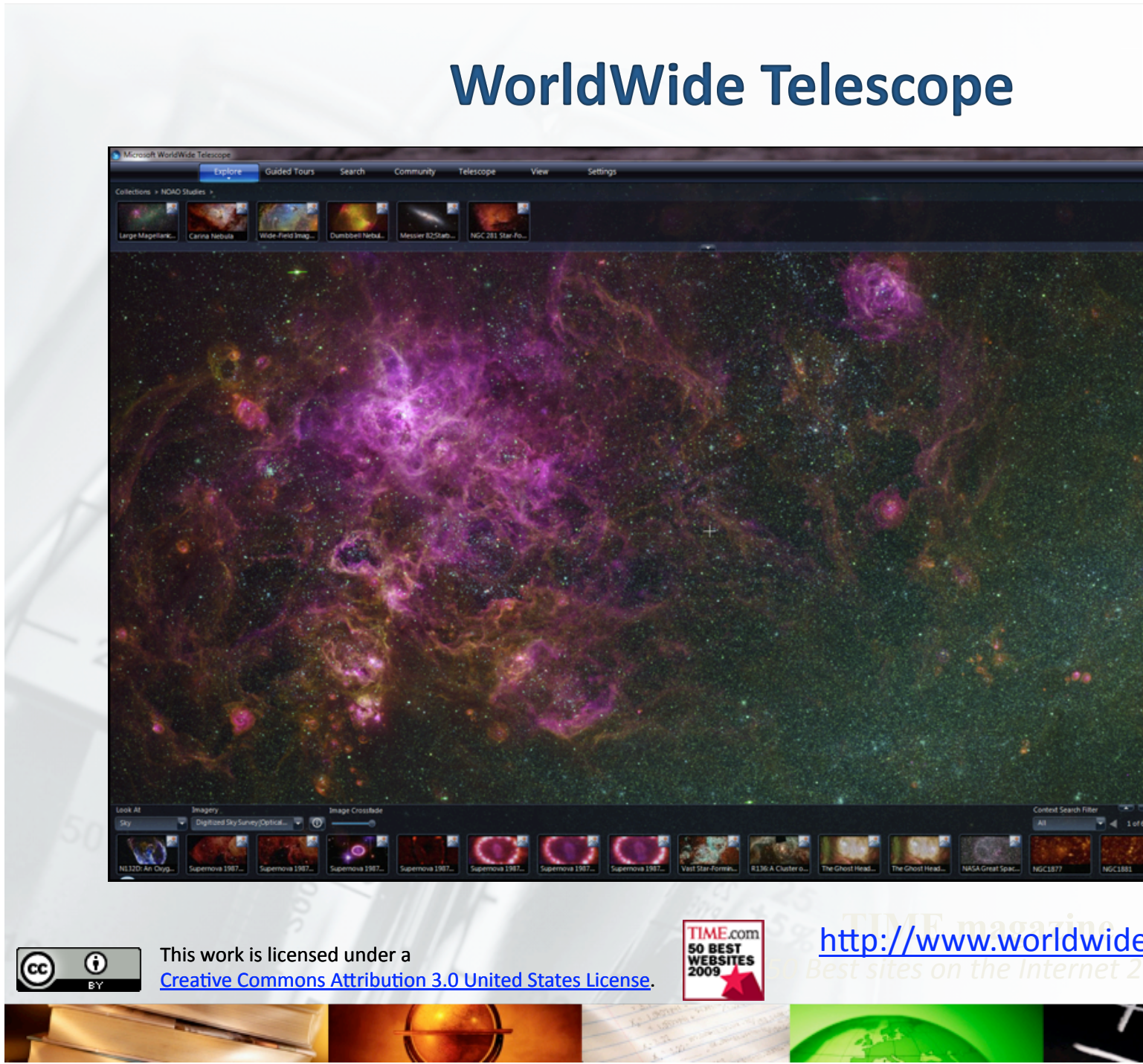


This work is licensed under a [Creative Commons Attribution 3.0 United States License](http://creativecommons.org/licenses/by/3.0/).

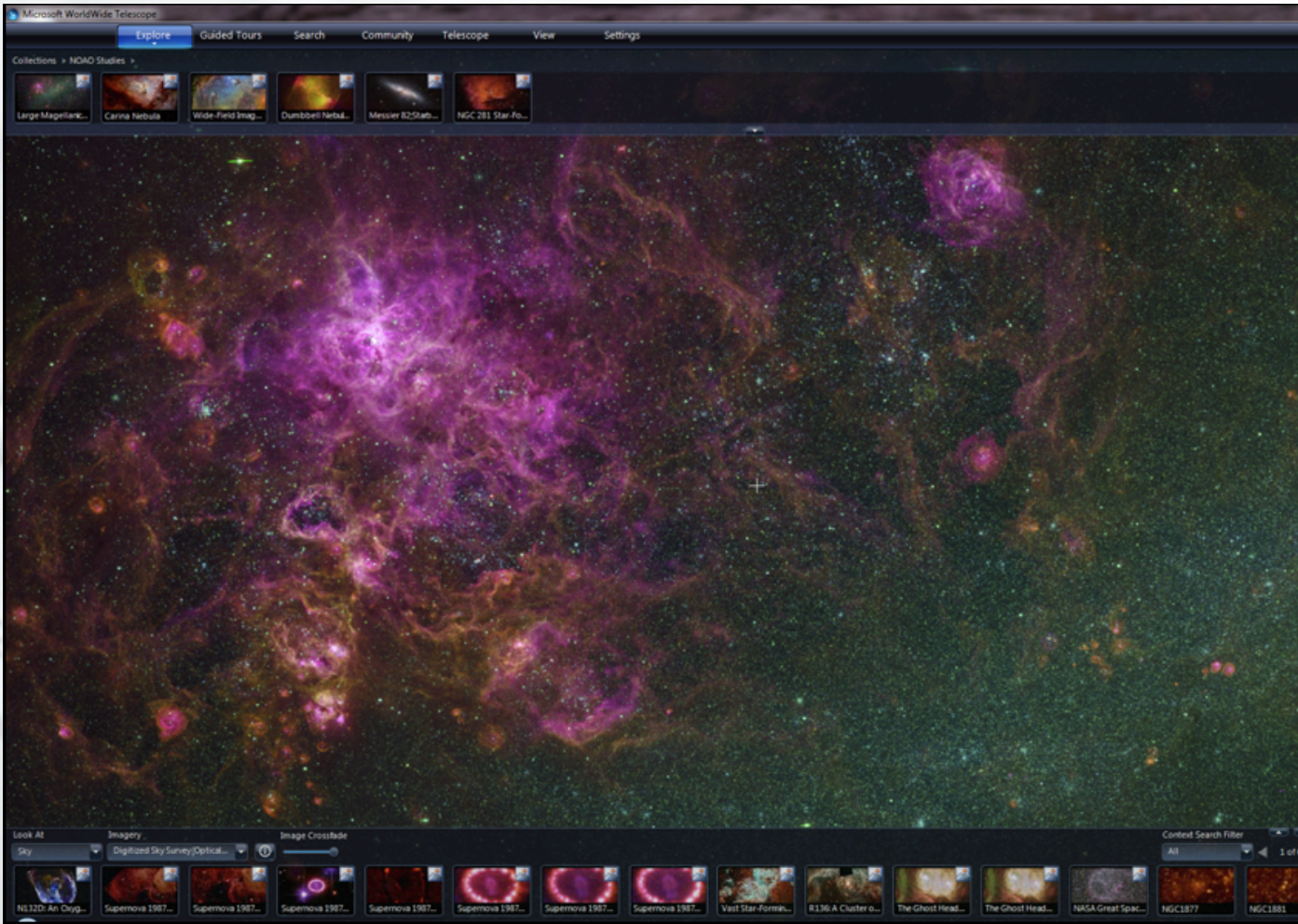
 

<http://www.worldwidetelescope.org>

TIME magazine
50 Best sites on the Internet 2009



WorldWide Telescope

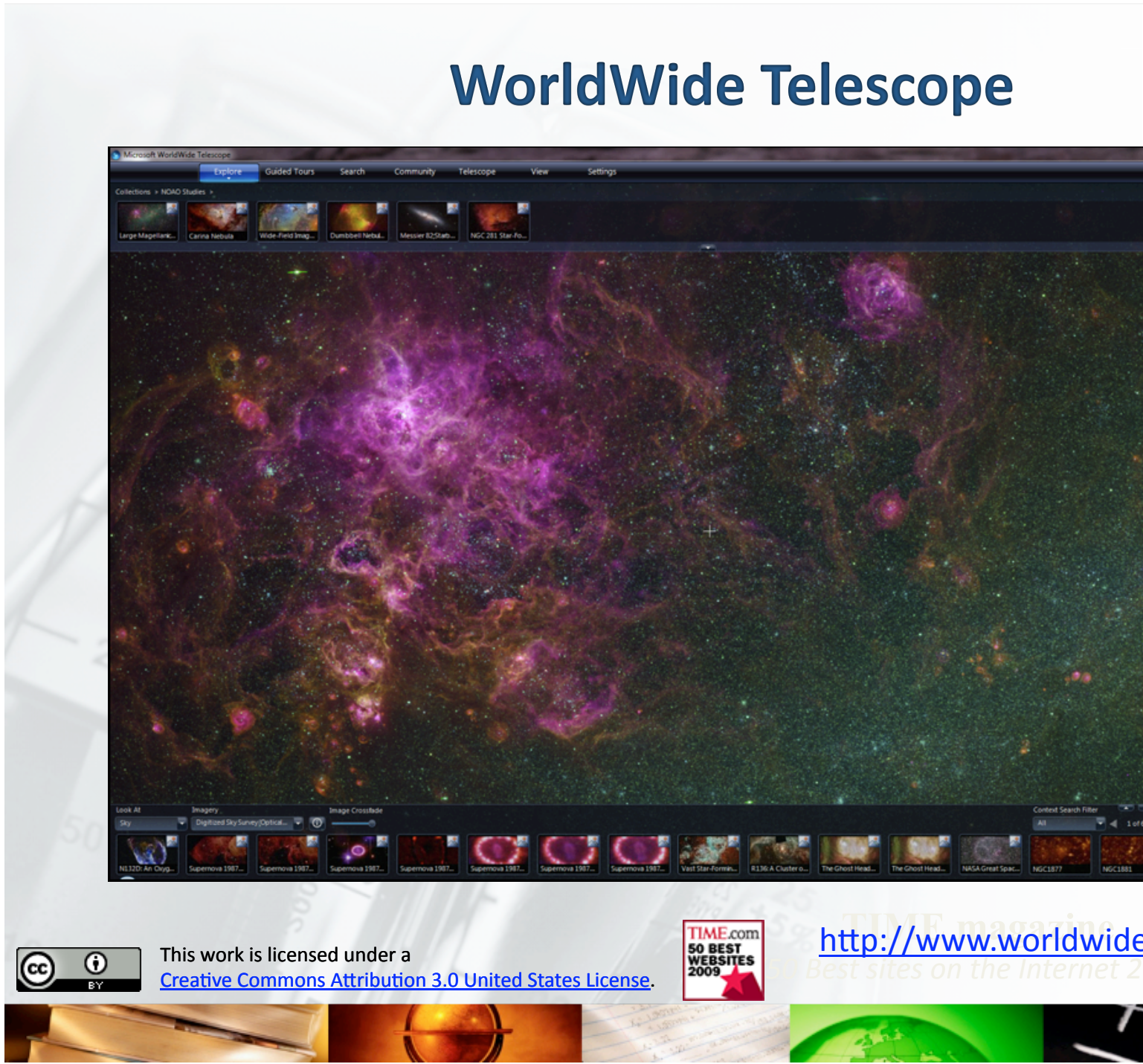


This work is licensed under a [Creative Commons Attribution 3.0 United States License](http://creativecommons.org/licenses/by/3.0/).

