



# Cyberinfrastructure Frameworks for Community Driven Science

Gwen Jacobs  
Director of Cyberinfrastructure

University of Hawai'i

# A new era of community driven science

- Driven by needs to
  - to collaborate across science domains, disciplines and distance
  - to discover, access and understand heterogeneous distributed data
  - for powerful data analysis, visualization and workflow tools
  - to educate next the generation of scientists
- Necessary cyberinfrastructure:
  - Institutional and national infrastructure: high speed networks, access to national research and computational resources
  - Software frameworks for data acquisition, management, discovery and access
  - Tools for data analysis, visualization, simulation, workflow
  - Professional staff scientists, developers and curators
  - Tools for collaboration

# Exemplar community driven efforts

- End to end CI research platform
  - iplant collaborative
- Data discovery portals
  - DataONE, NIF
- Community data sharing efforts
  - PDB, ADNI, NEON, Citizen Science

A stylized, monochromatic illustration of a plant with several large, pointed leaves and a cluster of small, round buds or flowers on a stem, set against a dark brown background on the left side of the slide.

# CI ENABLED RESEARCH PLATFORM

iPlant Collaborative




# The iPlant Collaborative

The iPlant Collaborative develops cyberinfrastructure and computational tools to solve Grand Challenges in plant science


## CHALLENGE      DISCOVER      LEARN      CONNECT

**iPlant Genotype to Phenotype (iPG2P)**



Mapping the links between genotypes and phenotypes

**iPlant Tree of Life (iPToL)**



Understanding the phylogenetic relationships between all plant life

**Seed Projects**



Supporting diverse cyberinfrastructure needs

**Discovery Environment**

Access iPlant tools through a single user-friendly interface

[MORE...](#)



**DNA Subway**

An educator-tailored interface for bringing iPlant to the classroom

[MORE...](#)



**Atmosphere**

An integrative, private, self-service cloud computing platform

[MORE...](#)



**Powered by iPlant**

Projects that make use of iPlant's CI to support their third party applications

[MORE...](#)



**Upcoming Events**

- iPlant Genomics in Education Barcoding Workshop @ SACNAS, Seattle  
October 10 2012
- iPlant Genomics in Education Workshop @ Schoolcraft College - Detroit, MI  
October 26 2012 - October 27 2012
- iPlant Genomics in Education Workshop @ University of Alaska, Anchorage  
October 30 2012 - October 31 2012

[MORE...](#)

**the iPlant Leaflet**



[Read the current issue](#)

**News and Announcements**

- iPlant Participating in 2013 Tucson Winter Institute in Plant Breeding
- RAxML-Light 1.0.9 is now available through the CIPRES Science Gateway Interface

[MORE...](#)

**People at iPlant**

Community driven science



**My-Plant.org**

iPlant social networking



**Find Us...**



#iPlant is participating in 2013 Tucson Winter Institute in Plant Breeding. Come check it out @BIO5I <http://t.co/cSOuX5Ob> 2 days ago



## Discovery Environment



The Discovery Environment (DE) is one of the ways users can interact with iPlant cyberinfrastructure. Rather than managing computing resource details, or learning new software for every type of analysis, the DE allows you to handle all aspects of your bioinformatics workflow (e.g., data management, analysis, sharing large datasets, etc.) in one space.

## Atmosphere



Atmosphere, the iPlant Collaborative's cloud infrastructure service platform, facilitates and addresses the growing need for highly configurable and cloud-enabled computational resources by the plant sciences research community.

## Data Store



The iPlant Data Store is where your data are stored. The Data Store is cloud-based and is the central repository from which data is accessed by all of iPlant's technologies.



## APIs: Agave / iPlant Foundation API

Underlying the rich, domain-specific middleware layer driving the iPlant cyberinfrastructure is a low-level, HTTP- and command-line level API that provides fine-grained access to the storage, authentication, data manipulation, and storage infrastructure maintained by iPlant. The iPlant IO service API enables the asynchronous movement of file data into and out of the iPlant cyberinfrastructure.

## DNA Subway



DNA Subway makes high-level genome analysis broadly available to students and educators and provides easy access to the types of data and informatics tools that drive modern biology. Using the intuitive metaphor of a subway map, DNA Subway organizes research-grade bioinformatics analysis tools into logical workflows and presents them in an appealing interface.

