

# How Data Science & Big Data are Reshaping Research Computing

*Christine Kirkpatrick*

September 2017



**I ♥ BIG DATA**  
SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO

# SDSC: Your Neighborhood Supercomputer Center

- Established as a *national supercomputer resource center* in **1985** by NSF
- Became an Organized Research Unit in **1997** serving the UC system
- Largest at UC San Diego:
  - grant revenues (~\$30M/yr)
  - people (~250)
- High proposal acceptance ratio
- *World leader in data-intensive computing and data management*





# Supercomputing: Then & Now



# Research Computing Services

## Traditional

- Storage
- Backups
- Pre-award support
- Web services
- Bleeding edge scale

## What's New

- (Really) large-scale storage (PBs)
- Cloud Consulting
  - Cloud Broker
- Data \_\_\_\_\_ as a Service
  - Management
  - Science
  - Compliance

# SDSC (Big) Data Platform: AWESOME



SDSC PI:  
*Amarnath Gupta*

## Example Science Drivers

- **Political Science**: Can we detect acts of Chinese Government's propaganda and information policing by observing the social media? Can we detect these activities in real-time?
- **Sociology**: Can we construct a statistical model to detect and predict political unrest by observing user behavior and actions on social media? Can we trace how political movements form by watching important players in a situation?
- **Social Medicine**: Can we identify individuals and communities who are at high-risk for HIV/AIDS? Can we use a combination of social media and transmission networks to make informed medical intervention (e.g., proactive vaccination)?
- **Social Media Analytics**: Can we develop a prediction model for identifying potential violence by watching social media? Can we identify and correlate disparate events that may lead to such violence?



# Pain Points: Social Media Analytics

- Data too large and heterogeneous

→ *Semistructured? Networks? Text? Images?  
AsterixDB, SciGraph, Fiona + Tools*

- Commercial cloud is expensive

→ *SDSC Storage starting at a 100 TB?*

- Regular, systematic collection of data a problem

- No collection mechanism
- Nowhere to store and perform “continual analysis”



- *A Data Lake with multiple streaming inputs over a high-bandwidth network?*
- *Facility to auto-compute some statistics?*
- *A visual way of “looking” at collected data?*

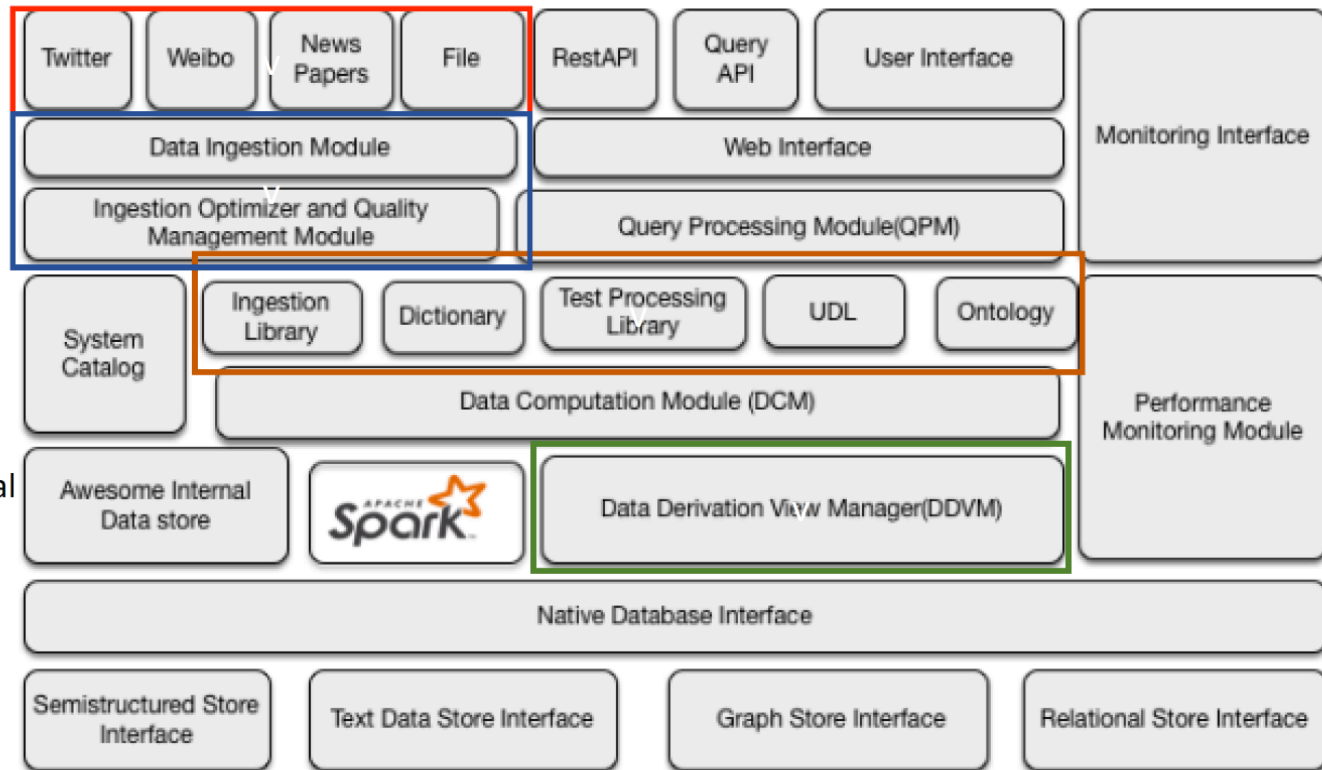
- Graduate researchers have limited facilities