<http://www.worldwidetelescope.org>

TIME magazine
50 Best sites on the Internet 2009



The Fourth Paradigm

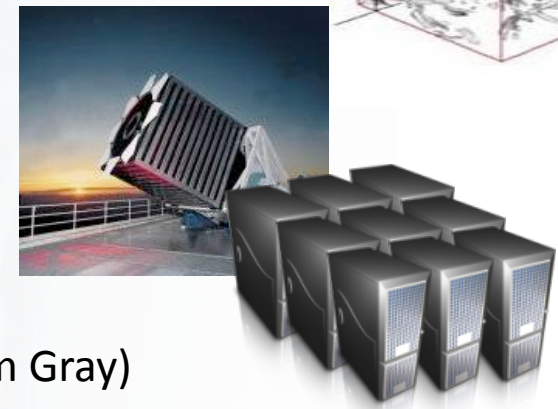
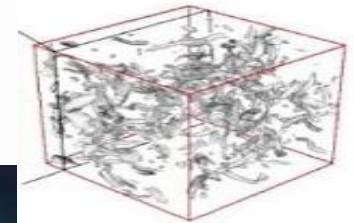


Emergence of a Fourth Research Paradigm

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 data federation and collaboration**
- 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)

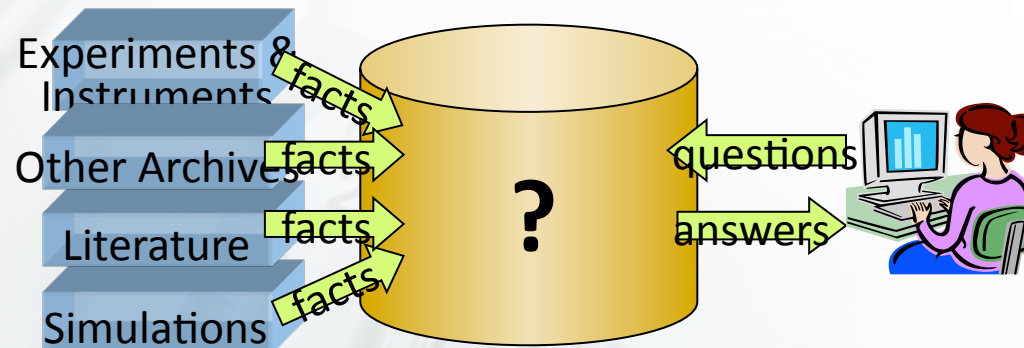


This work is licensed under a
[Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



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



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

With thanks to Jim Gray



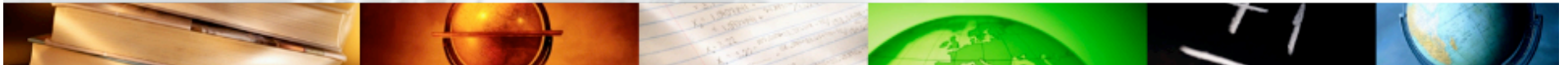
Jim Gray's Call to Action

Listed 7 key areas for action by Funding Agencies:

1. Fund both development and support of software tools
2. Invest at all levels of the finding 'pyramid'
3. Fund development of 'generic' Laboratory Information Management Systems
4. Fund research into scientific data management, data analysis, data visualization, new algorithms and tools



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



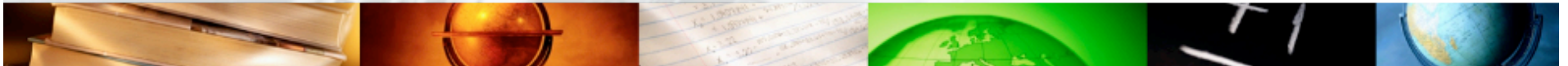
Jim Gray's Call to Action (continued)

Remaining three key areas for action relate to the future of Scholarly Communication and Libraries:

5. Establish Digital Libraries that support the other sciences like the NLM does for Medicine
6. Fund development of new authoring tools and publication models
7. Explore development of digital data libraries that contain scientific data (not just the metadata) and support integration with published literature



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





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



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



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



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



Some Examples of Data-Intensive Science

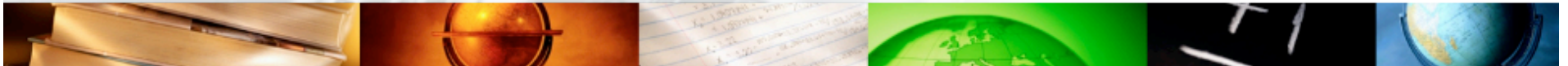


Statistical Machine Learning (ML)

- Computer Science challenges (Language, Speech, Vision) are not the only ones susceptible to data-driven ML approaches
- Anything with uncertainty, large amounts of data and a large search space can benefit, e.g.
 - SPAM
 - Search and Ad relevance
 - Traffic analysis and prediction
 - Large swaths of “science”



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





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.

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 wilyer than any of

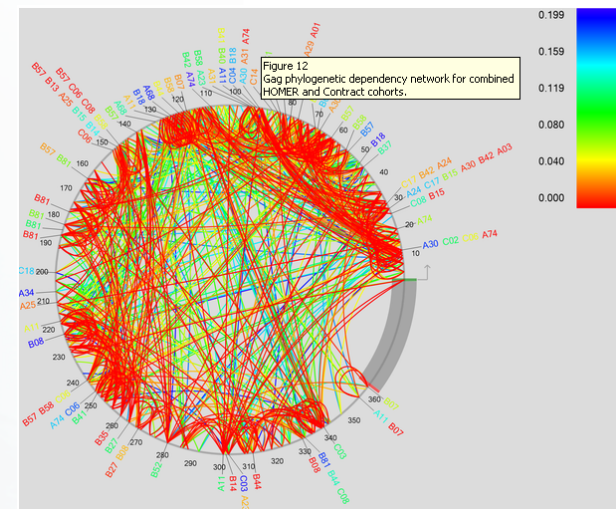
Microsoft's corporate foes. While Heckerman has high hopes that his tools will lead to vaccines that can be tested on humans within three years, his research

Similar mutations may crop up in computer and medical viruses



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
- Now used by dozens of HIV research groups
- Led to discovery of two key insights to fight HIV:
 - Our immune system attacks frameshift epitopes, which may be useful to include in a vaccine (*JEM*, 2010)
 - Natural killer cells directly attack HIV (*Nature Medicine*, in review)
- Typical job
 - 10 – 20 CPU hours with extreme jobs requiring 1K – 2K CPU hours
 - Requires a large number of test runs for a given job (1 – 10M tests)



PhyloD.Net on cover of *PLoS Comp Bio*, Nov 2008
Carlson, Kadie, & Heckerman et al.



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



The Swiss Experiment (EPFL, Marc Parlange)

- Climate change affects on the regional hydrologic cycle will have profound implications for the Alps and therefore Europe
- Need for field measurements remains crucial to test simulations and guide the design of new models used in warning networks.
 - There are known areas where predictability is poor yet potential implications are large
 - Larger and more diverse observational spatial scales with appropriate instrumentation
- Partnering with schools to get sensors deployed over a wide spatial area
 - 1000 children from 10-14
 - ‘touching’ the environment changes how people perceive environmental issues

“Our ability to estimate the hydrologic cycle at regional scales generally remains behind societal needs”

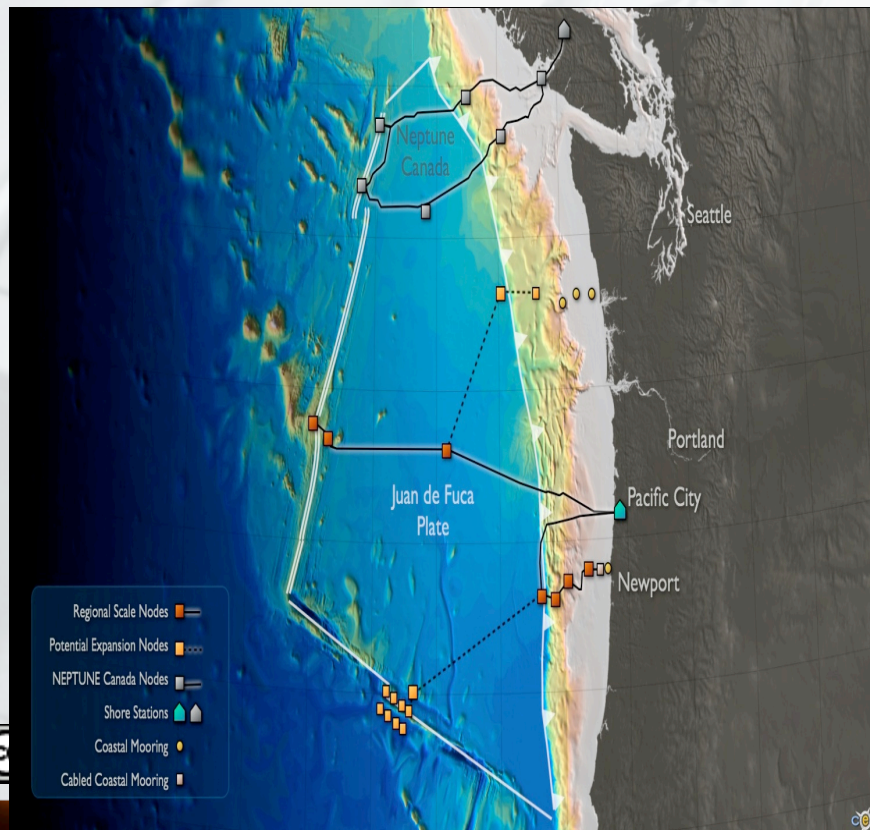


This work is licensed under a [Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



