

eScience: The Next Decade Will Be Exciting

Talk @ ETH, Zurich
29 May 2006

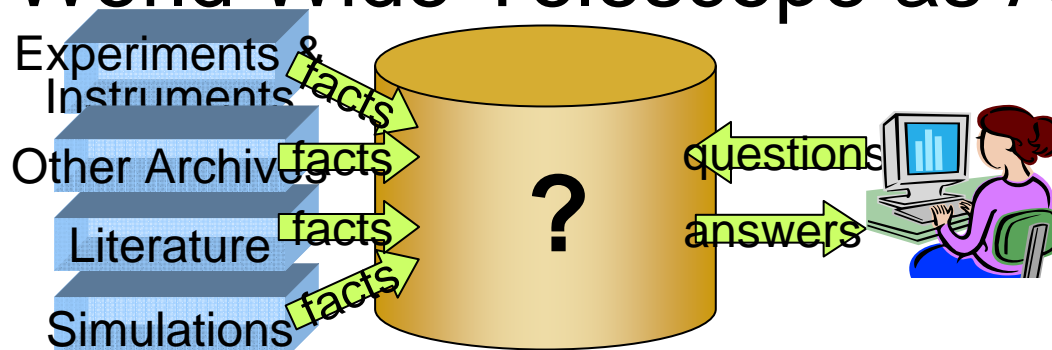
Jim Gray
Microsoft Research
Gray@Microsoft.com

Alex Szalay
Johns Hopkins University
Szalay@pha.JHU.edu

<http://research.microsoft.com/~gray/talks>

Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype



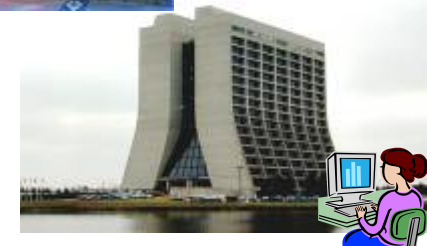
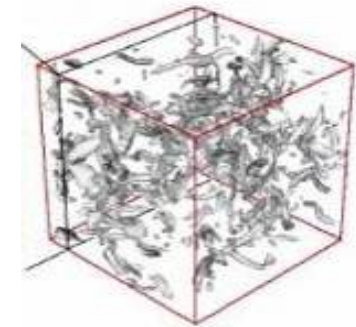
The Big Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute 4
 - Batch query scheduling

Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today:
data exploration (eScience)
unify theory, experiment, and simulation
using data management and statistics
 - Data captured by instruments
Or generated by simulator
 - Processed by software
 - Scientist analyzes database / files

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



Computational Science Evolves

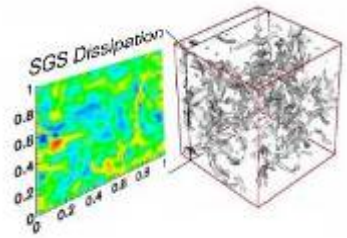
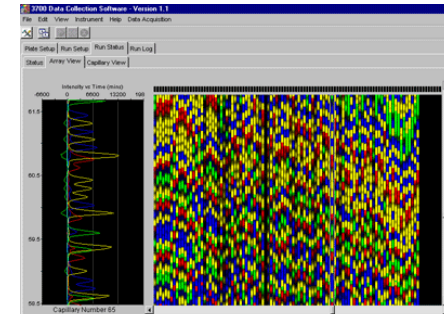


Image courtesy
C. Meneveau & A. Szalay @ JHU

- Historically, Computational Science = simulation.
- New emphasis on informatics:
 - Capturing,
 - Organizing,
 - Summarizing,
 - Analyzing,
 - Visualizing
- Largely driven by observational science, but also needed by simulations.
- Will comp-X and X-info will unify or compete?



P&E Gene Sequencer
From
<http://www.genome.uci.edu/>



BaBar, Stanford



Space Telescope

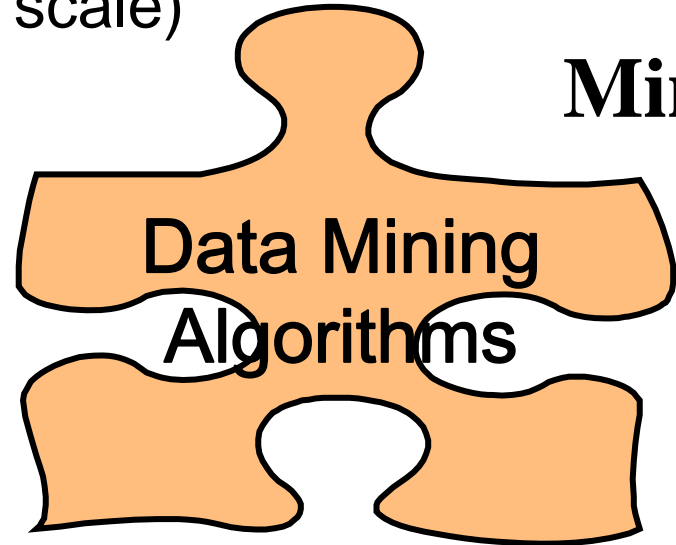
What X-info Needs from us (cs)

(not drawn to scale)

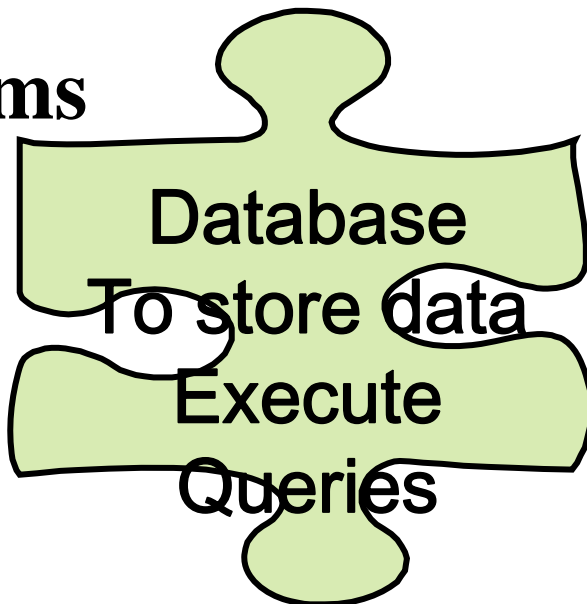
Scientists



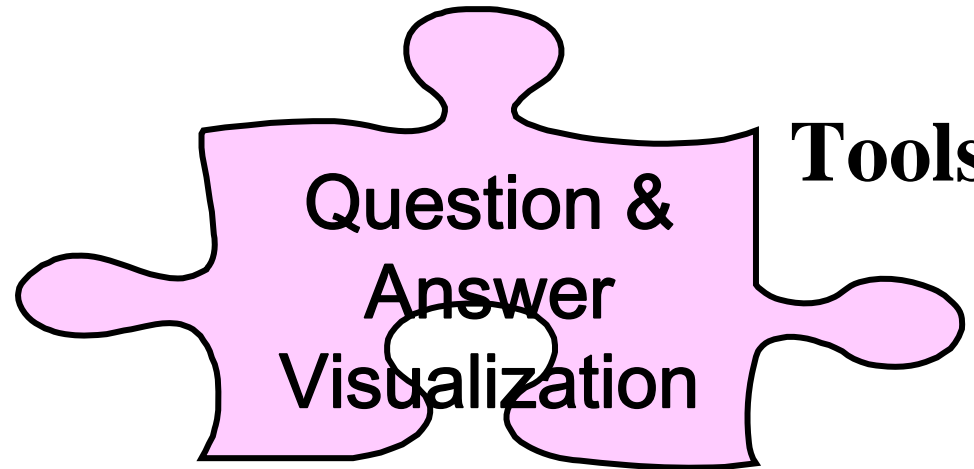
Miners



Systems



Tools



Experiment Budgets $\frac{1}{4} \dots \frac{1}{2}$ Software

Software for

- Instrument scheduling
- Instrument control
- Data gathering
- Data reduction
- Database
- Analysis
- Visualization

Millions of lines of code

Repeated for experiment
after experiment

Not much sharing or learning

Let's work to change this

Identify generic tools

- Workflow schedulers
- Databases and libraries
- Analysis packages
- Visualizers
- ...

