RESEARCH DATA ALLIANCE

Playing YOUR Part
RDA Third Plenary Meeting

26 – 28 March 2014
Croke Park Conference Centre, Dublin, Ireland

research data sharing without barriers
rd-alliance.org

Twitter: @resdatall, #RDAPlenary

Microsoft Research

Microsoft

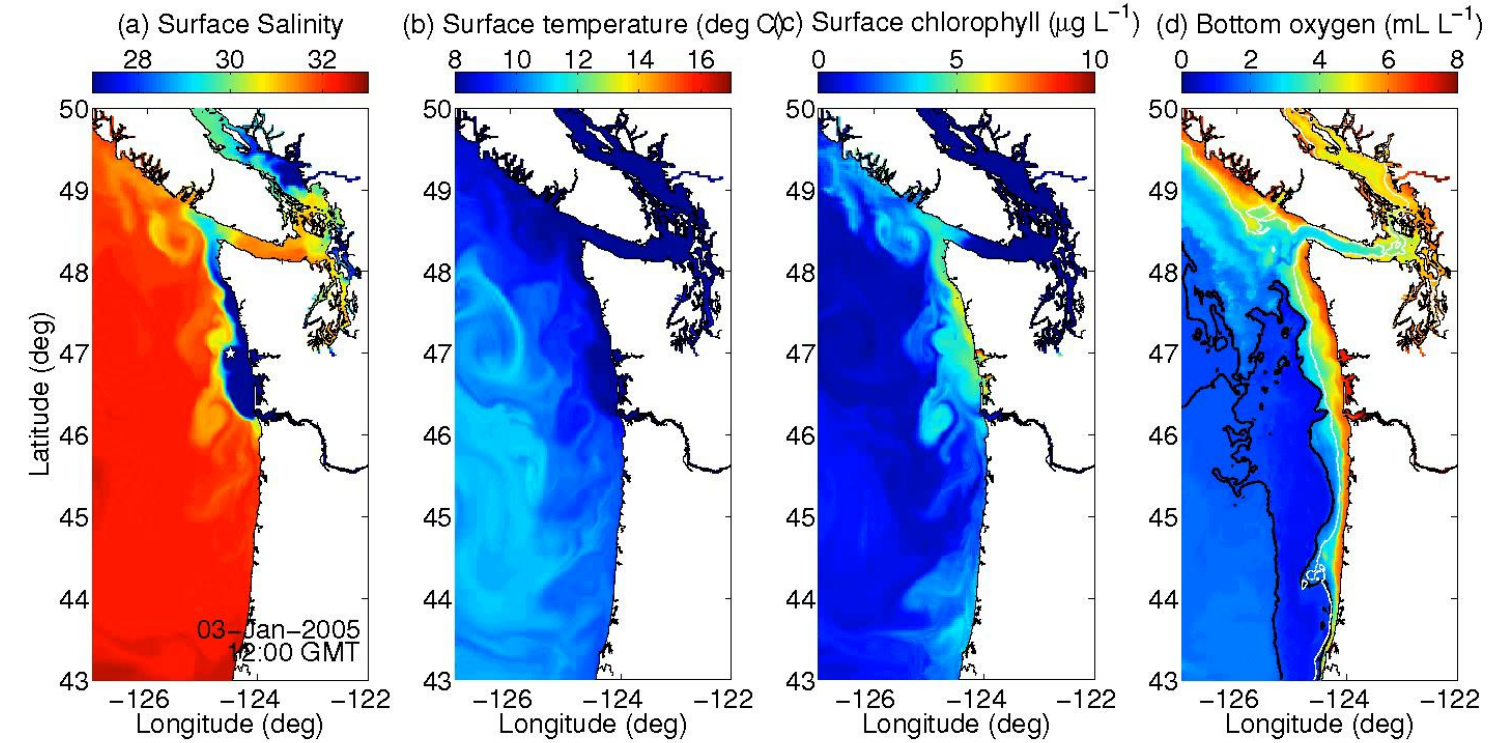
Beyond Open Access to Open Data

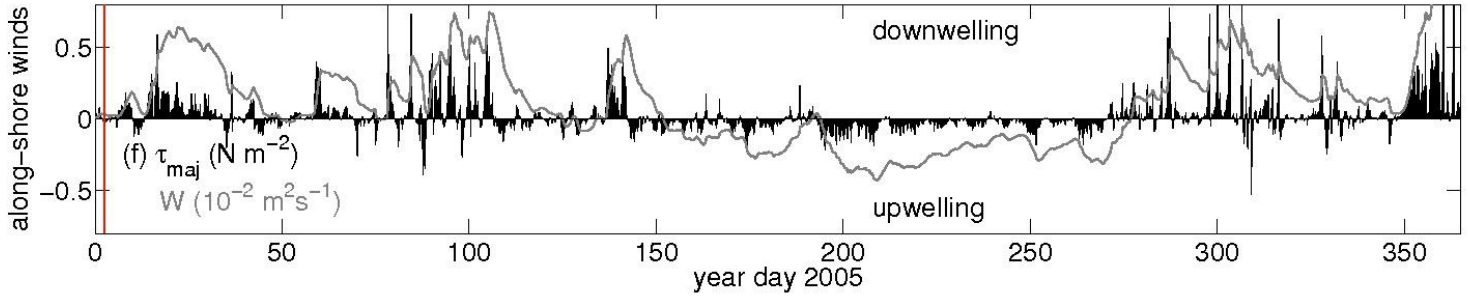
Tony Hey
Vice President
Microsoft Research

Ocean Modeling:
A Case Study from a Data Scientist

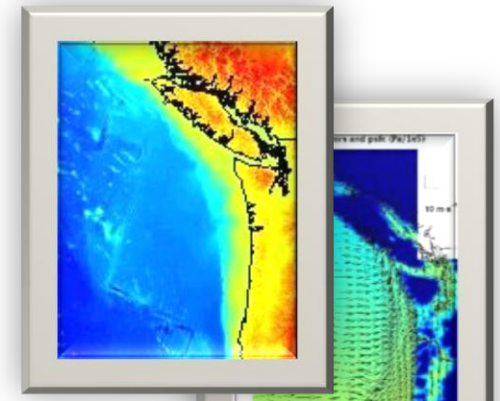
Parker MacCready
University of Washington
School of Oceanography
Coastal Modeling Group

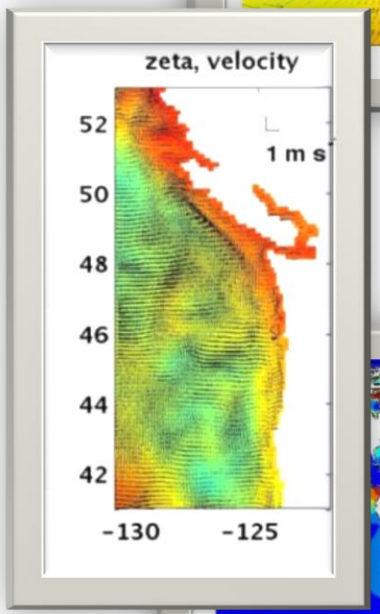
Realistic Regional Ocean Hindcasts

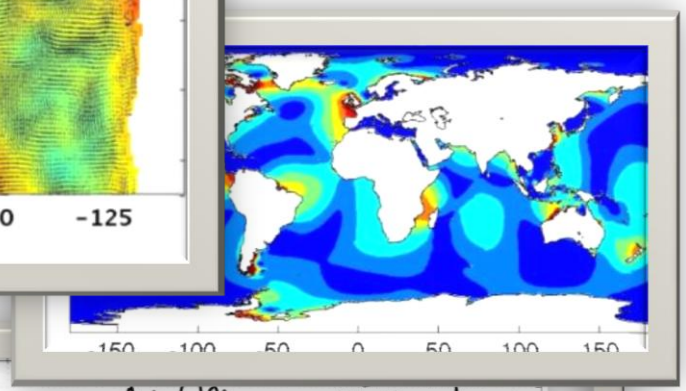


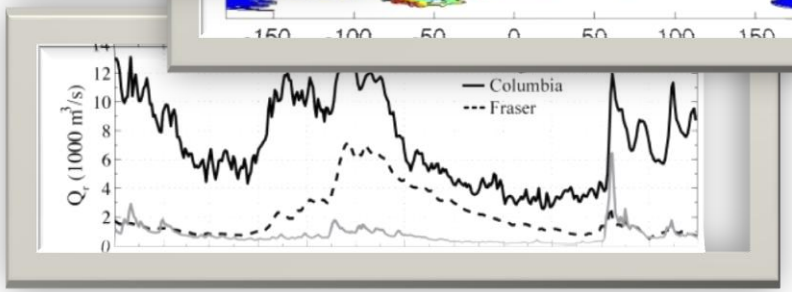


Workflow









ROMS

Cluster
200 cores
1 week/year

Raw Forcing Data → Forcing Data Processed into Standard Format → Model-specific Forcing Files → MODEL

Raw Observational Data → Observations Processed into Standard Format → SKILL TEST

MODEL → Output

SKILL TEST → Skill Result

Standard
Post Processing
1 week

2 TB/year

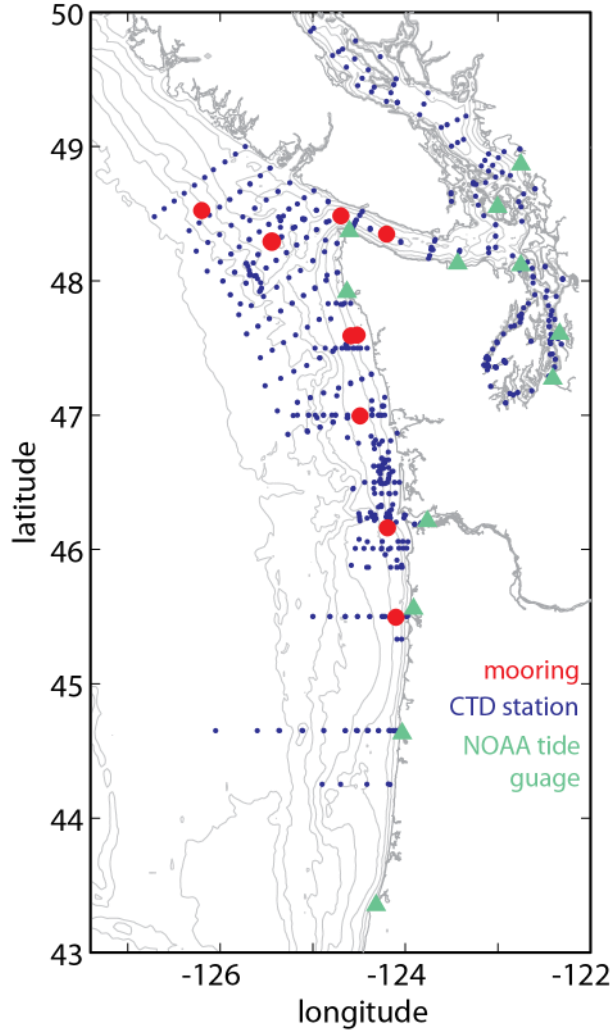
How do we know the model is working?

Comparisons were done to an extensive suite of in-situ observations

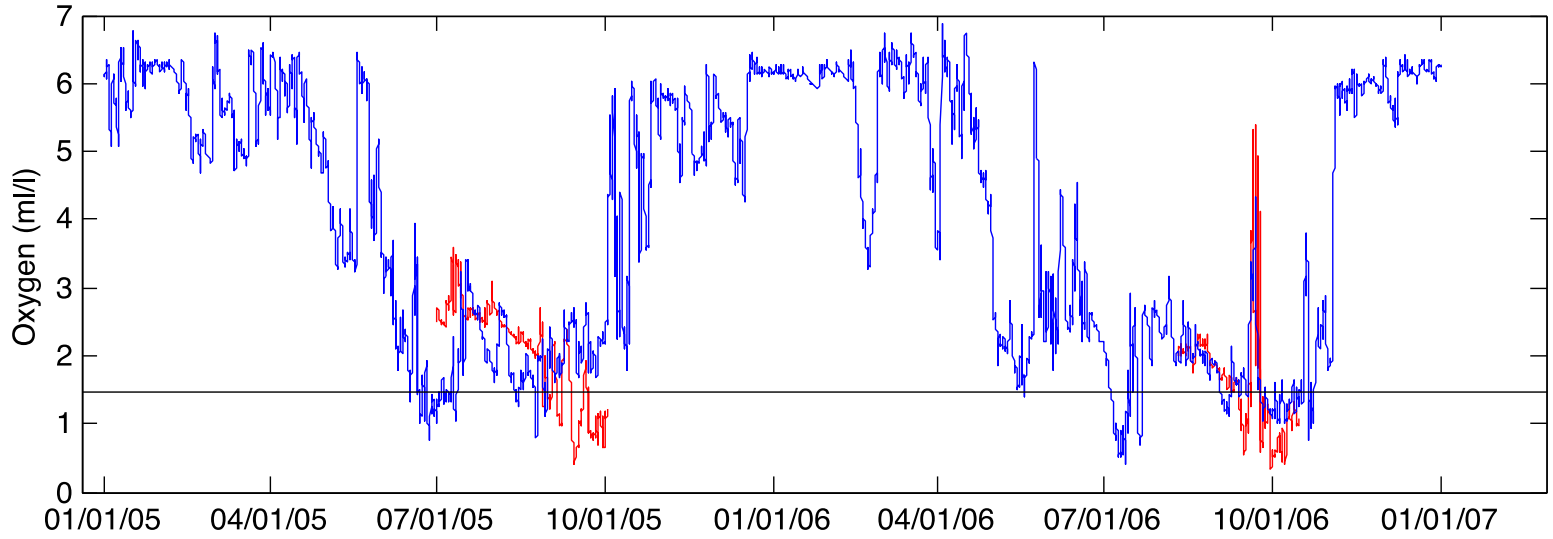
(see Sutherland et al. 2011)

