Microsoft® Research

Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Microsoft® Research
Faculty Summit 2010
Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Scaling Science to the Cloud

Tony Hey
Corporate Vice President
Microsoft Corporation

Scientific Discovery and Understanding

Huge opportunities for insight and innovation

*through ‘**scaling**’ our research capabilities*

Scaling science

- Length scales: from **infinitesimal** to **galactic**
- Research teams: from **individual** to **community**
- Timescales: from **instant** to **eon**
- Complexity: from **single source** to **webscale**
- Data: from **documents** to **digital libraries**

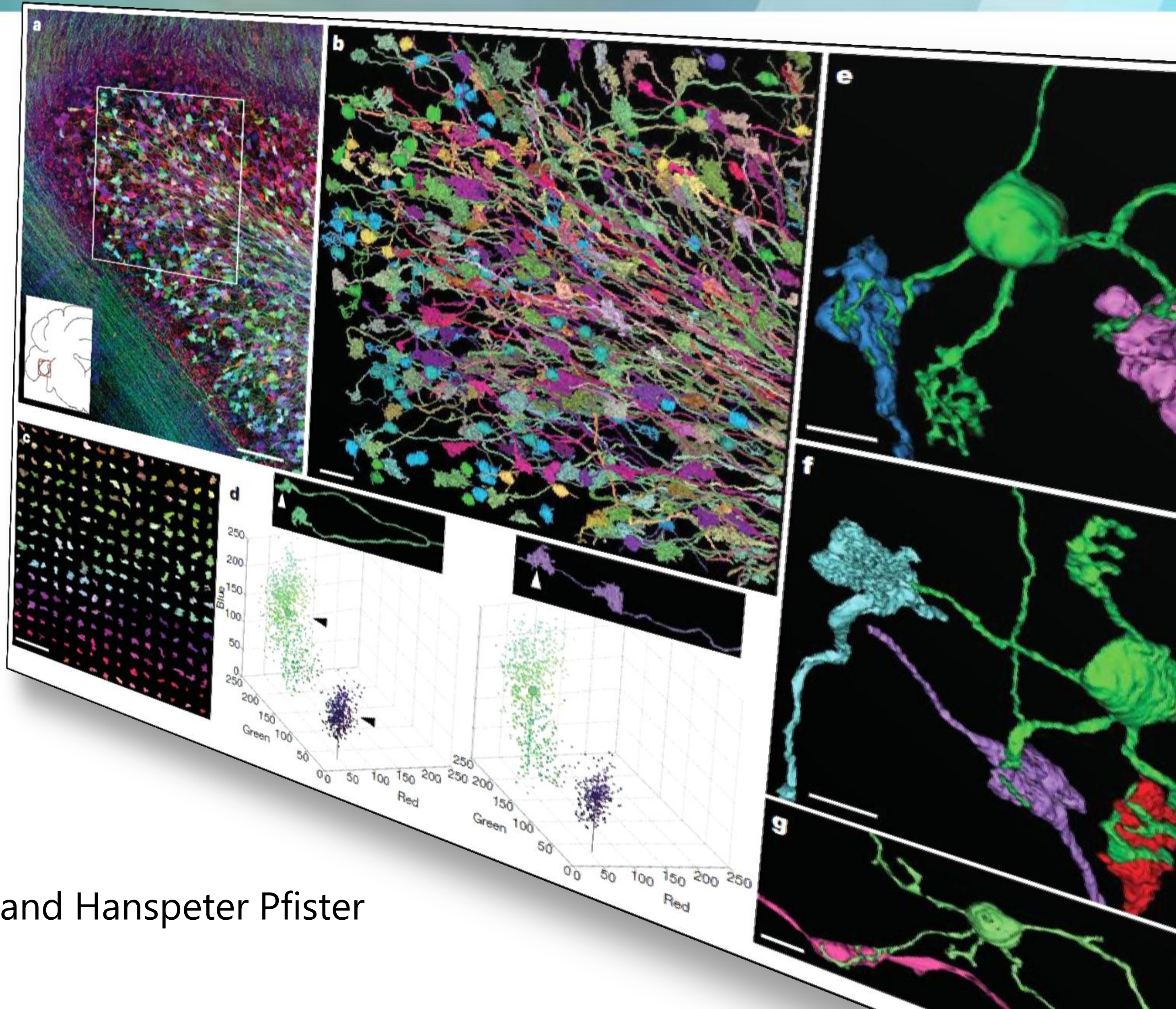
Length

from **infinitesimal** to **galactic**

Detailed neural circuitry of the brain - the infinitesimal

3D registration, segmentation, and complete neural circuit reconstruction

- Visualize and "fly through" 3D images of neural circuitry at least 100GBs
- Database design and management for storing and querying multi-TB datasets containing 3D microscopy images of neural circuits



Harvard University: Jeff Lichtman, Kenny Blum and Hanspeter Pfister
MSR: Michael Cohen

World Wide Telescope – the galactic

Seamless Rich Social Media Virtual Sky
Web application for science and education

- Science- Seamless integration of multi-wavelength, multiple telescope distributed image/data sets and one click contextual access to distributed web information/data sources
- Education- Easy as Powerpoint, rich social media authoring environment within the sky allowing astronomers, educators and kids to create and share rich narrated guided tours of the universe

www.worldwidetelescope.org

ID magazine *International Design Annual “Best in category; Interactive 2009”*

TIME magazine *“50 Best sites on the Internet 2009”*

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

Collaboration

from **individual** to **community**

Chasing HIV – to web scale analysis

InfoTech | Research

HECKERMAN at work in a lab at the University of Washington

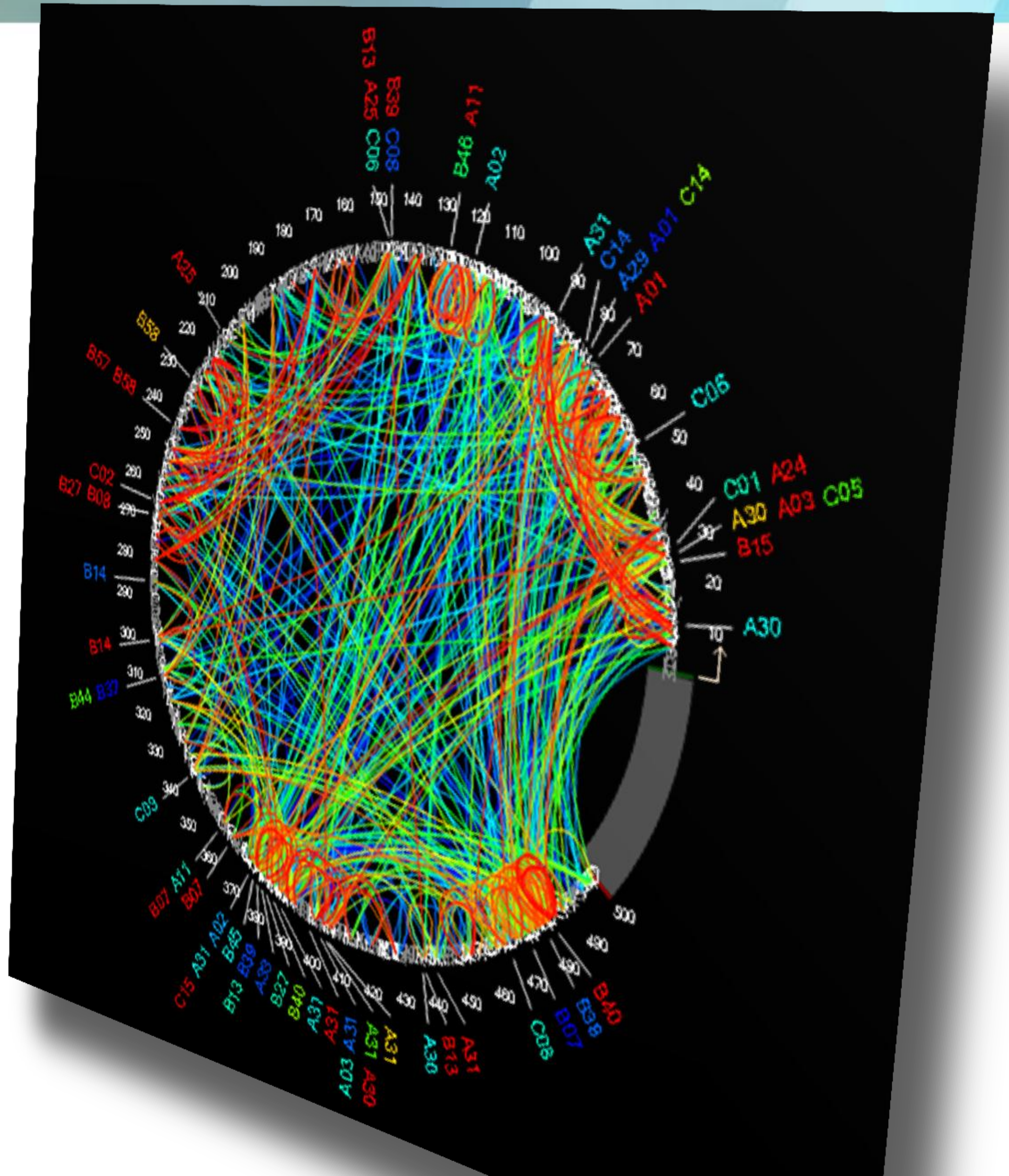
Using Spam Blockers To Target HIV, Too

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

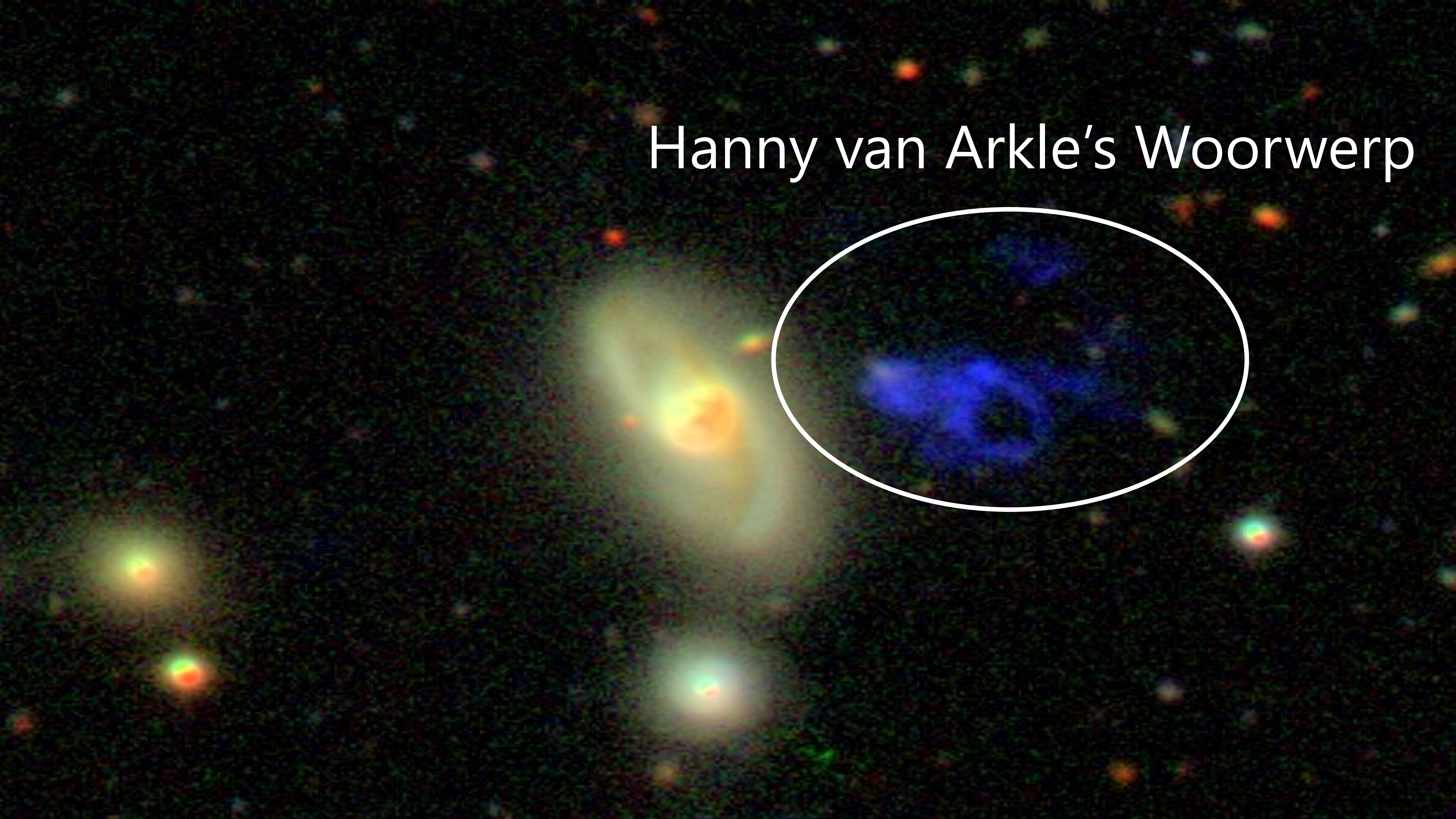
BY STEPHEN BAKER AND JAY GREENE

Similar mutations may crop up in computer and medical viruses

Tracking the evolution of HIV inside an individual using advanced machine-learning algorithms



http://www.galaxyzoo.org


Hanny van Arkle’s Woorwerp

“Green Peas”

Complexity

from **single source** to **webscale**

Carbo-Climate Synthesis

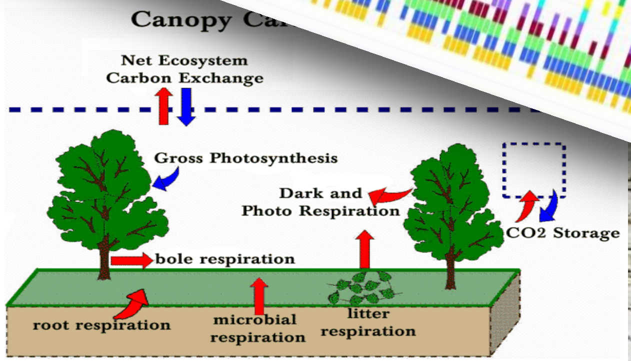
FLUXNET

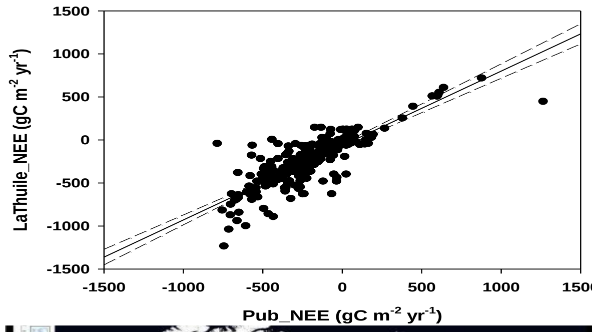
A Global Network

Understanding the global carbon cycle

- Measurements of CO2 in the atmosphere show 16-20% less than emissions estimates predict…the difference is either due to plants or ocean absorption.