Data Access Hitting a Wall

Current science practice based on data download (FTP/GREP)

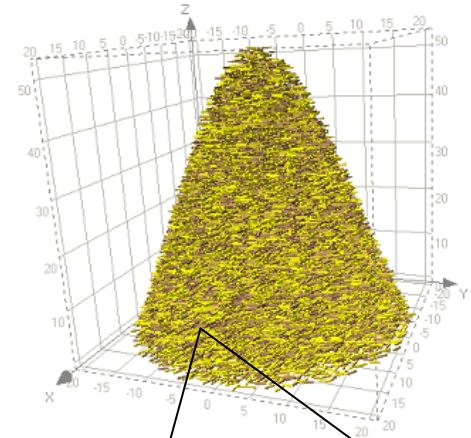
Will not scale to the datasets of tomorrow

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years.
- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1\$)
- ... 2 days and 1K\$
- ... 3 years and 1M\$
- Oh!, and 1PB ~5,000 disks
- At some point you need **indices** to limit search
parallel data search and analysis
- This is where databases can help



New Approaches to Data Analysis

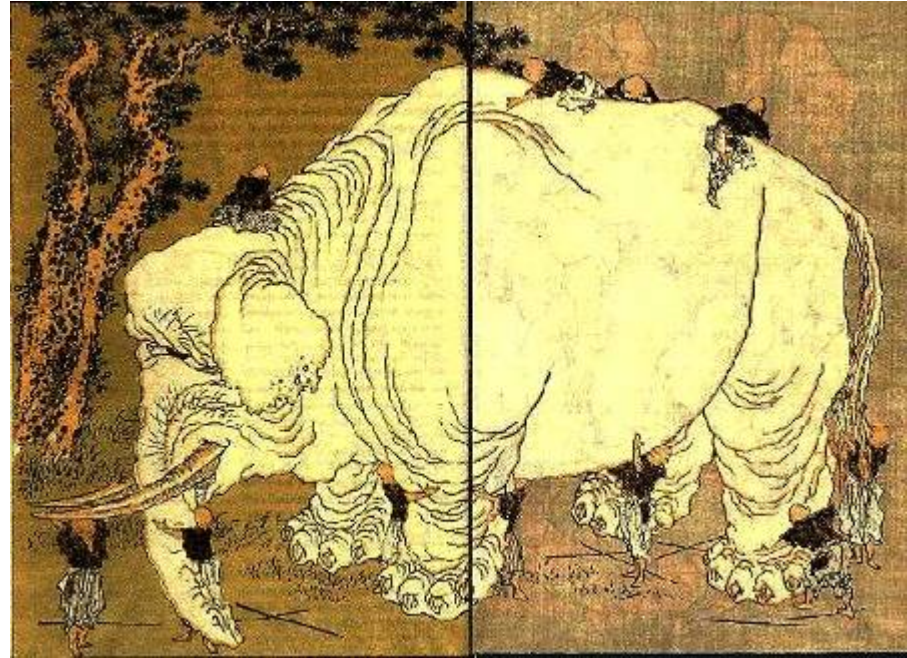
- Looking for
 - Needles in haystacks – the Higgs particle
 - Haystacks: Dark matter, Dark energy
- Needles are easier than haystacks
- Global statistics have poor scaling
 - Correlation functions are N^2 , likelihood techniques N^3
- As data and computers grow at same rate, we can only keep up with $N \log N$
- A way out?
 - Discard notion of optimal (data is fuzzy, answers are approximate)
 - Don't assume infinite computational resources or memory
- Requires combination of statistics & computer science



Analysis and Databases

- Much statistical analysis deals with

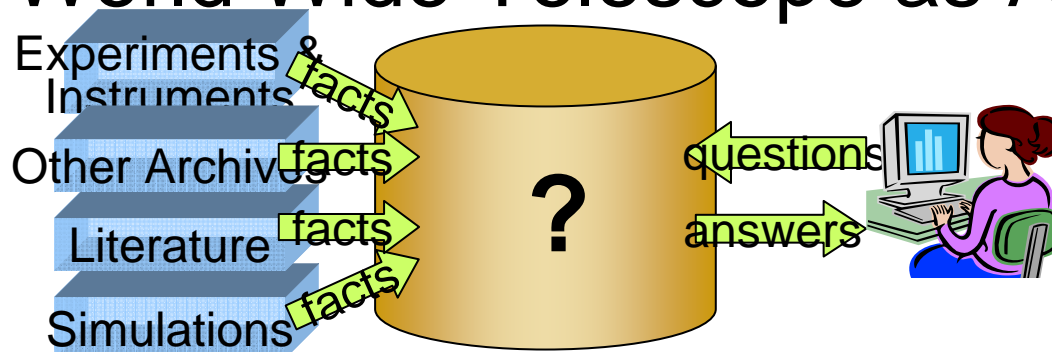
- Creating uniform samples –
- data filtering
- Assembling relevant subsets
- Estimating completeness
- Censoring bad data
- Counting and building histograms
- Generating Monte-Carlo subsets
- Likelihood calculations
- Hypothesis testing



- Traditionally these are performed on files
- Most of these tasks are much better done inside a database
- Move Mohamed to the mountain, not the mountain to Mohamed.

Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype



The Big Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute ¹³
 - Batch query scheduling

Peer-Reviewed Science Literature Is Coming Online

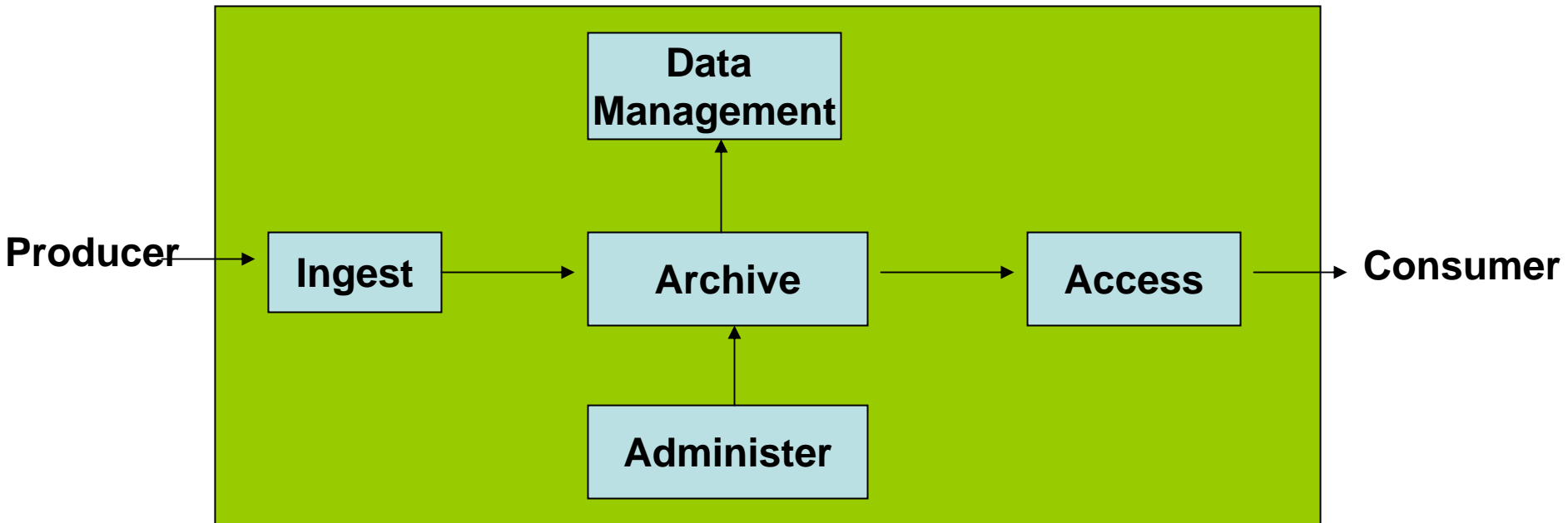
- Agencies and Foundations mandating research be public domain.
 - NIH (30 B\$/y, 40k PIs,...)
(see <http://www.taxpayeraccess.org/>)
 - Wellcome Trust
 - Japan, China, Italy, South Africa,.....
 - Public Library of Science..
- Other agencies will follow NIH
- Publishers will resist (not surprising)
- Professional societies will resist (amazing!)