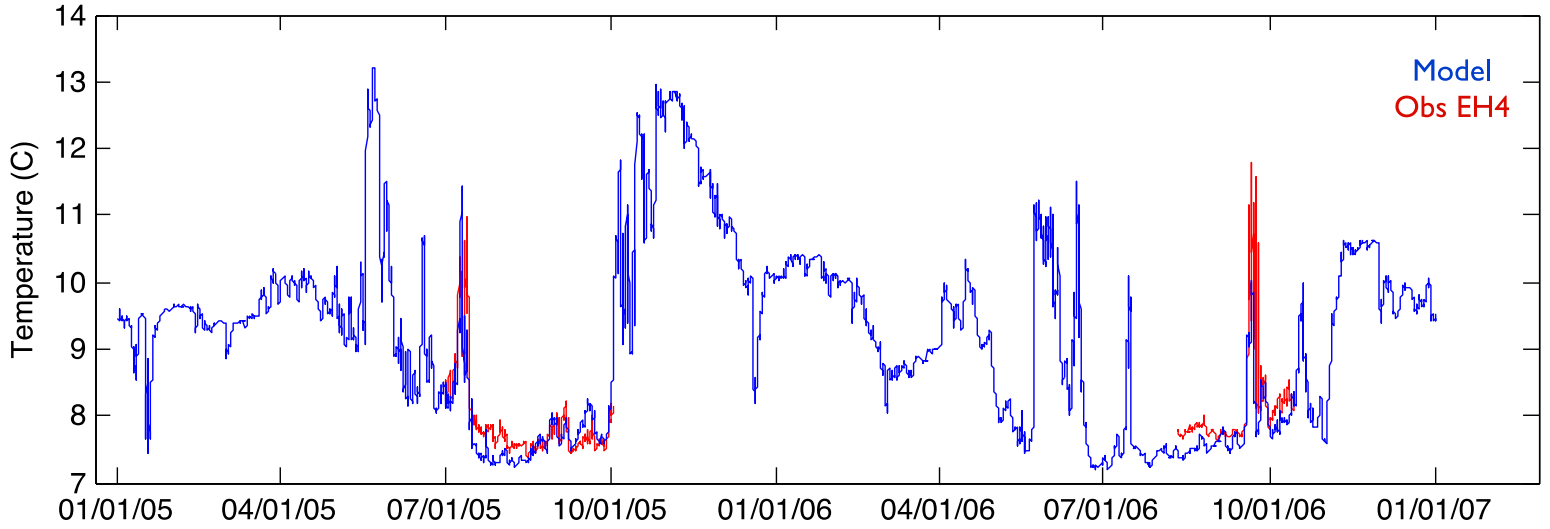
- ▲ sea surface height
  12 NOAA tide gauges
- • salinity and temperature
  over 2000 CTD casts from ECOHAB, RISE, DOE, PRISM, Hood Canal, IOS, King County, and NOAA
- ● velocity and moored S,T
  7 coastal ADCP / CTD moorings from the ECOHAB and RISE projects, 2 moorings from IOS

in addition: satellite and glider data



(Special thanks to R. Thomson, B. Peterson, B. Hickey, R. Kudela, N. Pelland, PRISM, ORCA, DOE, King County, Hood Canal citizen monitoring project, ORHAB, V. Trainer, and A. Odell for use of their data for model validation!!)





Model Validation – Dissolved Oxygen

Observations from Connolly et al, 2010

Need to work with many different data formats

Formats

- NetCDF
- .mat
- CSV
- XML
- Excel

Software

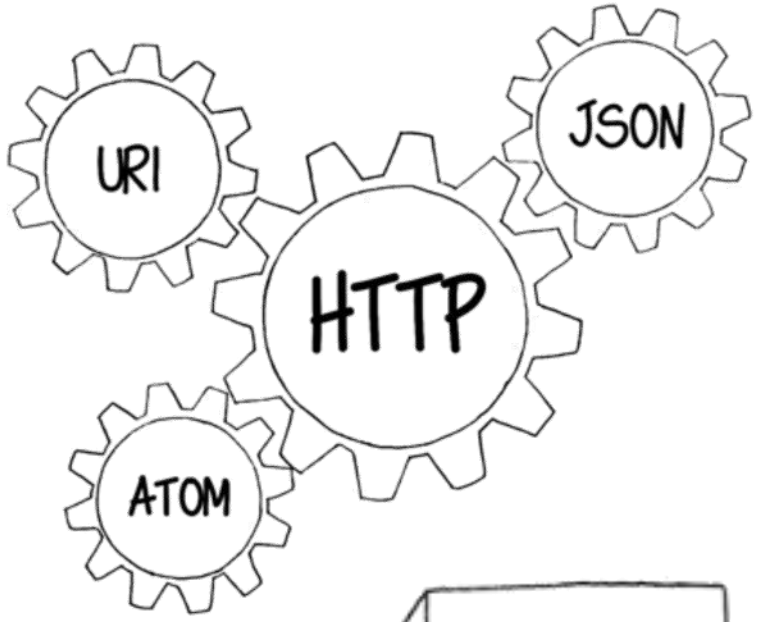
- MATLAB
- Python
- C#
- Java

Uses of the Data?

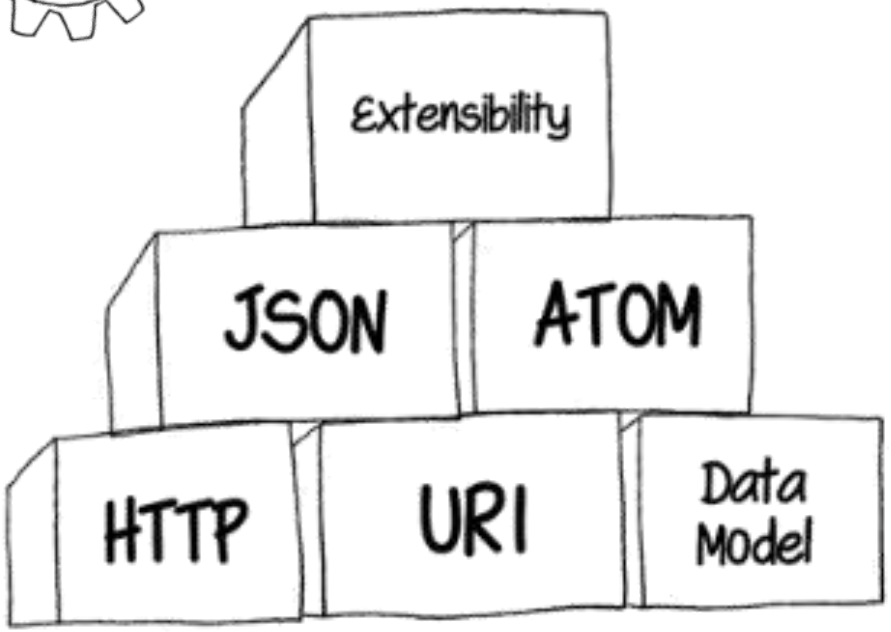
Physics: NSF
Harmful Algal Blooms: NSF+NOAA
Improve FW sources in CESM Climate Model: DOE
Shellfish Health: EPA
Ocean Acidification: WA State?

OData Protocol?

- OASIS standard protocol for creating and consuming data APIs.
- Builds on core protocols like HTTP and common methodologies like REST.
- Result is a uniform way to expose full-featured data APIs.



OData.org



Data Intensive Science

eScience and the Fourth Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

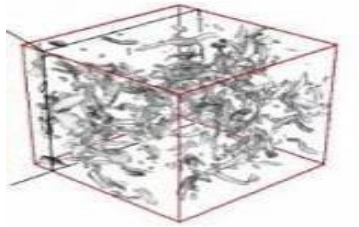
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

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



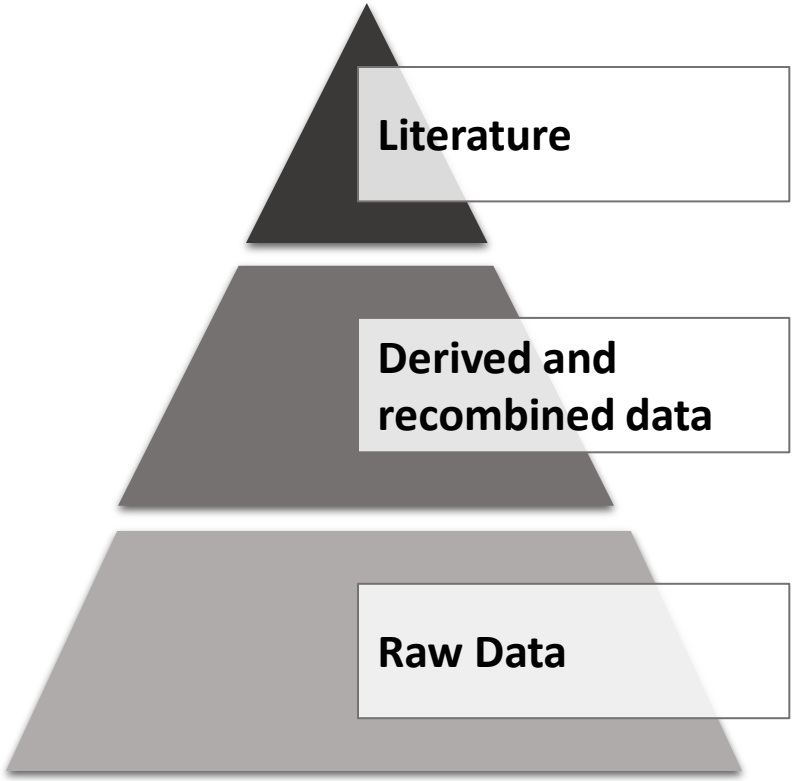
eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

(With thanks to Jim Gray)

All Scientific Data Online

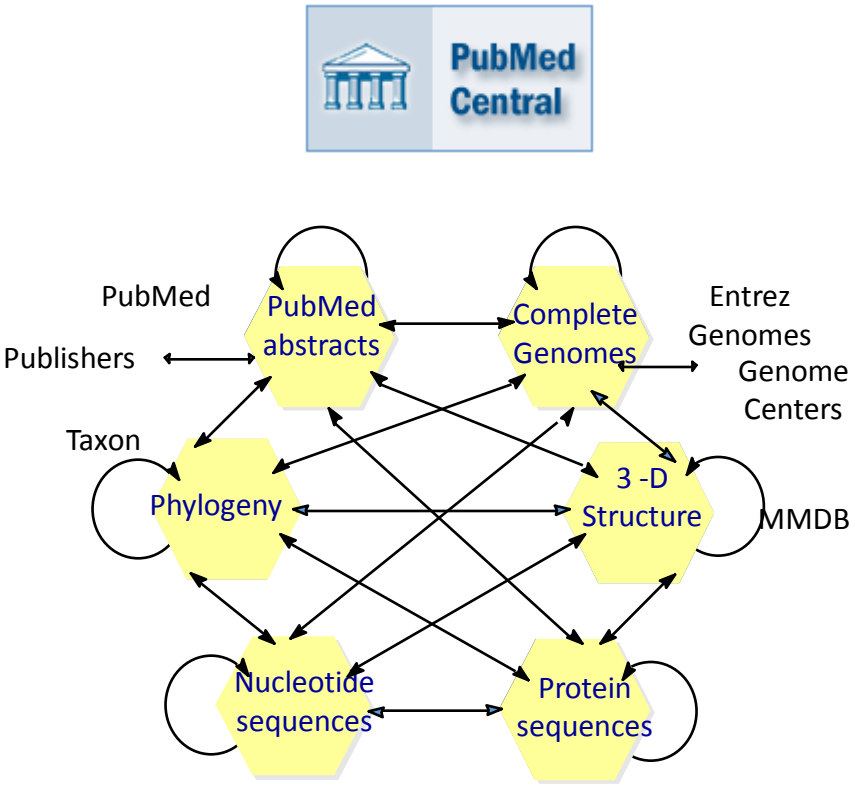
- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips –
  For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(*From Jim Gray's last talk)*

The US National Library of Medicine

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.



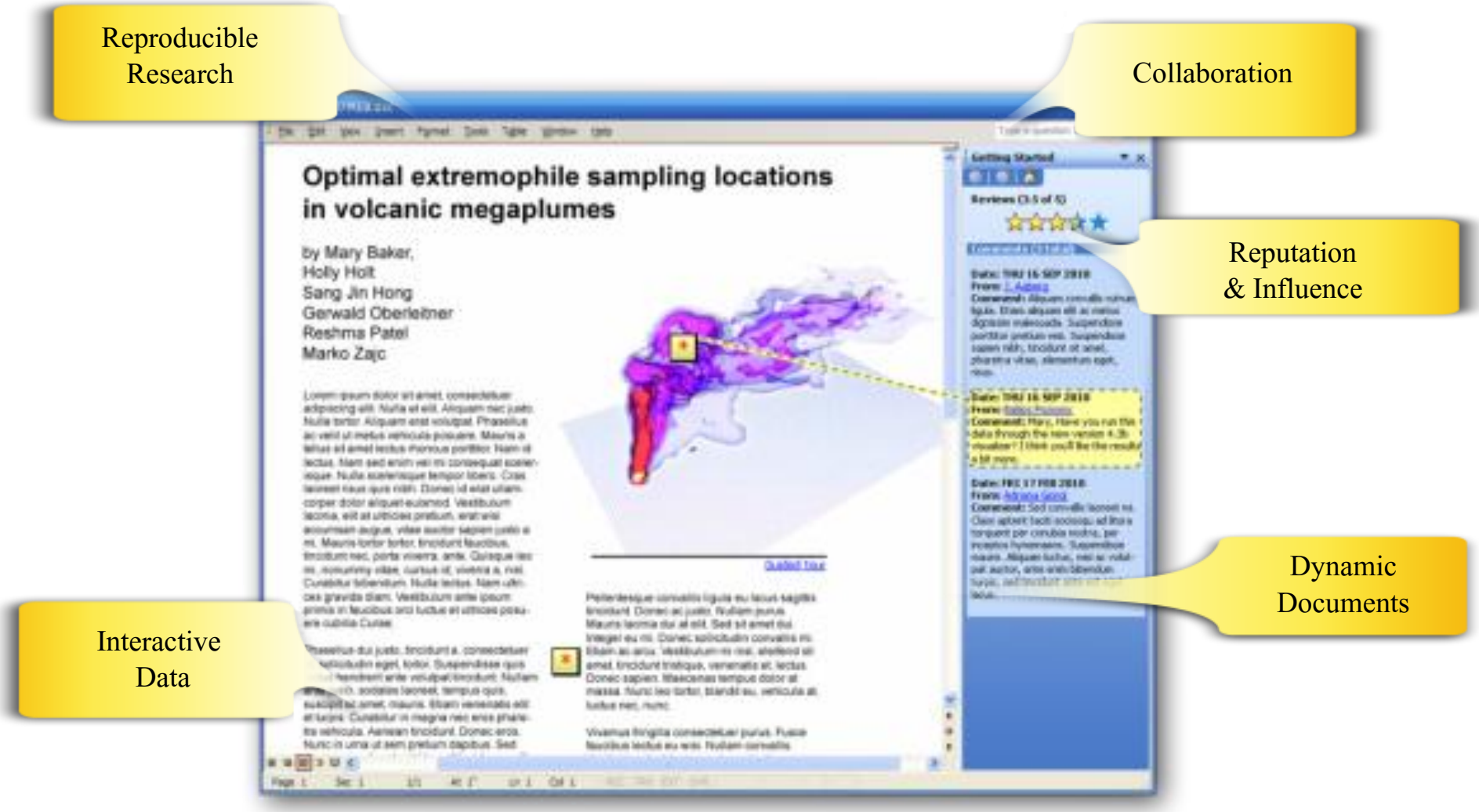
Entrez cross-database search

Open Data and Open Science

Open Access to Research Publications: The Tipping Point

- US White House Memorandum 26 February 2013
- Global Research Council Action Plan 30 May 2013
- G8 Science Ministers Joint Statement 12 June 2013
- University of California OA Mandate 2 August 2013
  - UC produces 40,000 publications per annum corresponding to about 2 – 3 % of all peer-reviewed articles in world each year
  - “The faculty remains committed to working with publishers to transform the publishing landscape in ways that are sustainable and beneficial to both the University and the public.”