- Cross site studies and integration with modeling increasingly important
  - 921 site-years of data
  - 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



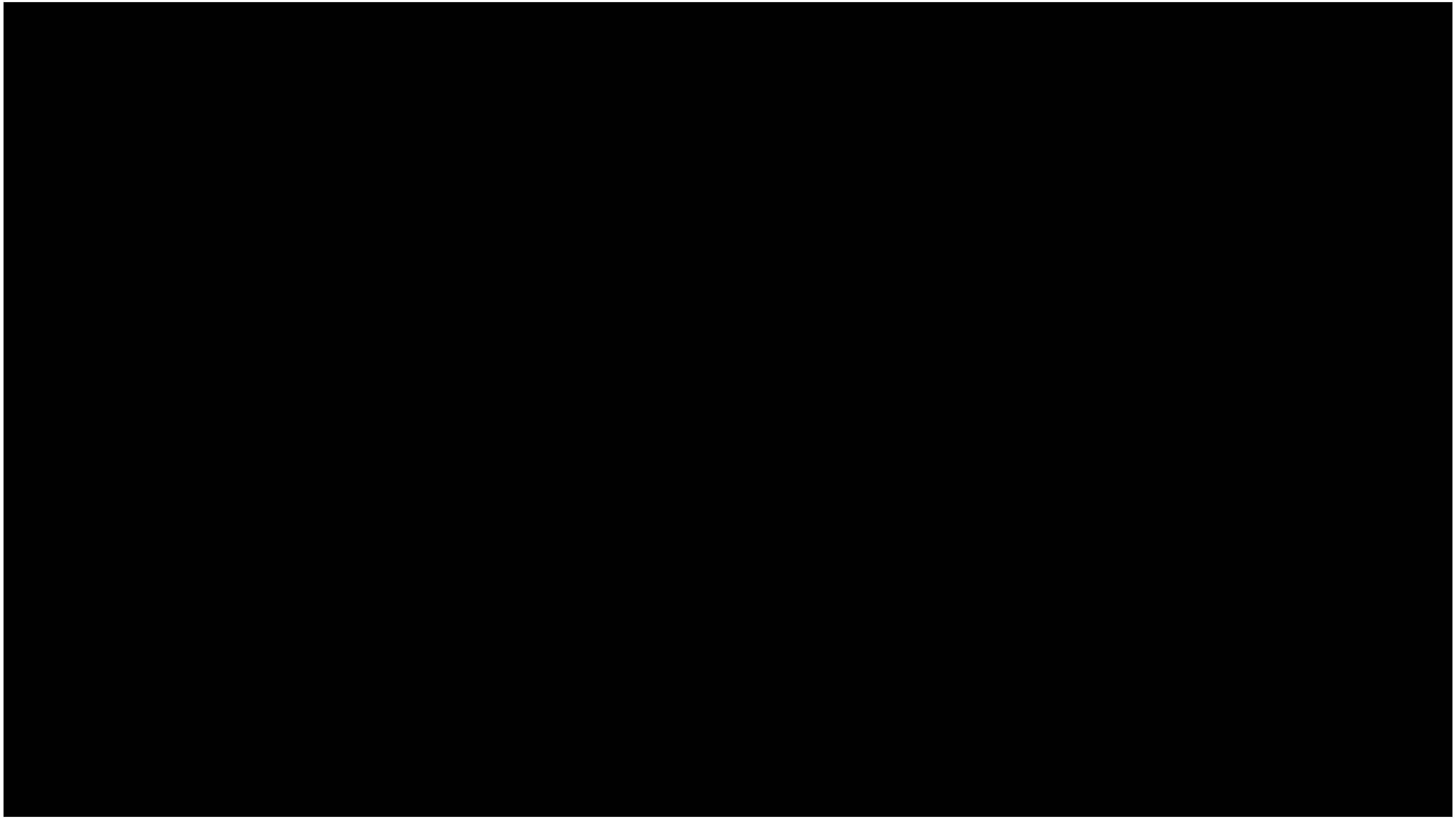


www.fluxnet.org

UC Berkeley: Dennis Baldocchi
Microsoft Research: Catharine van Ingen

Sensor Networks in Brazil

Sensor Networks in Brazil



Sensor Networks in Brazil

- Pilot project with **hundreds** of sensors
  - ‘Hostile environment’
  - Demonstrating good reliability
  - Key datasets for environmental science

- Now looking to scale this to **thousands**

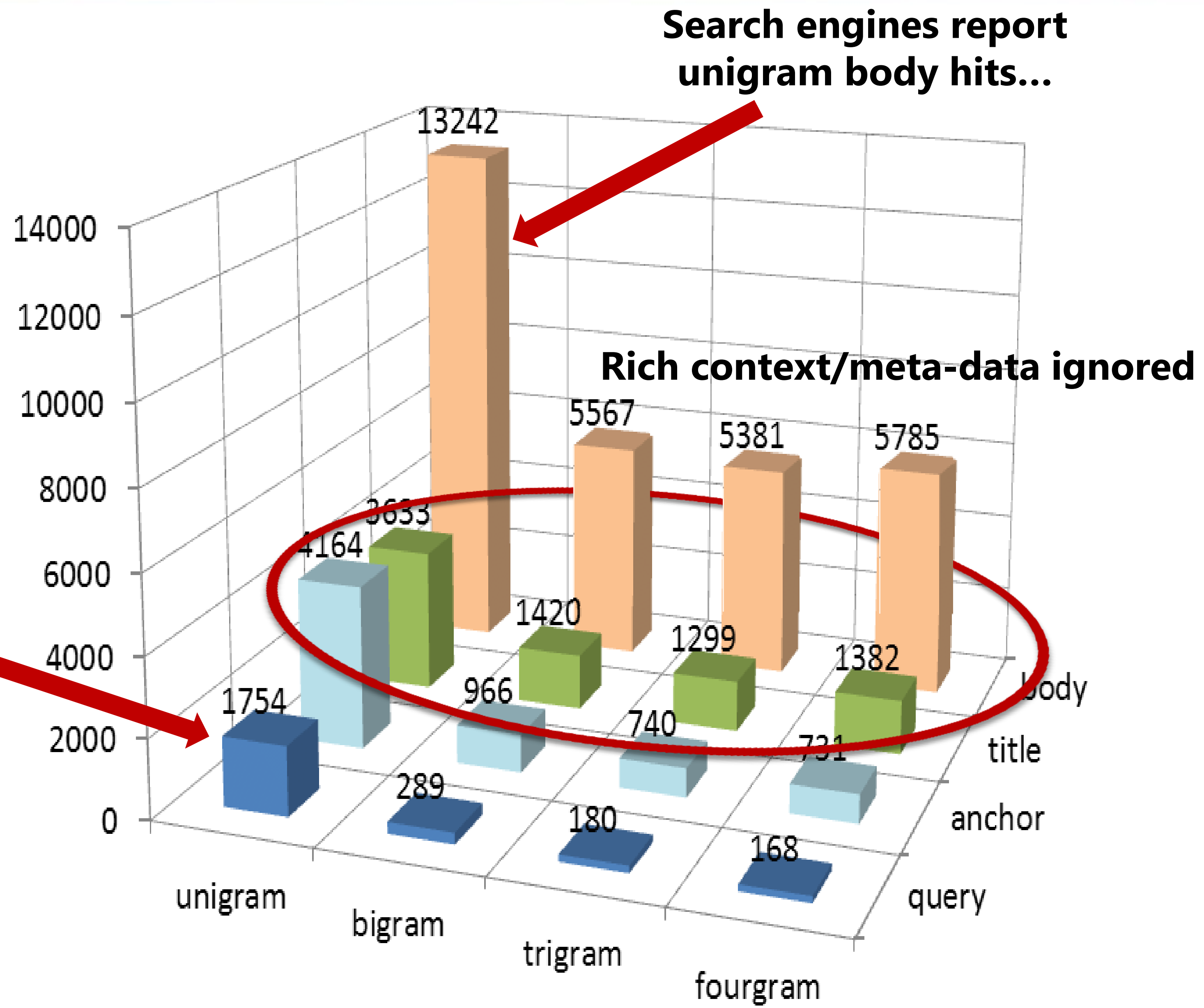
- Microsoft has experience of dealing with **webscale** data challenges…

Web N-Gram

Web data has structure…

…and that counts

(e.g. Body, Title, Anchor)



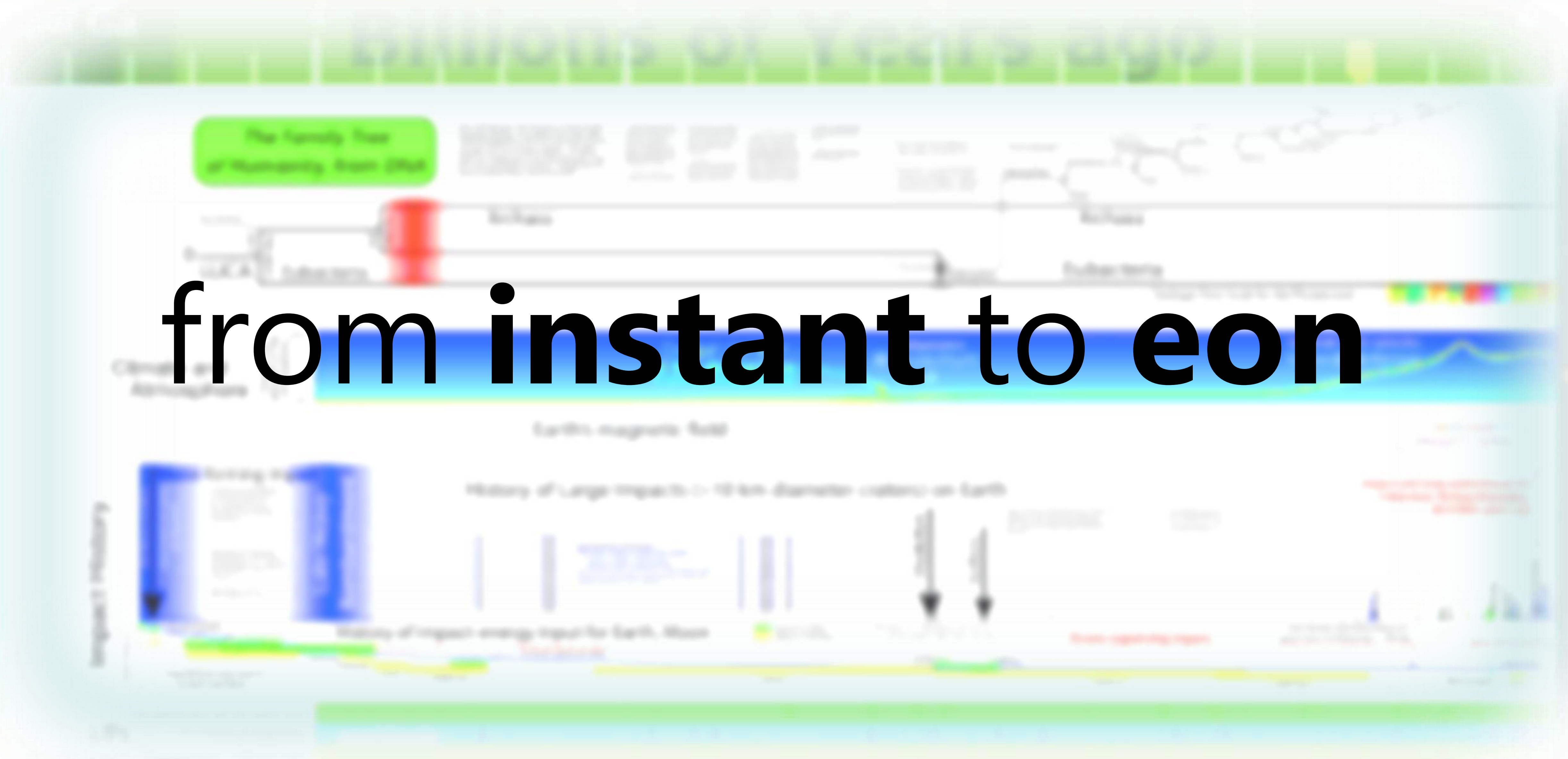
Multi-word Tag Cloud from Government Dataset Titles

Single Tag Cloud

Multi Tag Cloud

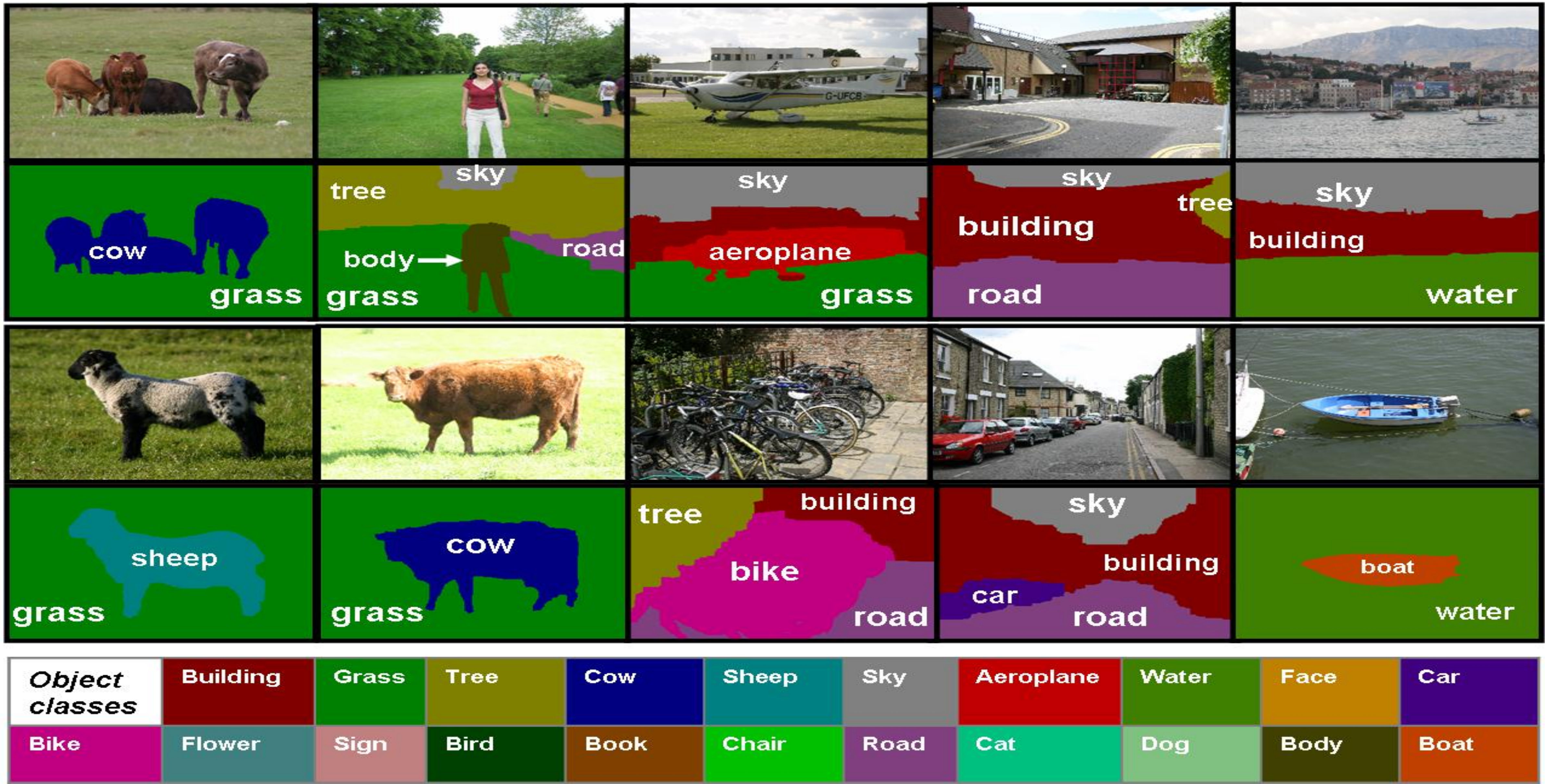
Ref: Dr. Li Ding, Rensselaer Polytechnic Institute