How Does the New Library Work?

- Who pays for storage access? (unfunded mandate).
 - Its cheap: 1 milli-dollar per access
- But... curation is not cheap:
 - Author/Title/Subject/Citation/.....
 - Dublin Core is great but...
 - NLM has a 6,000-line XSD for documents <http://dtd.nlm.nih.gov/publishing>
 - Need to capture document structure from author
 - Sections, figures, equations, citations,...
 - Automate curation
 - NCBI-PubMedCentral is doing this
 - Preparing for 1M articles/year
 - MUST be automatic.



The OAIS model (Open Archive Information System)



Access Challenges

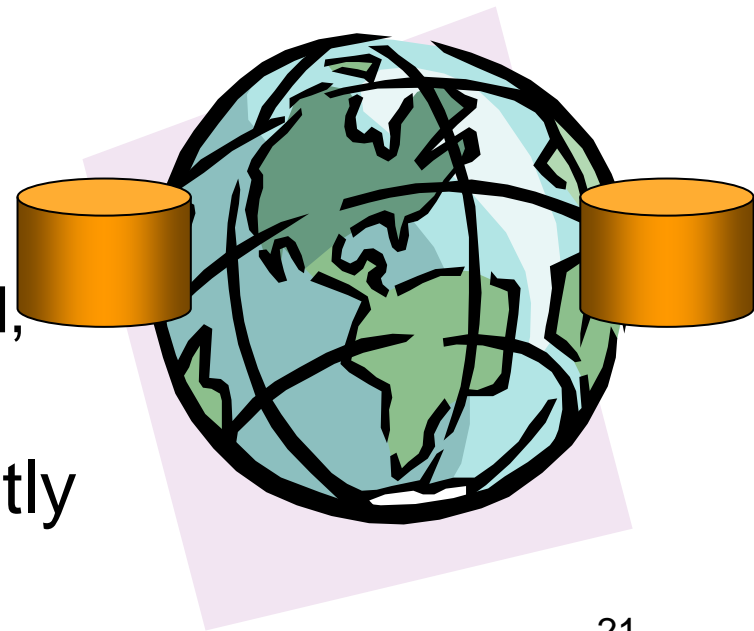
- Archived information “rusts” if it is not accessed. Access is essential.
- Access costs money – who pays?
- Access sometimes uses IP, who pays?
- There are also technical problems:
 - Access formats different from the storage formats.
 - migration?
 - emulation?
 - Gold Standards?

Archive Challenges

- Cost of administering storage:
 - Presently 10x to 100x the hardware cost.
- Resist attack: geographic diversity
- At 1GBps it takes 12 days to move a PB
- Store it in two (or more) places online (on disk).

A geo-plex

- Scrub it continuously (look for errors)
- On failure,
 - use other copy until failure repaired,
 - refresh lost copy from safe copy.
- Can organize the copies differently (e.g.: one by time, one by space)



Tangible Things (1)

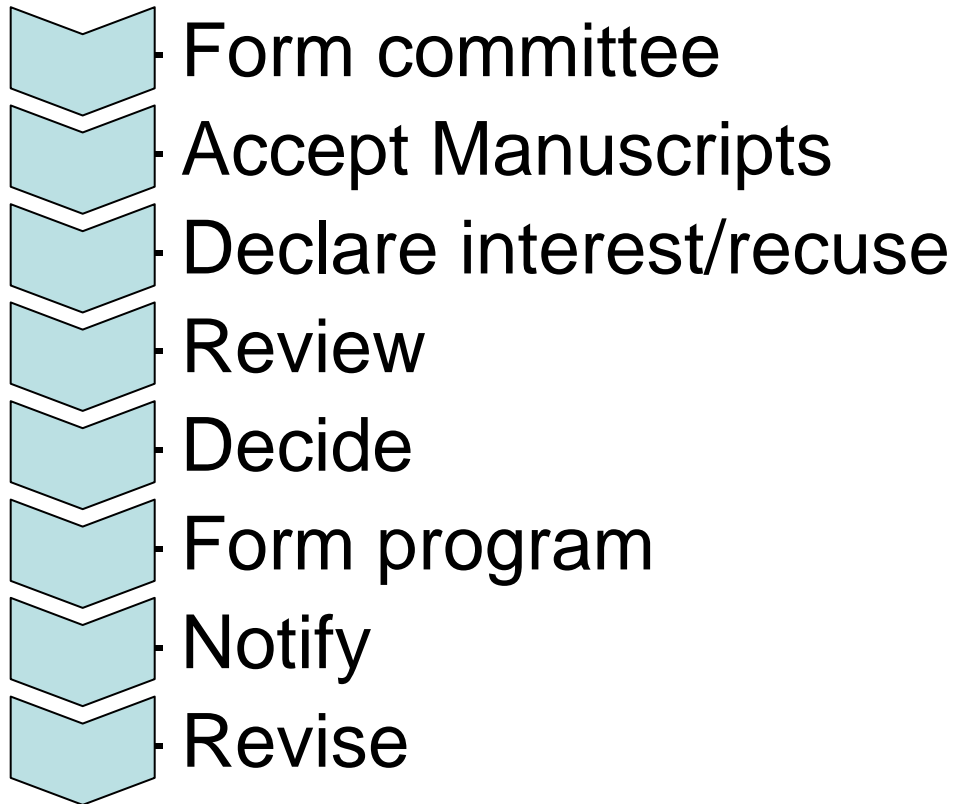


- “Information at your fingertips”
- Helping build PortablePubMedCentral
- Deployed US, China, England, Italy, South Africa, (Japan soon).
- Each site can accept documents
- Archives replicated
- Federate thru web services
- Working to integrate Word/Excel/... with PubmedCentral – e.g. WordML, XSD,
- To be clear: NCBI is doing 99% of the work.



Tangible Things (2)

- Currently support a conference **peer-review** system (~300 conferences)



Address http://insront.research.microsoft.com/faq/faq_author_error.asp

msn Search Web Highlight Viewer Blocked (29) Spaces

MSN Search: Conferen... Microsoft's Conferen...

Welcome to Microsoft's Conference Management Site

FAQ
Phases
Errors

Common Issues and Solutions

WARNING: CMT doesn't support Safari browser. Please use one of the browsers listed below.

If you continue to receive this error, please send email to cmt@microsoft.com along with the following information:

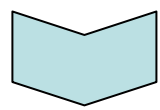
- What is the problematic URL?
- What is your Browser Type?
- What is your operating system?
- What is your CMT role? (e.g. Author, Reviewer, Chair)
- What is your CMT login email address?
- What are you trying to do?
- What time did this occur?
- What was your last action before seeing this error? (e.g. Click on the 'submit' button)?
- Additional information?

Microsoft Research

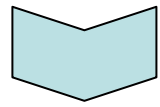
Powered by Conference Management Tool 1

Tangible Things (2)

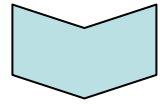
- Add publishing steps



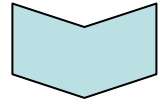
– Form committee



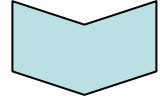
– Accept Manuscripts



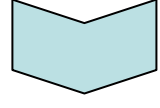
– Declare interest/recuse



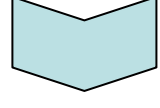
– Review



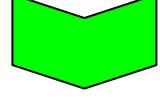
– Decide



– Form program



– Notify



– Publish

& improve

author-reader experience

- Manage versions
- Capture data
- Interactive documents
- Capture Workshop
 - presentations
 - proceedings
- Capture classroom
ConferenceXP
- Moderated discussions
of published articles
- Connect to Archives

Why Not a Wiki?