- Tools not equipped for large data volumes
- No query facilities to subset data or provide statistical previews for the large and complex data



- *R in main memory on large-memory machines?*
- *Community/Researcher tools as UDFs in AsterixDB?*
- *Data Integration engine?*
- *API + Access Portal over Data Facility?*
- *Analytics Interface?*

# AWESOME Architecture: Built on UC Technologies



Simple statistical data structures

JSON, XML (after conversion) streaming data feeds



Text + interval annotations



Cypher+ our indices



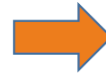
Relational Data



# Social Science Use Case

## Example Data

- Microblogs
- News articles/blogs/comments
- Political/geographic organization structures
- Biographies
- Message transmission networks
- Legal documents
- Economic surveys
- Demographic data



## Example Analysis

- Public View analysis
  - Topic modeling over
    - Microblog content of communities within a date range clustered around a set of hashtags or referring to political personalities in cities where protests are going on
    - Newspaper articles covering the same protests and comments on the same articles
  - Statistical analysis of “interesting” topic clusters and their covariates from both sources



# The China Lab at UC San Diego

- **UCSD Social Sciences**
  - 21<sup>st</sup> Century China
  - 12 Faculty Members
  - Wide diversity of data
- **Initial funding from Luce Foundation**
  - SDSC Cloud/AWESOME + analysis layer
  - Large Text Analytics Models running on 1M+documents
  - Machine learning, comparative analysis, integrative analysis

## Unveiling objectives of China's censorship

An overview of Assistant Professor Molly Roberts's research on 'reverse-engineering' censorship in China and how partnerships with GPS keep it fueling

Feb. 26, 2016 By Sarah Pfledderer | GPS News



Molly Roberts, assistant professor at the UC San Diego Department of Political Science, does not get easily intimidated. For the past seven years, Roberts has been devoted to the formidable task of understanding censorship in China—or, as she and her colleagues put it, 'reverse-engineering' censorship.

"Understanding the incentives behind governments and political entities to prioritize information is quite interesting," she explained. "It uncovers a lot about governments' intentions."

Since joining UC San Diego in 2014, she's leveraged the School of Global Policy and Strategy's (GPS) strengths in the Asia Pacific and big data, including partnerships with its [21st Century China Program](#) and [Policy Design and Evaluation](#)

Lab (PDEL), to peel back the layers even further on Chinese censorship.

She's answering the question on many minds that is: What is the Chinese government trying to hide?

## Discerning the strategy

In a PDEL workshop on Jan. 27, Roberts presented to her faculty peers and graduate students what she knows so far.

The project began when her Ph.D. adviser Gary King, Albert J. Weatherhead III University Professor at Harvard University's Department of Government, came to her with a collection of millions of Chinese blog posts.

Roberts, at the time fresh off earning her master's in statistics from Stanford University, saw the collection as an opportunity to do what she enjoys most—use text as data. She worked alongside fellow Ph.D. candidate Jennifer Pan.

"We thought we wanted to study public opinion online in China, but we realized a lot of the blog posts had gone missing," Roberts recalled. "We began focusing on that rather than public opinion. We wanted to measure what the government was censoring."

Together, King, Pan and Roberts conducted a large-scale observational and experimental study of censorship. This entailed collecting more than 3 million social media posts from around 1,400 social media sites and downloading them immediately after posting, before they were potentially censored. Through a random sample of 127,000 posts, they found that the Chinese government censors during events with "collective action potential." Surprisingly, the main focus is not criticism of the state. This first [paper](#) was published in 2013.