Time



from **instant** to **eon**

A moment in time – an instant



ChronoZoom – history in its broadest possible context

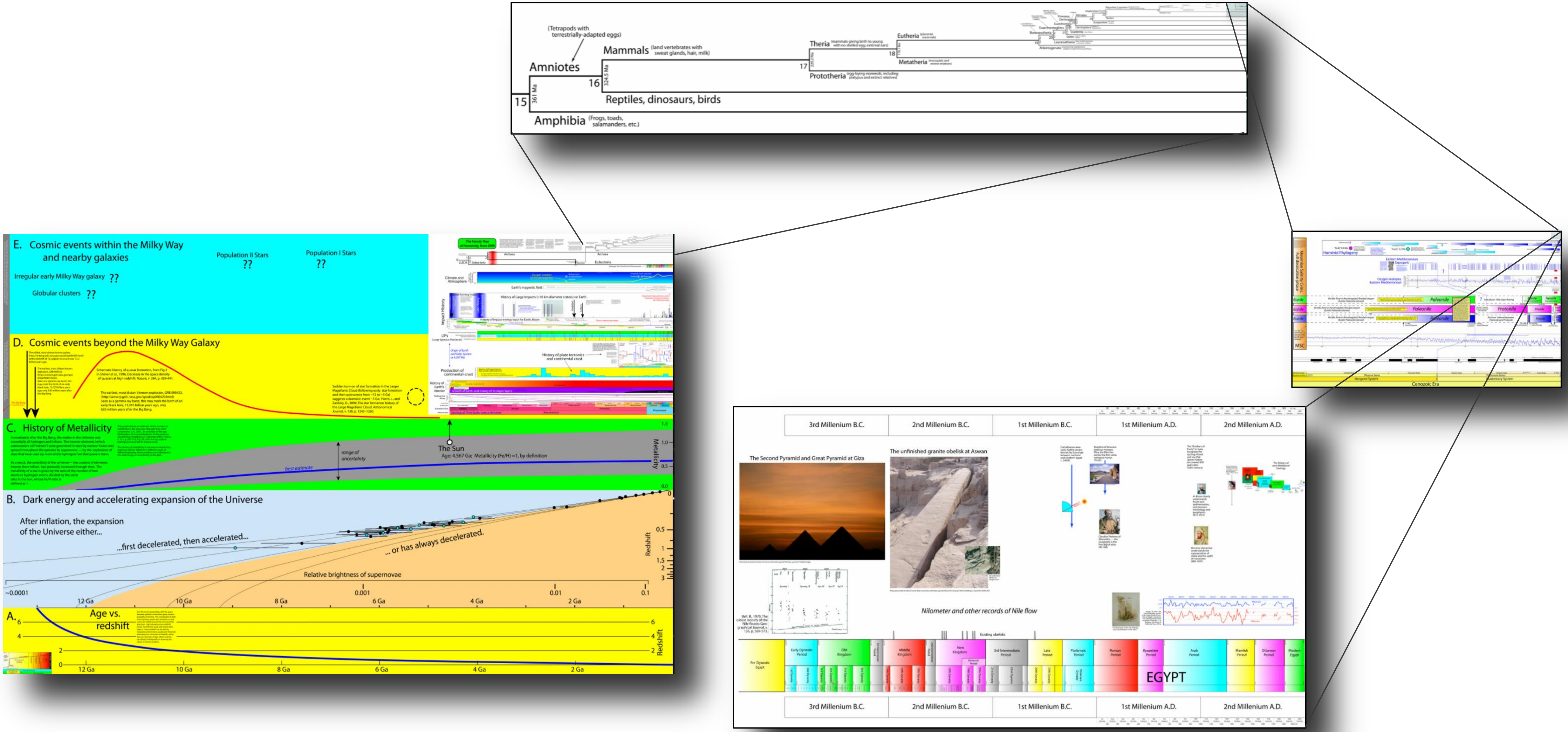
The challenge: exploration of all known time series, and smoothly transition from billions of years down to individual nanoseconds…

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 External Research and the Live Labs team.

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

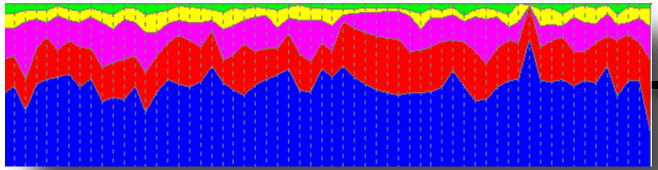
www.chronozoomtimescale.org

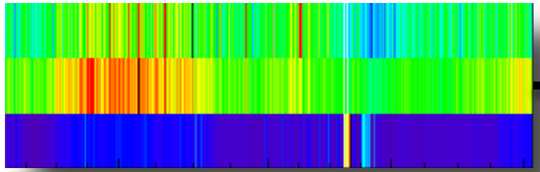
Given the history of the universe



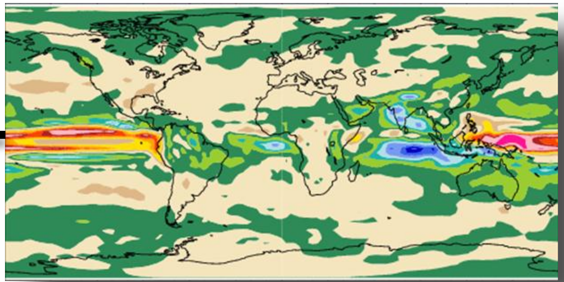
See the demo live at www.chronozoomtimescale.org Walter Alvarez with the support of Bill Crow and the Live Labs Seadragon team.

ChronoZoom: From the Dawn of Time to – Right Now





Insights & Innovation

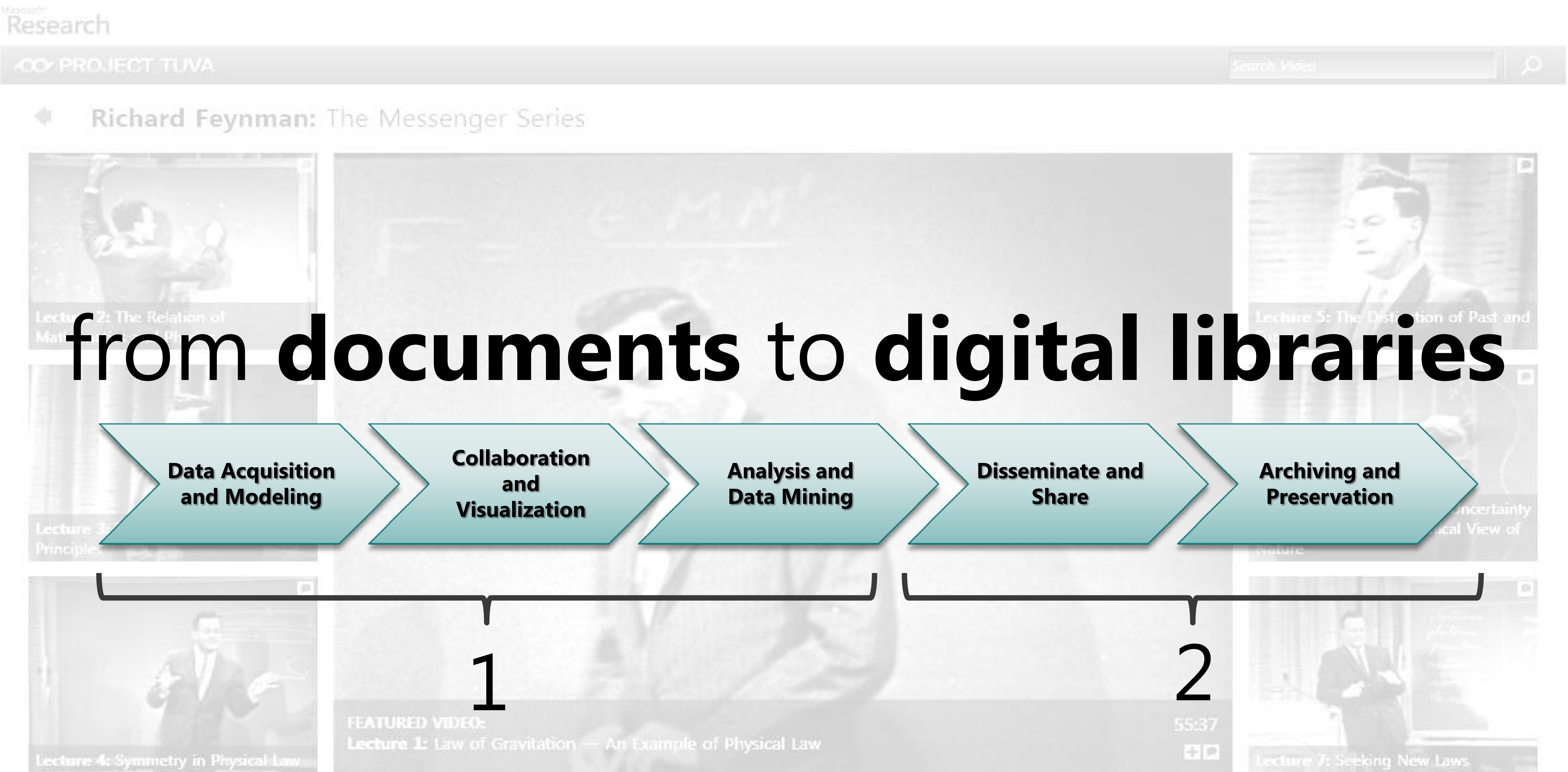


Our vision is to create an application that allows researchers to browse, overlay, and explore research both inside and outside of their specific expertise. Applications for such a tool are massively interdisciplinary and touch not only earth sciences but physics, genomics, astronomy, economics, history, biology, marketing, and just about any field that produces or consumes data that is somehow related to time.

Zoom [zoom]: an apt metaphor for the contextual age. Zoom is about the (dis)aggregation, exploration, and comparison of massive data sets at multiple scales. Zoom is the new search.

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

Data



from **documents** to **digital libraries**

Data Acquisition and Modeling → **Collaboration and Visualization** → **Analysis and Data Mining** → **Disseminate and Share** → **Archiving and Preservation**