## Powered by iPlant

Through the **Powered by iPlant** program, existing projects can use the iPlant cyberinfrastructure to provide services to their users by integrating with iPlant's authentication system, Data Store, job execution system, or semantic web services, or by using its servers for hosting their resources. Projects that are currently Powered by iPlant are listed below; if you would like to have your project included in this program, please email [support@iplantcollaborative.org](mailto:support@iplantcollaborative.org).



- Discovery environment
  - Bioinformatics workflow: data management, analysis, sharing
- Atmosphere
  - Cloud based computational resources
- Data Store
  - Cloud based data store
  - IRODS metadata
- APIs
  - Fine grained access to infrastructure
- DNA Subway
- Powered by iPlant

# PRAGMA

## PACIFIC RIM APPLICATIONS AND GRID MIDDLEWARE ASSEMBLY



10 year history of international research collaborations and student training opportunities, using grid enabled cloud computing to support research expeditions in bioscience, geoscience and biodiversity research.

### What's New

◆ [GLEON 14](#) in Mullanny, Co. Mayo, Ireland has concluded a successful week of student workshop on Team Science & Professional Development, Challenges for GLEON Science - now and into the future, Global Water Management public panel discussion, Working Group meetings and field trip to the Marino Institute by Lough Feegah and the Burnishoole Catchment. Thanks a million of the co-hosts Dundalk Institute of Technology and Marino Institute Newport station! Updates of presentations at G14 and follow-ups will be available on GLEON 14 website and via members list-serv in Nov 2012.

◆ A new issue of [GLEON GSA Newsletter \(vol. 1 Issue 2\)](#) edited by the GLEON GSA, contributed by both students and non-student GLEONites is out now! See what's happening in the community around the world and what to expect at the upcoming GLEON 14 Meeting in Ireland!

2012-10-27:  
[Postdoc position in New York City Water Supply Modeling Program](#)

2012-10-27:  
[Postdoc in Arctic Limnology](#)

2012-09-17:  
[PhD position at Eawag, Zürich, Switzerland](#)



## Global Lake Ecological Observatory Network

Lake Surkheng Zhui, Broghil Valley, Pakistan. Photo by: Ghulam Rasool

GLEON is a grassroots network of limnologists, ecologists, information technology experts, and engineers who have a common goal of building a scalable, persistent network of lake ecological observatories



A stylized, light-colored illustration of a plant with several large, rounded leaves and a cluster of small, round buds or flowers on a stem, positioned on the left side of the slide against a dark brown background.