Supporting Smart Sensors and Data Fusion

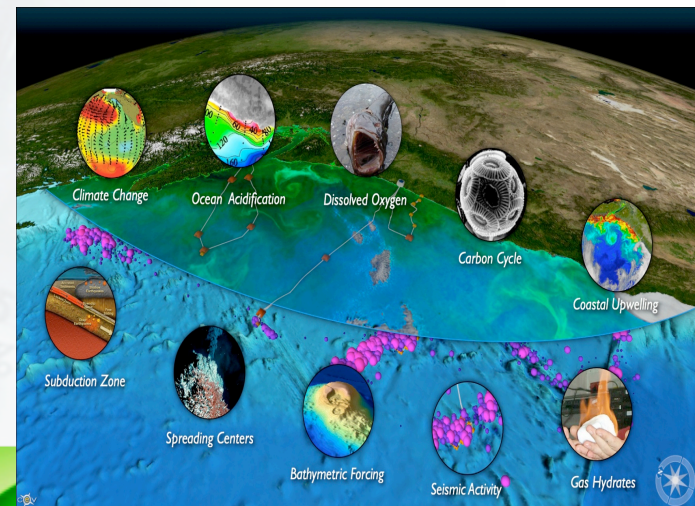
- The NSF Ocean Observatory Initiative
 - Hundreds of cabled sensors and robots exploring the sea floor
 - Data to be collected, curated, mined
 - OOI Architecture plan of record, store this data in the cloud



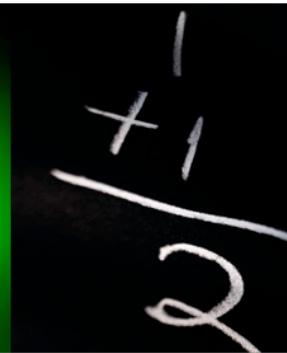
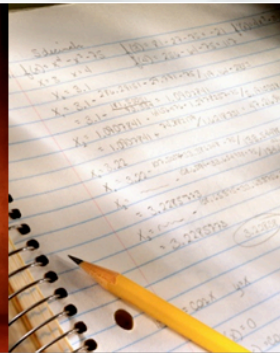
Data collected from:

- 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**.



Digital Humanities



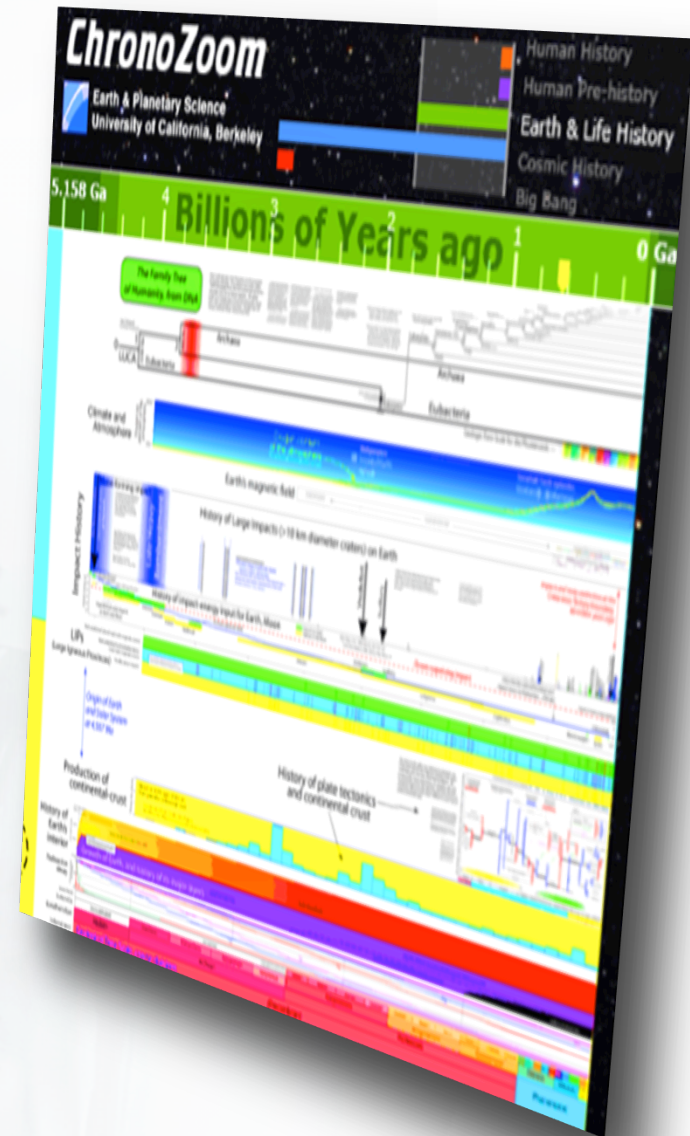
ChronoZoom – History in its broadest possible context ...

The challenge: exploration of all known time series, and 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. And he did it, with the help of Microsoft Research and the Live Labs team.

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

www.chronozoomtimescale.org

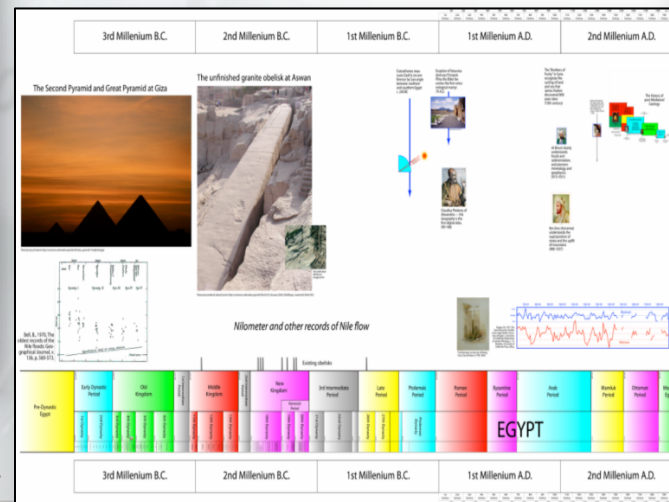
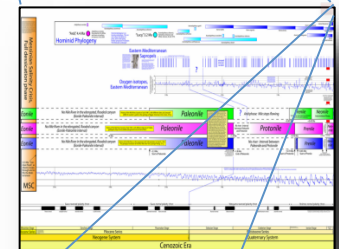
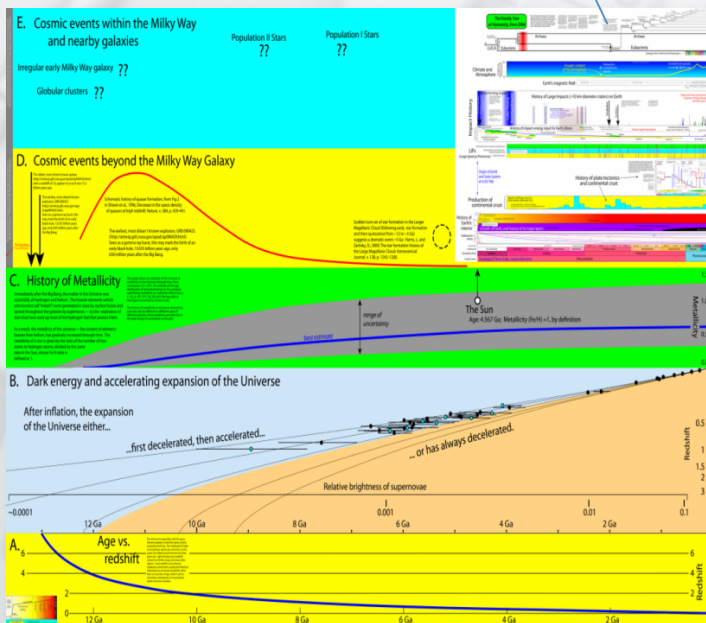
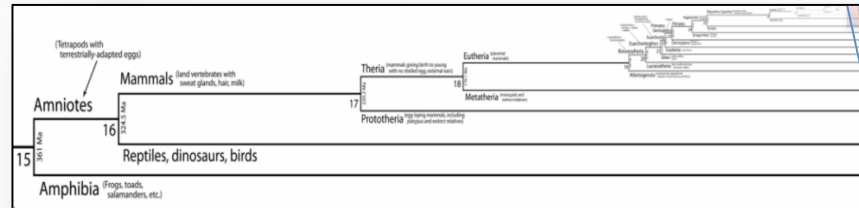


This work is licensed under a
[Creative Commons Attribution 3.0 United States License](http://creativecommons.org/licenses/by/3.0/us/).



New explorations of the history of the universe

See the demo live at www.chronozoomtimescale.org Walter Alvarez with the support of Bill Crow and the Live Labs Seadragon team.

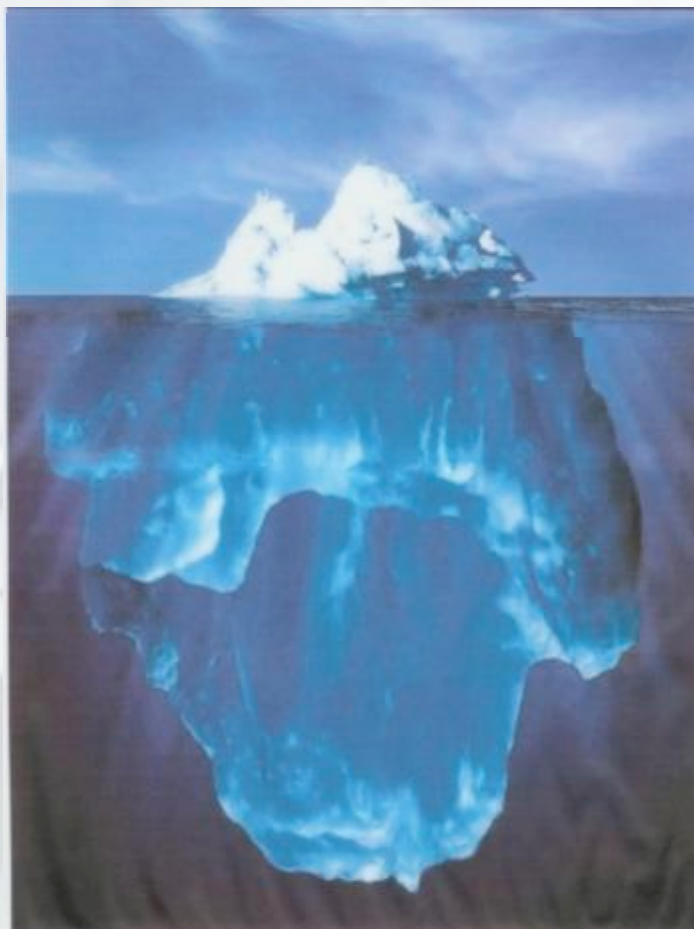


This work is licensed under a [Creative Commons Attribution 3.0 United States License](http://creativecommons.org/licenses/by/3.0/).