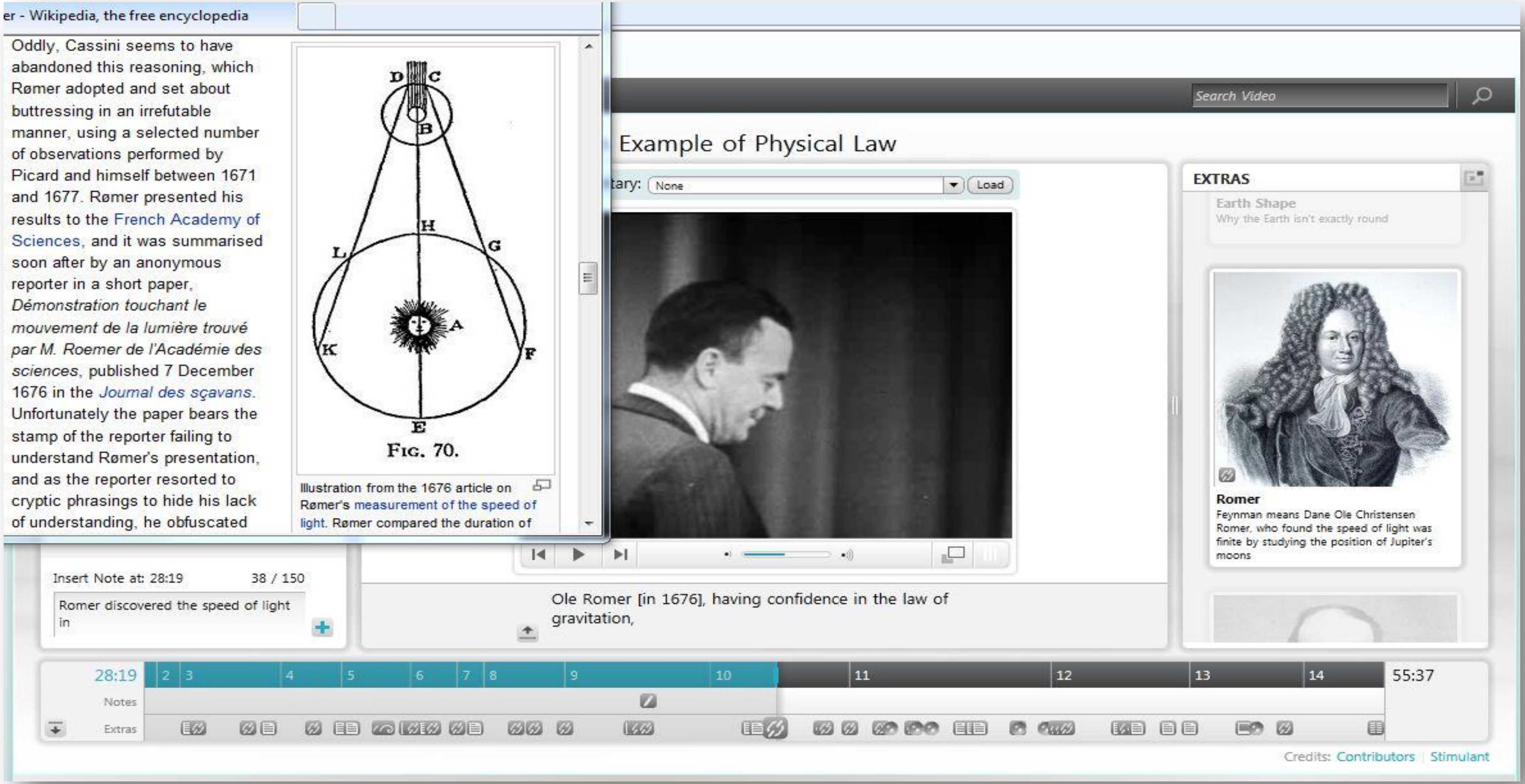
1

2

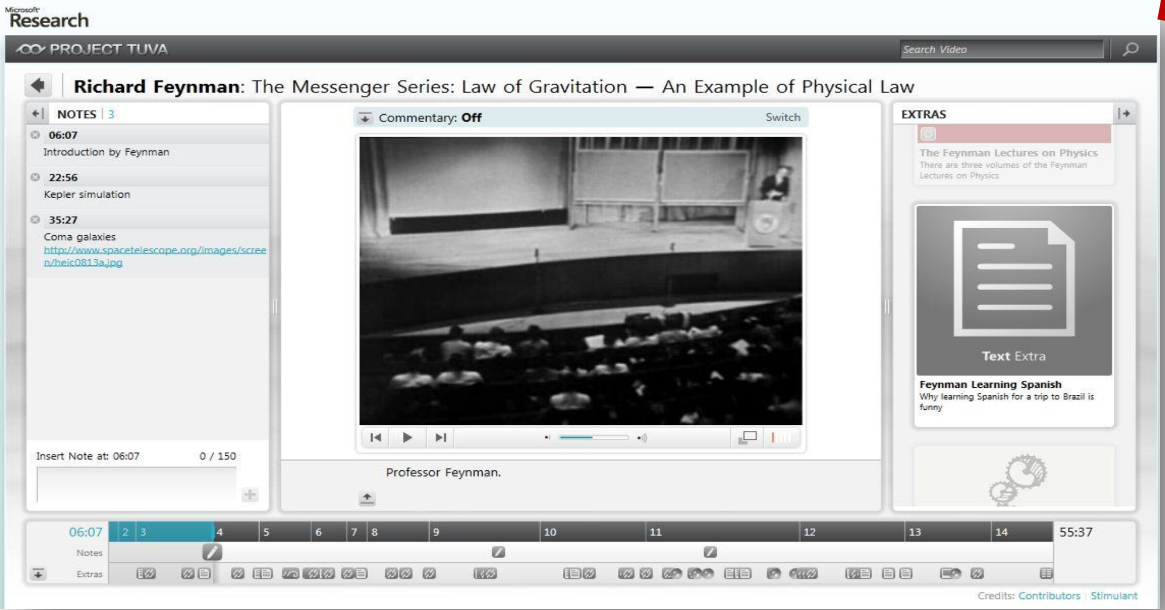
Project Tuva

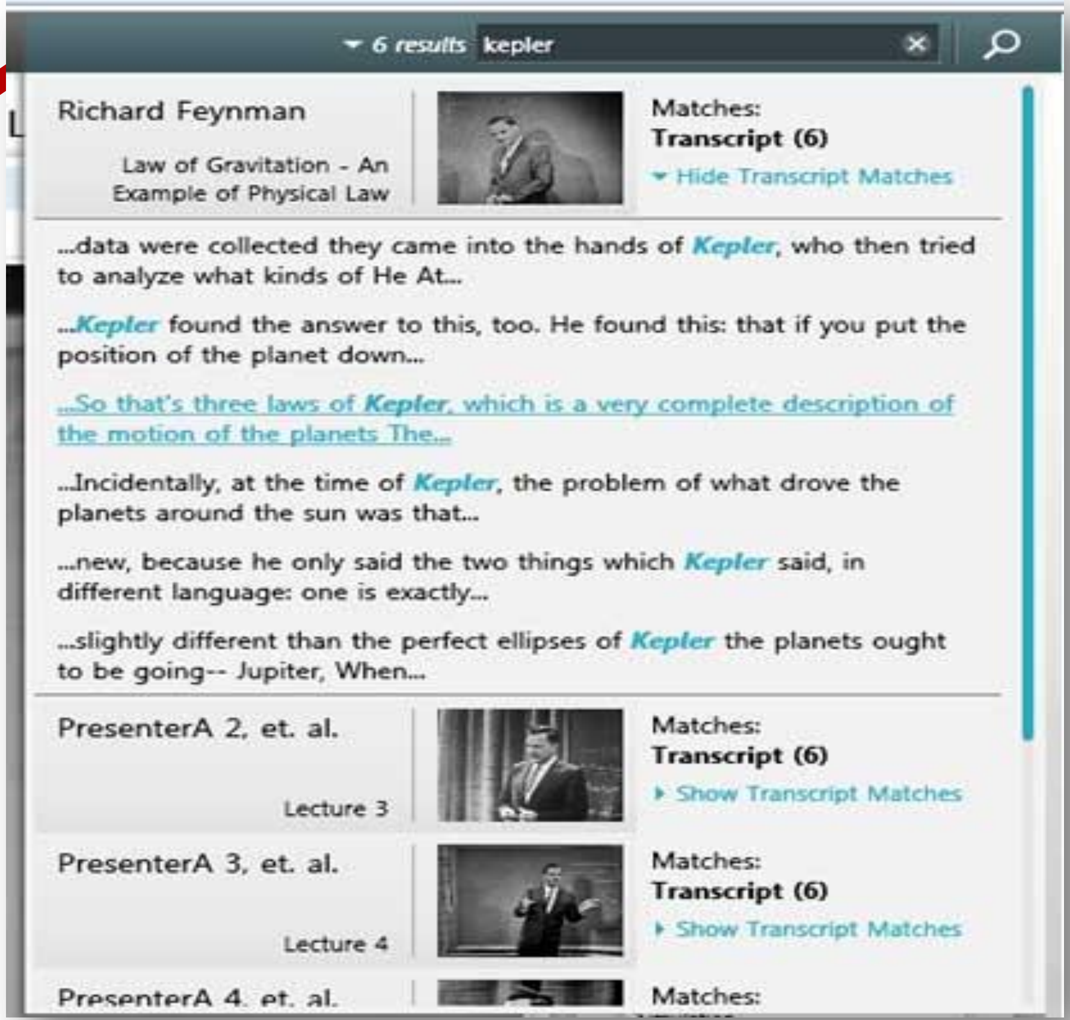
Messenger Series Lectures – Character of Physical Law, Cornell University 1964

Video hyperlinked to rich Web resources (e.g. Wikipedia)