# DATA DISCOVERY PORTALS

## Goals

Aggregate data and data resources

Develop search capabilities to discover data and publications

One stop shopping for data and tools

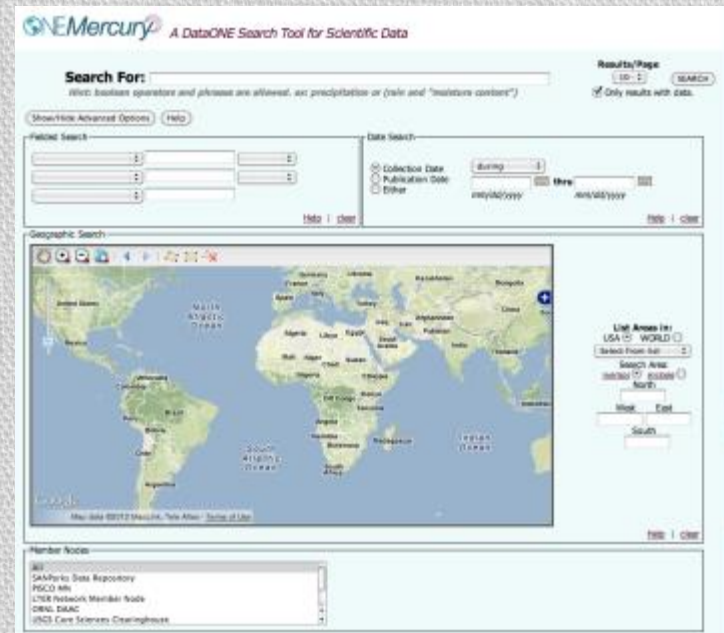
## Ecological and Earth Sciences

DataONE

## Neuroscience

Neuroscience Information Framework

- Goals of DataONE
- A distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust and secure access to well described and easily discovered Earth observational data.
- Coordinating nodes
  - University of New Mexico
  - UC Santa Barbara
  - University of Tennessee/Oakridge National Laboratory
- Member nodes
  - South African National Parks
  - Knowledge Network for Biocomplexity
  - Ecological Society for America
  - Dryad
  - Oakridge National Laboratories Distributed Active Archive Center
  - USGS
  - Long Term Ecological Research Network
  - Partnership for Interdisciplinary Studies of Coastal Oceans
  - California Digital library
  - More coming soon....
    - VOEIS – Virtual Observatory and Ecological Informatics



# DataONE Search Engine

The screenshot shows the DataONE search engine interface. At the top, there are navigation links: About, Participate, Resources, Education, and Data. Below this is the GNEMercury logo and the text "A DataONE Search Tool for Scientific Data". The main search area includes a "Search For:" field with a hint: "Hint: boolean operators and phrases are allowed, ex: precipitation or rain and 'moisture content'". There are also "Results/Page" and "SEARCH" buttons. Below the search field, there are sections for "Fielded Search" and "Geographic Search". The "Geographic Search" section features a world map with a "List Areas in" dropdown menu set to "USA" and "WORLD". The map shows various countries and regions, with a "Search Area" dropdown set to "overlaps" and "North".

The screenshot shows a search result page for a document. The document identifier is doi:10.5063/AA/connolly.116.10. The title is "Parallel effects of land-use history on species diversity and genetic diversity of forest herbs." The author is Mark Willard, from the Department of Ecology and Evolutionary Biology, Cornell University. The abstract discusses the effects of land-use history on species diversity and genetic diversity of forest herbs. The keywords include biodiversity, forest herbs, genetic diversity, land-use history, species diversity, Trillium grandiflorum, and Trillium grandiflorum.

Identifier	Type	Size	Download
doi:10.5063/AA/connolly.105.1	text/csv	4733	<a href="#">Data</a>
doi:10.5063/AA/connolly.106.1	application/octet-stream	48427	<a href="#">Data</a>
doi:10.5063/AA/connolly.104.1	text/csv	2141	<a href="#">Data</a>
doi:10.5063/AA/connolly.102.1	text/csv	6249	<a href="#">Data</a>
doi:10.5063/AA/connolly.116.10	eml://ecoinformatics.org/eml-2.0.0	69098	<a href="#">Metadata</a>

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
16	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	micro	95	128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Queries retrieve reports and data from multiple sources. EML is the metadata standard for most resources.

# Neuroscience Information Framework

The screenshot shows the NIF website homepage. At the top left is the NIF logo. The main header reads "NEUROSCIENCE INFORMATION FRAMEWORK" with the "incfnode" logo on the right. Below the header is a search bar with the text "Search for All Things Neuroscience" and a "Search Data Federation" button. A search results box contains the text "SEARCH TYPE | WHAT IS THIS? (example search: 'neuroscience', 'scholar:harvard', 'gene:grc1)". Below the search bar is a large banner with the text "Share Your Data Through NIF" and "Click to find out how." with a brain icon. To the right of the banner is a "Data Sharing with NIF" button. Below the banner is a "Community News & Events" section with a "Twitter" link. The right side of the page features a "NIF STATISTICS" section with the following data: NIF Version: 4.8, Ontology Version: 2.6, Level: 2.02.0 Resources: 180, Registry Entries: 9,247, Total Records: 386,197 ABZ. Below this is a "NIF NAVIGATOR" section with a "Powered by NIF" logo. The "NAVIGATOR" section lists various categories: LITERATURE (PLMEd 5116161), NIF DATA FEDERATION (DATA TYPE: Anesthesia (20048), Antibodies (20060), Atlas (200), Disposition (10053), Brain Activation Foci (20081), Clinical Trials (21010), Connective (20449), Database (2017), Disease (10113), Drugs (200100), Genes (2100007), Image (200400), Molecules (21200000), Models (100), Multimedia (21170), Negative Data (1627002), Pathways (201000), People (207), Phenomena (20031), Regimens (2000), Software (2000)). Below the "NAVIGATOR" section is a "NEUROSCIENCE SYSTEM LEVELS" section with the following data: Brain Region (10270), Cellular Level (20100), Cortex (20100), Molecular Level (20000), Multi-Level (20000), Nervous System Function (20000). At the bottom of the page is a "NIF REGISTRY (5114)" section. The footer contains the text "Support where affordable used, this work is licensed under a Creative Commons Attribution 3.0 License. | Privacy Policy | Terms of Service" and "SUPPORT: FAQ | General Help | Feedback | Contact Us | IMPROVING: 2014 | 2013 | 2012 | 2011".