Vision for a New Era of Research Reporting



(*Thanks to Bill Gates SC05*)

The Berlin Declaration 2003

- ‘To promote the Internet as a functional instrument for a global scientific knowledge base and for human reflection’
- Defines open access contributions as including:

  ‘original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material’

EUROPEAN COM

MEMO

Commission welcomes Parliam
Open Data rules

Today the European Parliament formally adopte
public sector information, thereby completing the

The EU Committee of Member States' Perman
endorsed the new rules on 10 April and they will
Council on 20t June. The agreement on the new
showing how all EU institutions are committed to
goldmine.

European Commission Vice-President Neelie Kr
efforts to bring government data closer to citi
finally getting the much needed legal framewor
jobs.

After this final endorsement, the Commission will
the most relevant elements addressed in the D
charging arrangements

Member States will have 24 months from the
Directive to transpose it into national laws Once
the data market in Europe by making all the gen
available for re-use. Developers, programmers,
able to get and re-use public sector data at zero
also have access to more exciting and inspirat
museums, libraries and archives now fall under th

G8 Open Data Charter will 'increase transparency' and 'fuel innovation'

Five key principles outlines how governments should release datasets for economic and social benefits

JAMES VINCENT | WEDNESDAY 19 JUNE 2013

Send mmend 16 Tweet 35 Share 5 +1 2 PRINT | EMAIL | AAA

News in pictures

Life & Style blogs

Why government must act to

http://www.independent.co.uk/life-style/gadgets-and-tech/g8-open-data-charter-will-increase-transparency-and-fuel-innovation-8665696.html

Collaboration and Sharing of Data is Expected and Growing

… expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.

NIH reaffirms its support for the concept of data sharing. We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health … The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.

A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government.

NSF Data Sharing Policy 2010

"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing."

All future grant proposals now require a two-page Data Management Plan that addresses the above requirement and the Plan will be subject to peer review.

Key driver from a UK Research Council

EPSRC Policy Framework on research data (May 2011)

- "all institutions in receipt of their funding should develop a clear roadmap for research data management, which should be implemented by May 1st 2015"
- "organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years"

PLOS' New Data Policy: Public Access to Data

By Liz Silva
Posted: February 24, 2014

UPDATE: A flurry of interest has arisen around the revised PLOS data policy that we announced in December and which will come into effect for research papers submitted next month. We are gratified to see a huge swell of support for the ideas behind the policy, but we note some concerns about how it will be implemented and how it will affect those preparing articles for publication in PLOS journals. We'd therefore like to clarify a few points that have arisen and once again encourage those with concerns to check the details of the policy or our FAQs, and to contact us with concerns if we have not covered them.

Is the policy about what to share, or about how and where to share it?

There is nothing new in the policy about what types and forms of data should be shared. As we said in December, "PLOS journals have requested data be available since their inception, but we believe that providing more specific instructions for authors regarding appropriate data deposition options, and providing more information in the published article as to how to access data, is important for readers and users of the research we publish." As we have further clarified, "the Data Policy states the 'minimal dataset' consists "of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. This does not mean that authors must submit all data collected as part of the research, but that they must provide the data that are relevant to the specific analysis presented in the paper." The 'minimal dataset'

Linking Publications to Data: The State of the Art

Astrophysics Data System ADS

· Find Similar Abstracts (with default settings below)

Toggle Highlighting

· Custom Format
· Electronic Refereed Journal Article (HTML)
· Full Refereed Journal Article (PDF/Postscript)
· FIND IT @ HARVARD
· arXiv e-print (arXiv:astro-ph/0412451)
· On-line Data
· References in the article
· Citations to the Article (84) (Citation History)
· Refereed Citations to the Article
· SIMBAD Objects (3)
· NED Objects (1)
· Also-Read Articles (Reads History)
·
· Translate This Page

Links to e-resources

Links to data

Links to objects

Title: Bow Shock and Radio Halo in the Merging Cluster A520

Authors: Markevitch, M.; Govoni, F.; Brunetti, G.; Jerius, D.

Affiliation: AA(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; Space Research Institute, Russian Academy of Sciences, 84/32 Profsoyuznaya Street, Moscow 117997, Russia. maxim@head.cfa.harvard.edu), AB(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AC(Istituto di Radioastronomia del CNR, via Gobetti 101, 40129 Bologna, Italy.), AD(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138 maxim@head.cfa.harvard.edu)

Publication: The Astrophysical Journal, Volume 627, Issue 2, pp. 733-738. (ApJ Homepage)

Publication Date: 07/2005

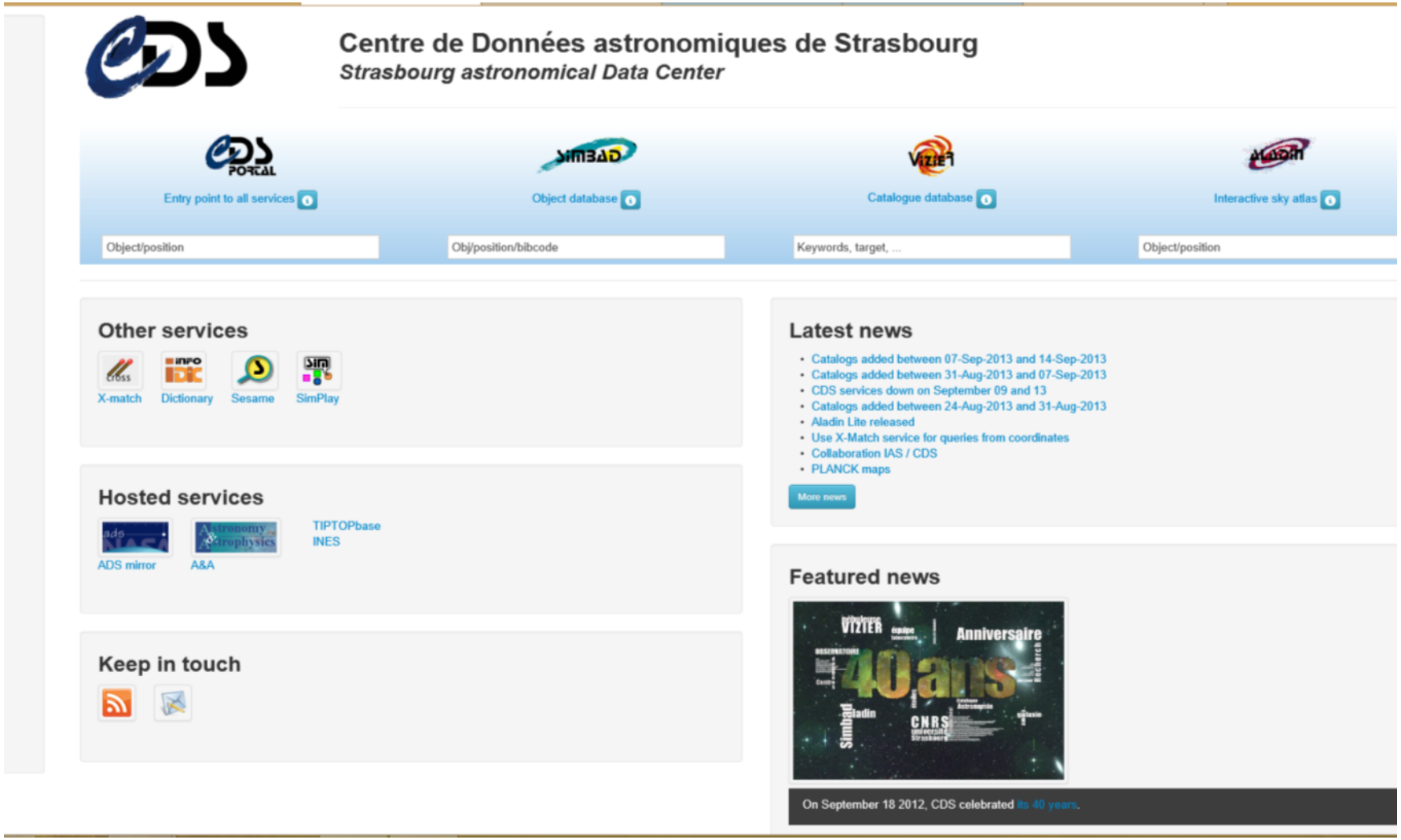
Origin: UCP

Astronomy Keywords: Galaxies: Clusters: Individual: Alphanumeric: A520, Galaxies: Intergalactic Medium, Radio Continuum: General, X-Rays: Galaxies: Clusters

DOI: 10.1086/430695

Bibliographic Code: 2005ApJ...627..733M

Strasbourg CDS Datasets



"Seamless Astronomy" (Tools)

Literature

ads labs

World Wide Telescope

Astrometry.net

flickr

TOPCAT

ds9

WIKIPEDIA
The Free Encyclopedia

AstroBetter

Blogs, Wikis, etc.

Courtesy of Alyssa Goodman

Data

VizieR

SIMBAD

AIDA Astronomical Infrastructure for Data Access

"Registries"

DataScope

Disclaimer: *This slide shows key excerpts from within the astronomy community & excludes more general s/w that is used, such as Papers, Zotero, Mendeley, EndNote, graphing & statistics packages, data handling software, search engines, etc.*

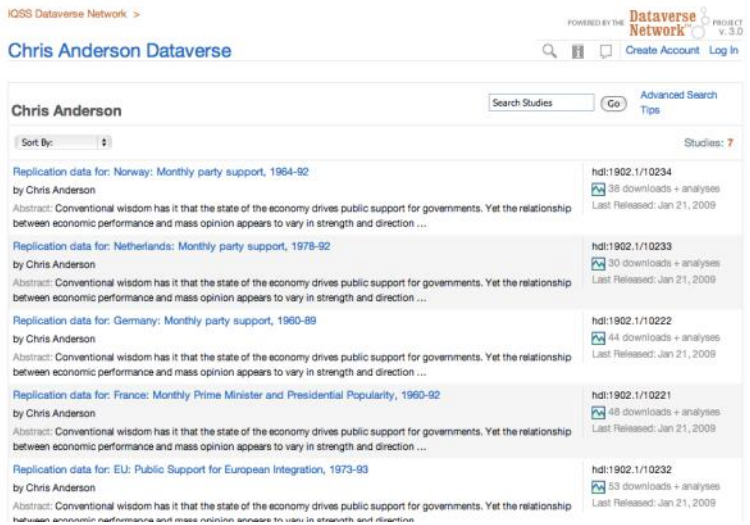
Reinforcing the Link between Research Publications and Research Data

The Dataverse project at Harvard has been awarded an Alfred Sloan Foundation grant for the next 2 years to enhance the link between journals and data

Open Journal System

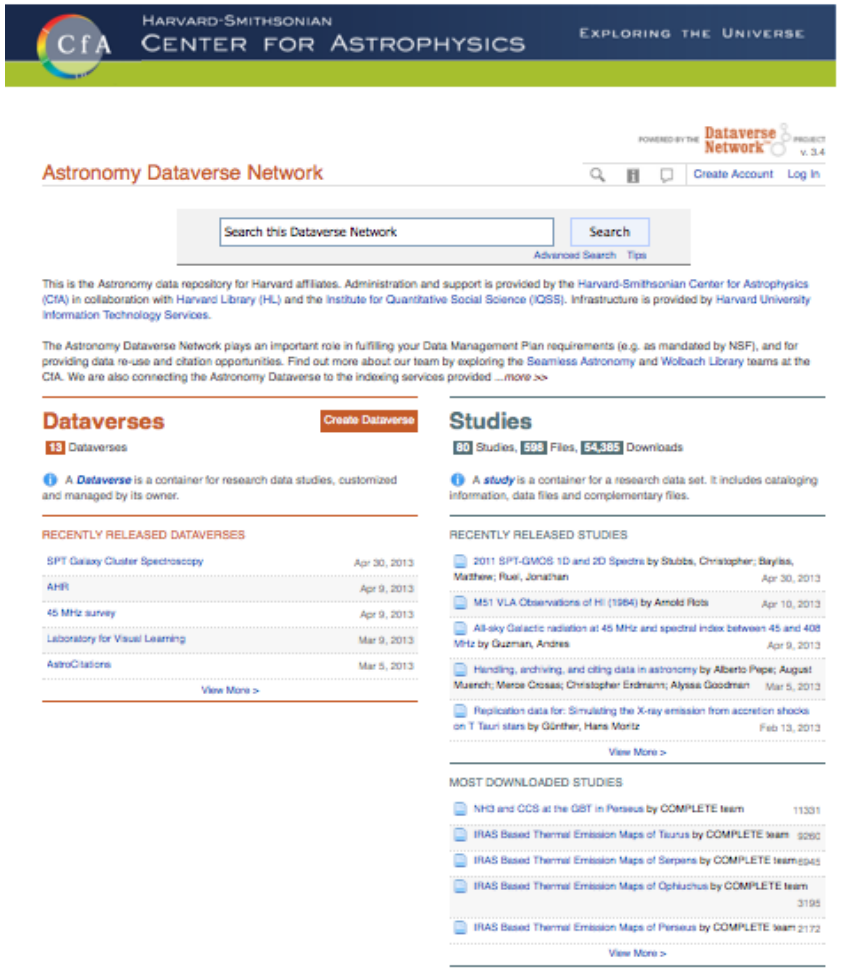
Seamless integration between the two systems:

Deposit Data to Dataverse through standard API (based on SWORD)



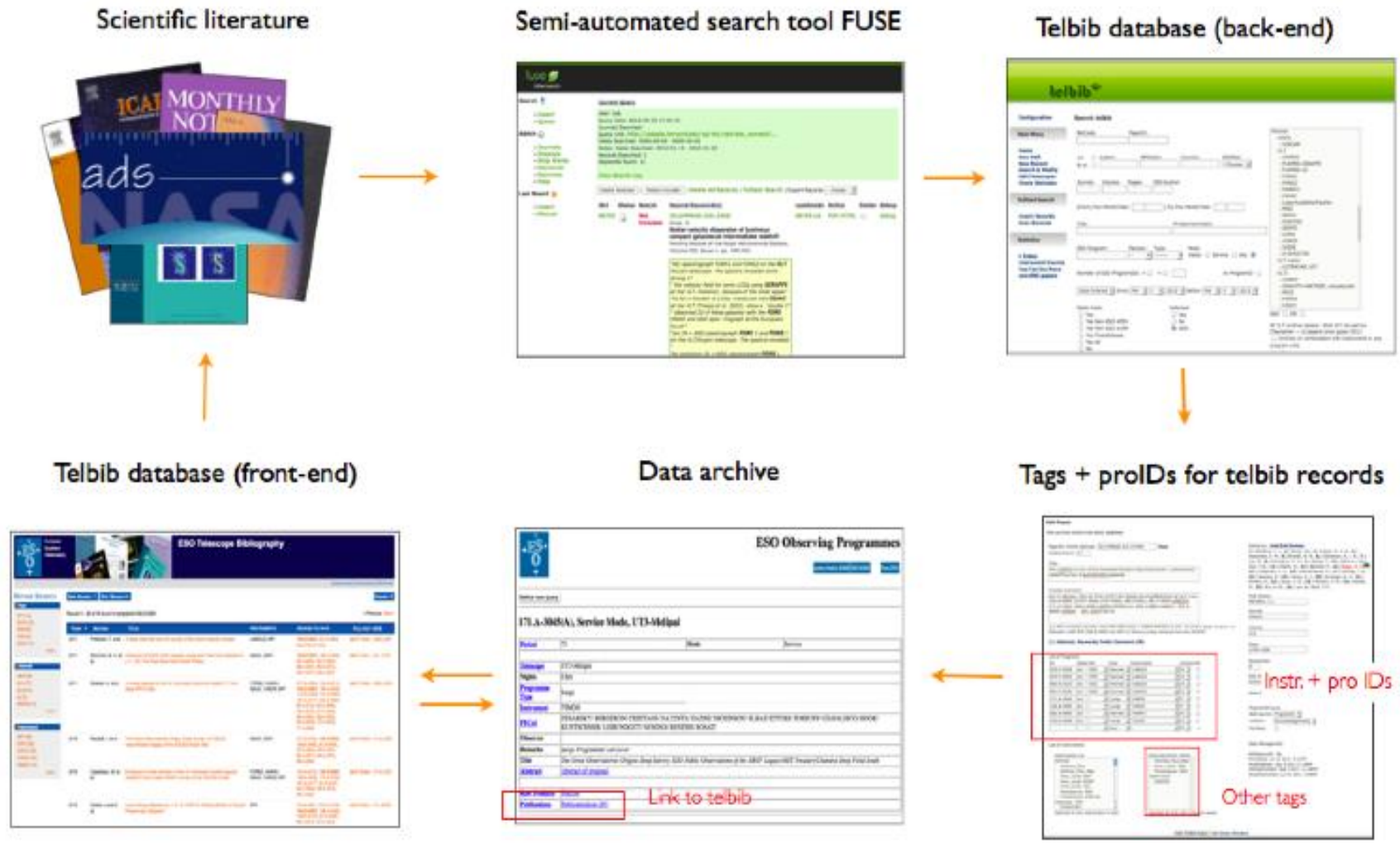
The Dataverse Network

Linking to All Data, Big and Small



Slide courtesy of Christopher Erdmann

AstroCurator: Telescope Bibliographies



Meakins & Grothkopf, 2011: Linking Publications and Observations: The ESO Telescope Bibliography. ASP Conf. Proc. 461, 767 http://adsabs.harvard.edu/abs/2012ASPC..461..767M

Slide courtesy of Christopher Erdmann

Sustainability of Data Links?

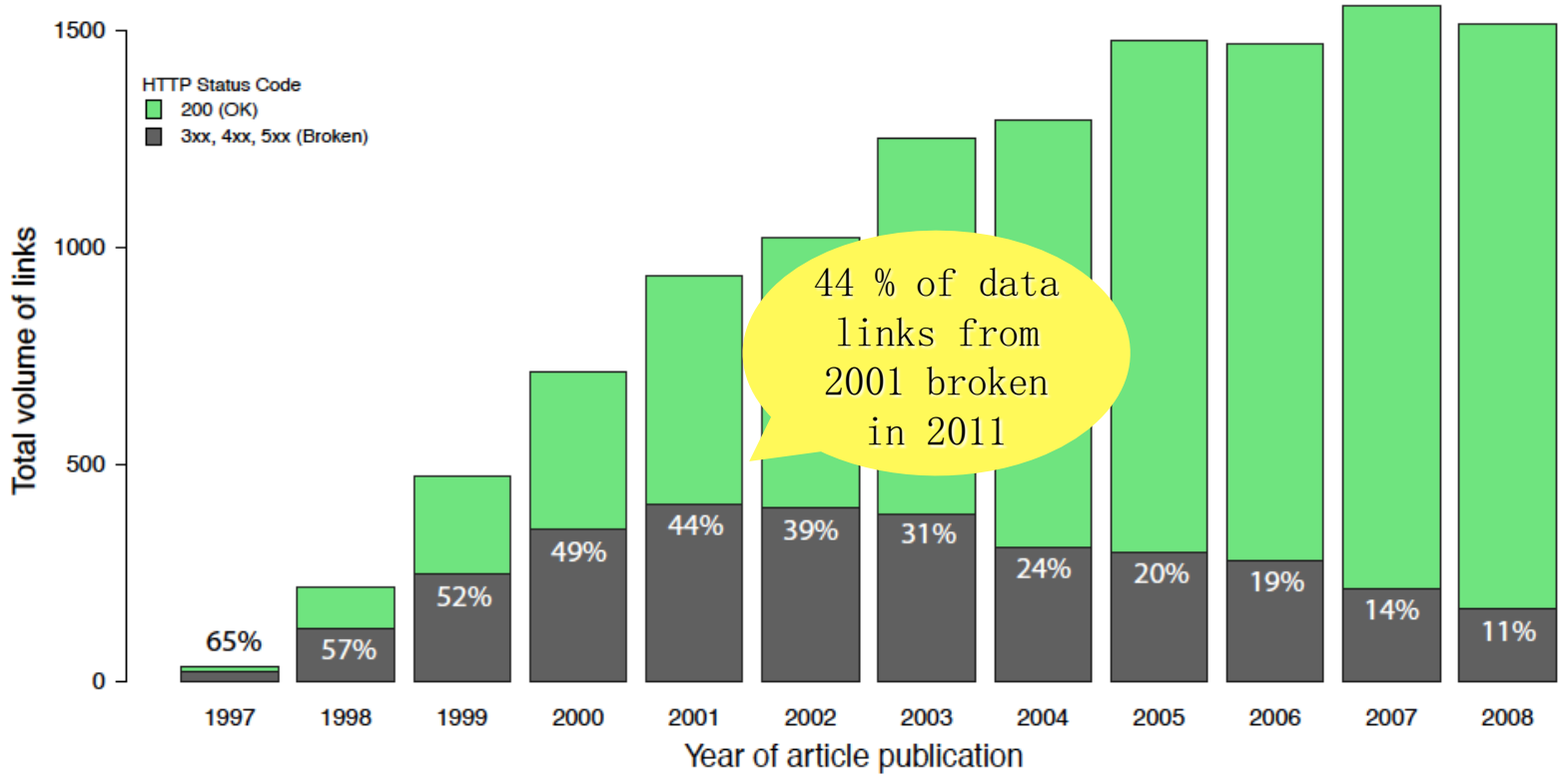
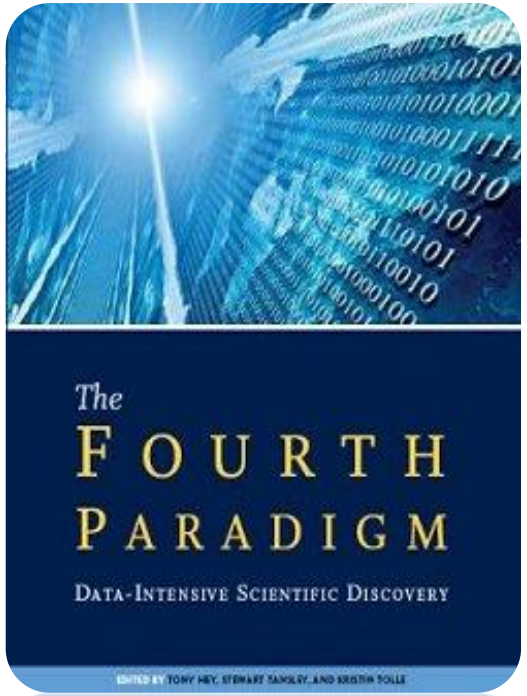


Figure 1. Volume of potential data links in astronomy publications. Total volume of external links in all articles published between 1997 and 2008 in the four main astronomy journals, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. .