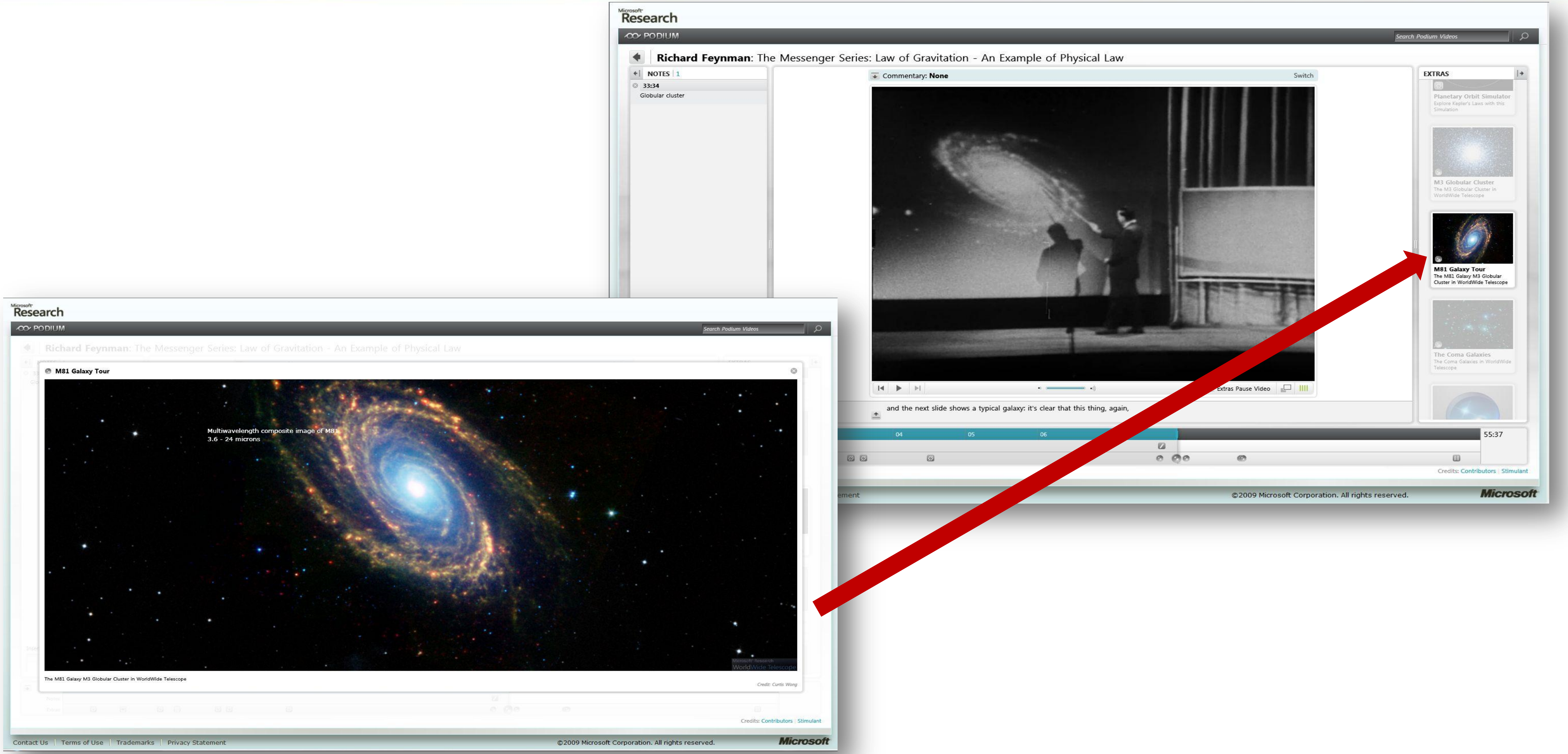


Full text search with contextual results linked to video



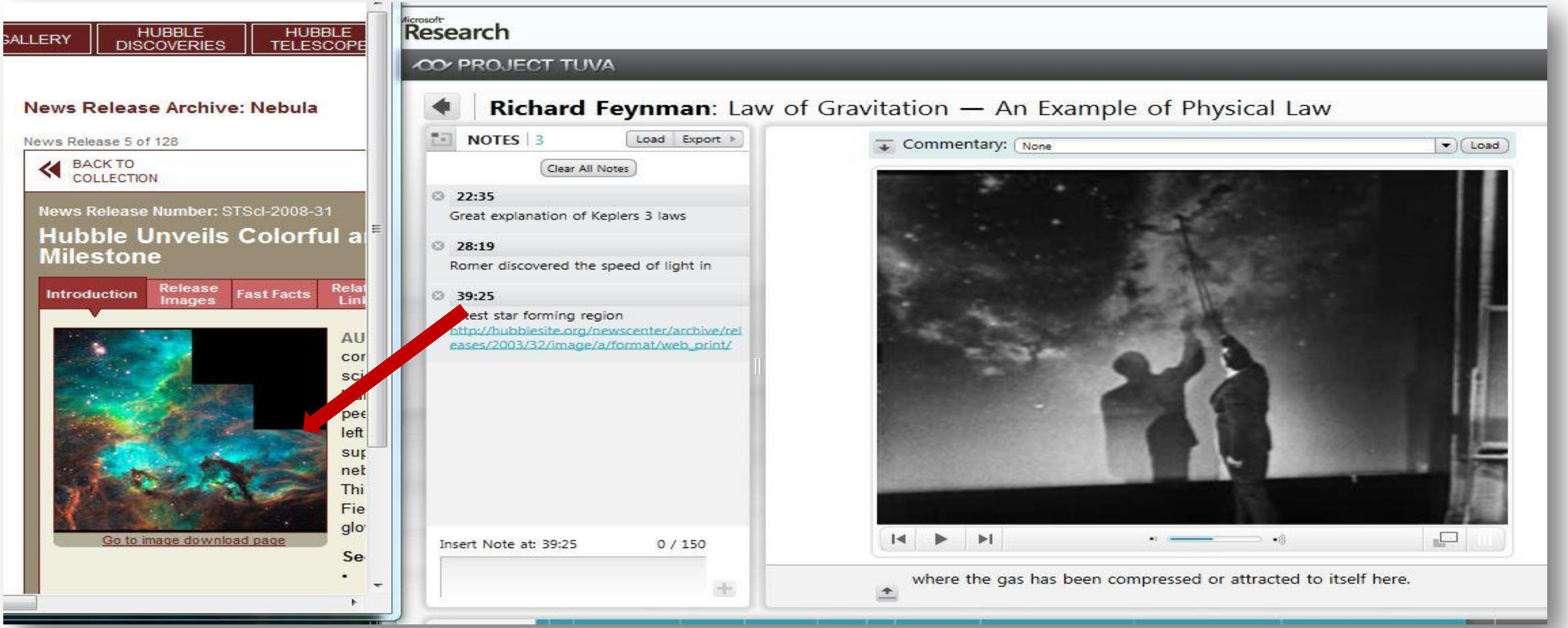


Linked Interactive Simulations and Visualizations

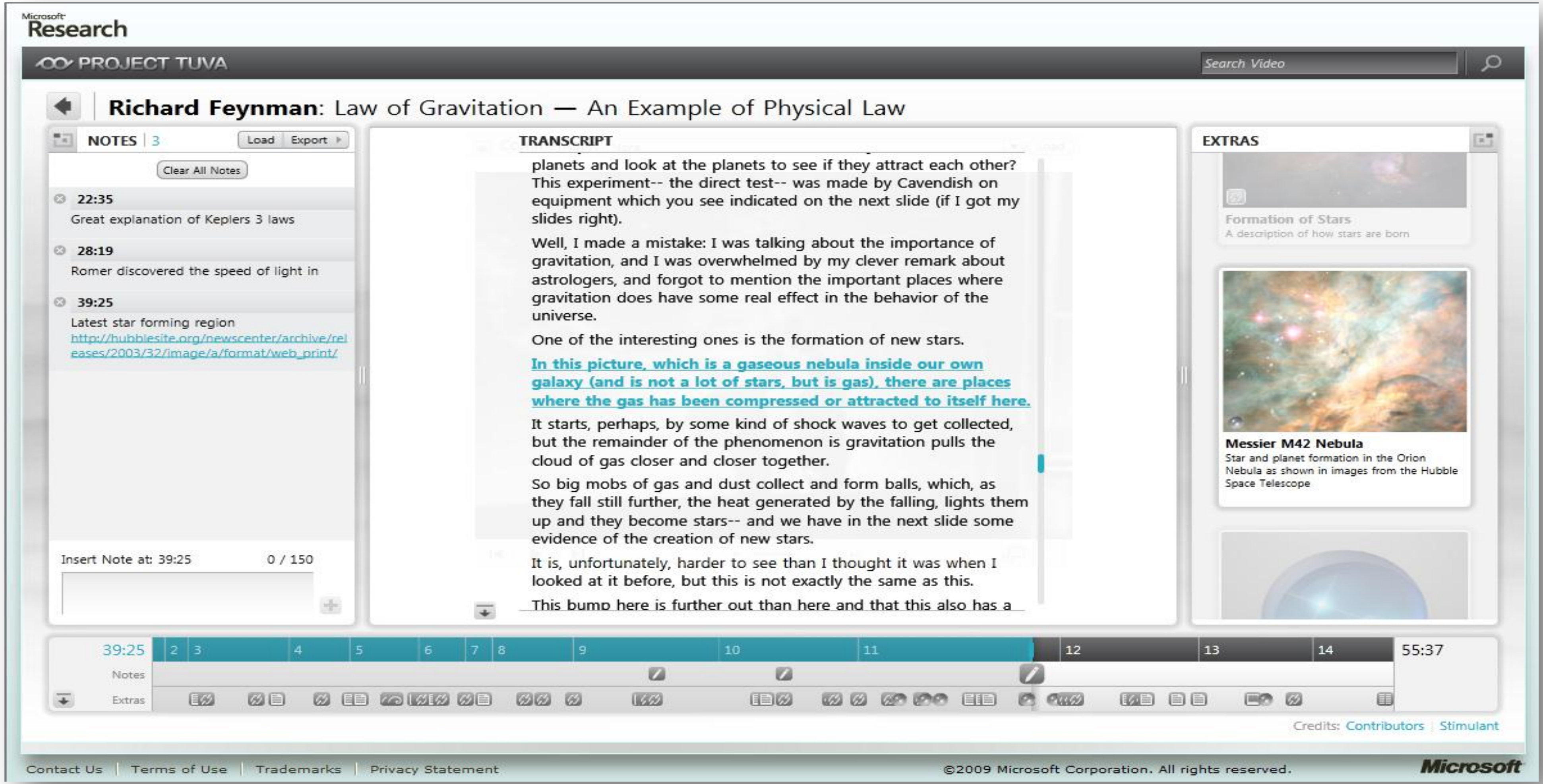


Project Tuva

Create and Share User Generated Notes with Hyperlinks

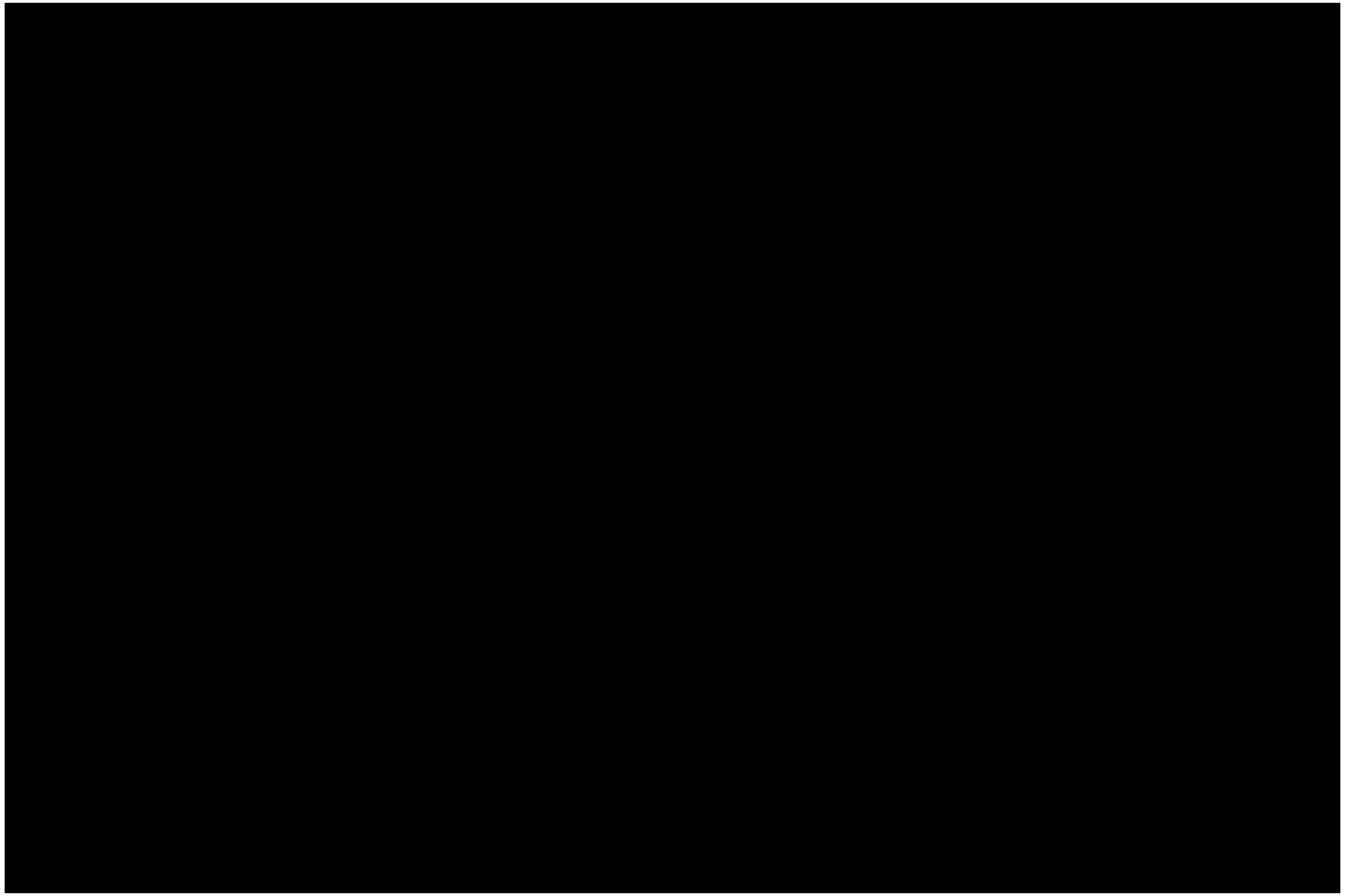


Fully Hyperlinked Videos Transcript



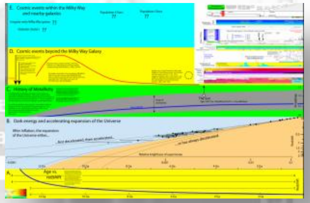
The Garibaldi Project at Brown University

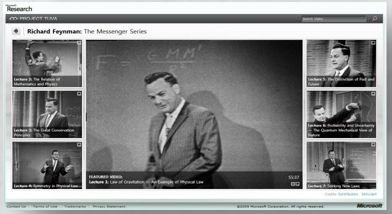
The Garibaldi Project at Brown University

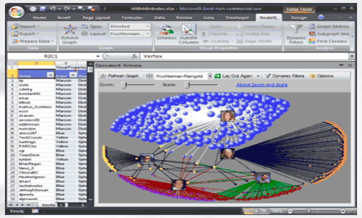


Data

- Novel ways to visualize and explore data, digital assets, relationships, meaning, sharing, archiving





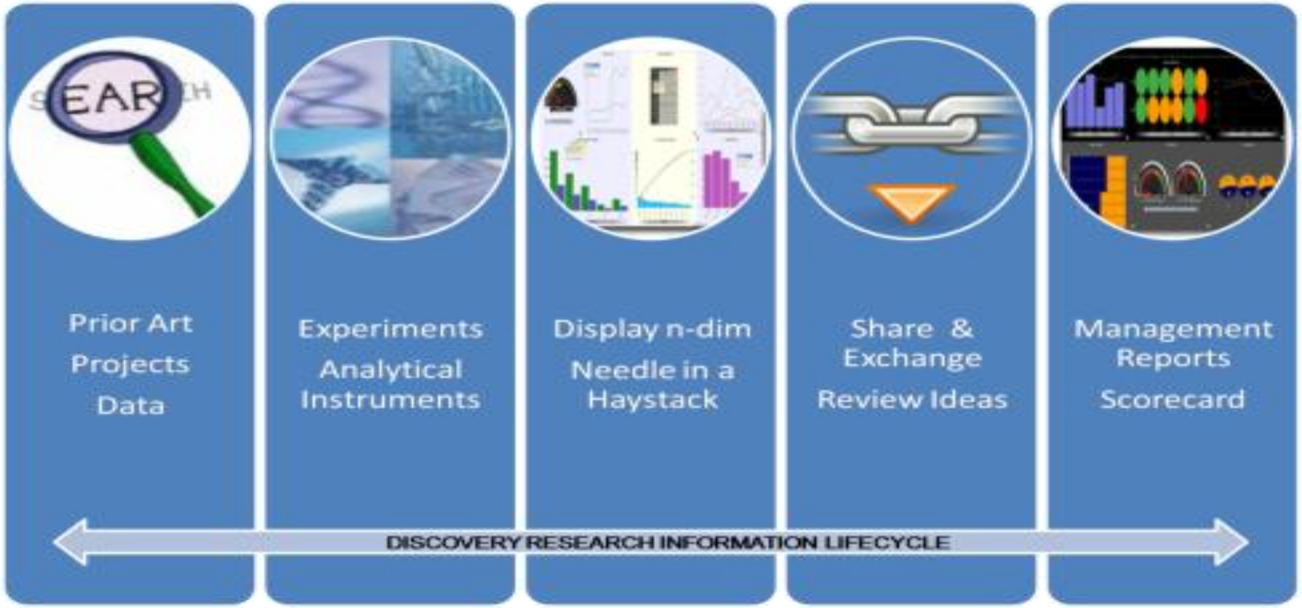


Data Acquisition and Modeling → Collaboration and Visualization → Analysis and Data Mining → Disseminate and Share → Archiving and Preservation

2

Facilitating the move from static summaries to rich information vehicles

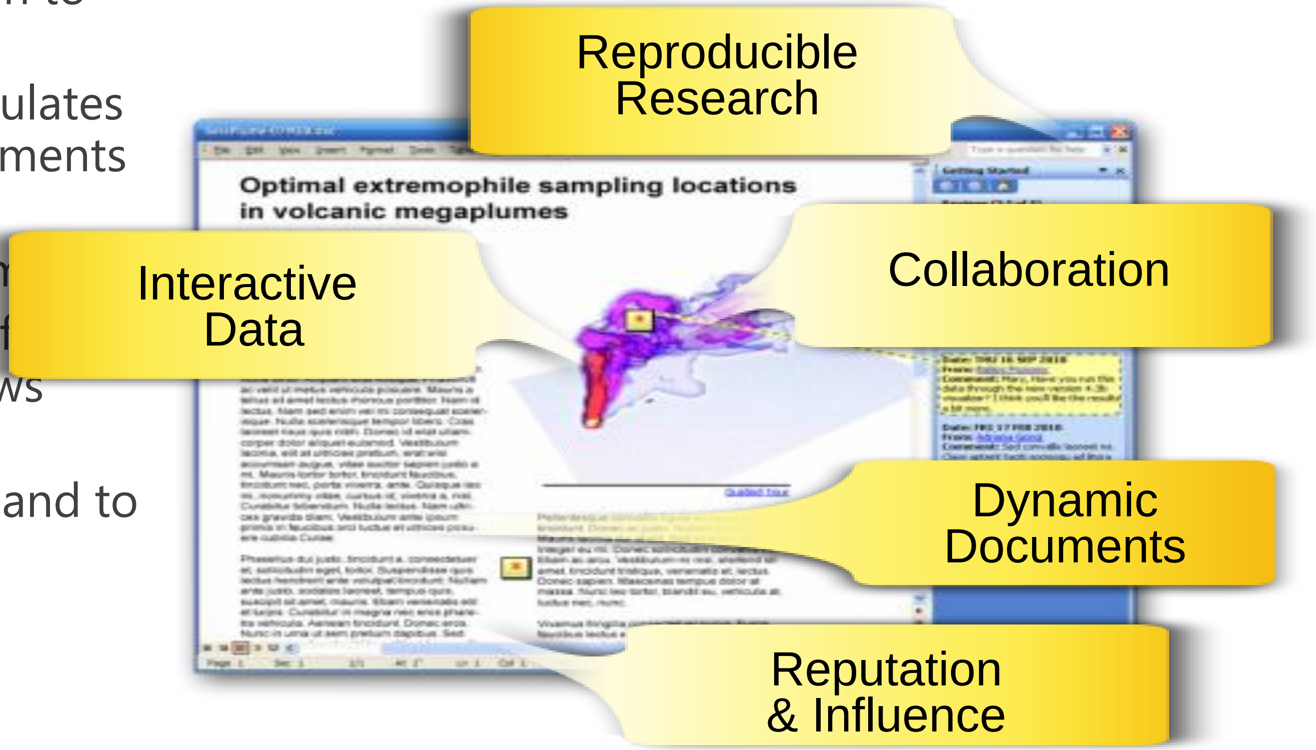
- Pace of science is picking up…rapidly
- The status quo is being challenged and researchers are demanding more
- Why can't a research report offer more …



Envisioning a New Era of Research Reporting

Imagine…

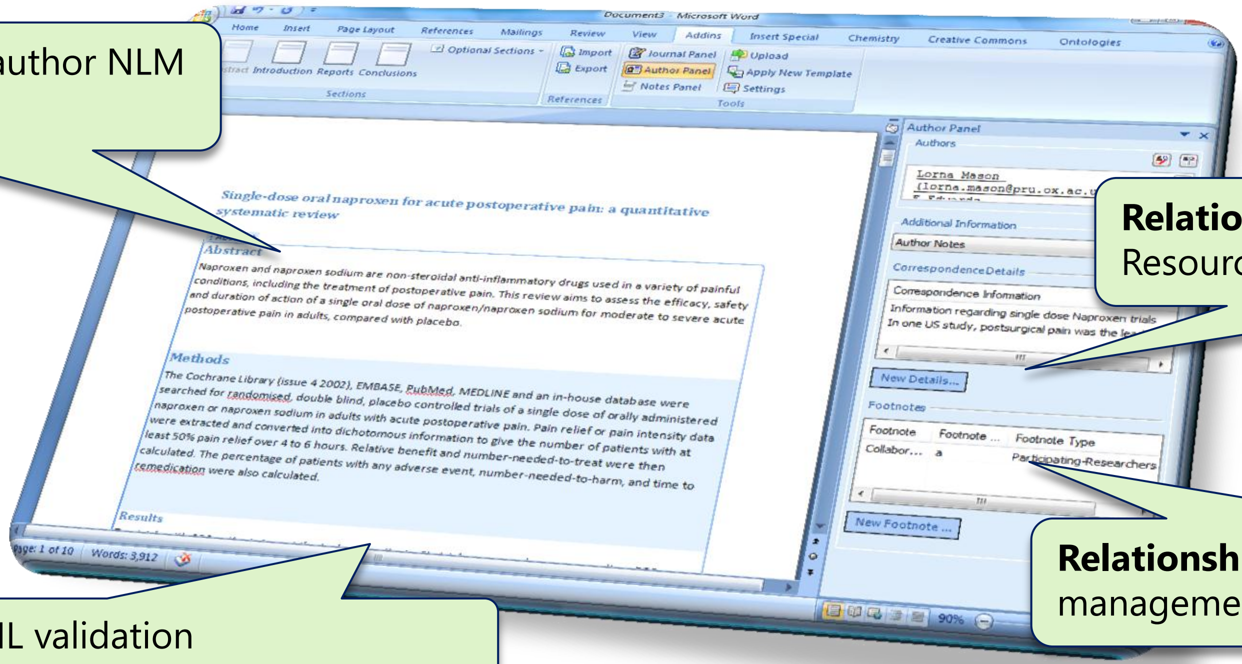
- Live research reports that had multiple end-user 'views' and which could dynamically tailor their presentation to each user
- An authoring environment that absorbs and encapsulates research workflows and outputs from the lab experiments
- A report that can be dropped into an electronic lab workbench in order to reconstitute an entire experim
- A researcher working with multiple reports on a Surf 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