- Data portal for all things neuroscience
- Resources curated against a community ontology – NeuroLex
- Search engine – ontology based
- Resource registration level facilitates access to data
- Wide resource variety
  - Data – data bases
  - Analysis Tools
  - Animals, reagents
  - Literature – via Textpresso

# NIF Search

Search the NIF

Preferences Report a problem ?

parkinson's disease

AND terms  OR terms [Search tips](#)

View / Edit Query



(Disease^5.0 OR disease^5.0) AND  
(("Parkinsons disease"^5.0 OR Parkinson OR PD OR "Parkinson disease" OR "Parkinson syndrome" OR  
"Paralvisis Aoitans" OR "Parkinson's svndrome" OR Parkinson's OR "Parkinson's disease" OR "Parkinsonian

**Search Options**  Synonyms

- Disease
- disease
- parkinson's (Parkinsons disease)

Data Federation (473,623) NIF Registry (133) Literature (101,581) Grants (11)

- Categories**
- Data Type**
- Atlas (258)
  - Images (285)
  - Antibodies (96)
  - Negative Data (8522)
  - Clinical Trials (2470)
  - Biospecimen (66)
  - Drugs (81)
  - Multimedia (227)
  - Microarray (452400)
  - Gemma: Microarray (451381)**
  - GeneNetwork: Info (345)
  - GEO: Gene Expression Omnibus (673)
  - ODE: GeneInfo (1)
  - Registries (159)
  - Grants (8094)
  - Models (6)
  - People (1)
  - Animals (59)

**Gemma**  is a database and software system for the meta-analysis of gene expression data. Gemma contains data from hundreds of public microarray data sets, referencing hundreds of published papers. Users can search, access and visualize coexpression and differential expression results.  [tutorial](#)

Page 1 of 45139 | Displaying 1 - 10 of 451381 | Page size: 10  Strict | [Export](#) | [Load in a new window](#)

Gene Symbol	Tissue	Organism	Experimental Fa...	Exp vs Control	Gene Expression	Description	Source	Array Platform
<a href="#">Ssx2jp</a>	<a href="#">Brain</a>	<a href="#">mouse</a>	<a href="#">Disease state</a>	<a href="#">Parkinson's Disease vs. wildtype</a>	Increased expression	Multiplex three dimensional brain gene expression mapping in a mouse model of Parkinson's Disease. Voxelation... <a href="#">More</a>	<a href="#">GEO:GSE30</a>	
<a href="#">Timm17a</a>	<a href="#">Brain</a>	<a href="#">mouse</a>	<a href="#">Disease state</a>	<a href="#">Parkinson's Disease vs. wildtype</a>	Increased expression	Multiplex three dimensional brain gene expression mapping in a mouse model of Parkinson's Disease. Voxelation... <a href="#">More</a>	<a href="#">GEO:GSE30</a>	
<a href="#">Zfp664</a>	<a href="#">Brain</a>	<a href="#">mouse</a>	<a href="#">Disease state</a>	<a href="#">Parkinson's Disease vs. wildtype</a>	Decreased expression	Multiplex three dimensional brain gene expression mapping in a mouse model of Parkinson's Disease. Voxelation... <a href="#">More</a>	<a href="#">GEO:GSE30</a>	

# NIF – accessible resources world wide



A stylized, light-colored illustration of a plant with several leaves and a cluster of small, round fruits or buds, positioned on the left side of the slide against a dark brown background.