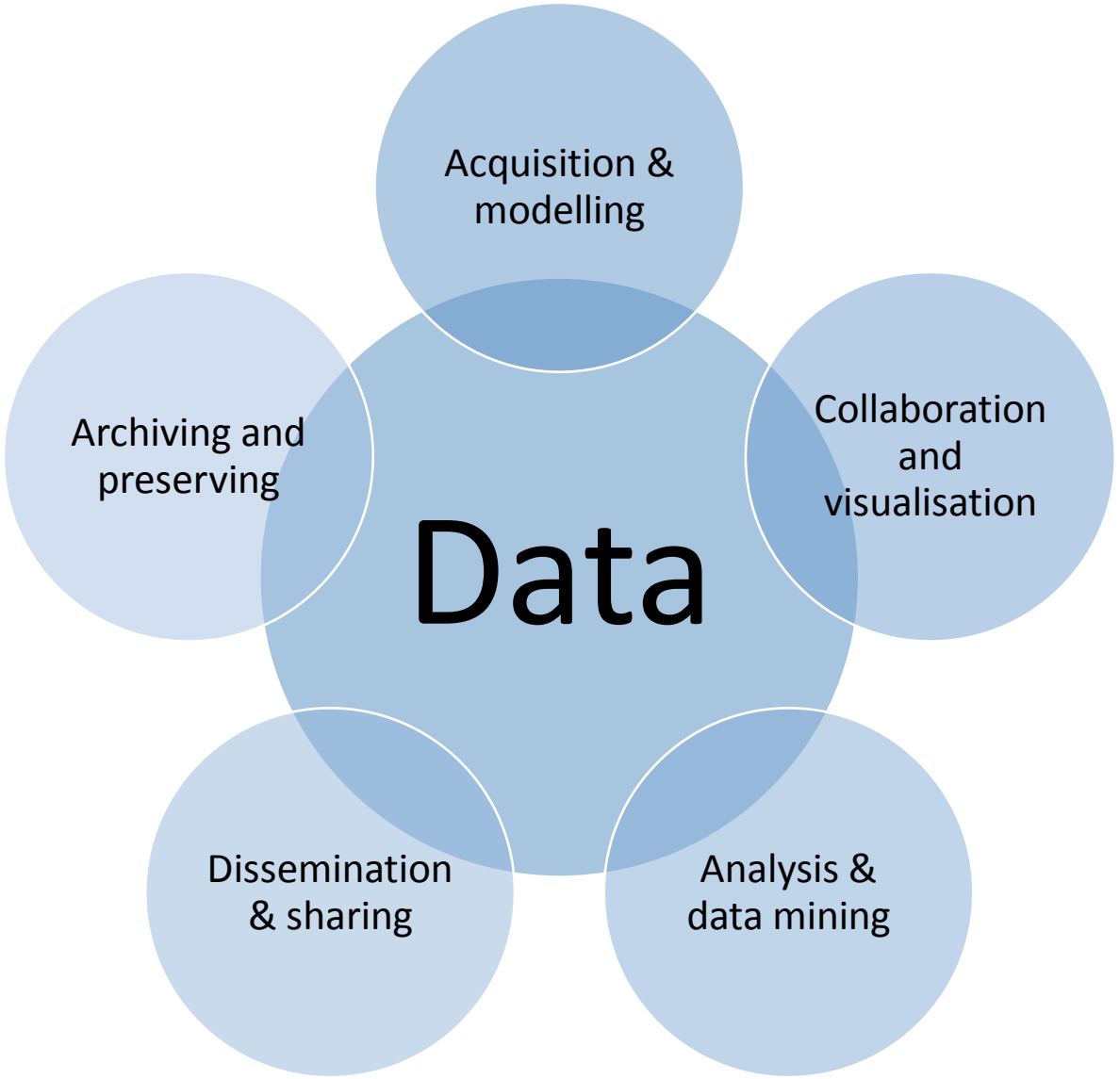
Pepe et al. 2012

Data Science Tools

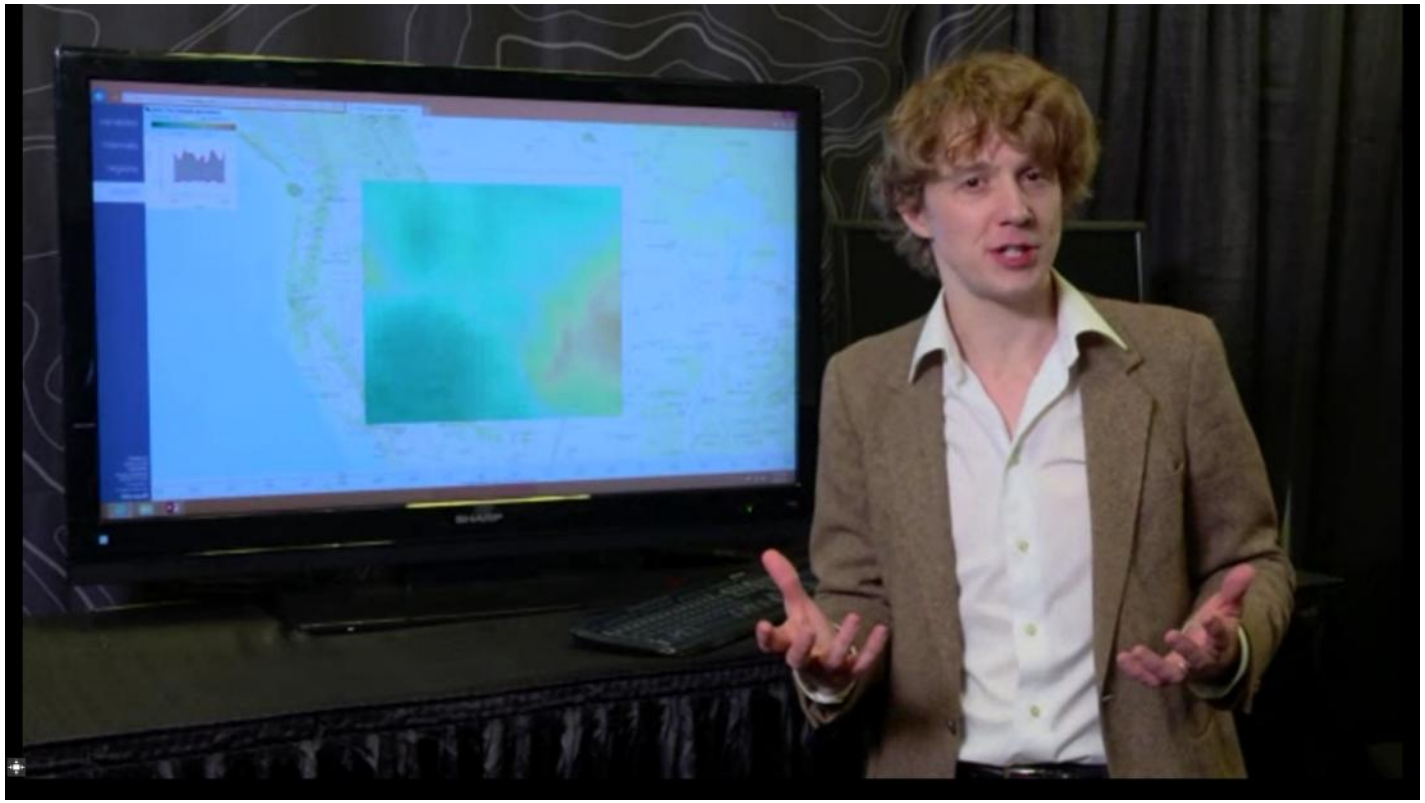
Tools to Support the Entire Data Life Cycle



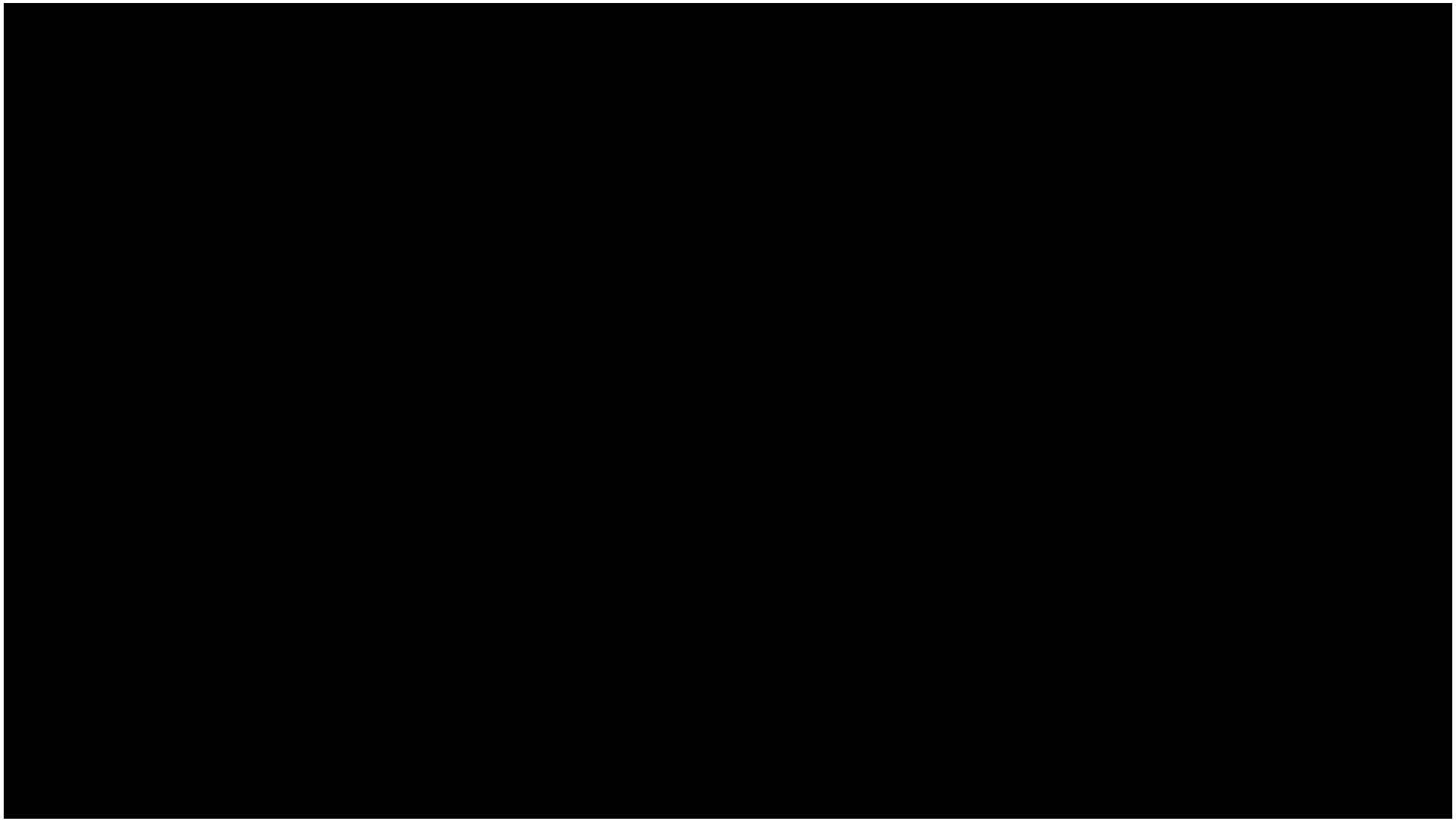
fourthparadigm.org



From Data to Decisions

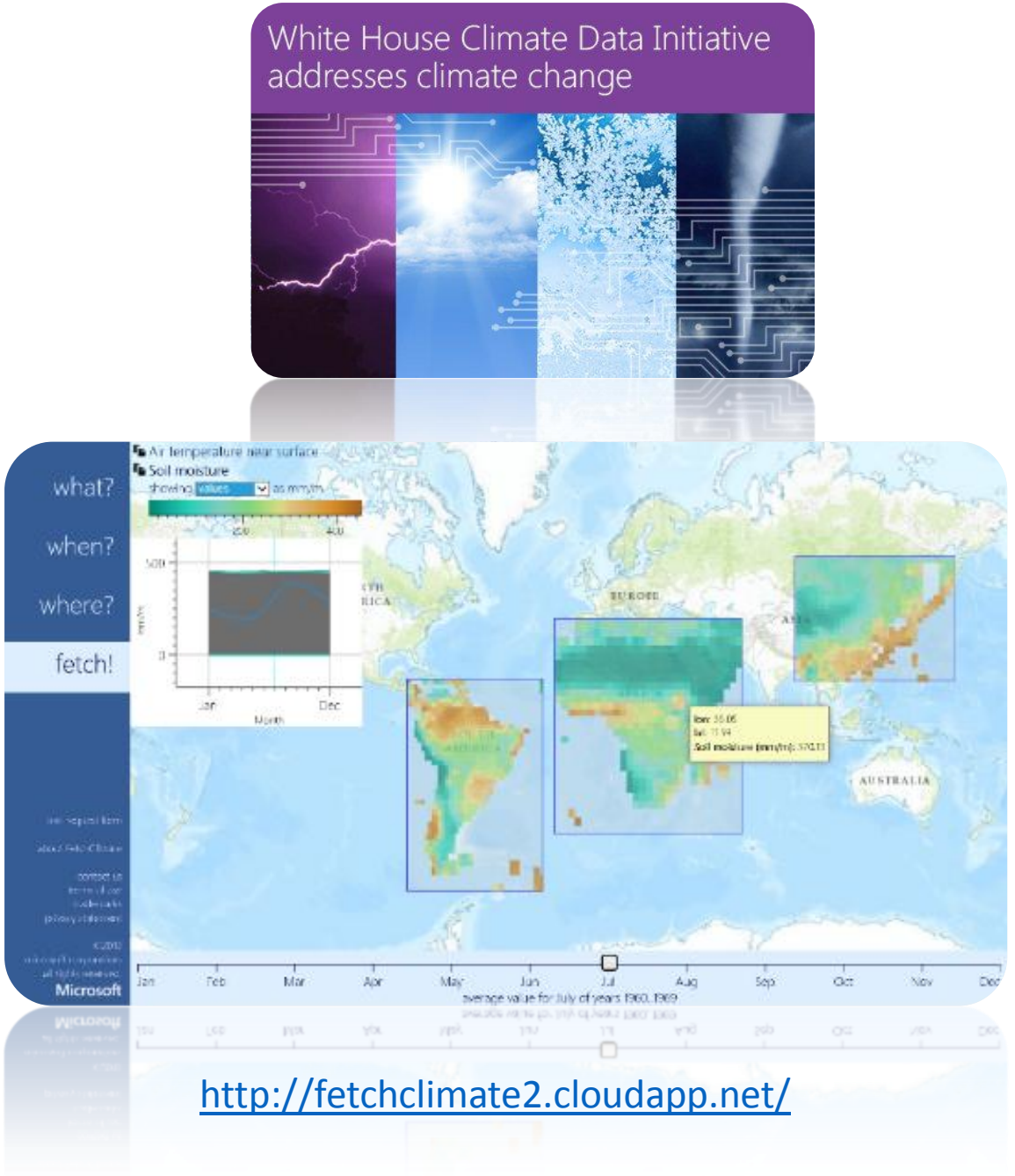


From Data to Decisions



FetchClimate

- Intelligent environmental information service
- Automatically:
  - Selects best data source to answer the query
  - Re-grids results
  - Calculates uncertainty
- Windows Azure grants for FetchClimate



http://fetchclimate2.cloudapp.net/


FloodAlerts
Home 20+
Help
About
Contact
Community Page
Find a Location:
Find
Add to my Apps
Map Layers
Save View
Hide Pins
Games Your Friends Are Play
Recommended Games
Sponsored
SurveyMonkey
surveymonkey.com
Create Su
Answers.
John Lewis
johnlewis.com
Micro Sco
Pull, Blue
Gas bill increased?
santander-products.co.uk
Use our c
see if San
save you
househol
In need of some culture?
eurotunnel.com
Discover
with Euro
Shuttle. T
to France
Chat (20)
Like 16k
My Location Open
Currently Monitoring:
Please login to enable
All Flood Alerts Open
2 Severe Flood Warning
14 Flood Warning
89 Flood Alert
Hide Menu
20 miles
25 km
© 2014 Microsoft Corporation Earthstar Geographics SIO © 2013 Nokia
FloodAlerts
Status: Good. Last refreshed on Wed Fe
26 21:31:37 UTC 2014
Created by Shoothill (Copyright © 2012-2014)
Environment Agency
created by shoothill

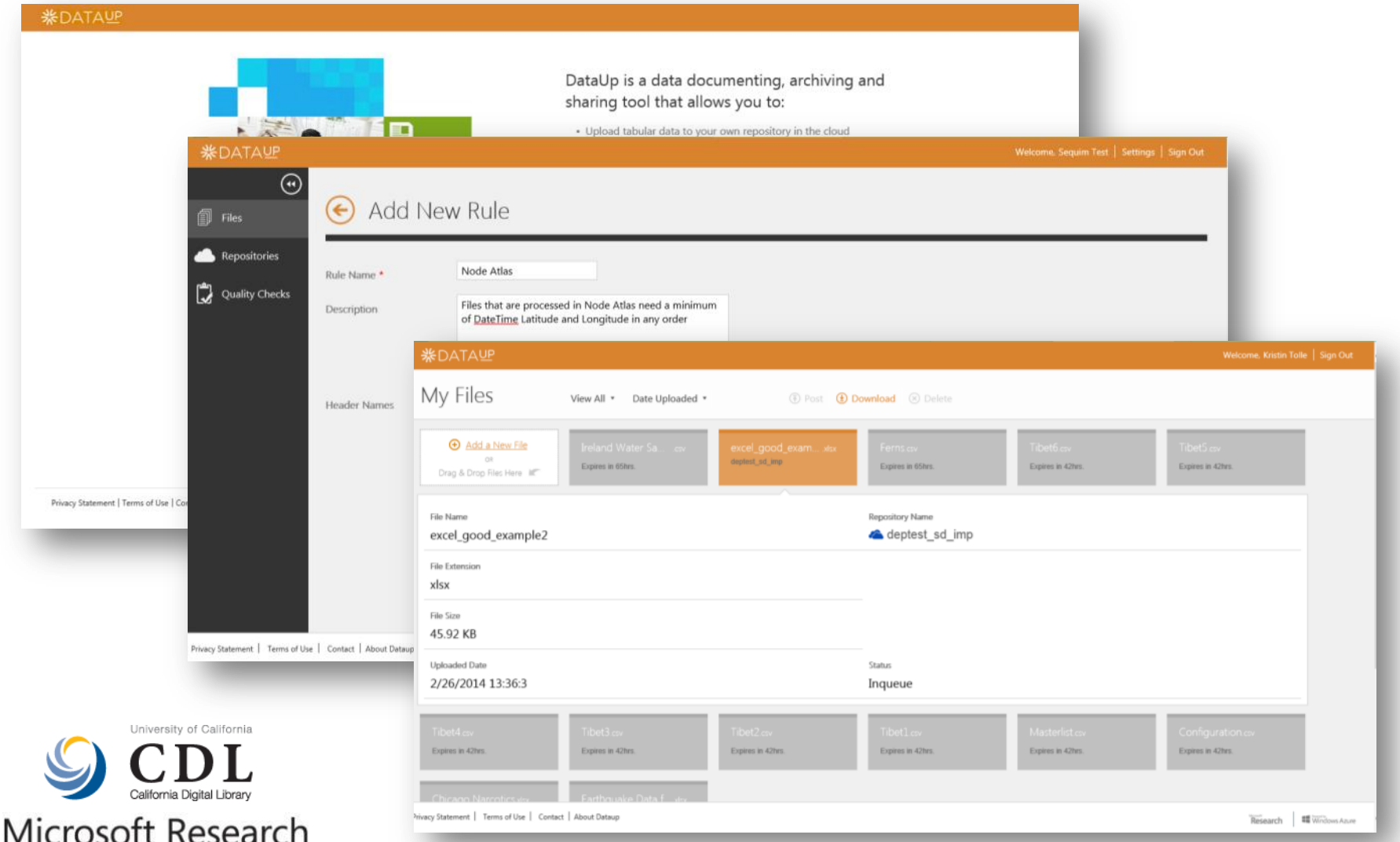
Open UK Weather Forecast Data

- Daily forecast + 5 days
- 3 hourly forecast + 5 days
- Observation data

Web search:
“open weather data azure”