Article Authoring Add-in for Word

Services: repository deposit via SWORD

Structure: Read, convert, and author NLM XML documents



Relationships: ORE Resource Map creation

Relationships: Citation lookup and reference management

Structure: Client-side XML validation

http://research.microsoft.com/authoring/

Chemistry add-in for Word

Data Acquisition and Modeling → Collaboration and Visualization → Analysis and Data Mining → **Disseminate and Share** → Archiving and Preservation

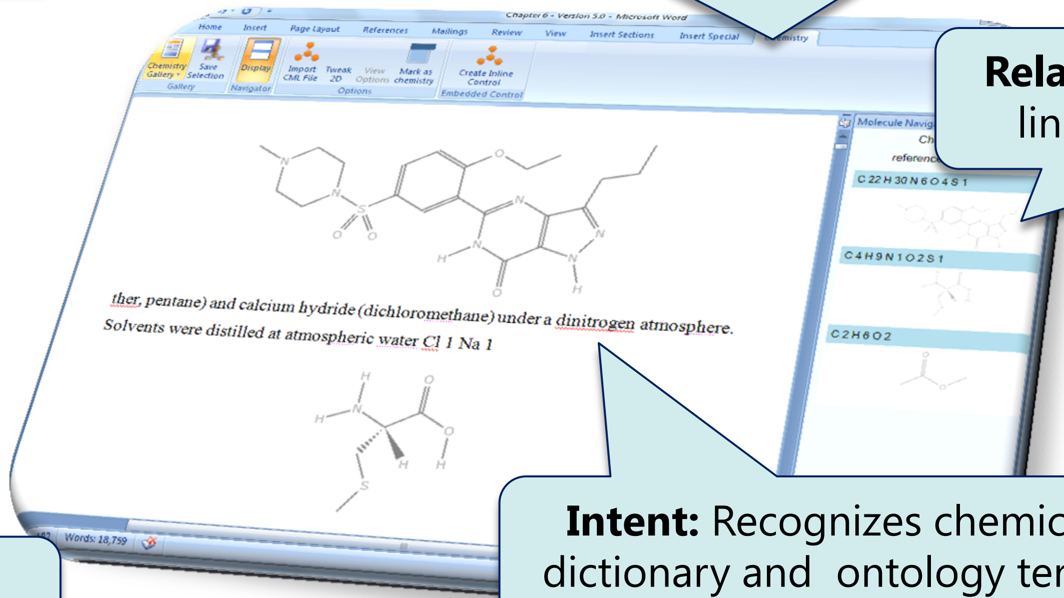
Data: Semantics stored in Chemistry Markup Language

```
<?xml version="1.0" ?>
<cml version="3" convention="org-synth-report" xmlns="http://www.xml-cml.org/schema">
 <molecule id="m1">
  <atomArray>
   <atom id="a1" elementType="C" x2="-2.9149999618530273" 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="-2.914999961853028" y2="-0.7700000190734864" />
   <atom id="a7" elementType="H" x2="-4.248679083681063" y2="-1.907348645691087E-8" />
   <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>
</cml>
```

Intelligence: Verifies validity of authored chemistry

Author/edit: 1D and 2D chemistry to change chemical layout styles.

Relationships: Navigate and link referenced chemistry



Intent: Recognizes chemical dictionary and ontology terms

Cambridge University: Peter Murray-Rust, Joe Townsend
Microsoft Research: Lee Dirks, Alex Wade

Ontology Plug-in for MS Word

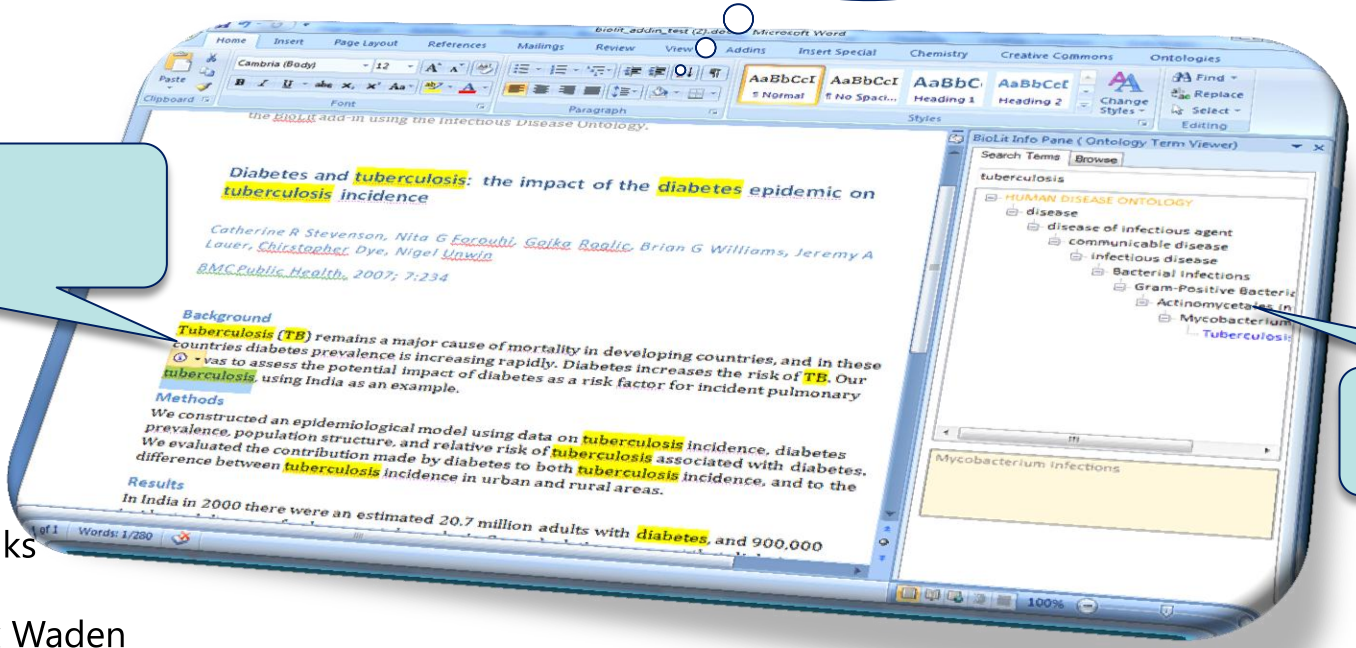
Data Acquisition and Modeling → Collaboration and Visualization → Analysis and Data Mining → **Disseminate and Share** → Archiving and Preservation

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

science commons

Services: Ontology download web service

University of California San Diego

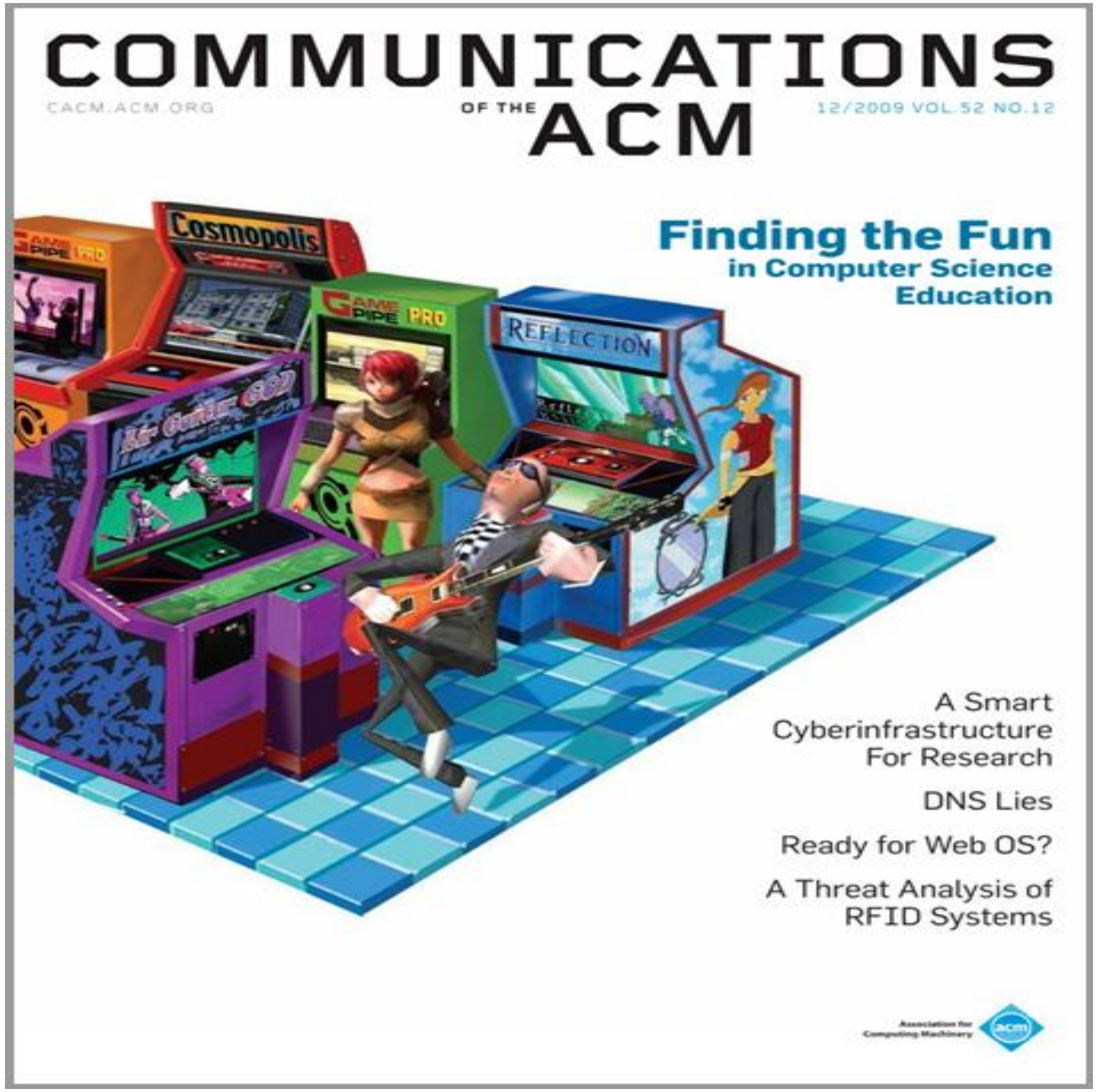


Intent: Term recognition & disambiguation

Relationships: Ontology browser

CC Science Commons: John Wilbanks
UCSD: Phil Bourne, Lynn Fink
Microsoft Research: Lee Dirks, Alex Waden

A “Smart” Cyberinfrastructure for Research”



V viewpoints

DOI:10.1145/1610252.1610267

Savas Parastatidis, Evelyne Viegas, and Tony Hey

Viewpoint

A “Smart” Cyberinfrastructure for Research

A view of semantic computing and its role in research.

THE WEB HAS emerged as the largest distributed information repository on the planet. Human knowledge is captured on the Web in various digital forms: Web pages, news articles, blog posts, digitized books, scanned paintings, videos, podcasts, lyrics, speech transcripts, and so forth. Over the years, services have emerged to aggregate, index, and enable the rapid searching of all this digital data but the full meaning of that data may only be interpretable by humans. In the common case, machines are incapable of understanding or reasoning about the vast amounts of data available on the Web. They are not able to interpret or infer new information from the data and this has been a topic of active research interest for decades within the artificial intelligence community. While the dream of artificial intelligence—machines capable of human-level reasoning and understanding—may still not be within our grasp, we believe semantic technologies hold the promise that machines will be able to meaningfully process, combine, and infer information from the world's data in the not-too-distant future (see Figure 1).

The Web ecosystem of simple formats and protocols is an example of how we can effectively manage, share, access, and represent large amounts of data. Companies like Microsoft and Google are building large-scale services (such as search and cloud services) leveraging the existing hardware and software infrastructures. Schema languages, XML, Entity Data Models, Microformats, RSS, Atom, RDF (see http://www.w3.org/RDF/), OWL (see http://www.w3.org/2007/OWL/), and other technologies are being used to capture the information in data while machine learning, entity extraction, neural networks, clustering, and latent semantics are approaches to extracting information from that data and help reason about it. The field is an active area of research and experimentation and is still rapidly evolving (see the sidebar “Semantic Computing” vs. “Semantic Web”).

Data Mesh

At the center of our discussion is the concept of a “data mesh,” a term we use to refer to the various information and knowledge representation techniques/technologies that have been developed over the years (see Figure 2). In its simplest form, a data mesh looks like a directed graph in which the nodes represent data/information captured in well-known formats and the edges capture a relationship, characterized by a predicate and perhaps other information, between the linked data. For example, “Jane listens to Santana every day” is a relationship, in which “Jane” (the subject) and “Santana” (the object) are the nodes, “listens to” is the edge (the predicate), and “every day” is an attribute of the edge. Other tuples could add further information to the data mesh (for example, “Santana is an artist,” “Santana plays the guitar,” “Santana makes music,” “Jane met Santana in 1995” and so forth). Semantic Web's RDF is one, but not the only, technology that can be used to represent such graphs or knowledge bases. Indeed, Cyc,[3] Semantic Networks, WordNet,[7] MultiNet[4] are examples of other such technologies/approaches. Scaling to the same level as the Web still remains a challenge for these approaches.

We believe there is an opportunity to involve users, who are now equally producers as they are consumers of information on the Web, and not just the very few experts in producing

Data — Information — Knowledge
Current technologies/infrastructure
Opportunities for innovation

Figure 1. Data, information, knowledge: While we are good at data management at scale (for example, Google, Amazon) we are still far away from supporting information representation and reasoning. Knowledge management at scale is a great opportunity for innovation.