- Peer-Review is
 - It is very structured
 - It is moderated
 - There is a degree of confidentiality
- Wiki is egalitarian
 - It's a conversation
 - It's completely transparent
- Don't get me wrong:
 - Wiki's are great
 - SharePoints are great
 - But.. Peer-Review is different.
 - And, incidentally: review of proposals, projects,... is more like peer-review.

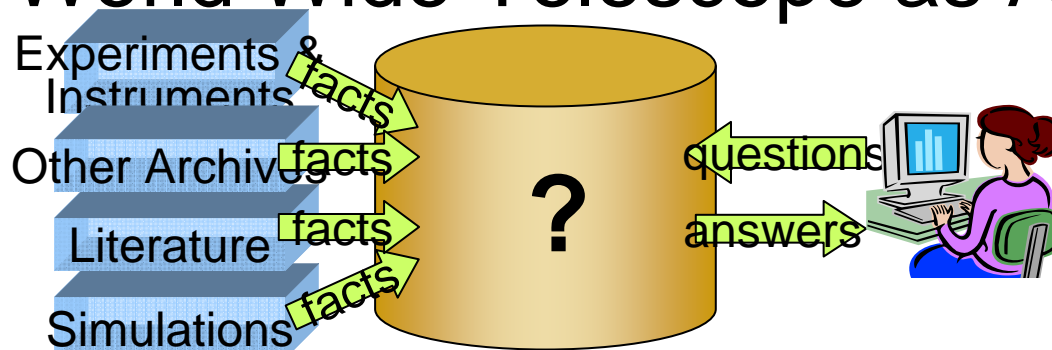


Why Am I Telling You This?

- “Library Science” has challenging problems (not all of them are social/economic).
- “Library Science”
is central to the way we do science:
 - Teaching & research
 - Review & evaluation
 - Search & access
- Increasingly Library Science
is Computer Science
- Its Info-Info in the X-info model
- Its not just search.

Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype

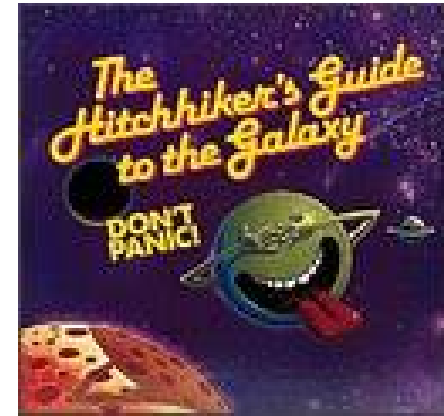


The Big Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute ²⁷
 - Batch query scheduling

So... What about Publishing Data?

- The answer is **42**.
- But...
 - What are the units?
 - How precise? How accurate $42.5 \pm .01$
 - Show your work
data *provenance*

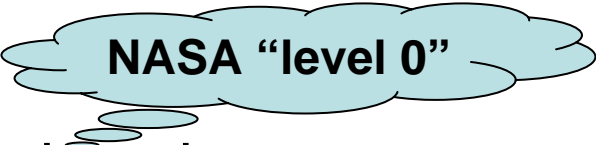


Publishing Data

<i>Roles</i>	<i>Traditional</i>	<i>Emerging</i>
Authors	Scientists	Collaborations
Publishers	Journals	Project www site
Curators	Libraries	Bigger Archives
Consumers	Scientists	Scientists

- Exponential growth:
 - Projects last at least 3-5 years
 - Data sent to deep archive at project end
 - Data will **never** be centralized
- More responsibility on projects
 - Becoming Publishers and Curators
 - Often no explicit funding to do this **(must change)**
- Data will reside with projects
 - Analyses must be close to the data (see later)
- Data cross-correlated with Literature and Metadata²⁹

Data Curation Problem Statement

- Once published, scientific data needs to be available forever, so that the science can be reproduced/extended.
- What does that mean?
 - **Data** can be characterized as  **NASA “level 0”**
 - **Primary Data:** could not be reproduced
 - **Derived data:** could be derived from *primary* data.
 - **Meta-data:** how the data was collected/derived **is primary**
 - Must be preserved
 - Includes design docs, software, email, pubs, personal notes, teleconferences, ...

Thought Experiment

- You have collected some data and want to publish science based on it.
- How do you publish the data so that others can read it and reproduce your results in 100 years?
 - Document collection process?
 - How document data processing (scrubbing & reducing the data)?
 - Where do you put it?

The Vision: Global Data Federation

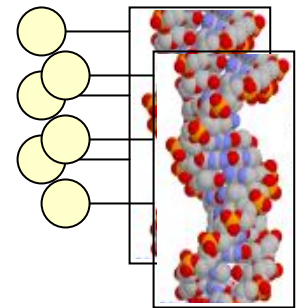
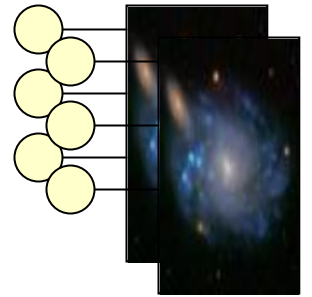
- Massive datasets live near their owners:
 - Near the instrument’s software pipeline
 - Near the applications
 - Near data knowledge and curation
- Each Archive publishes a (web) service
 - Schema: documents the data
 - Methods on objects (queries)
- Scientists get “personalized” extracts
- Uniform access to multiple Archives
 - A common global schema



Federation

Objectifying Knowledge

- This requires agreement about
 - **Units:** cgs
 - **Measurements:** who/what/when/where/how
 - **CONCEPTS:**
 - What's a planet, star, galaxy,...?
 - What's a gene, protein, pathway...?
- **Need to objectify science:**
 - what are the objects?
 - what are the attributes?
 - What are the methods (in the OO sense)?
- This is mostly Physics/Bio/Eco/Econ/...
But CS can do generic things



Objectifying Knowledge

- This requires agreement about

Warning!

Painful discussions ahead:

– **CONCEPTS:**

- What's a planet, star, galaxy,...?

The “O” word: Ontology

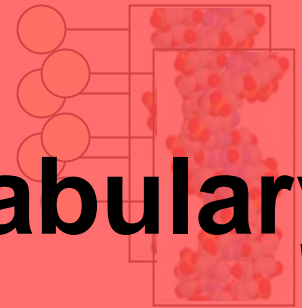
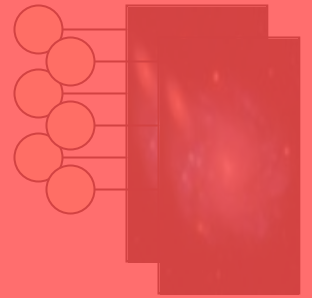
The “S” word: Schema

The “CV” words:

Controlled Vocabulary

Domain experts do not agree

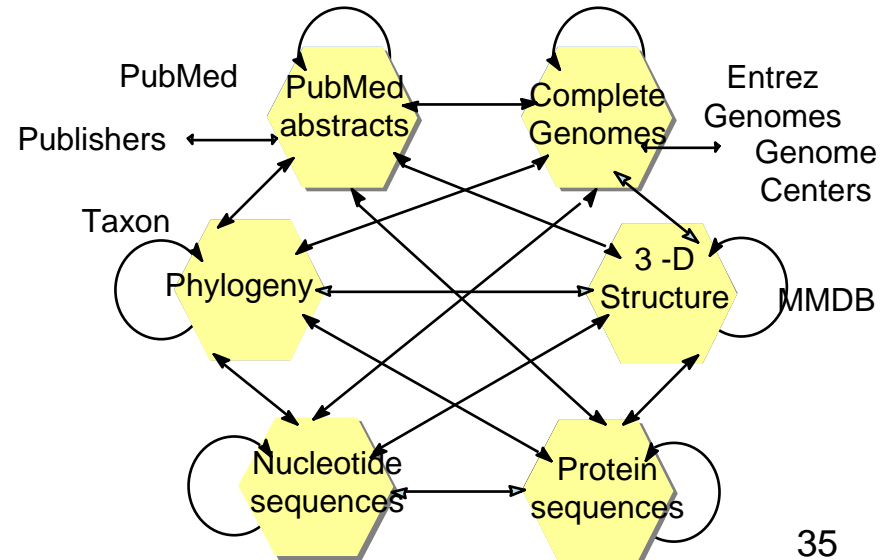
But CS can do generic things



The Best Example: Entrez-GenBank

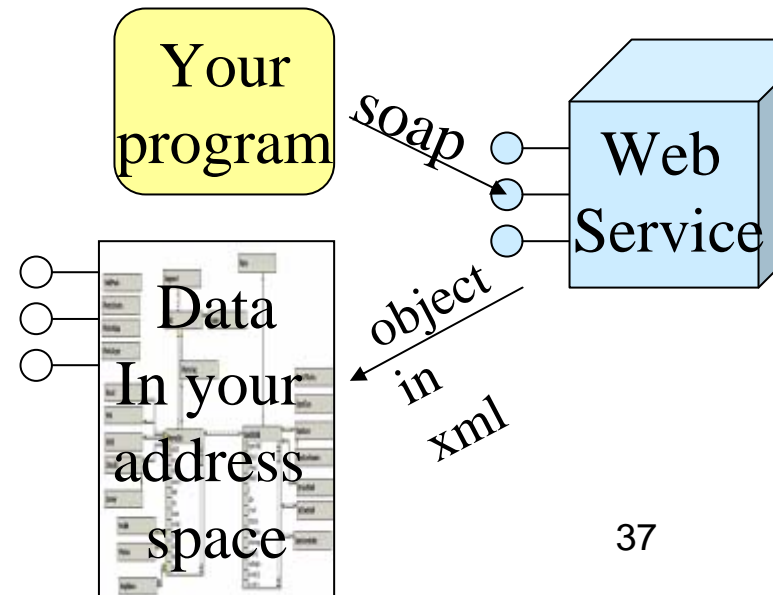
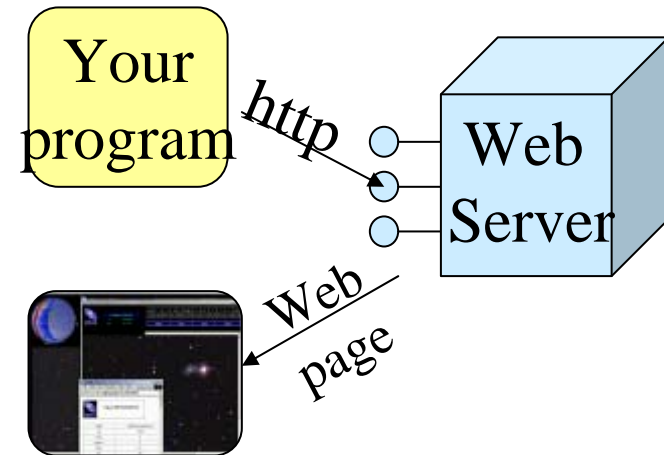
<http://www.ncbi.nlm.nih.gov/>

- Sequence data deposited with Genbank
- Literature references Genbank ID
- BLAST searches Genbank
- Entrez integrates and searches
 - PubMedCentral
 - PubChem
 - Genbank
 - Proteins, SNP,
 - Structure,...
 - Taxonomy...



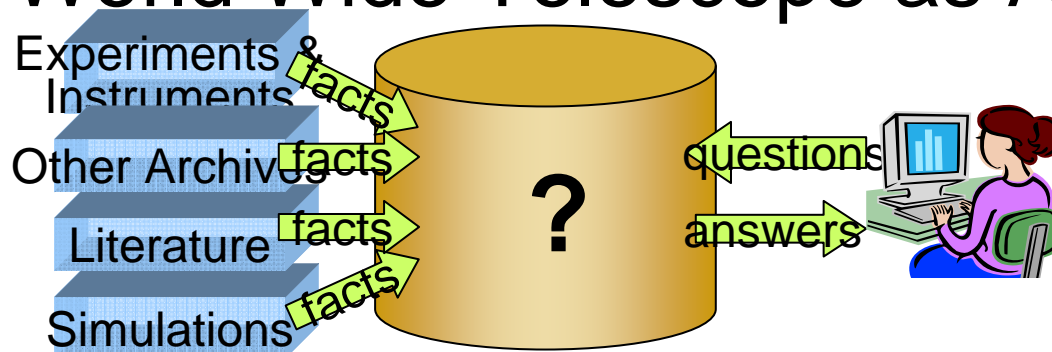
Web Services: Enable Federation

- **Web SERVER:**
 - Given a url + parameters
 - Returns a web page (often dynamic)
- **Web SERVICE:**
 - Given a XML document (soap msg)
 - Returns an XML document
 - Tools make this look like an RPC.
 - $F(x,y,z)$ returns (u, v, w)
 - Distributed objects for the web.
 - + naming, discovery, security,...
- **Internet-scale distributed computing**
- **Now: Find object models for each science.**



Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype



The Big Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute ³⁸
 - Batch query scheduling

World Wide Telescope Virtual Observatory

<http://www.us-vo.org/>

<http://www.ivoa.net/>

- Premise: Most data is (or could be online)
- So, the Internet is the world's best telescope:
 - It has data on every part of the sky
 - In every measured spectral band: optical, x-ray, radio..
 - As deep as the best instruments (2 years ago).
 - It is up when you are up.
The “seeing” is always great
(no working at night, no clouds no moons no..).
 - It's a smart telescope:
links objects and data to literature on them.



Why Astronomy Data?

- **It has no commercial value**

- No privacy concerns
- Can freely share results with others
- Great for experimenting with algorithms

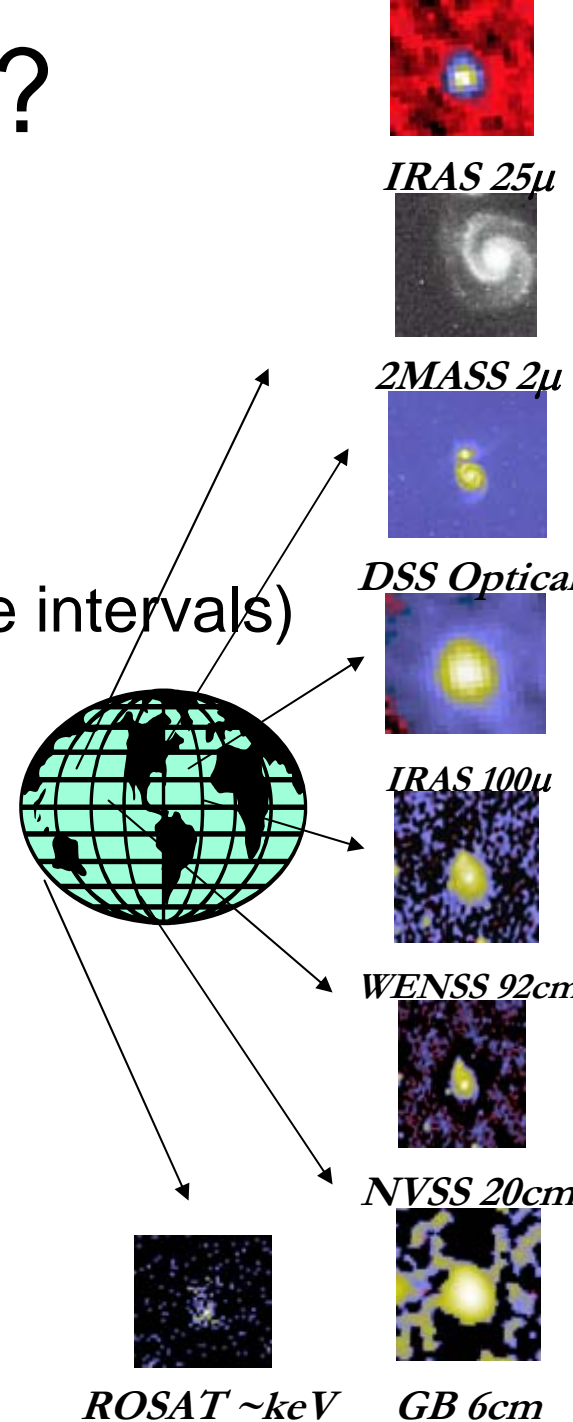
- **It is real and well documented**

- High-dimensional data** (with confidence intervals)
- Spatial data**
- Temporal data**