DATAUP

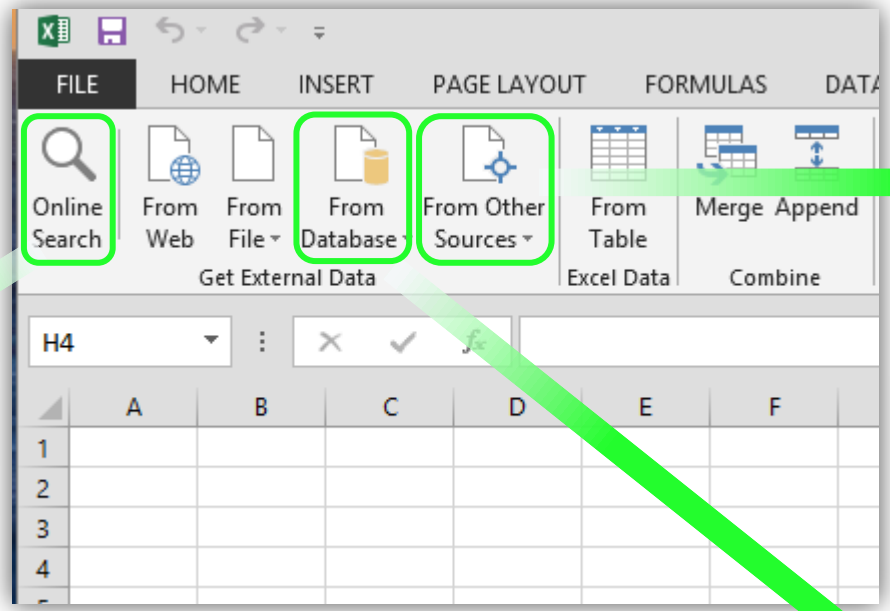
Describe, manage & share your data

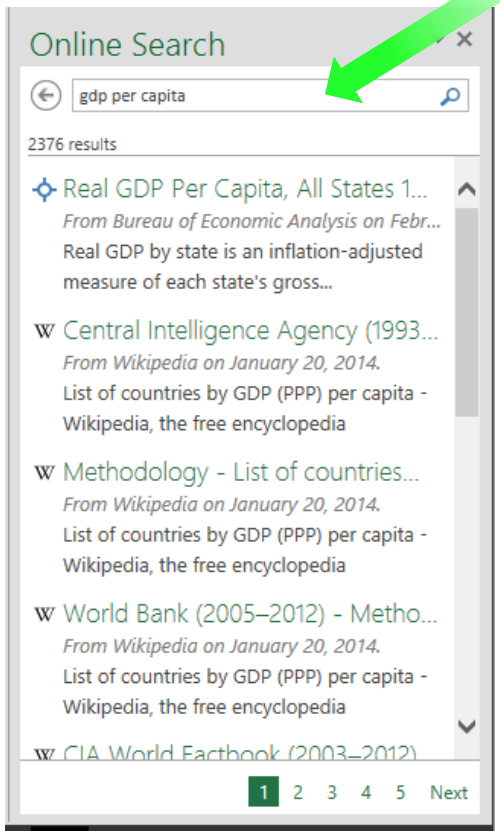


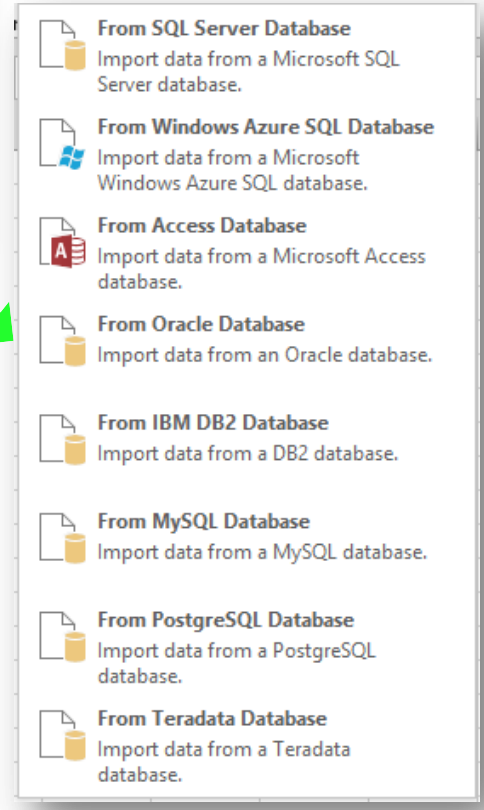
University of California
CDL
California Digital Library

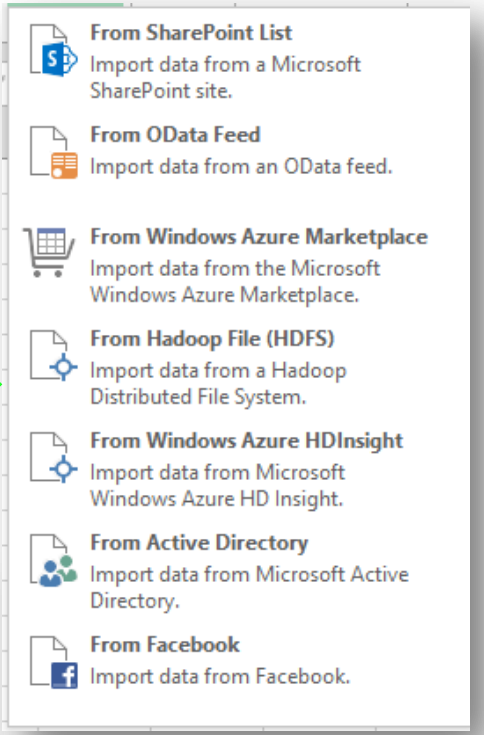
Microsoft Research

Power Query (discover & access data)









http://www.microsoft.com/powerbi/

Power Map for Excel

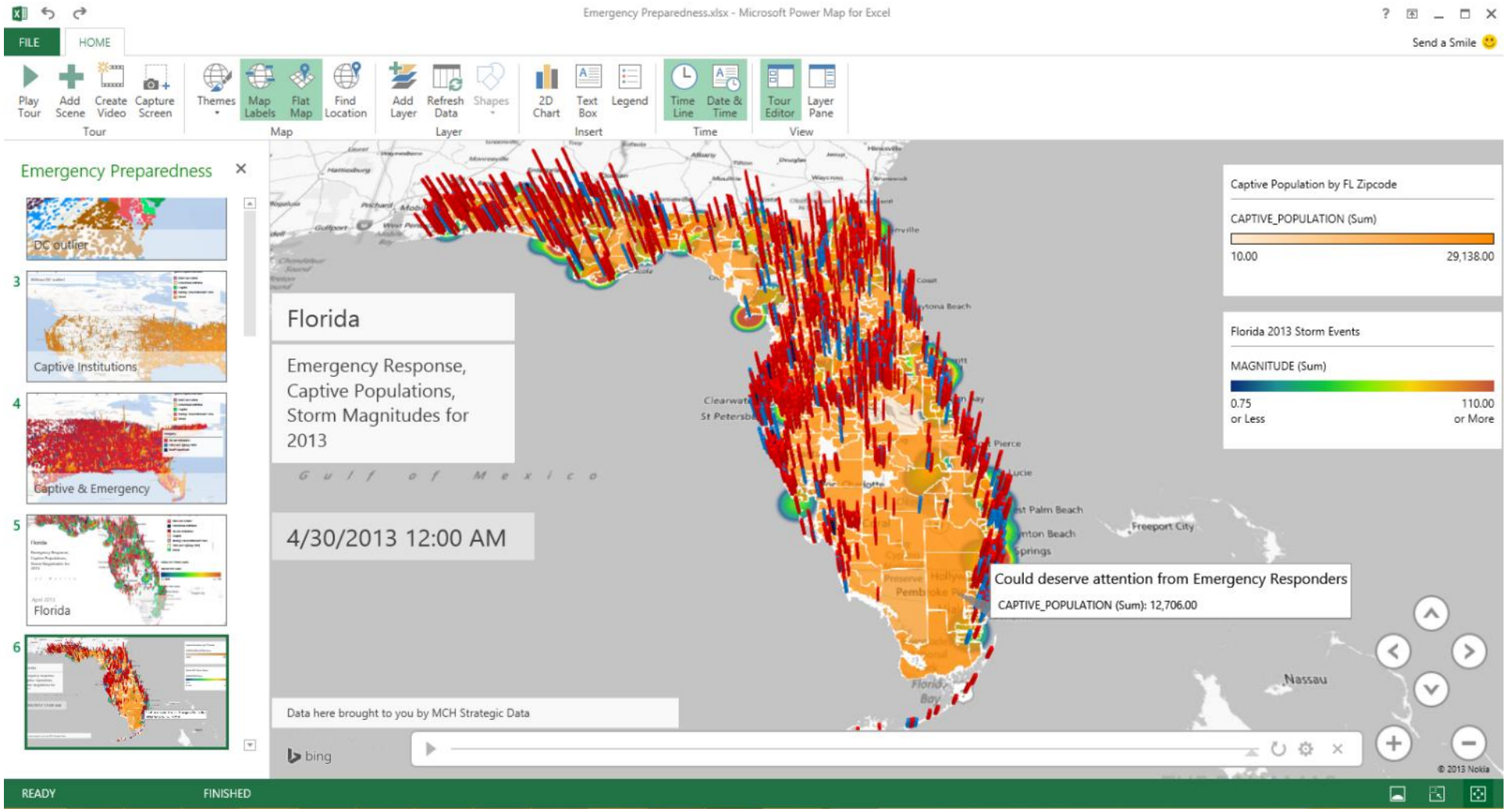
Explore geospatial data with Power Map

Power Map for Excel is a three-dimensional (3D) data visualization tool for Excel 2013 which provides a powerful method for people to look at information in new ways. It enables the data discoveries that might not be seen in traditional two-dimensional (2D) tables and charts. With Power Map you can plot geographic and temporal data visually, analyze that data on a 3D globe and over time, and create visual tours to share with others.

See Power Map in action

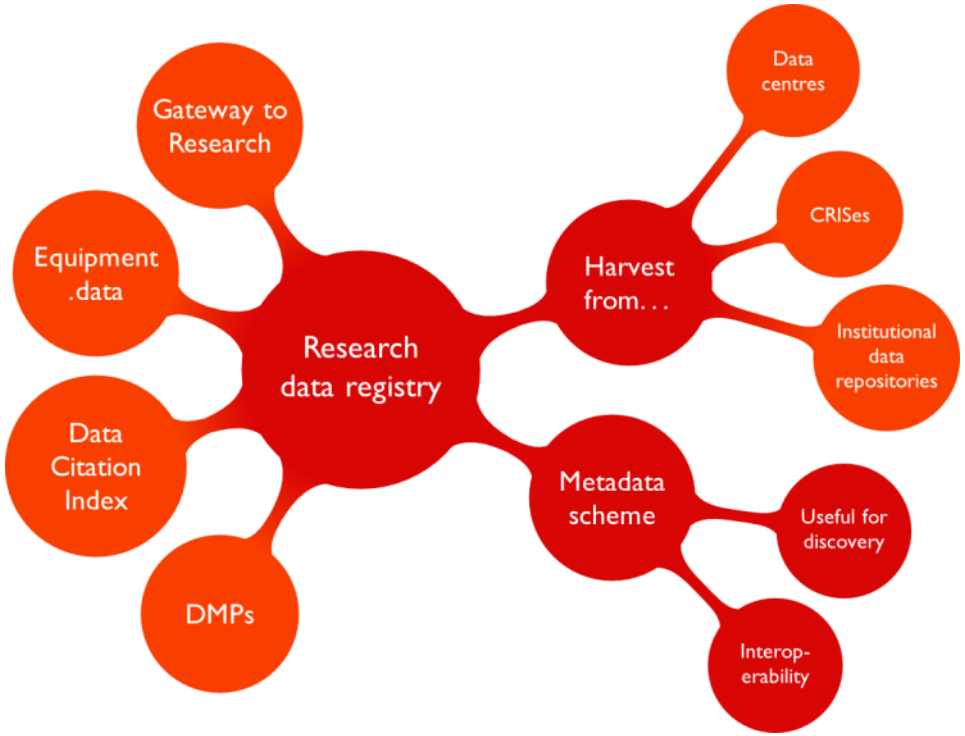
http://www.microsoft.com/powerBI/solutions/demo/business-demos.aspx

Excel PowerBI



Research Data Registry and Discovery Service

- UK pilot project
  - Stand up working system
  - Explore metadata harvesting
  - Test metadata harvesting
  - Collect feedback
- UK Data Archive, NERC Data Catalogue, nine universities
- Based on ANDS platform and modified



http://www.dcc.ac.uk/projects/research-data-registry-pilot

Powered by Windows Azure

because good research needs good data

Data Repository Registries

Purdue University
603 repositories
http://databib.org

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Humboldt-Universität zu Berlin
603 repositories
http://re3data.org