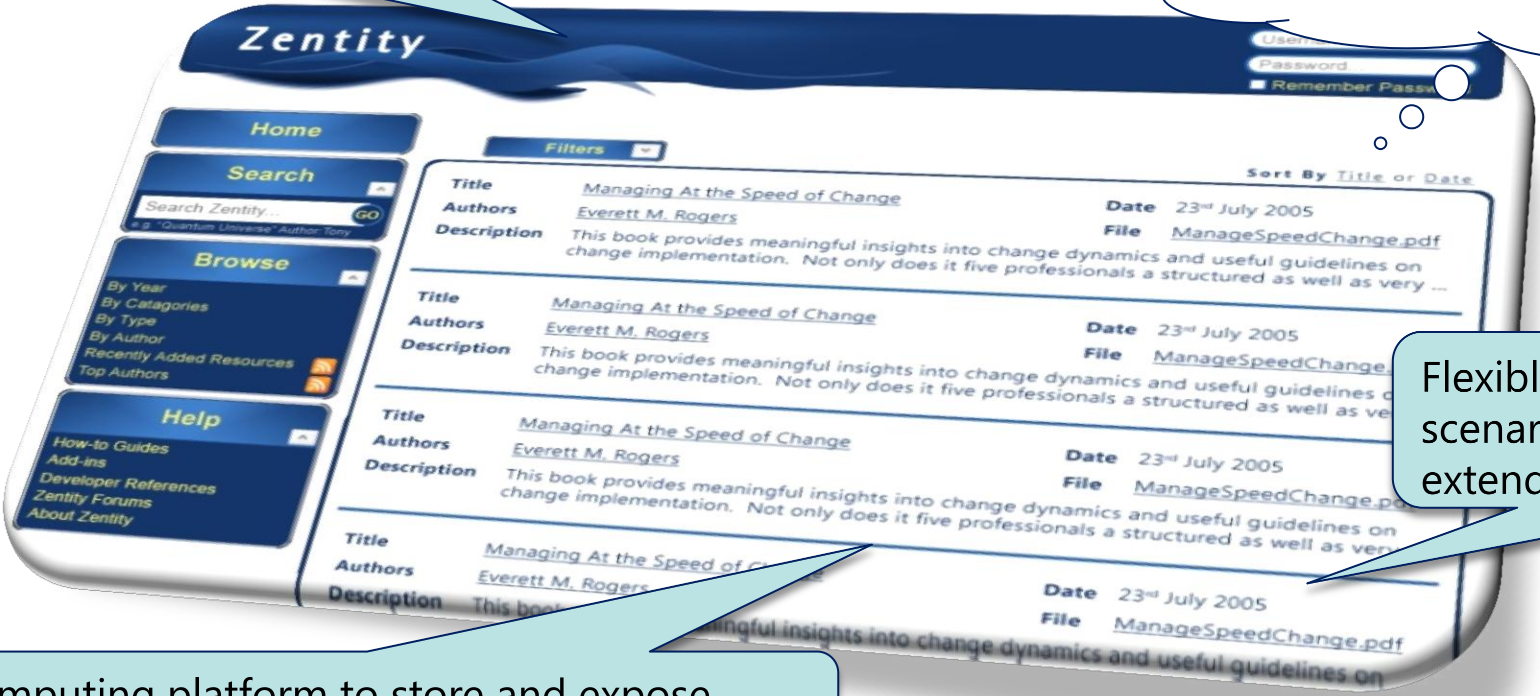
Handling semantic relationships

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

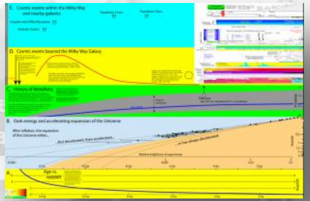
Native support for RSS, OAI-PMH, OAI-ORE, AtomPub and SWORD

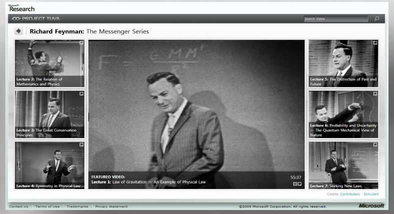


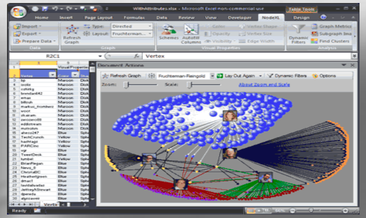
Flexible data model enables many scenarios and can be easily extended over time

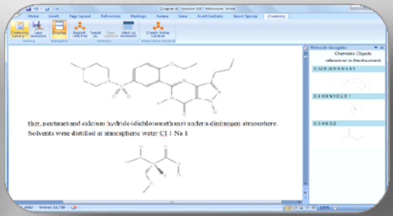
A semantic computing platform to store and expose relationships between digital assets

Data









Data Acquisition and Modeling → **Collaboration and Visualization** → **Analysis and Data Mining** → **Disseminate and Share** → **Archiving and Preservation**

Scaling science

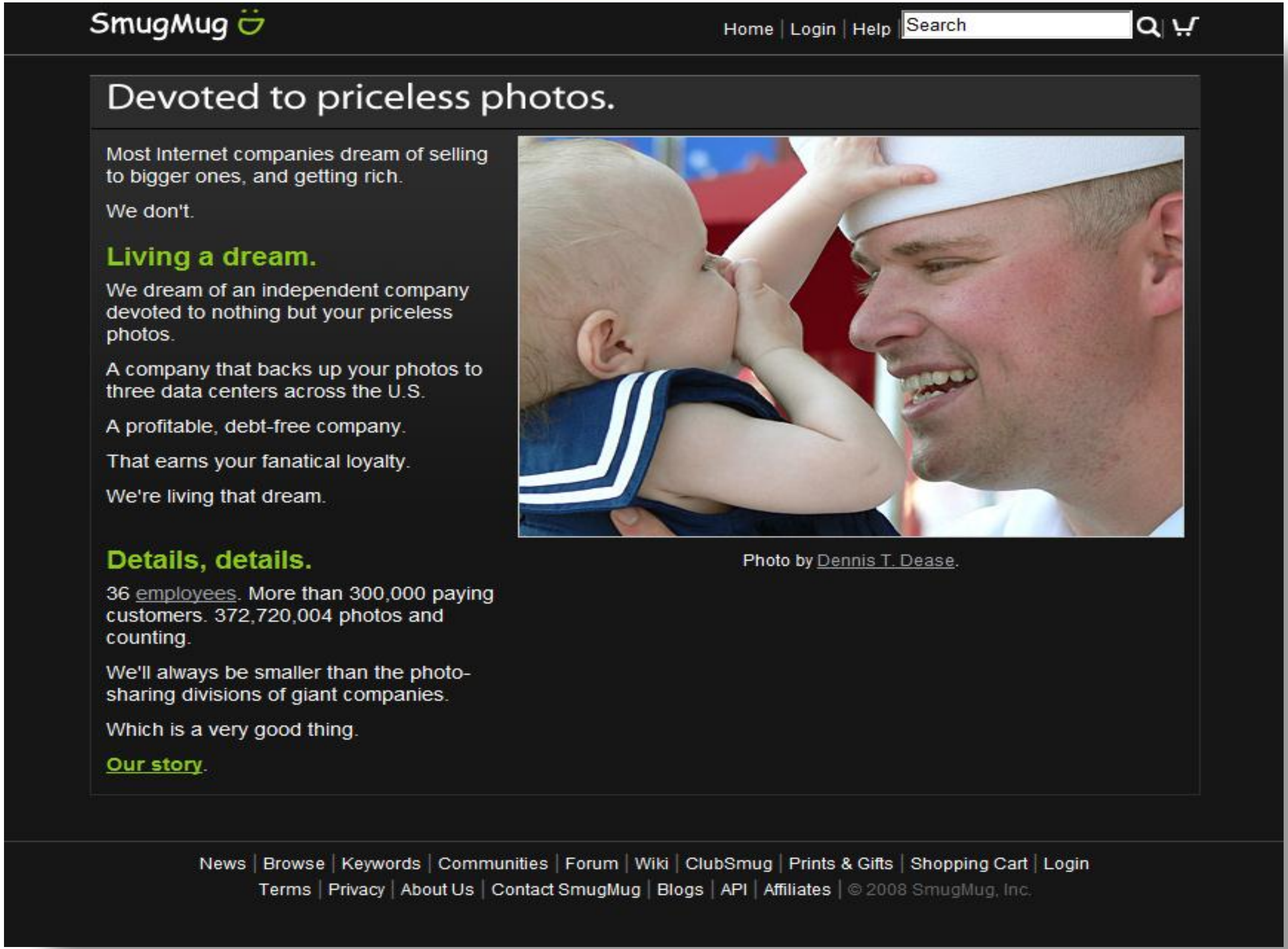
- *Length scales:* from **infinitesimal** to **galactic**
- *Research:* from **individual** to **community**
- *Timescales:* from **instant** to **eon**
- *Complexity:* from **single source** to **webscale**
- *Data:* from **documents** to **digital libraries**

Enabled/powered/accelerated by Cloud Computing

The Cloud

- A model of computation and data storage based on “pay as you go” access to “unlimited” remote data center capabilities
- Provides a framework to manage scalable, reliable, on-demand access to applications
- The “invisible” backend to many of our mobile applications
- Historical roots in today’s Internet apps
  - Search, email, social networks
  - File storage (Live Mesh, MobileMe, Flickr, …)

A Cloud Service : www.smugmug.com



Why Cloud Computing could be in your future

A 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 application
- Reliable means on-demand; 5 "nines" available right now

The Data Center Landscape

Range in size from "edge" facilities to mega scale.

Unprecedented economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K 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
Admin	~140 servers/ Admin	>1000 Servers/ Admin	7.1

Each data center is
11.5 times
the size of a football field

Advances in Data Center Deployment

Conquering complexity

- Building racks of servers & complex cooling systems all separately is not efficient.
- Package and deploy into bigger units
- 3 Sockets: Power, Cooling, Bandwidth

Programming the Cloud

Infrastructure as a Service (IaaS)

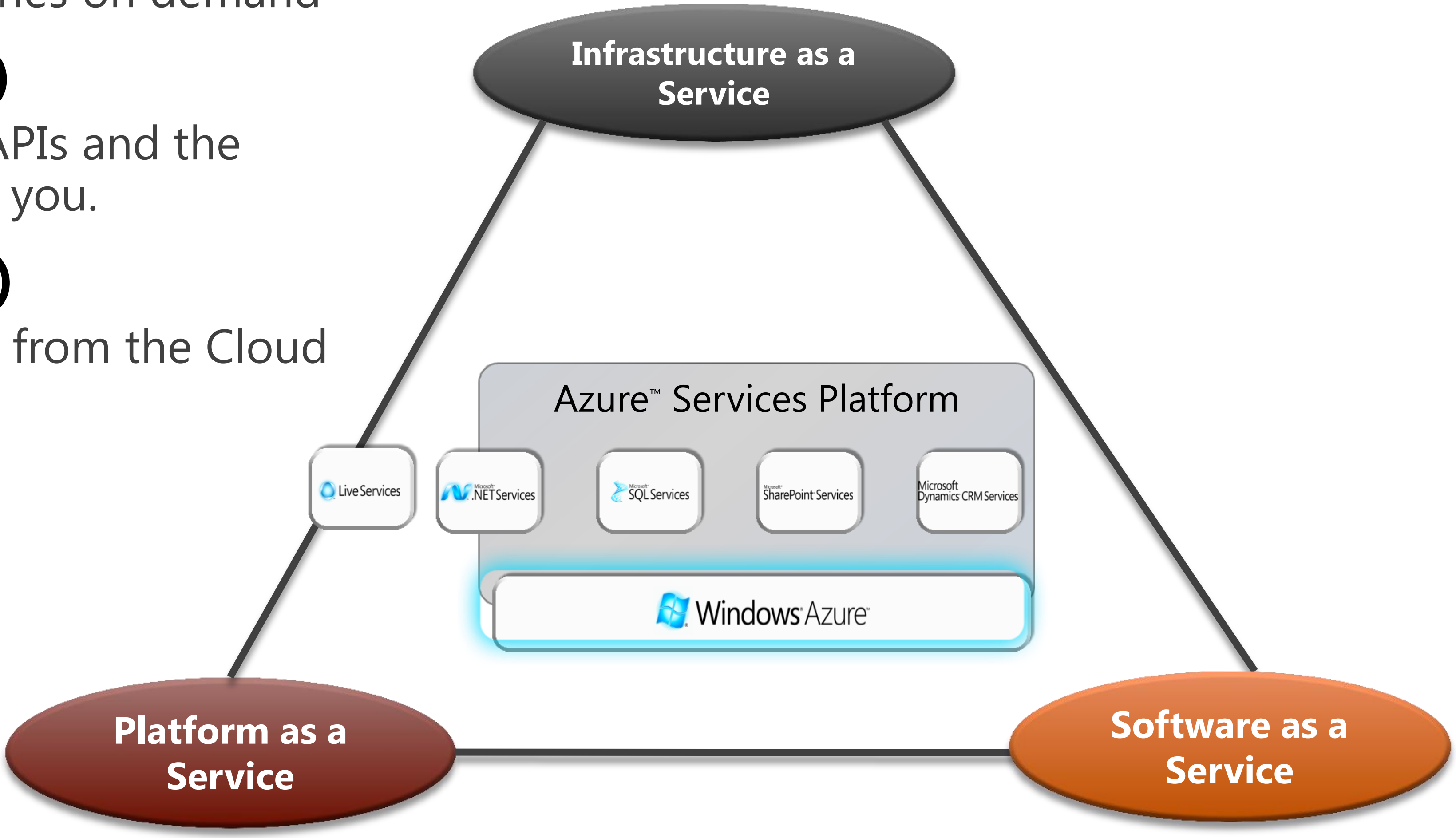
Provide a way to host virtual machines on demand

Platform as a Service (PaaS)

You write an Application to Cloud APIs and the platform manages and scales it for you.