- **Many different instruments** from many **different places** and many **different times**

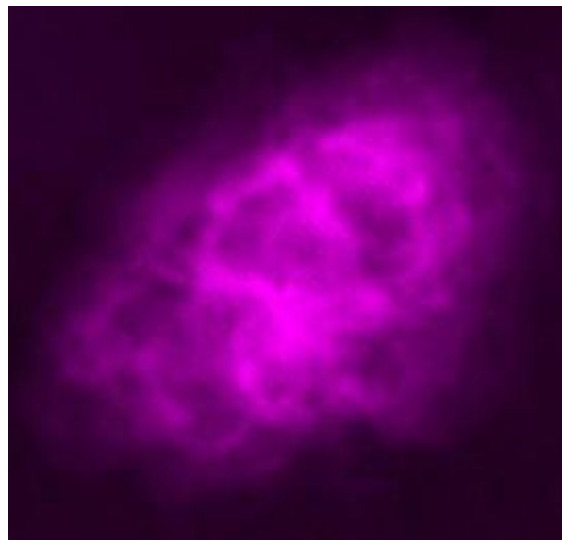
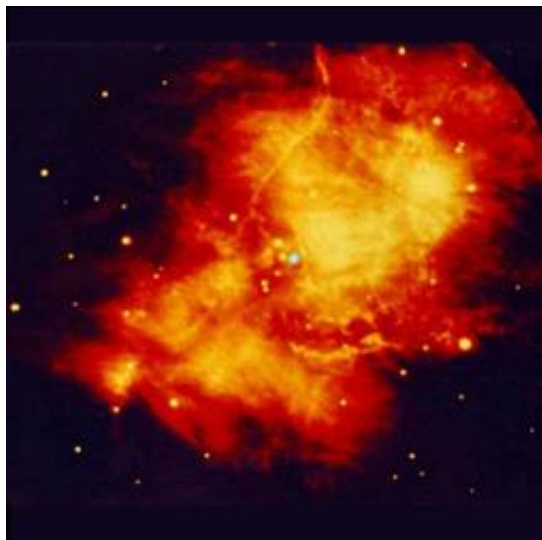
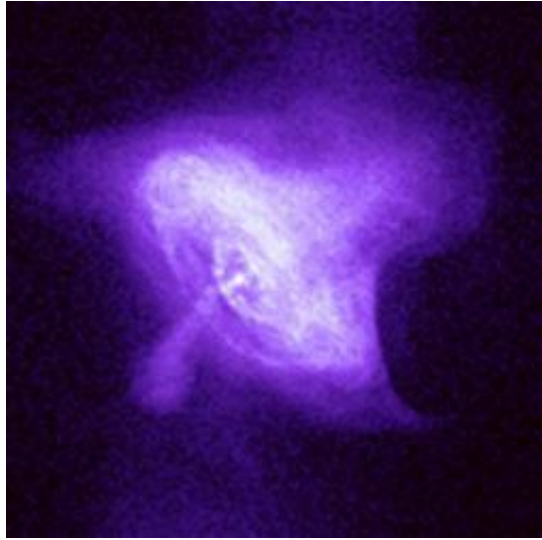
- **Federation is a goal**

- There is a lot of it (petabytes)



Time and Spectral Dimensions

The Multiwavelength Crab Nebulae



X-ray,
optical,
infrared, and
radio

views of the nearby
Crab Nebula, which is
now in a state of chaotic
expansion after a
supernova explosion
first sighted in 1054
A.D. by Chinese
Astronomers.

SkyServer.SDSS.org

- A modern archive
 - Access to Sloan Digital Sky Survey Spectroscopic and Optical surveys
 - Raw Pixel data lives in file servers
 - Catalog data (derived objects) lives in Database
 - Online query to any and all
- Also used for education
 - 150 hours of online Astronomy
 - Implicitly teaches data analysis
- Interesting things
 - Spatial data search
 - Client query interface via Java Applet
 - Query from Emacs, Python,
 - Cloned by other surveys (a template design)
 - Web services are core of it.



SkyServer

SkyServer.SDSS.org

- Like the TerraServer, but looking the other way: a picture of $\frac{1}{4}$ of the universe
- Sloan Digital Sky Survey Data: Pixels + Data Mining
- About 400 attributes per “object”
- Spectrograms for 1% of objects

The screenshot displays the SkyServer Object Explorer interface. The main content area shows the object ID **SDSS J121755.52+002623.87** and its classification as a **GALAXY** with coordinates **ra=184.481364, dec=0.4399658** and **ObjId = 2255030989160697**. A table of flags is visible, including **STATUS** (TARGET PRIMARY OK, STRIP1 OK, SCANLINE PSEGMENT RESOLVED OK, RUN GOOD, OCT) and **Flags** (BINNED1, SATURATED INTERP, COSMIC_RAY, NOPESTRO, NOBLEND, CHILD, BLENDED). Below this is a table of photometric data:

u	g	r	i	z	reddening_r	petroRad_r
17.57	15.88	15.52	15.21	15.43	0.07	25.108

Another table shows **SpecObjId= 81006758046203904** with a table of spectroscopic data:

plate	md	rowid	z	zErr	zConf	specClass	ra	dec	fileMag_r	objid
287	02023	631	0.100	0.00006	9.93E-1	GALAXY	184.48137	0.43999	19.57	2255030989160697

The interface also features a sidebar with navigation options like **Search by** (ObjId, Ra, dec, 5 point SDSS, Plate, Filter, SpecObjId), **Summary**, **PhotoObj** (Field, Frame, PhotoZ, Neighbors, Navigate, FITS), **SpecObj** (SpecLine, SpecI, modelIndex, X-Cred Shift, E-Red Shift, Spectrum, Plate, FITS), **NED search**, **Virtual Sky**, **Save in Notes**, **Show Notes**, and **Print Page**. A spectrogram plot is visible at the bottom, and the **Cross-identifications** section is partially shown.

Demo of SkyServer

- Shows standard web server
- Pixel/image data
- Point and click
- Explore one object
- Explore sets of objects (data mining)