# COMMUNITY DATA SHARING EFFORTS

PDB, ADNI, NEON, Citizen Science

# Protein Data Bank – pdb.org

The screenshot shows the PDB website interface. At the top, the PDB logo and 'PROTEIN DATA BANK' are visible. A search bar contains the text 'e.g., PDB ID, molecule name, author'. Below the search bar, there are navigation tabs for 'All Categories', 'Author', 'Macromolecule', 'Sequence', and 'Ligand'. The main content area is titled 'Biological Macromolecular Resource' and includes a 'Full Description' section. This section features 'Featured Molecules' with a 'Molecule of the Month' section for 'Vitamin D Receptor' and a 'Protein Structure Initiative Featured System' section for 'Cytochrome Oxidase'. The sidebar on the left contains links for 'Customize This Page', 'Available on the App Store', 'PDB-101', 'MyPDB', 'Home', and 'Deposition'. The right sidebar contains links for 'New Structures', 'New Features', 'RCR PDB News', and a survey prompt.

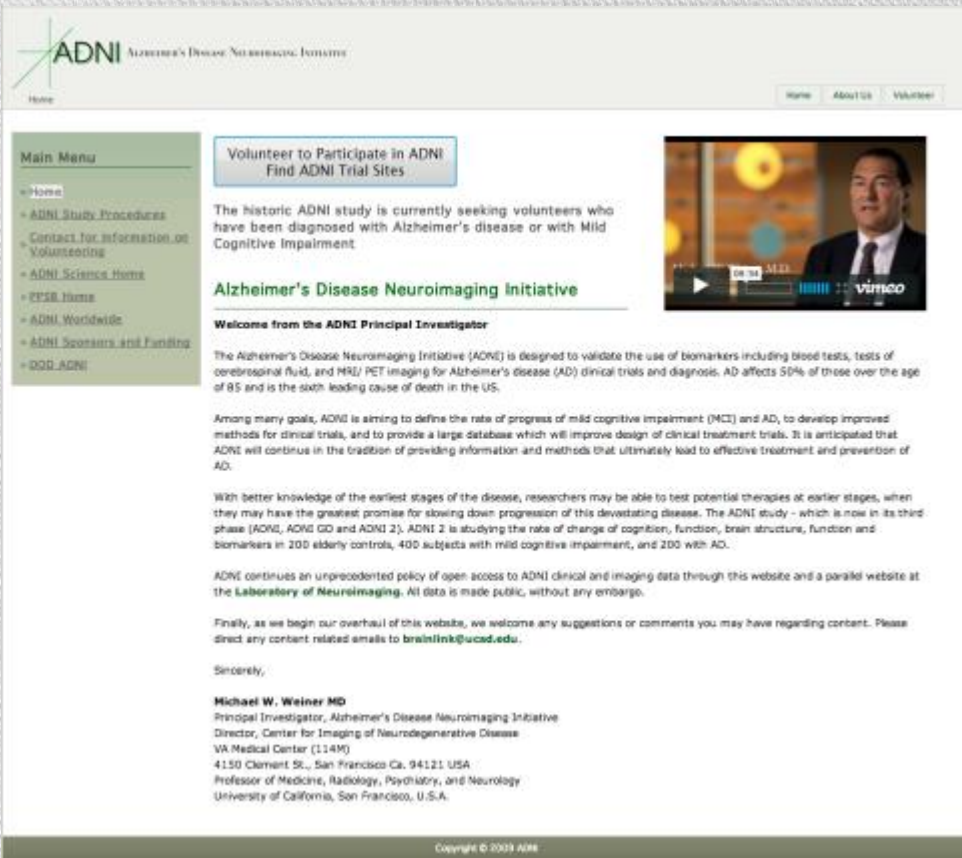
- Data base of protein structures: USCD and Rutgers
- Vibrant user community, CASP: Critical Assessment of Techniques for Protein Structure Prediction
- Multiagency support platform: NSF, DOE, NIGMS, NLM, NINDS, NCI, NIDDK