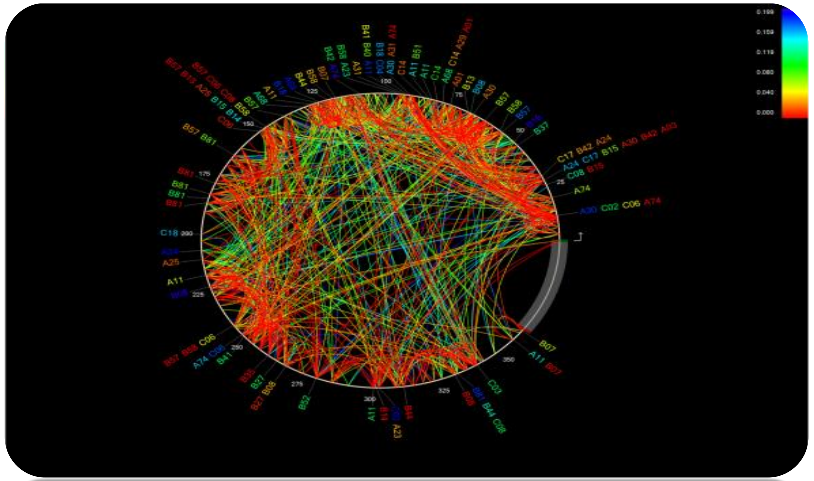
Software as a Service (SaaS)

Delivery of software to the desktop from the Cloud



PhyloD as an Azure Service

- Statistical tool used to analyze DNA of HIV from large studies of infected patients
- PhyloD was developed by Microsoft Research and has been highly impactful
- Small but important group of researchers
  - *100's of HIV and HepC researchers actively use it*
  - *1000's of research communities rely on results*



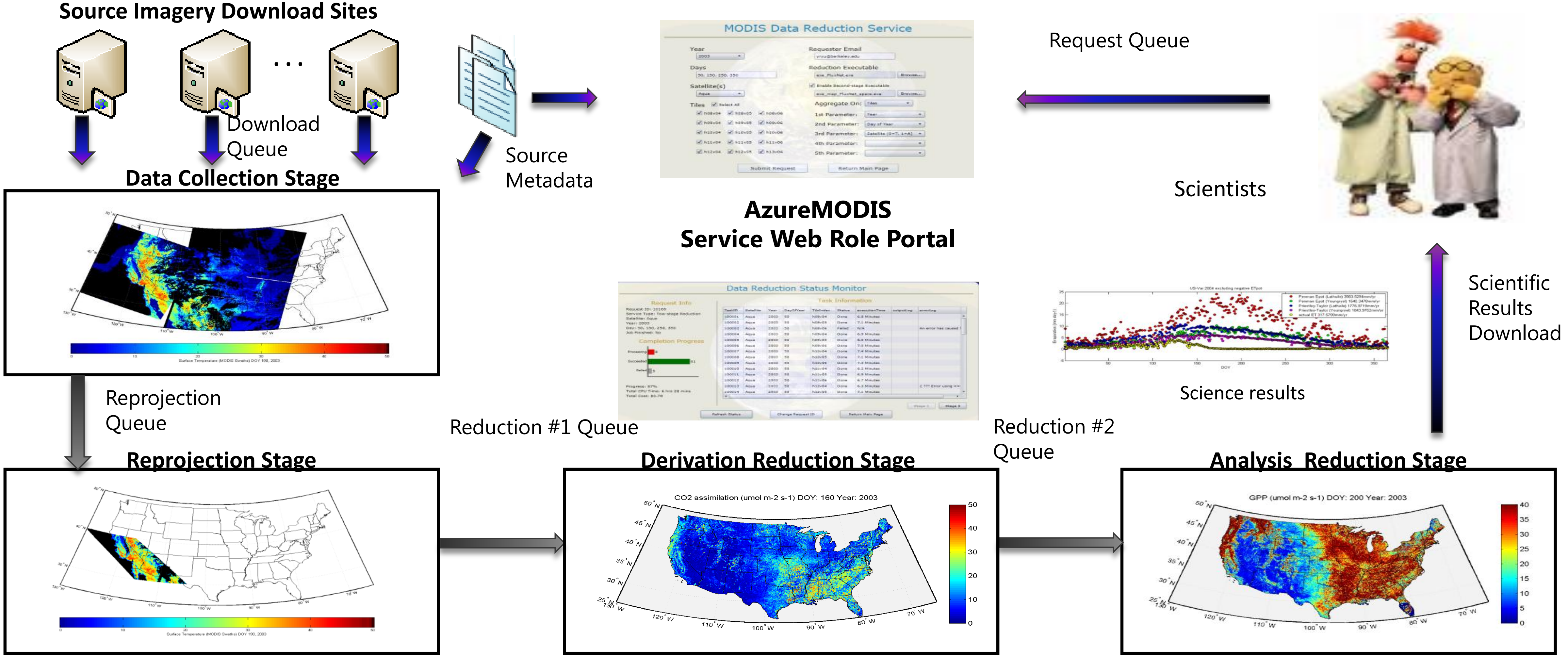
Cover of PLoS Biology November 2008

Typical job: 10 – 20 CPU hours; Extreme jobs: 1K – 2K CPU hours

- Large number of test runs for a given job (1 – 10M tests)
- Highly compressed data per job (~100 KB per job)

ModisAzure

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)



Scaling science

- *Length scales:* from **infinitesimal** to **galactic**
- *Research:* from **individual** to **community**
- *Timescales:* from **instant** to **eon**
- *Complexity:* from **single source** to **webscale**
- *Data:* from **documents** to **digital libraries**

Enabled/powered/accelerated by Cloud Computing

Live Services

Microsoft .NET Services

Microsoft SQL Services

Microsoft SharePoint Services

Microsoft Dynamics CRM Services

Today…

Computers are great **tools** for

storing

computing

managing

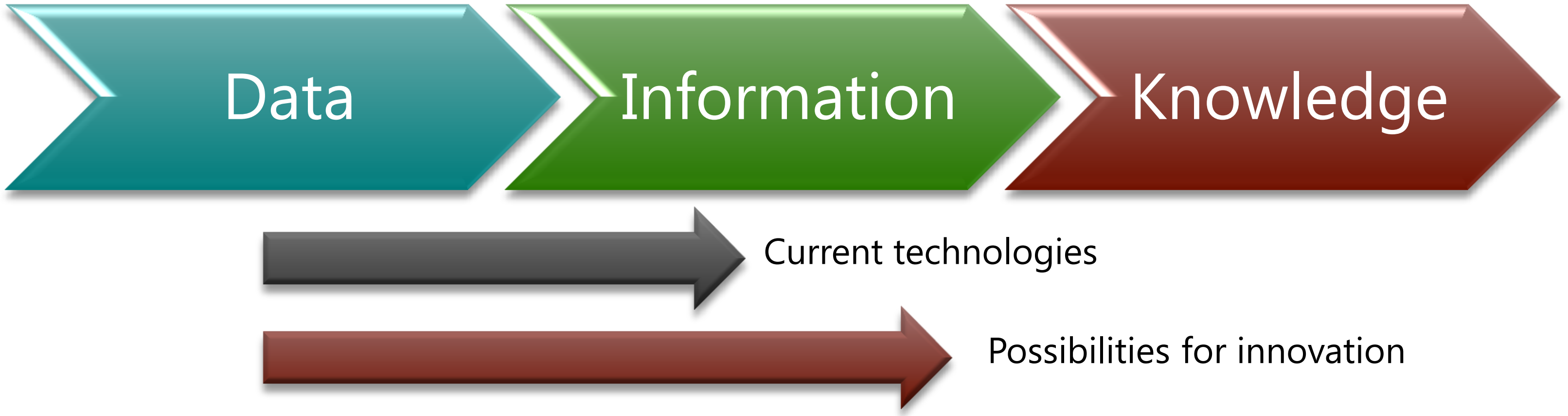
indexing

huge amounts of **data**

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

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

storing | computing | managing | indexing

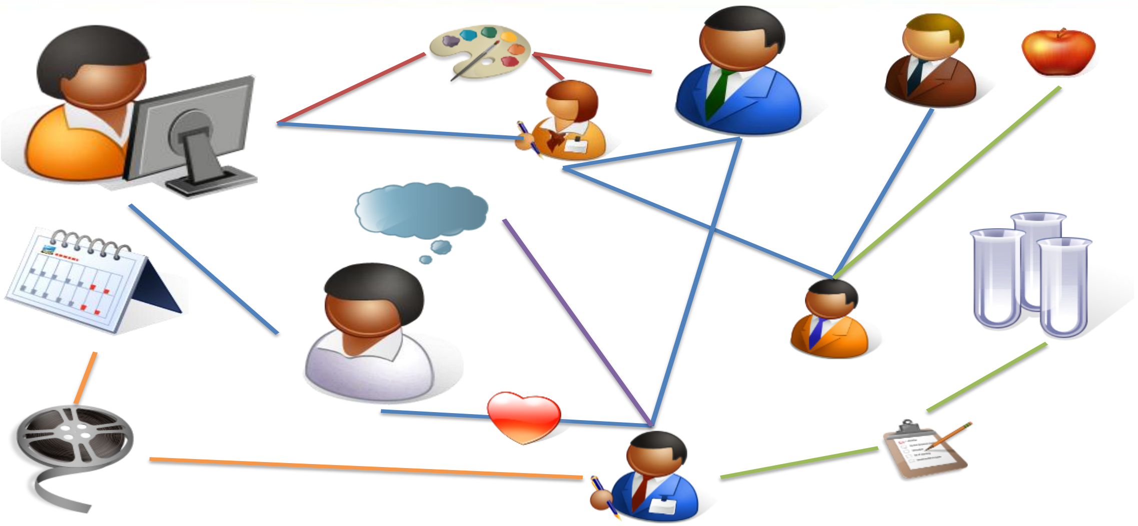
huge amounts of **data**

We would like computers to also help with the **automatic**

acquisition | discovery | aggregation | organization | correlation | analysis | interpretation | inference

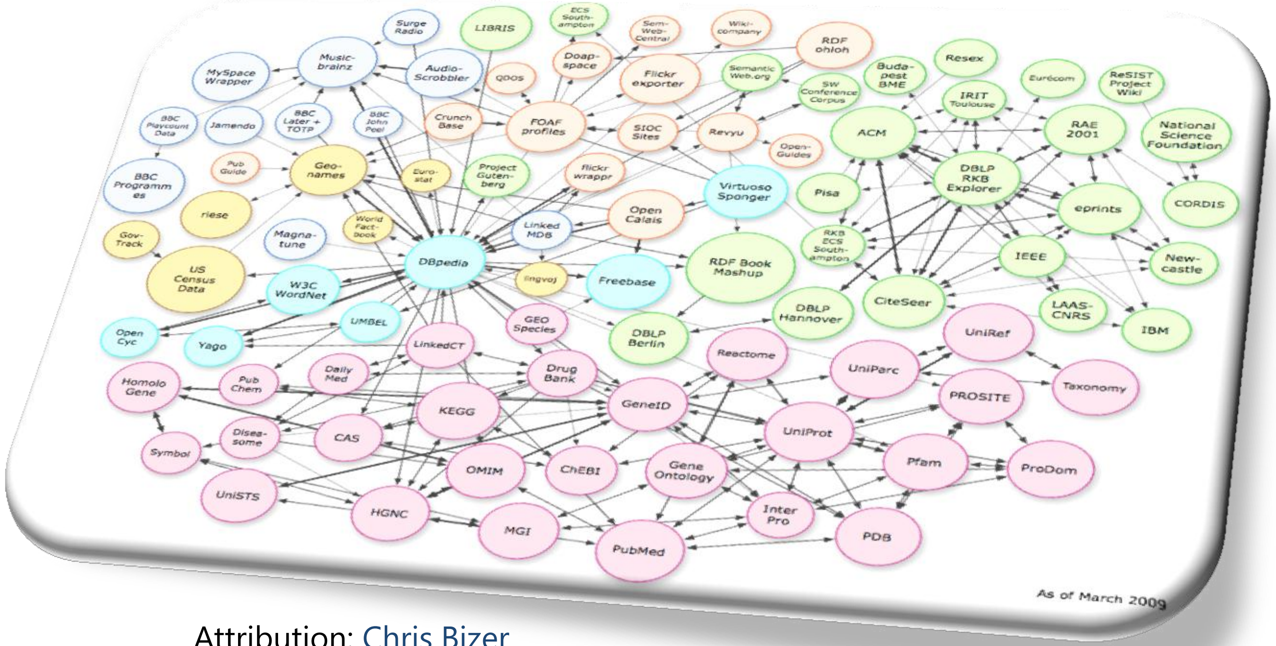
of the world's **information**

Moving to a world where all data is linked …



- Data/information is inter-connected through machine-interpretable information (e.g. paper X **is about** star Y)
- Social networks are a special case of 'data meshes'

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



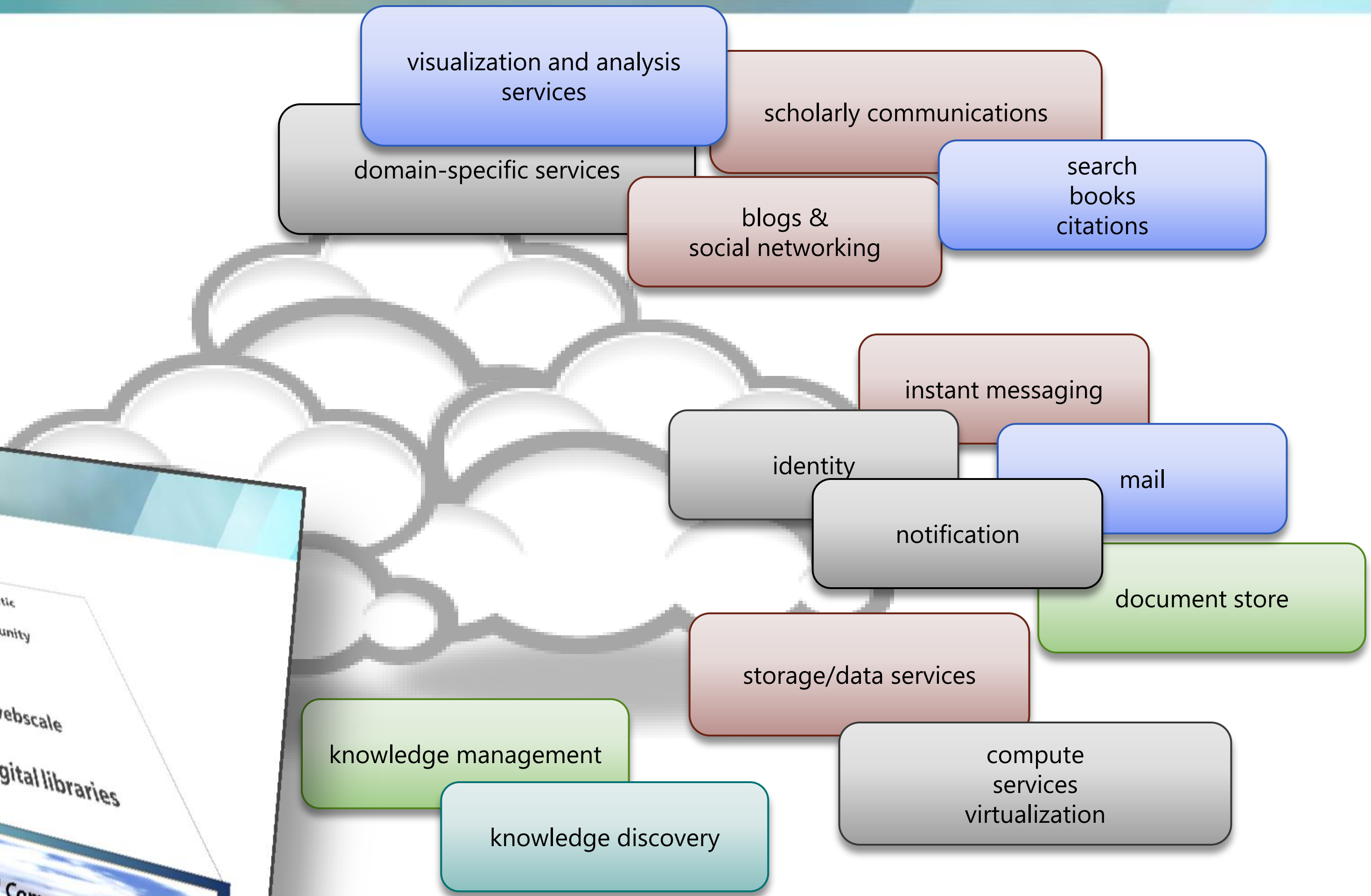
Attribution: [Chris Bizer]()

… and can be stored/analyzed in the Cloud

Vision of Future Research: scaling cyber-infrastructure with Client + Cloud

Reference management

Project management



http://research.microsoft.com/collaboration

PowerPoint Guidelines

- Font, size, and color for text have been formatted for you in the Slide Master
- This template uses Segoe UI, a standard Windows Vista and Office 2007 font
- Hyperlink color: www.microsoft.com
- Use the color palette shown below

Sample Fill	Sample Fill	Sample Fill
Sample Fill	Sample Fill	Sample Fill

Microsoft® Research

Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Video Title

video