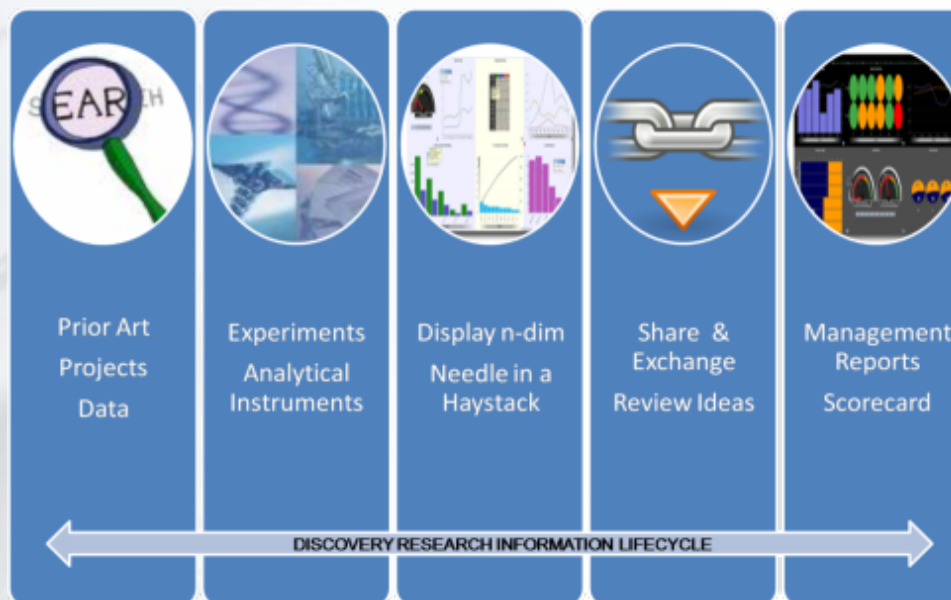
The Future of Scholarly Communications



Facilitating the move from static summaries to rich information vehicles



- Pace of science is picking up...rapidly
- The status quo is being challenged and researchers are demanding more
- Why can't a research report offer more ...



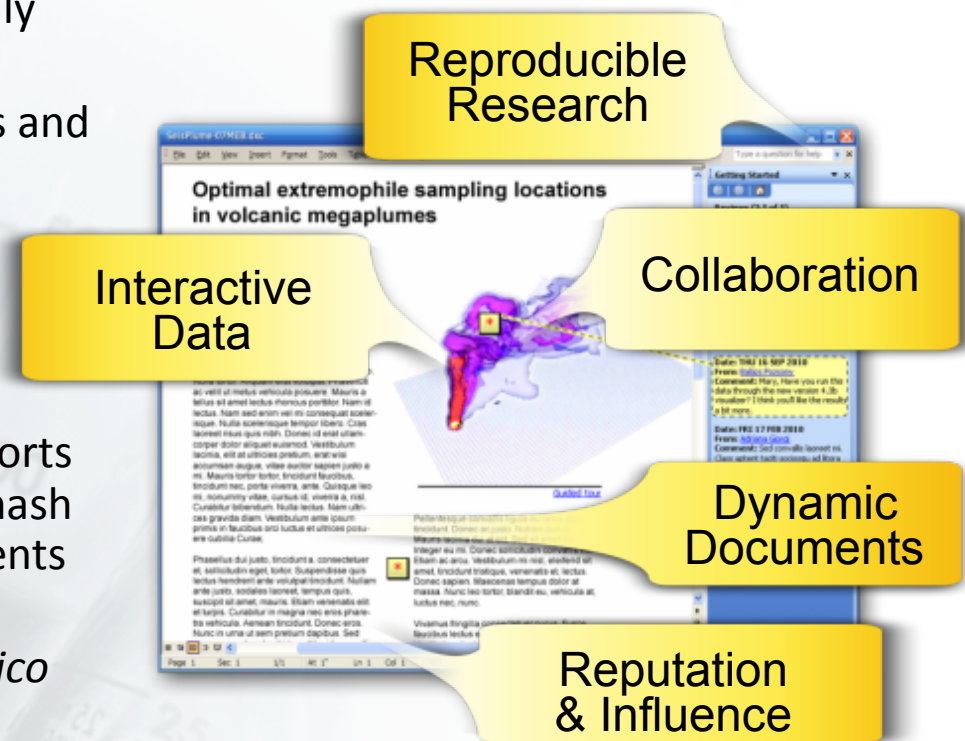
This work is licensed under a [Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



Envisioning a New Era of Research Reporting

Imagine...

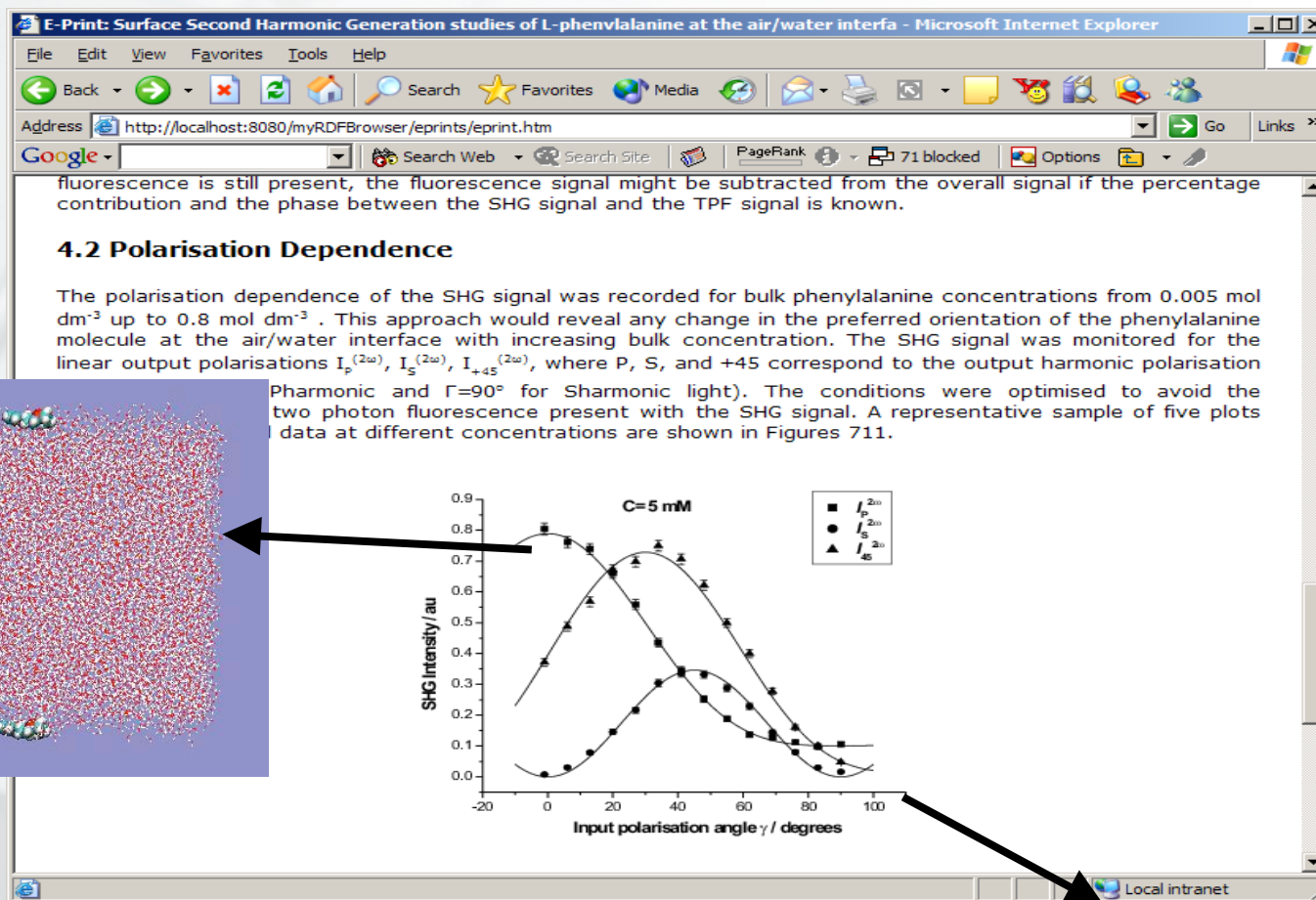
- Live research reports that had multiple end-user 'views' and which could dynamically tailor their presentation to each user
- An authoring environment that absorbs and encapsulates research workflows and outputs from the lab experiments
- A report that can be dropped into an electronic lab workbench in order to reconstitute an entire experiment
- A researcher working with multiple reports on a Surface and having the ability to mash up data and workflows across experiments
- The ability to apply new analyses and visualizations and to perform new *in silico* experiments



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



Publications as Live Documents



Link to simulation software and data in archive

Link to data, follow links back to the raw data archive



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



New Forms of Peer Review

[Home](#) [Feedback](#) [Support](#) [Log on / Register](#)

 **FACULTY OF 1000 BIOLOGY**
MAJOR ADVANCES. EXPERT OPINIONS.

[Tell your librarian what you think of F1000 >](#) [Customize this site to your areas of interest >](#) [Receive regular email alerts >](#)

[Subscription Info](#) [Institutional Access](#) [Free Trial](#) [Take a Tour](#) [Hidden Jewels](#) [Advanced Search](#) [About](#)

What is Faculty of 1000 Biology?

Faculty of 1000 Biology is a new online research tool that highlights the most interesting papers in biology, based on the recommendations of over 1000 leading scientists.