# Alzheimer's Disease Neuroimaging Initiative

- ADNI: <http://adni-info.org>

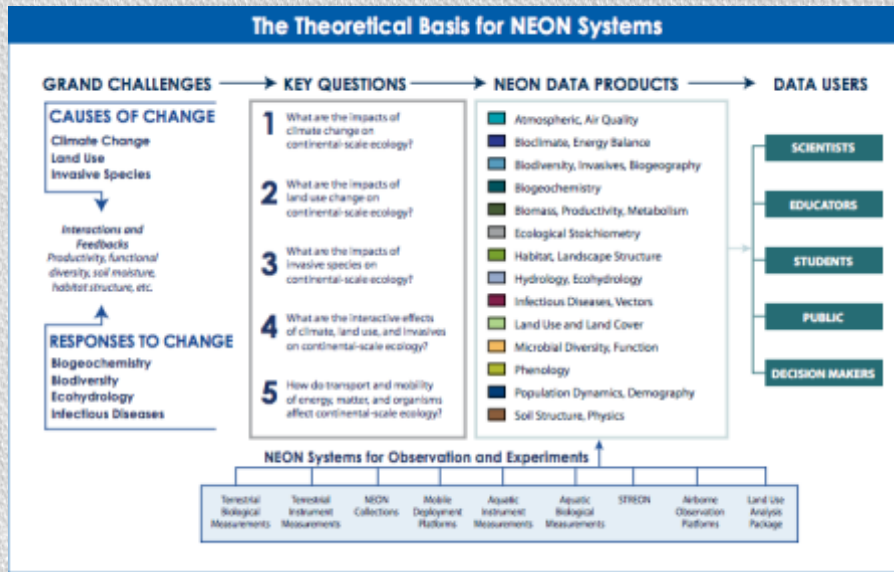
- Landmark study to identify factors in progression of Alzheimer's
- ADNI GO; ADNI 2 2012
- Identify biomarkers
  - Serum factors
  - CSF biomarkers
  - MRI/PET Scans
  - Psychometric testing
- Goal – identify biomarkers to initiate treatment earlier
  - Volunteer study – over 300 participants
- ALL DATA available without embargo
- Funding: NIH and Pharma



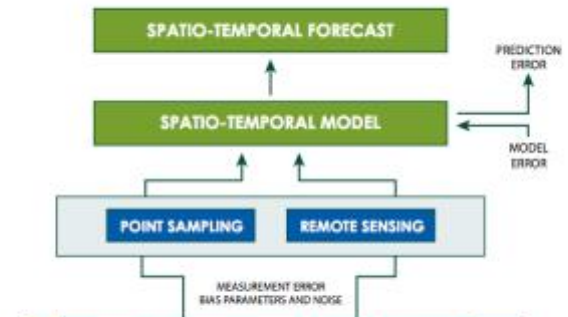
The screenshot shows the homepage of the Alzheimer's Disease Neuroimaging Initiative (ADNI). The header includes the ADNI logo and navigation links for Home, About Us, and Volunteer. A main menu on the left lists various sections like ADNI Study Procedures and Contact for information. The main content area features a video player with a man speaking, a 'Volunteer to Participate in ADNI Find ADNI Trial Sites' button, and a 'Welcome from the ADNI Principal Investigator' section. The text describes the study's goals, including validating biomarkers and providing a large database for clinical trials. It also mentions the study's phases (ADNI 1, ADNI GO, and ADNI 2) and the number of participants. The footer includes contact information for Michael W. Weiner MD, Principal Investigator, and a copyright notice for 2008 ABB.

# National Ecological Observatory Network

## A 30 year continental-scale observatory



The Flow of Information from Observations to Forecasts



# Citizen Science: strength in numbers

**Project BudBurst** Timing in every detail [Report Now](#)

Help Improve Project BudBurst by Entering the *Miscellaneous & Discover Magazine* Citizen Science Contest!

Recent Reports: First Leaf on Nov 14, 2012 (Reported on 11/14/12), 20% Leaf Fall on Nov 7, 2012 (Reported on 11/07/12), 20% Color on Nov 11, 2012 (Reported on 11/11/12)

Go Mobile with a BudBurst App

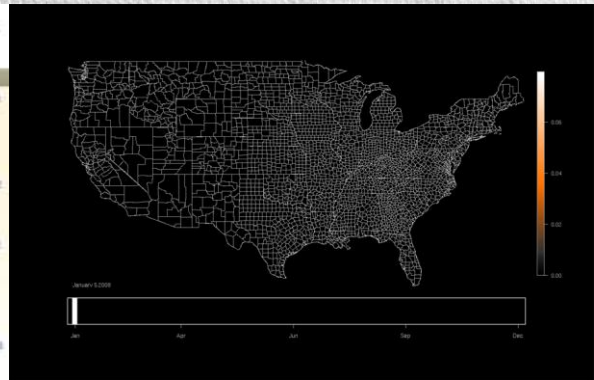
**eBird**

About eBird

Global tools for birders, critical data for science

- Record the birds you see
- Keep track of your bird data
- Explore dynamic maps and graphs
- Share your sightings and join the eBird community
- Contribute to science and conservation

Overview: A real-time, online checklist program, eBird has revolutionized the way that the birding community reports and accesses information about birds.