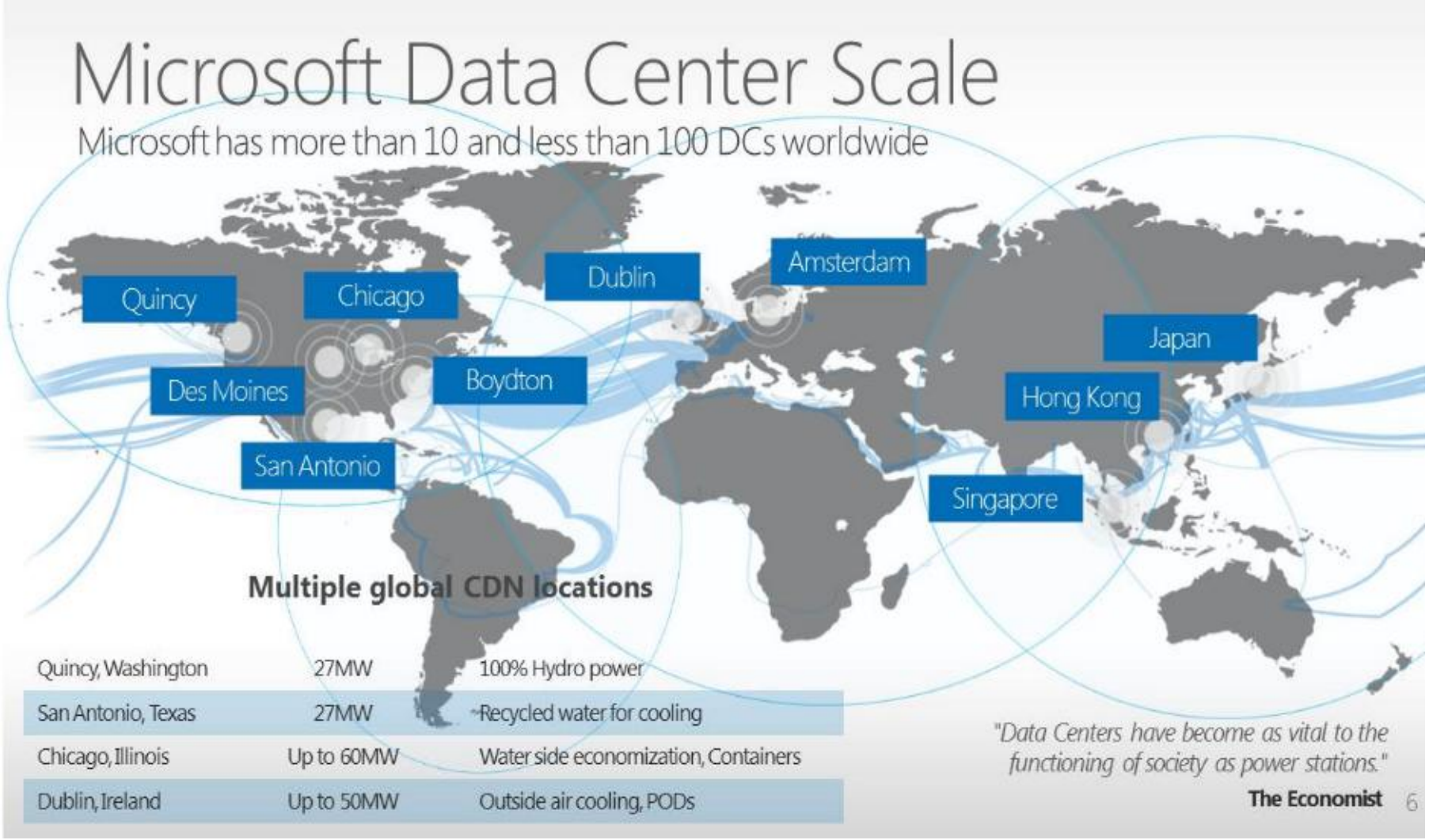
Identify and locate online repositories of research data:

- What repositories are appropriate for a researcher to submit his or her data to?
- How do users find appropriate data repositories and discover datasets that meet their needs?
- How can librarians help patrons locate and integrate data into their research or learning?

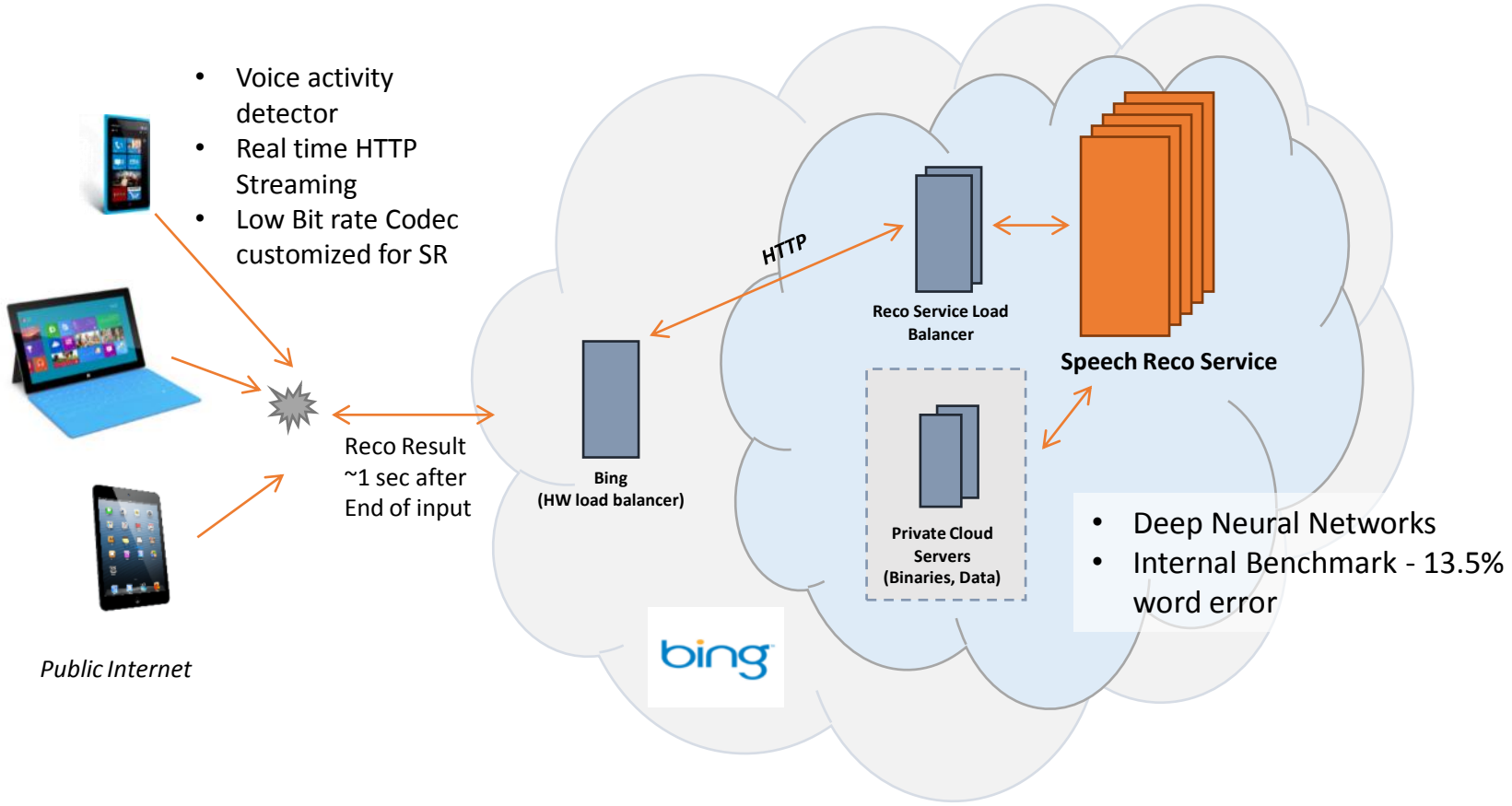
The goal of re3data.org is to create a global registry of research data repositories. The registry will cover research data repositories from different academic disciplines. re3data.org will present repositories for the permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions. In the course of this mission re3data.org aims to promote a culture of sharing, increased access and better visibility of research data.

The Role of the Cloud?

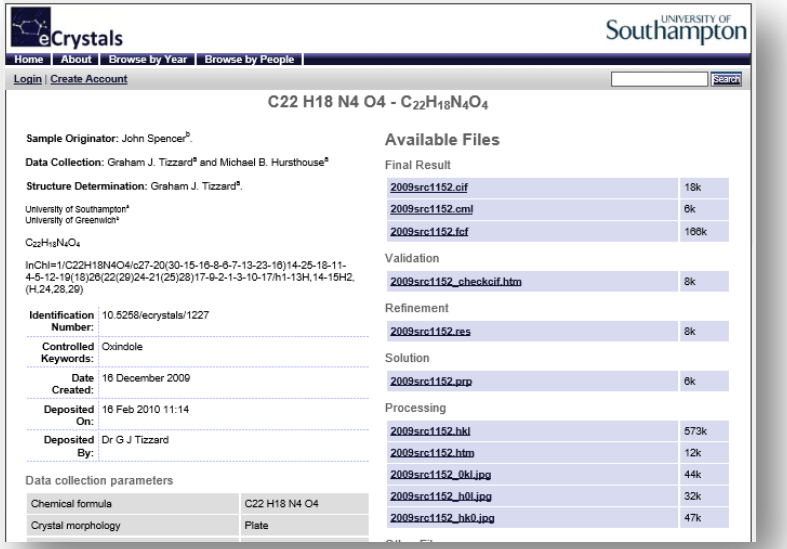
Industry is building out massive Cloud Infrastructure



Bing Speech Recognition Service: The Cloud Changes the Game



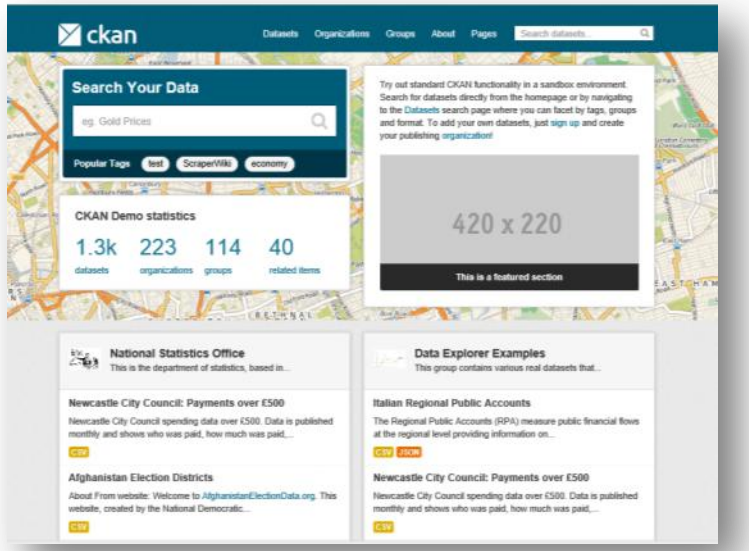
Data Repository software examples



EPrints (Southampton)

http://eprints.org

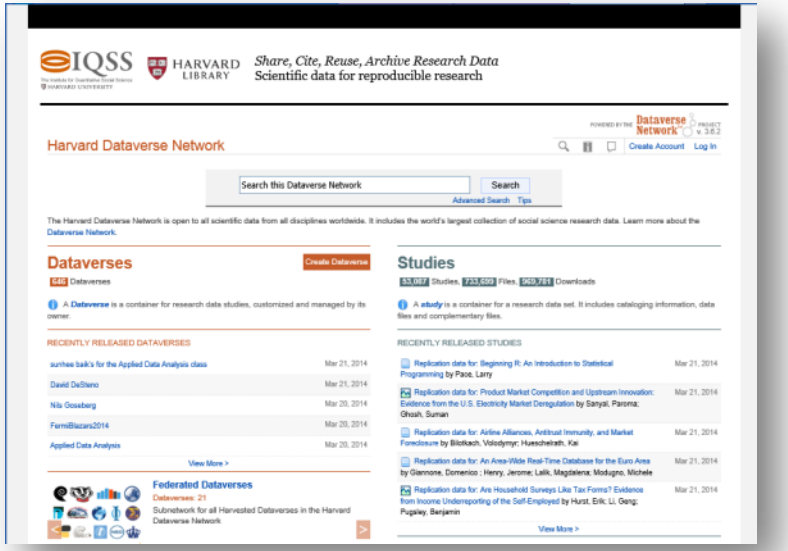
eCrystals example



CKAN (OKFN)

http://ckan.org

On Azure via VM Depot



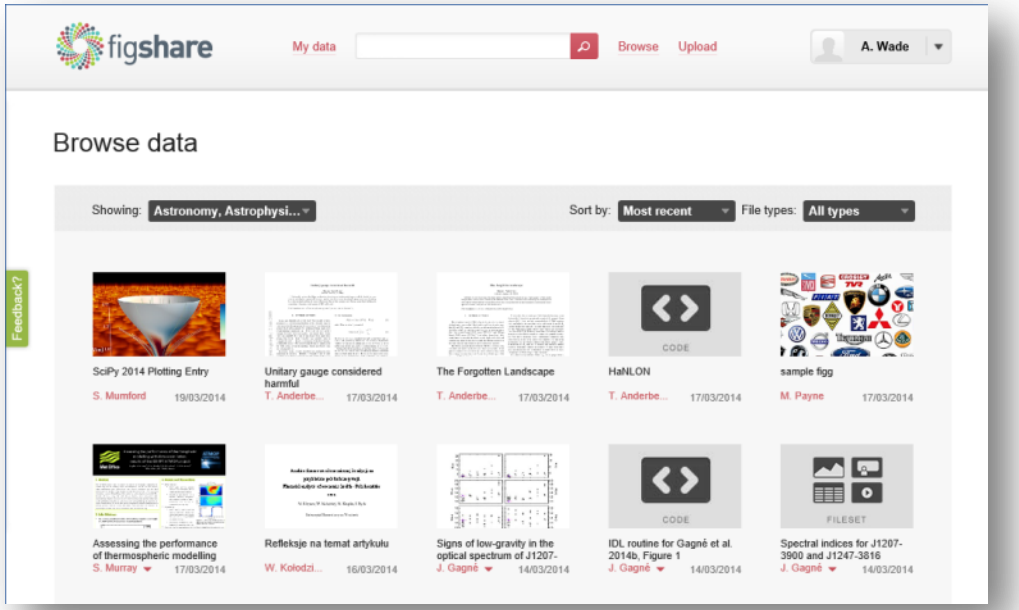
Dataverse (Harvard IQSS)

http://thedata.org

Coming soon to Azure

Hosted data repositories

figshare
credit for all your research



figshare is a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner.

http://figshare.com

Windows Azure Marketplace

Windows Azure Marketplace is an online market for buying and selling finished Software as a Service (SaaS) applications and premium datasets.

http://datamarket.azure.com

Windows Azure for Research

Accelerate the Speed of Scientific Discovery

Windows Azure provides researchers with the power and scalability of cloud computing for collaboration, computation, and data-intensive processing. This open and flexible global cloud platform supports any language, tool, or framework.