5 reasons to subscribe to F1000 Biology

- 1 F1000 Factor
- 2 Selection by peers
- 3 Systematic coverage
- 4 Expert opinion
- 5 Hidden Jewels

Find out more by taking a tour of F1000 Biology

sample evaluation


F1000 Factor 6.6

New Finding

1 **Cadherin-mediated cell sorting not determined by binding or adhesion specificity.**
Niessen CM, Gumbiner BM
J Cell Biol 2002 **156**(2):389-399 [[abstract on PubMed](#)]
[[order article](#)]

2 **Selected by** | Vance Lemmon / Kathleen Green / Nick Brown
First evaluation 30 Jan 2002 | Latest evaluation 12 Feb 2002

3 **Relevant Sections**
CELL BIOLOGY > Cell adhesion
DEVELOPMENTAL BIOLOGY > Developmental molecular mechanisms
NEUROSCIENCE > Neurodevelopment | Neuronal & glial cell biology
STRUCTURAL BIOLOGY > Cell signalling & trafficking structures

Faculty Member **4** **Comments**

Vance Lemmon
Case Western Reserve University, United States
CELL BIOLOGY
This paper challenges long held ideas in the cadherin field and consequently calls for a reevaluation of our concepts of tissue sorting. Cadherins are generally believed to have only one or two binding partners, yet this paper shows that cells expressing different classic cadherins bind equally well to different cadherin substrates. Nonetheless, they showed different sorting behaviors. So it appears that, like axonal growth and guidance, cell sorting is not necessarily mediated by the strength or ligand binding specificities of cell adhesion molecules.
Evaluated 12 Feb 2002

Kathleen Green
Northwestern University Medical School, United States
CELL BIOLOGY
This interesting paper challenges the existing dogma that most type I classical cadherins exhibit homophilic adhesive specificity. Using adhesion flow assays, the authors demonstrated that CHO lines expressing various classical cadherins exhibit comparable adhesion to purified cadherin extracellular domains regardless of their identity.

[Log on..](#) Log On
 **Free Trial**
 Subscriptio
[ac.uk](#)
[l.edu](#)
[ac.nz](#) Institution
Receive email
To receive ema
and news about
1000 Biology, e
email into the b

Tagging for Researchers

The screenshot displays the Connotea web application. At the top, a search bar is set to 'This collection' with a 'Find results' button. The Connotea logo and tagline 'Organize. Share. Discover.' are prominent. A navigation menu includes links for Home, Latest News, About Connotea, Site Guide, and Community pages. On the right, there are buttons for 'Sign up' and 'Log in'. The main content area is divided into several sections:

- crmc's tags:** A sidebar with a 'Find:' input field and a list of tags including [algorithms](#), [AMD](#), [applications](#), [architecture](#), [aurora07](#), [book](#), [cache](#), [CCR](#), [cell](#), [ClearSpeed](#), [compiler](#), [CRMC](#), [cyclops](#), [DAMP2007](#), [dataflow](#), and [dataparallel](#).
- Free online reference management for clinicians and scientists:** A featured article with a 'Sign up now' button.
- Bookmarks by "crmc":** A list of bookmarks with a 'Sign up now' button and links to 'Learn more' and 'Watch a short video (2m 41s)'.
- Toolbox:** A sidebar with links for 'Add a bookmark', 'Create a new group', 'Create a tag note', 'Rename a tag', 'Import from local file', 'Export my library', and 'Report a problem'.
- crmc's bookmarks:** A section showing that the user 'crmc' does not have a profile yet, with a link to 'Community Pages' and a pagination control for 'Number of bookmarks per page' (10 | 25 | 50 | 100).
- General-Purpose Computing on Graphics Processing Units:** A Wikipedia review of programming graphics unit, posted by 'crmc' to 'StreamProgramming CRMC gpu' on 'Tue Feb 06 2007' at 22:11 UTC.
- Multi-Core to the masses:** A link to a post by Justin Rattner.

The bottom of the image shows a Windows taskbar with the 'start' button, several open applications (Connotea, Microsoft Office), a search bar, and system icons including a 97% battery level and a clock showing 2:30 PM.

Lab Notebooks as Blogs?



A record of an experiment that failed...

- Publishable?
- Useful?

With thanks to Jeremy Frey



New Tools: Disseminate and Share

Data
Acquisition
and Modeling

Collaboration
and
Visualization

Analysis and
Data Mining

Disseminate
and Share

Archiving and
Preservation

Chem4Word

- Chemistry Drawing in Word
- Created in collaboration with University of Cambridge; Peter Murray-Rust, et.al.

Intent: Recognizes chemical dictionary and ontology terms

Data: Semantics stored in Chemistry Markup Language

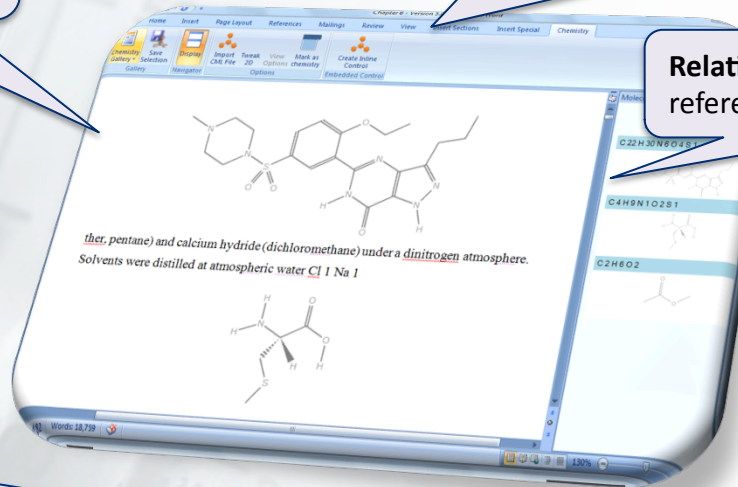
```
<?xml version="1.0" ?>
<cmi version="3" convention="org-synth-report" xmlns="http://www.xml-cml.org/schema">
  <molecule id="m1">
    <atomArray>
      <atom id="a1" elementType="C" x2="2.9149999818530273" y2="0.7699999809265137" />
      <atom id="a2" elementType="C" x2="1.5813208400249916" y2="1.5399999809265137" />
      <atom id="a3" elementType="O" x2="0.24764171819695613" y2="0.7699999809265134" />
      <atom id="a4" elementType="O" x2="1.5813208400249912" y2="3.0799999809265137" />
      <atom id="a5" elementType="H" x2="4.248679083681063" y2="1.5399999809265137" />
      <atom id="a6" elementType="H" x2="2.914999981853028" y2="-0.7700000190734864" />
      <atom id="a7" elementType="H" x2="4.248679083681063" y2="-1.9073486456910878" />
      <atom id="a8" elementType="H" x2="1.0860374036310796" y2="1.5399999809265132" />
    </atomArray>
    <bondArray>
      <bond atomRefs2="a1 a2" order="1" />
      <bond atomRefs2="a2 a3" order="1" />
      <bond atomRefs2="a2 a4" order="2" />
      <bond atomRefs2="a1 a5" order="1" />
      <bond atomRefs2="a1 a6" order="1" />
      <bond atomRefs2="a1 a7" order="1" />
      <bond atomRefs2="a3 a8" order="1" />
    </bondArray>
  </molecule>
</cmi>
```

Author/edit 1D and 2D chemistry.
Change chemical layout styles.

Relationships: Navigate and link referenced chemistry

Intelligence: Verifies validity of authored chemistry

States license.



New Tools: Archiving and Preservation

Data
Acquisition
and Modeling

Collaboration
and
Visualization

Analysis and
Data Mining

Disseminate
and Share

Archiving and
Preservation

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

Zentity

Native support for RSS, OAI-PMH,
OAI-ORE, AtomPub and SWORD



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



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



Research Repositories

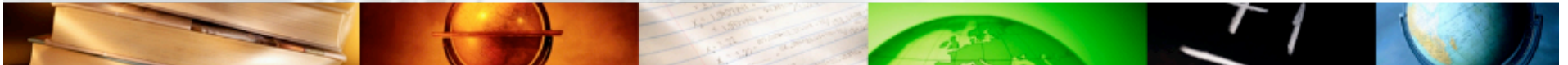


The Michigan Daily, September 24, 2007

- 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 percent per year since 2003. National social science journals increased 9.2 percent and national science journals increased by 8.3 percent



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



Current Scholarly Publishing Model is in Crisis

- Journal subscriptions rising faster than library budgets
 - Cancelling subscriptions, no freedom for new journals in new and emerging fields
- Web technology and digital media now make dissemination of knowledge ‘easy’ and ‘free’ without the traditional paper journals
 - Similar dilemma to that of the music industry with MP3 and ‘free’ digital copies



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

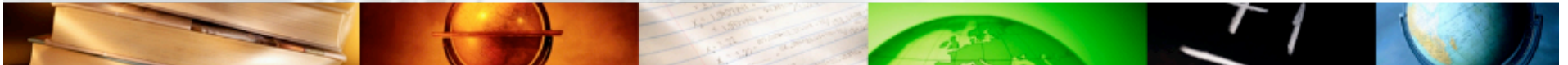


Open Access Repositories and OA Policies

- **Subject Repositories**
 - PubMedCentral, arXiv, RePEc, DBLP ...
 - Proposed Federal Research Public Access act
- **Institutional Repositories and OA Policies**
 - Registry of OA Repositories <http://roar.eprints.org/>
 - Over 100 OA policies
 - Harvard, MIT, University of California ...
 - Over 22M items in over 1000 OA repositories worldwide
- **Open Access Models**
 - Gold (author pays) – e.g. PLoS
 - Green (self deposit) – e.g. arXiv
- **Protocols**
 - OAI-PMH, OAI-ORE, SWORD, etc.