Lazuli Bunting Observations



**USGS** science for a changing world

Earthquake Hazards Program

Did You Feel It?

Found 12 matching results (Events - Last 24 Hours)

MW	Mag	Location	Event Time	Event ID	Intensity
5.1	5.1	COQUIMBO, CHILE	2012-11-14 19:02:05 UTC 2012-11-14 19:02:05 UTC	USC0000390	88
5.2	5.2	SEA REGION	2012-11-14 18:22:31 UTC 2012-11-14 18:22:31 UTC	USC0000545	0
4.7	4.7	BICHUAN-YUNNAN BORDER REGION, CHINA	2012-11-14 11:43:01 UTC 2012-11-14 11:43:01 UTC	USC0000597	0
4.5	4.5	TONGA	2012-11-14 11:42:59 UTC 2012-11-14 11:42:59 UTC	USC0000589	0
2.2	2.2	CENTRAL CALIFORNIA	2012-11-14 10:41:50 UTC 2012-11-14 10:41:50 UTC	USC0000588	4
4.8	4.8	HERMAIOS ISLANDS REGION	2012-11-14 09:13:31 UTC 2012-11-14 09:13:31 UTC	USC0000581	0
4.7	4.7	EASTERN MEDITERRANEAN SEA	2012-11-14 07:24:26 UTC 2012-11-14 07:24:26 UTC	USC0000580	2
3.6	3.6	REGIONS FUS. ETHIOPIA	2012-11-14 05:21:43 UTC 2012-11-14 05:21:43 UTC	USC0000569	101
4.0	4.0	ATACAMA, CHILE	2012-11-14 03:08:06 UTC 2012-11-14 03:08:06 UTC	USC0000586	3

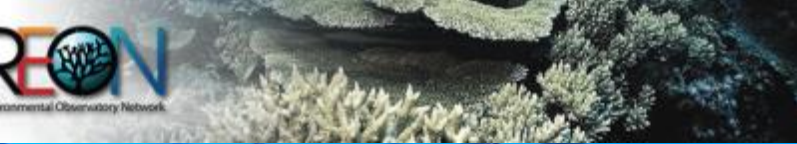
**M6.1 - Coquimbo, Chile**  
Wednesday, November 14, 2012 at 19:02:05 UTC  
Wednesday, November 14, 2012 at 18:02:05 Local

USGS Community Internet Intensity Map  
COQUIMBO, CHILE

Nov 14 2012 04:02:05 PM local (28 10585 71 2719W M6.1 Depth: 81 km @ChicoR00ds)

0 responses in 20 cities (Max CFI = 4)

Intensity Legend: I (None), II (Weak), III (Light), IV (Moderate), V (Very Light), VI (Weak), VII (Violent), VIII (Extreme), IX (Severe), X (Very Severe)



# References and useful links

- Neuroscience Information Framework
  - <http://neuinfo.org>
- Iplant Collaborative
  - <http://iplantcollaborative.org>
- NEON
  - The NEON 2011 Science Strategy
  - <http://neoninc.org>
- DataONE
  - <http://www.dataone.org>
- CREON
  - <http://www.coralreefon.org>
- DataTurbine
  - <http://dataturbine.org>
- iRODS
  - <http://irods.org>
- VOEIS
  - <http://voeis.msu.montana.edu>
- 3DSlicer
  - <http://www.slicer.org>
- Informatics for Integrating Biology and the Bedside
  - <http://www.i2b2.org>
- USGS Earthquake Hazards
  - <http://earthquake.usgs.gov/earthquakes/dyfi>
- eBIRD
  - [ebird.org](http://ebird.org)
- Project BudBurst
  - <http://neoninc.org/budburst>

Contact info:  
[gwen@montana.edu](mailto:gwen@montana.edu)