The Windows Azure for Research program:

- Free access to Windows Azure cloud computing and storage (submit proposals for Windows Azure Research Awards)
- Windows Azure for Research training classes
- Support and technical resources

Apply the power of cloud computing to your computational and data challenges. Experiment at **azure4research.com.**

Data Science in the Future?

254,000 RESULTS

The **Data Scientist** role is a role of the future!
www.**datascientists**.net
The **Data Scientist** role is a role of the future! Future proof your career and start transitioning today.

Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...
jobs.aol.com/articles/2011/08/10/**data-scientist**-the-hottest-job...
Aug 10, 2011 · Data **scientists** are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

LinkedIn's Monica Rogati On "What Is A **Data Scientist**?" - Forbes
www.forbes.com/.../linkedins-monica-rogati-on-what-is-a-**data-scientist**
Nov 27, 2011 · To continue our series on the emerging role of the **data scientist** in today's data-driven organizations, we spoke with Monica Rogati, Senior Data ...

Related searches for **"data scientist"**

Data Scientist **Seattle**
Data Scientist **Salary**
Data Scientist **Interview Ques...**
Data Scientist **Fortune**
Data Scientist **Jobs**
Introduction to Data **Science**

Data scientist: The hot new gig in tech - Fortune Tech
tech.fortune.cnn.com/2011/09/06/**data-scientist**-the-hot-new-gig-in-tech
Sep 06, 2011 · Companies that want to make sense of all their bits and bytes are hiring so-called data **scientists** - if they can find any. FORTUNE -- The unemployment rate ...

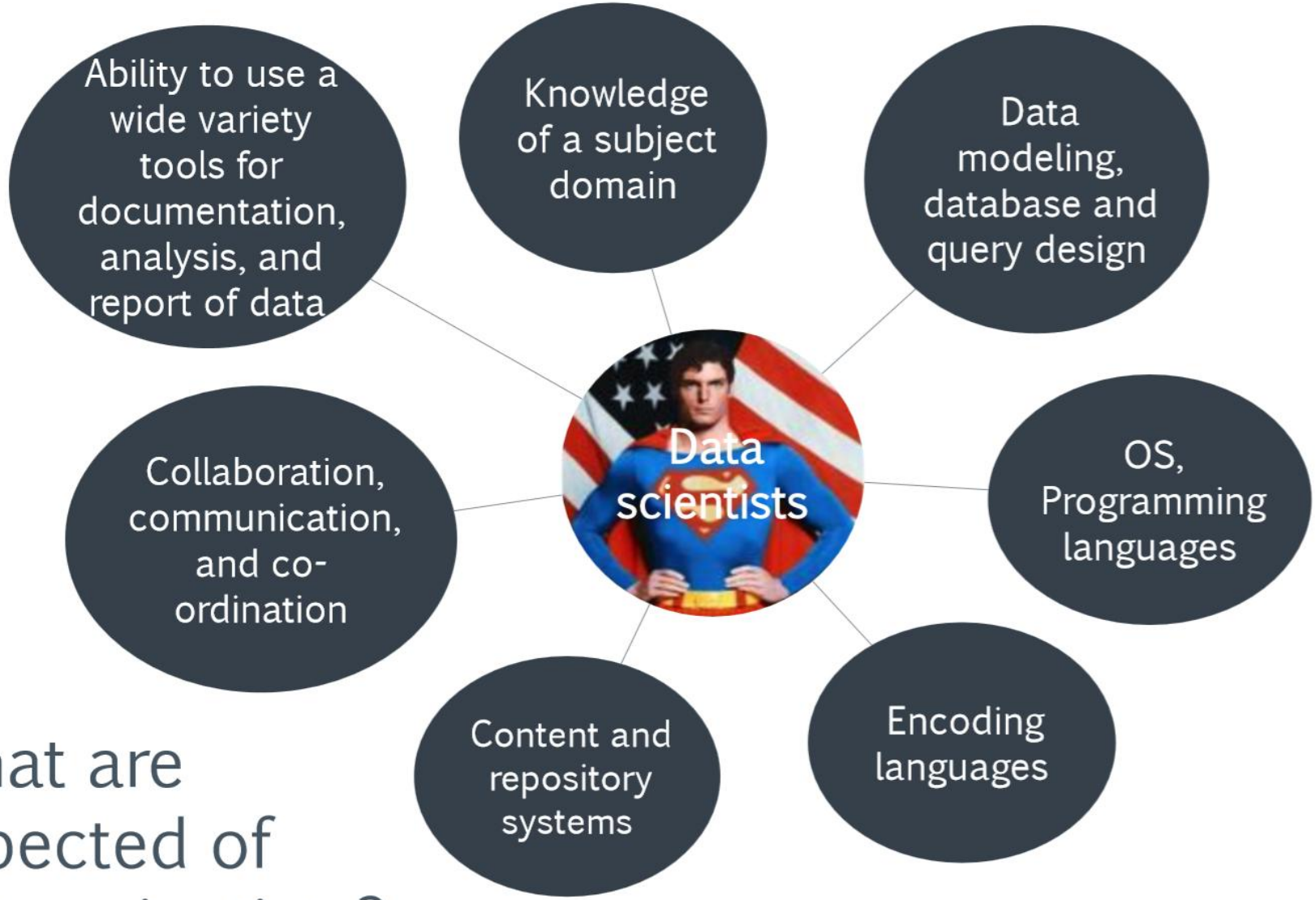
The **Data Scientist** | Mine, Visualize, and Learn
www.**thedatascientist**.com
As I jumped from room to room on Turntable.fm last night my eyes caught a glimpse of a rare room titled "AOKIxSOLREPUBLIC" . I clicked it with a fury.

What is a Data Scientist?

Data Engineer 	**People who are expert at** • Operating at low levels close to the data, write code that manipulates • They may have some machine learning background. • Large companies may have teams of them in-house or they may look to third party specialists to do the work.
Data Analyst 	**People who explore data through statistical and analytical methods** • They may know programming; May be an spreadsheet wizard. • Either way, they can build models based on low-level data. • They eat and drink numbers; They know which questions to ask of the data. Every company will have lots of these.
Data Steward 	**People who think to managing, curating, and preserving data.** • They are information specialists, archivists, librarians and compliance officers. • This is an important role: if data has value, you want someone to manage it, make it discoverable, look after it and make sure it remains usable.

What is a data scientist? Microsoft UK Enterprise Insights Blog, Kenji Takeda
http://blogs.msdn.com/b/microsoftenterpriseinsight/archive/2013/01/31/what-is-a-data-scientist.aspx

What are expected of data scientists?



Slide courtesy of Jian Qin

Some Resources

- Microsoft Research
  - http://research.microsoft.com
- Microsoft Research Connections
  - http://research.microsoft.com/en-us/collaboration/
- Science at Microsoft
  - http://www.microsoft.com/science
- Scholarly Communications
  - http://www.microsoft.com/scholarlycomm
- Azure Cloud for Research
  - http://research.microsoft.com/en-us/projects/azure/default.aspx
- Outercurve Foundation
  - http://www.outercurve.org/
- Tony Hey on eScience
  - http://tonyhey.net/

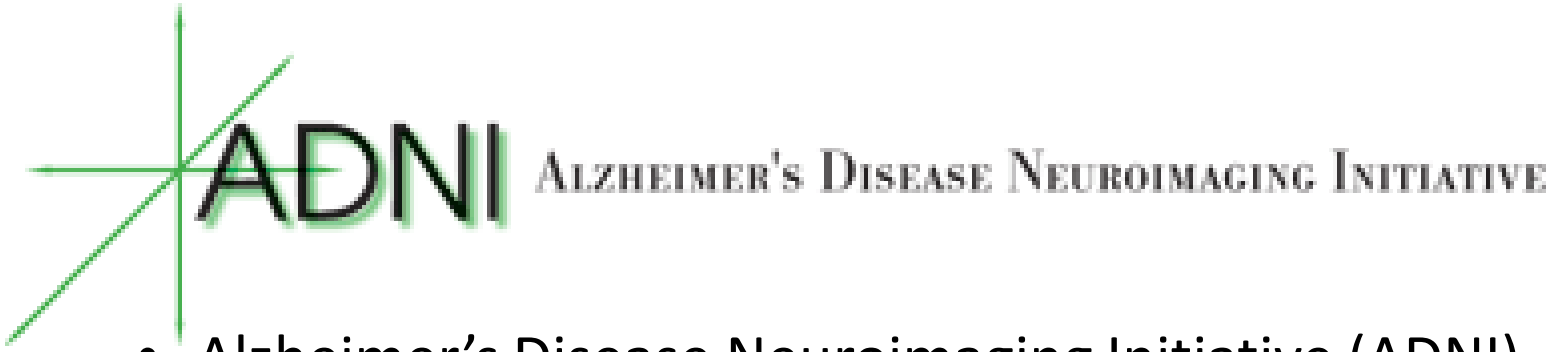
Appendix

The arXiv Sustainability Model

- Operation of arXiv is currently funded by Cornell University Library.
- In 2010, Cornell broadened funding support for arXiv by asking institutions to make an annual contribution based on the amount downloaded by each institution.
- Annual donations vary in size between $2,300 to $4,000, based on usage.
- As of February 2010, 27 institutions have pledged support on this basis.
- The annual budget for arXiv was $400,000 for 2010.

Problems of Research Reproducibility

- A detailed review of 2,047 retracted articles indexed in PubMed conducted in May of 2012 by Fang, Streen and Casadevall concluded that barely 21.3% were retracted because of errors, while 67.4% were retracted because of scientific misconduct, which included fraud or suspected fraud (43.4%), duplicate publication (14.2%) and plagiarism (9.8%).
- Studies carried out by the pharmaceutical companies Bayer (Germany) and Amgen (USA) concluded that between 60% and 70% of studies in the field of biomedicine may include non-reproducible results.



- Alzheimer's Disease Neuroimaging Initiative (ADNI) launched in 2004 specifically to improve clinical trials by different centers agreeing to share data
  - Data fro the 14 different centers involved in the initiative be combined and compared
  - Data is typically made publicly available within a week of being collected
- Hundreds of scientists have made tens of thousands of downloads from the ADNI website

➢ Of several dozen papers that have so far been published using ADNI data, a significant number were authored by researchers who are not even directly funded by the project.

http://www.adni-info.org/

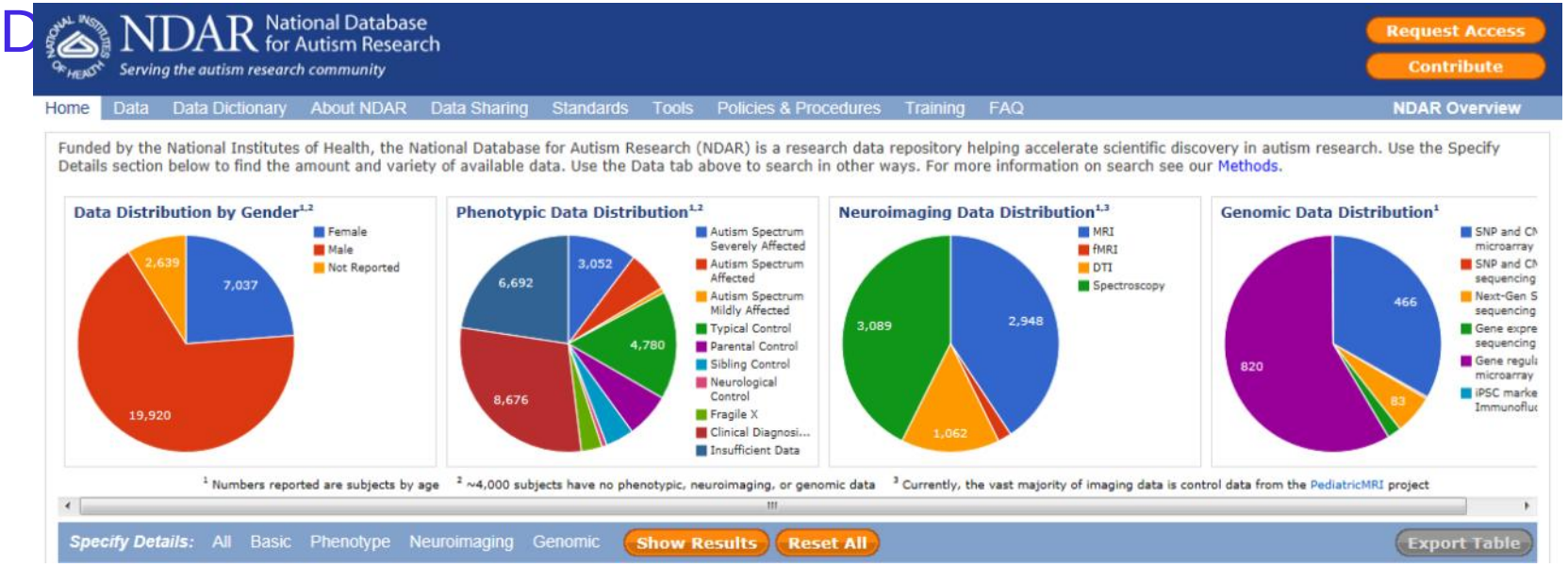
National Database for Autism Research

Federated sources of data, tools, & specimens from major US autism research funders and investigators

- partners adopting NDAR standards, e.g.:
  - Global Unique Identifier (GUID)
  - Data dictionary (29,000 elements defined)
  - Data definition & validation tools
  - Authentication scheme

"Community Science"

- D



RESEARCH DATA ALLIANCE

Playing YOUR Part
RDA Third Plenary Meeting

26 – 28 March 2014
Croke Park Conference Centre, Dublin, Ireland

research data sharing without barriers
rd-alliance.org

Twitter: @resdatall, #RDAPlenary