Get
SDSS
Cut Out
Image

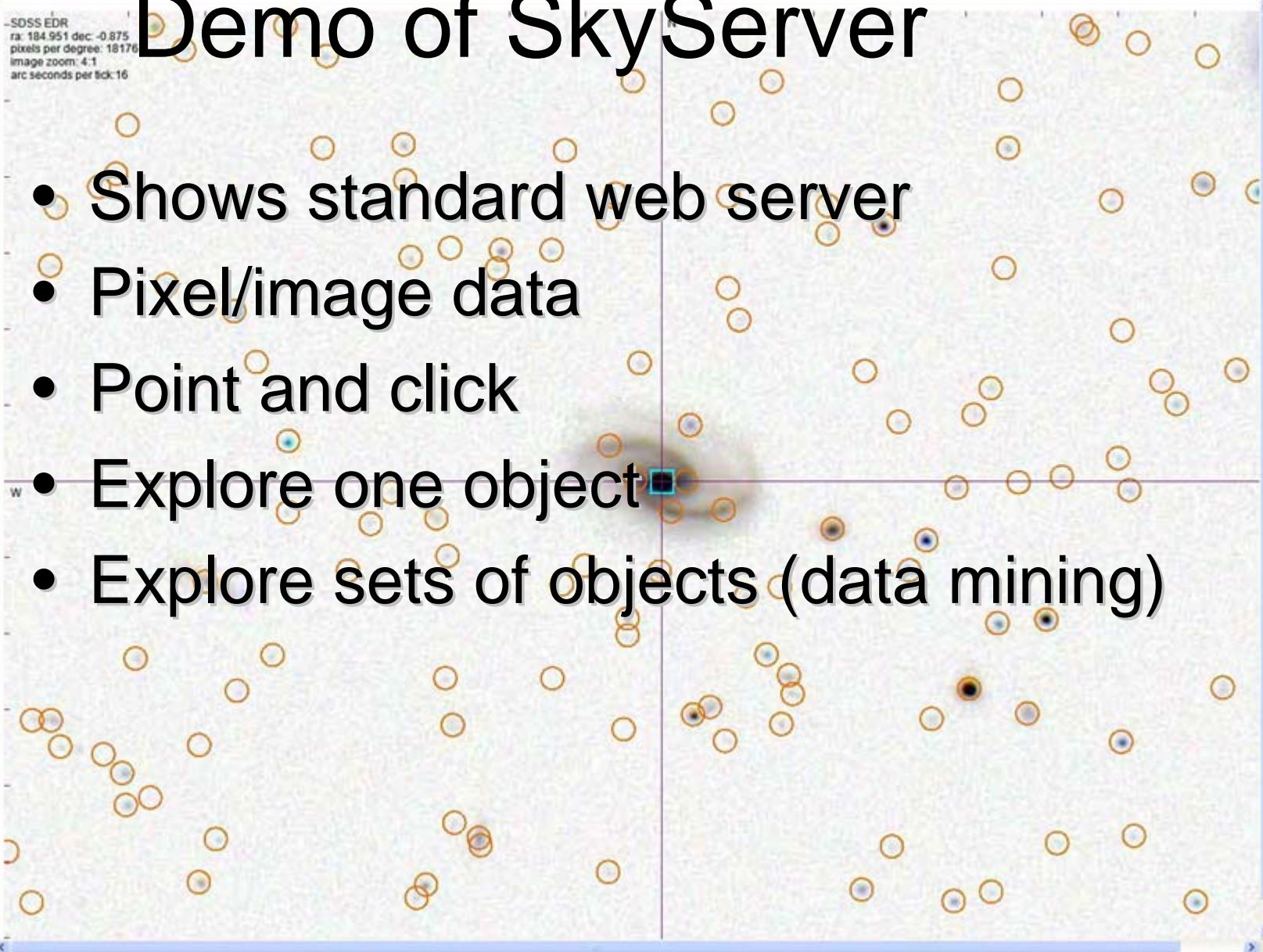
Get Jpeg

ra	184.9511 deg
dec	-0.8754 deg
zoom	-1 [-5..0..5]
pixels per degree	18176 (9088 == 1:1) range is 100 ... 1,000,000
width	1400 pixels
height	1000 pixels

- Grid
- PhotoObjs
- SpecObjs
- Invert Image

There are also HTTP and SOAP interfaces to this web service. See: Sds:Cutout.aspx?WSDL. A brief description of the service and its parameters is at: [intro](#).

-SDSS EDR
ra: 184.951 deg
dec: -0.8754 deg
pixels per degree: 18176
image zoom: 4.1
arc seconds per tick: 16



SkyQuery (<http://skyquery.net/>)

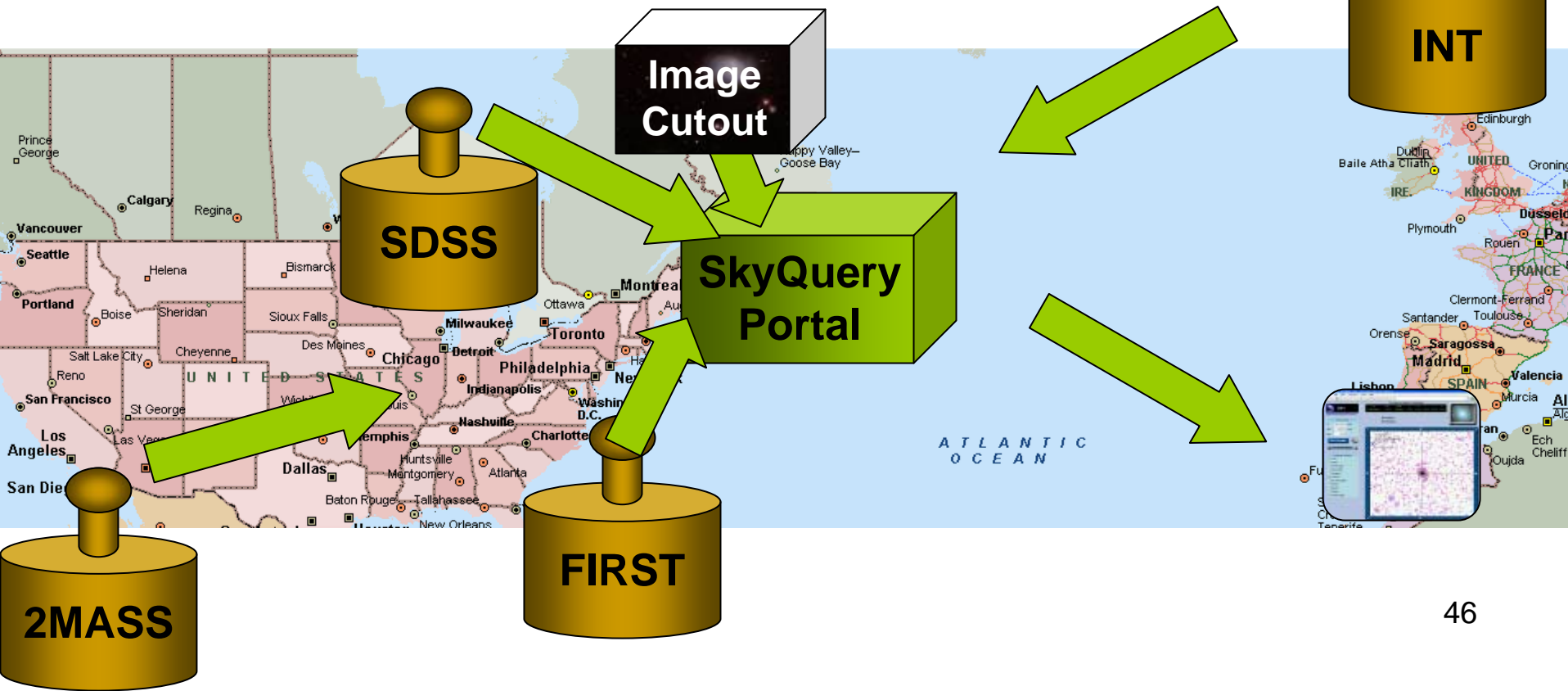
- Distributed Query tool using a set of web services
- Many astronomy archives from Pasadena, Chicago, Baltimore, Cambridge (England)
- Has grown from 4 to 15 archives, now becoming international standard
- WebService Poster Child
- Allows queries like:

```
SELECT o.objId, o.r, o.type, t.objId
FROM SDSS:PhotoPrimary o,
      TWOMASS:PhotoPrimary t
WHERE XMATCH(o,t)<3.5
      AND AREA(181.3,-0.76,6.5)
      AND o.type=3 and (o.I - t.m_j)>2
```

The screenshot shows the SkyQuery.net web interface. On the left, there is a list of surveys including ROSAT, CALEX, INTWFS, RCI, ULS, TWODF, TWOOZ, SDSS, HSTEP, HDEN, HDFO, GOODSN, COODSS, UDF, TWOMASS, PSZ, IRAS, NVSS, FIRST, and AGC. The main area is titled 'Table Info' and contains a search box and a 'Search' button. Below this is a 'SkyQL query' input field with the following SQL query: `SELECT o.objId, o.r, o.type, t.objId, t.g, t.i FROM SDSS:PhotoPrimary o, TWOMASS:PhotoPrimary t WHERE XMATCH(o,t)<3.5 AND AREA(181.3,-0.76,6.5)`. Below the query field is a 'Submit' button and a row of buttons numbered 1 to 12. At the bottom, there is a table with columns: s_objid, s_g, s_j, l_objid, l_g, l_j, class, x_ra, x_dec, match. The first row of data is: 58210188667706581|20.39947|19.56215|13008918|20.472|19.255|0.1895|355.51461|0.00402|1. On the right side of the interface, there is a star field visualization with a grid and a bright star.