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

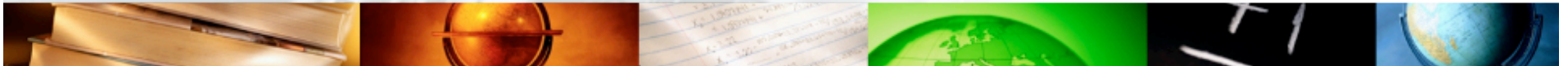


Research Repositories

- Repositories will contain not only full text versions of research papers but also 'grey' literature such as workshop papers, presentations, technical reports and theses
 - In the future repositories will also contain data, images and software
- Federated databases of scientific information and cross database search tools
 - NIH National Library of Medicine
 - WorldWideScience.org



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



Virginia Tech: Electronic Theses and Dissertations

Requests for ETDs grew from around 220,000 in 1997/98 to nearly 20M by 2006/07



VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

10,595 Electronic Theses and Dissertations

including available, restricted and withheld ETDs



<http://scholar.lib.vt.edu/theses/>

FOR ETD AUTHORS

- GUIDES
 - [Copyright](#)
 - [ETD Examples](#)
 - [General Information for Authors](#)
 - [Graduate School Catalog](#)
 - [Graduate School Guidelines](#)
 - [Preparing your VT ETD](#)
- LEARNING TOOLS
 - [Adobe PDF Tutorial VT Workshop](#)

VT ETDs

- [Browse VT ETDs](#)
- [Search VT ETDs](#)
- [Federated Search](#) search multiple ETD libraries simultaneously through the ETD OAI Union Catalog (not maintained by DLA)

Find out about ETDs

- [Facts, Data, Information](#) including archiving, servers, staffing, use, etc.
- [Survey Results: Graduates Whose ETDs Are Accessible Worldwide](#) (556 kb)
- [Presentations and Publications about VT ETDs](#)

User Survey Form

Share your opinions

NDLTD

*Networked Digital Library of
Theses and Dissertations*



This work is licensed under a
[Creative Commons Attribution 3.0 United States License](#).



Commentary

This is one of a series of commentaries on the future of scientific publishing. For a listing of the other commentaries, see <http://www.jneurosci.org/cgi/content/full/26/36/9077>.

As We May Read

Paul Ginsparg

Departments of Physics and Information Science, Cornell University, Ithaca, New York 14853

The e-print arXiv (<http://arXiv.org/>), initiated in August 1991, has effectively transformed the research communication infrastructure of multiple fields of physics and could play a prominent role in a unified set of global resources for physics, mathematics, and computer science. It has grown to contain >375,000 articles (as of July 2006), with >50,000 new submissions expected in calendar year 2006 and >40,000,000 full-text downloads per year. It is an international project, with dedicated mirror sites in 17 countries and

ders of magnitude. Even with the majority of science research journals now on-line, researchers continue to enjoy both the benefits of the rapid availability of the materials, even if not yet reviewed, and open archival access to the same materials, even if held in parallel by conventional publishers. The methodology works within copyright law, as long as the depositor has the authority to deposit the materials and assign a nonexclusive license to distribute at the time of deposition, because such a li-

helps ensure that the arXiv remains a forum for communication among research professionals, not a mechanism for outsiders to communicate to that community. Additionally, a small group of volunteer "moderators," consisting of interested experts from around the world, cursorily prescans new submissions, typically only at the level of title and abstract, for appropriateness to the proposed primary subject area.

The arXiv repository functions are flexible enough either to coexist with the

Paul Ginsparg is the creator of arXiv, an open access repository for pre-publication of much of the physics and astronomy literature

Published in the Journal of Neuroscience,
September 20, 2006



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



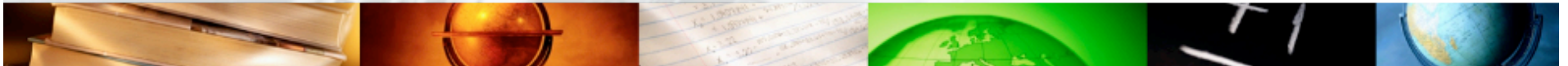
Ginsparg's Conclusions?

“On the one-decade time scale, it is likely that more research communities will join some form of global unified archive system without the current partitioning and access restrictions familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge.”

“Ironically, it is also possible that the technology of the 21st century will allow the traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research Enterprise.”



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



Cloud Computing



Cloud Computing: A Microsoft View

Another Definition:

- Cloud Computing means using a remote data center to manage scalable, reliable, on-demand access to applications
- Providing Applications and Infrastructure over the Internet
- Scalable means:
 - Possibly millions of simultaneous users of the app.
 - Exploiting thousand-fold parallelism in the app.
- Reliable means on-demand, continuous availability
- Applications span the continuum from the Client to the Cloud



Three New Aspects to Cloud Computing:

- Illusion of infinite computing resources available *on demand*
- Elimination of an upfront commitment by cloud users
- Ability to *pay for use* of computing resources on a short-term basis as needed



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



Cloud Data Centers: Economies of Scale

- Data Centers range in size from “edge” facilities to megascale (100K to 1MK servers)
- Offer real economies of scale
 - Approximate costs for a small size center (1K servers) and a larger, 400K server center.

Technology	Cost in small-sized Data Center	Cost in Large Data Center	Ratio
Network	\$95 per Mbps/ Month	\$13 per Mbps/ month	7.1
Storage	\$2.20 per GB/ Month	\$0.40 per GB/ month	5.7
Administration	~140 servers/ Administrator	>1000 Servers/ Administrator	7.1



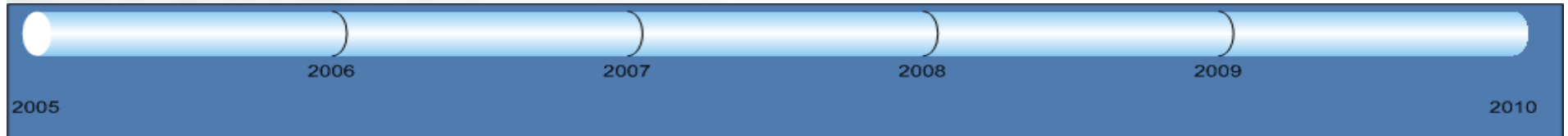
Data Center estimates from James Hamilton



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



Microsoft's Datacenter Evolution



Datcenter Co-Location
Generation 1

Quincy and San Antonio
Generation 2

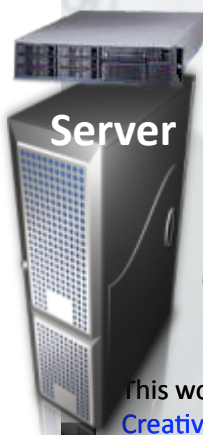
Chicago and Dublin
Generation 3

Modular Datacenter
Generation 4



Facility PAC

Deployment Scale Unit



Server

Capacity



Rack

*Density
and Deployment*



Containers

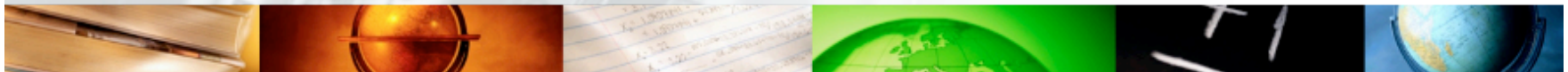
*Scalability and
Sustainability*



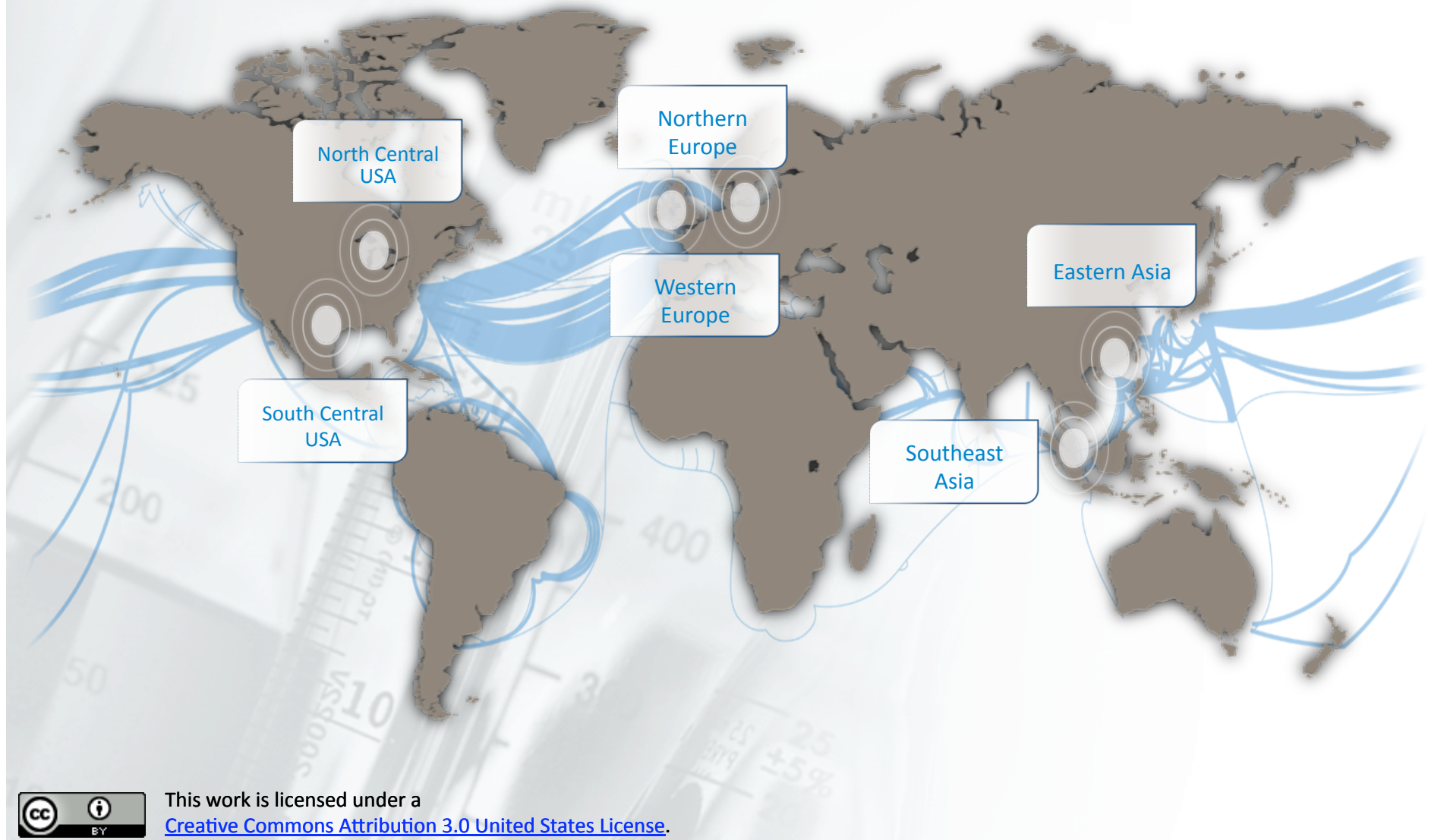
IT PAC

*Time to Market
Lower TCO*

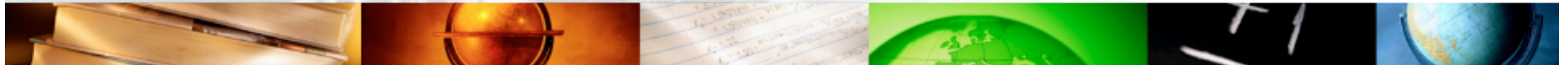
This work is licensed under a [Creative Commons Attribution-United States License](#).



Windows Azure Platform Availability

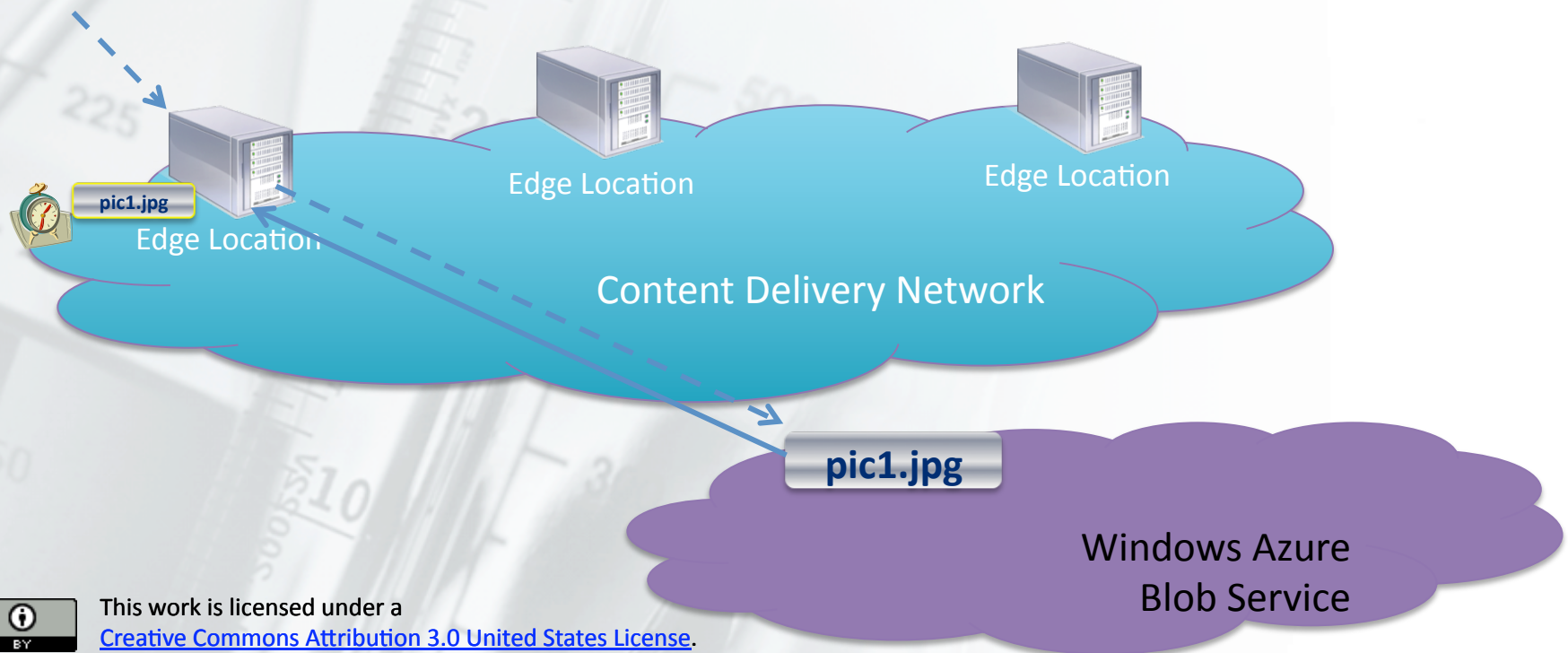


This work is licensed under a
[Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).

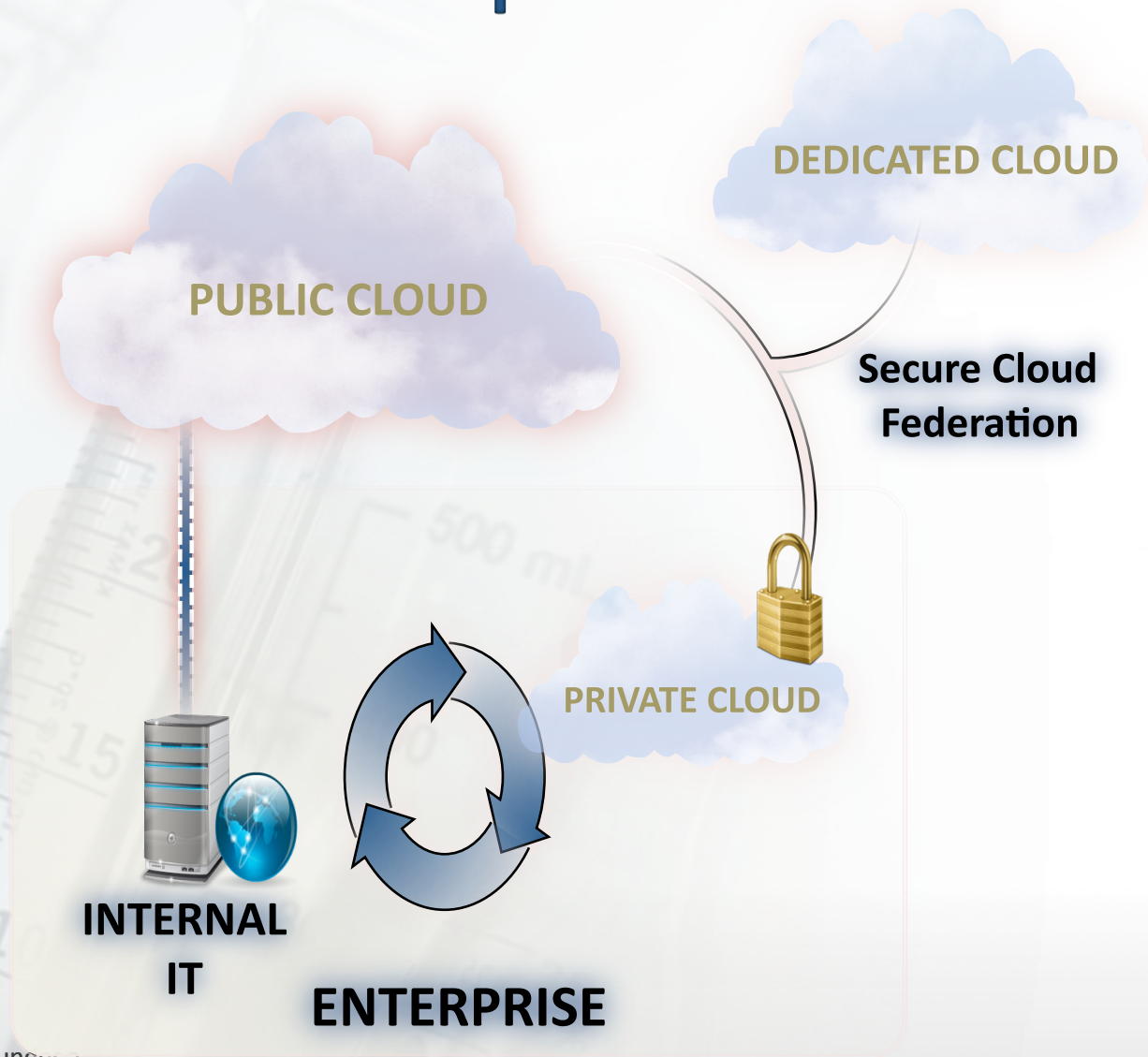


Azure Content Delivery Network

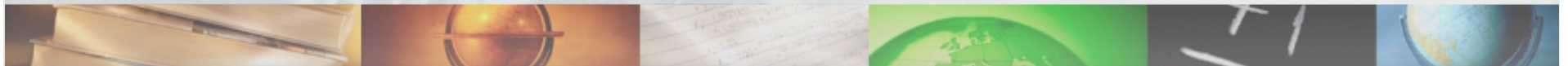
- Better performance and user experience by caching Azure blobs at strategically placed locations
 - 18 locations globally (United States, Europe, Asia, Australia and South America) and growing



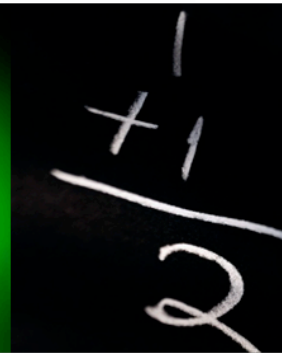
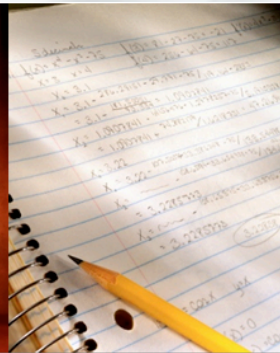
Cloud Options



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



Towards a Semantic Future



Today...

Computers are
great **tools** for

storing

computing

managing

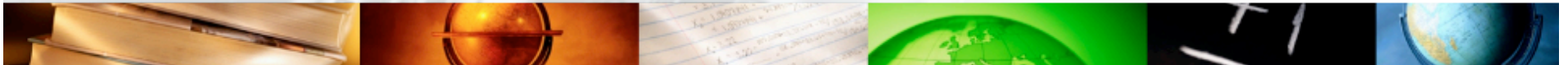
indexing

huge amounts
of **data**

For example, Google and Microsoft both have
copies of the entire Web for indexing purposes

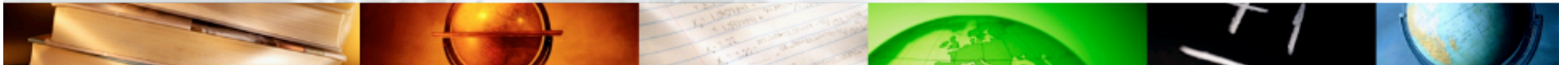
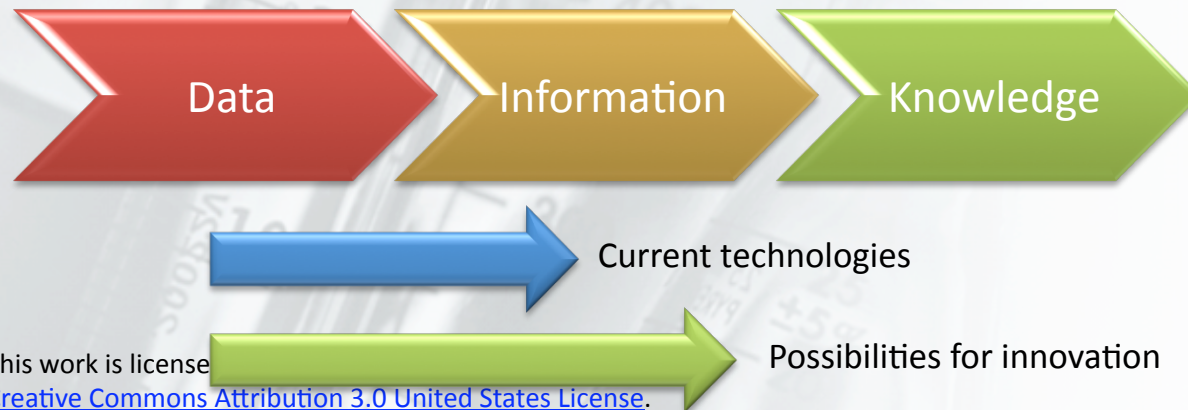


This work is licensed under a
[Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



Need for Semantic Computing?