SkyQuery Structure

- Each SkyNode publishes
 - Schema Web Service
 - Database Web Service
- Portal is
 - Plans Query (2 phase)
 - Integrates answers
 - Is itself a web service



SkyNode Basic Web Services

- Metadata information about resources
 - Waveband
 - Sky coverage
 - Translation of names to universal dictionary (UCD)
- Simple search patterns on the resources
 - Cone Search
 - Image mosaic
 - Unit conversions
- Simple filtering, counting, histogramming
- On-the-fly recalibrations

Portals: Higher Level Services

- Built on Atomic Services
- Perform more complex tasks
- Examples
 - Automated resource discovery
 - Cross-identifications
 - Photometric redshifts
 - Outlier detections
 - Visualization facilities
- Goal:
 - Build custom portals in days from existing building blocks (like today in IRAF or IDL)

SkyServer/SkyQuery Evolution

MyDB and Batch Jobs

Problem: need multi-step data analysis (not just single query).

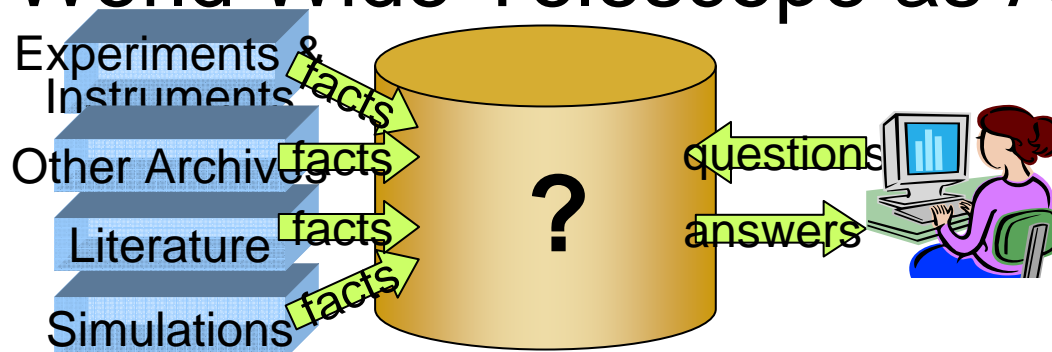
Solution: Allow personal databases on portal

Problem: some queries are monsters

Solution: “Batch schedule” on portal. Deposits answer in personal database.

Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype



The Big Problems

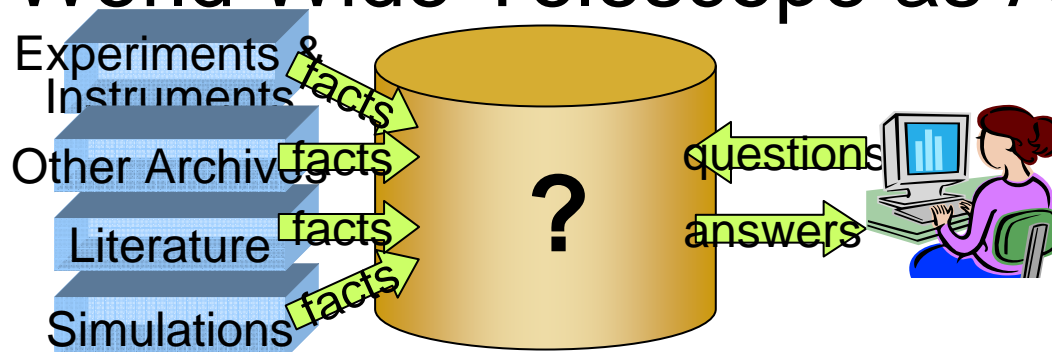
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute ⁵¹
 - Batch query scheduling

Call to Action

- X-info is emerging.
- Computer Scientists can help in many ways.
 - Tools
 - Concepts
 - Provide technology consulting to the community
- There are great CS research problems here
 - Modeling
 - Analysis
 - Visualization
 - Architecture

Outline

- The Evolution of X-Info
- Online Literature
- Online Data
- The World Wide Telescope as Archetype



The Big Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *reorganize* it
- How to coexist with others
- Query and Vis tools
- Integrating data and Literature
- Support/training
- Performance
 - Execute queries in a minute ⁷¹
 - Batch query scheduling

References

<http://SkyServer.SDSS.org/>

<http://research.microsoft.com/pubs/>

<http://research.microsoft.com/Gray/SDSS/> (download personal SkyServer)

• **Data Mining the SDSS SkyServer Database**

Jim Gray; Peter Kunszt; Donald Slutz; Alex Szalay; Ani Thakar; Jan Vandenberg; Chris Stoughton Jan. 2002 40 p.

An earlier paper described the Sloan Digital Sky Survey's (SDSS) data management needs [Szalay1] by defining twenty database queries and twelve data visualization tasks that a good data management system should support. We built a database and interfaces to support both the query load and also a website for ad-hoc access. This paper reports on the database design, describes the data loading pipeline, and reports on the query implementation and performance. The queries typically translated to a single SQL statement. Most queries run in less than 20 seconds, allowing scientists to interactively explore the database. This paper is an in-depth tour of those queries. Readers should first have studied the companion overview paper "The SDSS SkyServer – Public Access to the Sloan Digital Sky Server Data" [Szalay2].

• **SDSS SkyServer–Public Access to Sloan Digital Sky Server Data**

Jim Gray; Alexander Szalay; Ani Thakar; Peter Z. Zunszt; Tanu Malik; Jordan Raddick; Christopher Stoughton; Jan Vandenberg November 2001 11 p.:

[Word](#) 1.46 Mbytes [PDF](#) 456 Kbytes *The SkyServer provides Internet access to the public Sloan Digital Sky Survey (SDSS) data for both astronomers and for science education. This paper describes the SkyServer goals and architecture. It also describes our experience operating the SkyServer on the Internet. The SDSS data is public and well-documented so it makes a good test platform for research on database algorithms and performance.*

• **The World-Wide Telescope**

Jim Gray; Alexander Szalay August 2001 6 p.: [Word](#) 684 Kbytes [PDF](#) 84 Kbytes

All astronomy data and literature will soon be online and accessible via the Internet. The community is building the Virtual Observatory, an organization of this worldwide data into a coherent whole that can be accessed by anyone, in any form, from anywhere. The resulting system will dramatically improve our ability to do multi-spectral and temporal studies that integrate data from multiple instruments. The virtual observatory data also provides a wonderful base for teaching astronomy, scientific discovery, and computational science.

• **Designing and Mining Multi-Terabyte Astronomy Archives**

Robert J. Brunner; Jim Gray; Peter Kunszt; Donald Slutz; Alexander S. Szalay; Ani Thakar

June 1999 8 p.: [Word](#) (448 Kbytes) [PDF](#) (391 Kbytes)

The next-generation astronomy digital archives will cover most of the sky at fine resolution in many wavelengths, from X-rays, through ultraviolet, optical, and infrared. The archives will be stored at diverse geographical locations. One of the first of these projects, the Sloan Digital Sky Survey (SDSS) is creating a 5-wavelength catalog over 10,000 square degrees of the sky (see <http://www.sdss.org/>). The 200 million objects in the multi-terabyte database will have mostly numerical attributes in a 100+ dimensional space. Points in this space have highly correlated distributions.

• **There Goes the Neighborhood: Relational Algebra for Spatial Data Search,**

with Alexander S. Szalay, Gyorgy Fekete, Wil O'Mullane, Aniruddha R. Thakar, Gerd Heber, Arnold H. Rots, MSR-TR-2004-32,

• **Extending the SDSS Batch Query System to the National Virtual Observatory Grid,**

Maria A. Nieto-Santisteban, William O'Mullane, Jim Gray, Nolan Li, Tamas Budavari, Alexander S. Szalay, Aniruddha R. Thakar, MSR-TR-2004-12.

Explains how the astronomers are building personal databases and a simple query scheduler into their astronomy data-grid portals.