- Semantic computing combines concepts and technologies that
 - Enable data modeling
 - Capture relationships
 - Allow communities to define ontologies
 - Exploit machine learning
- Will empower computers to reason about the data



Tomorrow...

Computers will still
be great **tools** for



huge amounts
of **data**

We would like
computers to also
help with the
automatic



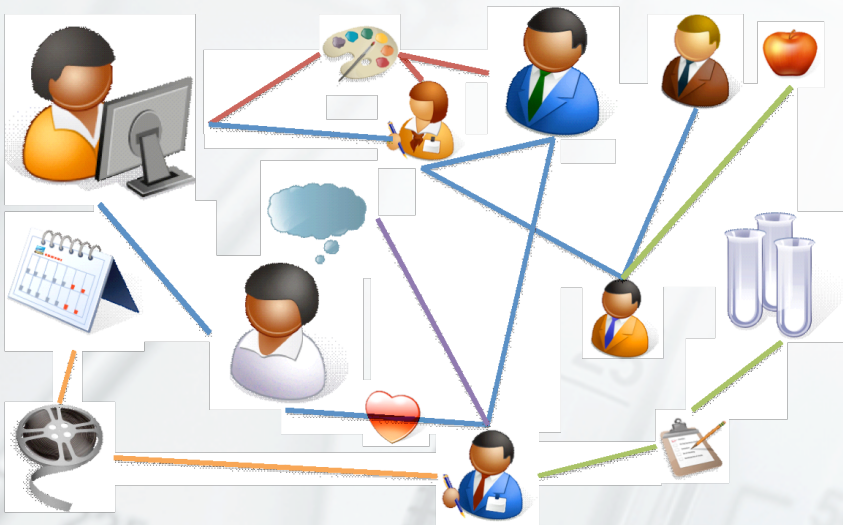
of the world's
information



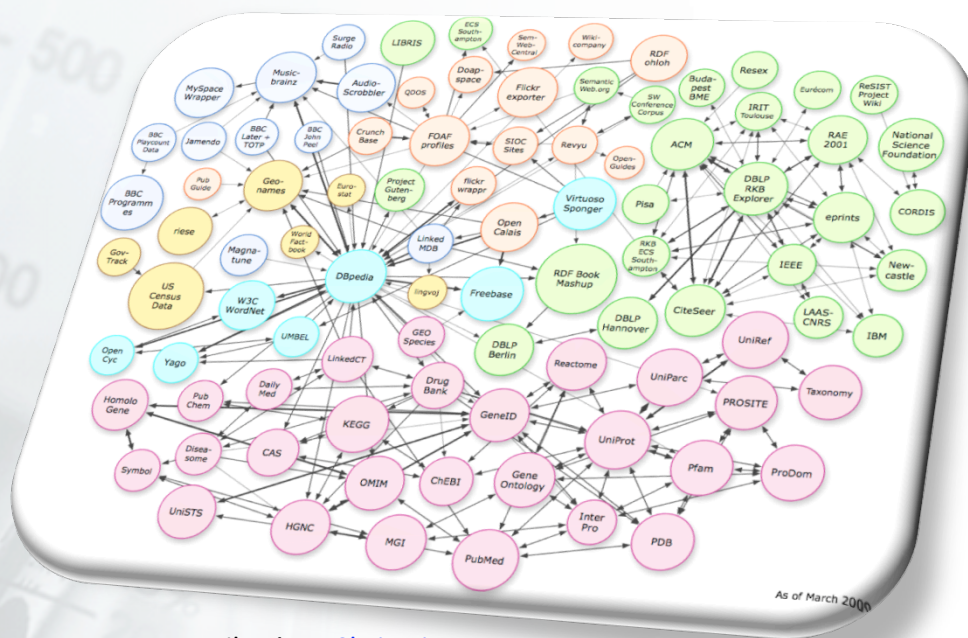
This work is licensed under a
[Creative Commons Attribution 3.0 United States License](https://creativecommons.org/licenses/by/3.0/us/).



Moving to a world where all data is linked ...



- A knowledge ecosystem:
 - A richer authoring experience
 - An ecosystem of services
 - Semantic storage
 - Open, Collaborative, Interoperable, and Automatic



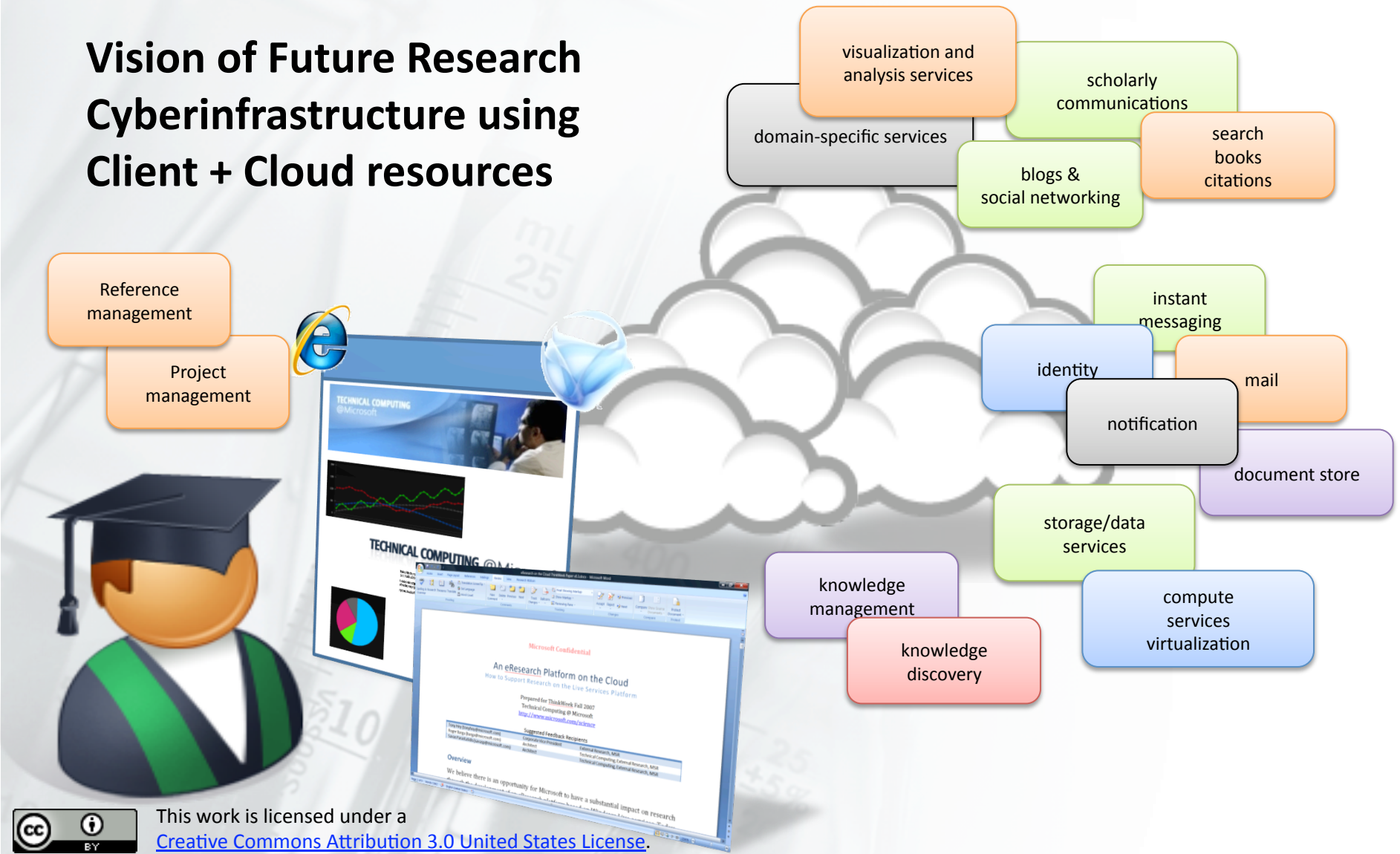
This work is licensed under a [Creative Commons Attribution 3.0 United States License](#).

Attribution: [Chris Bizer](#)



... and can be stored/analyzed in the Cloud

Vision of Future Research Cyberinfrastructure using Client + Cloud resources



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



research.microsoft.com/en-us/collaboration/tools

Dryad; DryadLINQ

Windows HPC release

Computational Biology Toolkit

Enables and accelerates fundamental advances in biology

F#

Collaboration with the academic and research community on F#'s typed functional and object-oriented programming on the .NET platform

Plug-ins for Office

Ontology Add-in for Word

Article Authoring Add-in for Word

Chem4Word – Chemistry Drawing in Word

Microsoft Electronic Journals Service

Open XML Document Viewer

Software Engineering Tools

Spec#: Program verifier for C# extended with design by contract

VCC: Program verifier for Concurrent C

PEX: automatic unit testing tool for .NET

CHESS: Unit testing tools for concurrent Win32 executable and .NET

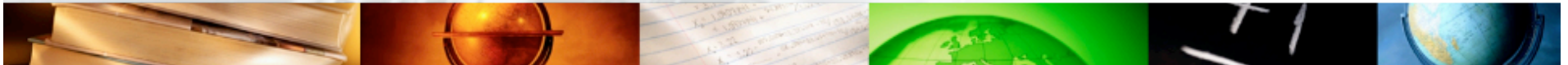


Windows Azure:



This work is licensed under a

<http://creativecommons.org/licenses/by/3.0/> United States License.





Microsoft[®]

Your potential. Our passion.[™]



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