# UC Institute for Predictive Technologies

**Started with UC seed funding.**  
*Thanks, President Napolitano!*

- Tyson Condee, Wei Wang, Sean Young, UCLA
- Mike Carey, Chen Li, UCI
- Vagelis Hristidis, UCR
- Amarnath Gupta, Christine Kirkpatrick, SDSC
- Natasha Martin, Stephanie Strathdee, UCSD



# Results 1: All Election-Related Events are Detected

In the news



Donald Trump revokes Washington Post press credentials - Jun. 13, 2016

CNN - 4 mins ago

Donald Trump said Monday that he is "revoking" the Washington Post's press access at his ...

Topic: 2

```
[('trump', 97), ('trump2016', 20), ('orlando', 13), ('prayforlilwayne', 10), ('p2', 10), ('uslatino', 10)]  
[('washingtonpost', 78), ('sexuaitalk', 44), ('usatoday', 39), ('anncoulter', 37), ('realdonaldtrump', 36), ]  
[('trump', 1949), ('donald', 585), ('washington', 213), ('post', 183), ('press', 171), ('credentials', 159), ]
```

US & Canada

US Election 2016

## Orlando shooting: Obama to console victims

🕒 33 minutes ago | US & Canada

🔗 Share


Topic: 6

```
[('orlando', 44), ('rt', 32), ('follow', 30), ('topstories', 22), ('news', 13), ('worstpresidentever', 11), ('trump', 7)]  
[('sheriffclarke', 62), ('abc', 26), ('realdonaldtrump', 17), ('writeintrump', 15), ('pincopallina93', 13), ('potus', 12)]  
[('obama', 1005), ('orlando', 248), ('president', 188), ('victims', 138), ('terror', 131), ('shooting', 112), ('console', 69)]
```

# Results 2: Early Events / News Item Prediction

 **palestina**  
@itsdatnunu 23 May

When you just trying to eat your ice cream but trump supporters won't let you live <pic.twitter.com/rW3MI8a2pn>

 **palestina**  
@itsdatnunu [Follow](#)

Shoutout to Andrews Ice cream in Orange. They were so sensitive and supportive as all Americans should be. 😊

10:54 PM - 23 May 2016

🔄 1,238 ❤️ 5,479

*Event just begins to start becoming popular*

Donald Trump surrogate is quietly courting Muslims to downplay ...  
[www.dailymail.co.uk/.../Trump-surrogate-quietly-courting-Muslims-downpla...](http://www.dailymail.co.uk/.../Trump-surrogate-quietly-courting-Muslims-downpla...) Daily Mail ▾  
May 23, 2016 - Trump called for a temporary ban on non-American Muslims entering the United States wearing a leather vest and black yoga pants as she feasts on ice cream in Saint Tropez ...

Ice cream shop closes on Frankfort Ave. - The Courier-Journal  
[www.courier-journal.com/story/news/.../ice-cream.../84777110/](http://www.courier-journal.com/story/news/.../ice-cream.../84777110/) ▾ The Courier-Journal ▾  
May 23, 2016 - Homemade Ice Cream kitchen on Frankfort Avenue has closed; new store to open on Dixie Highway.

Sale on Ben & Jerry Ice Cream at CVS!!! - The Killeen Daily Herald  
[kdhnews.com/.../ice-cream.../article\\_0866331e-2132-11e6-ab0b-a7...](http://kdhnews.com/.../ice-cream.../article_0866331e-2132-11e6-ab0b-a7...) Killeen Daily Herald ▾  
May 23, 2016 - Ben & Jerry Ice Cream is on sale for \$3.99 this week at CVS! (originally \$5.99). If you have the 5/22 Redplum inserts, there's a manufacturer coupon for a \$1 off ...

Trump on '90s suicide of top Clinton aide: 'Very fishy' | TheHill  
[thehill.com/blogs/.../280999-trump-on-suicide-of-top-clinton-aide-very-fishy](http://thehill.com/blogs/.../280999-trump-on-suicide-of-top-clinton-aide-very-fishy) ▾ The Hill ▾  
May 23, 2016 - "He had intimate knowledge of what was going on," Trump said of Foster's relationship with the Clintons. "He knew ... The Rubes will lap it up like it's chocolate ice cream. only tj \* 3 ... DHS ordered me to scrub records of Muslims with terror ties.

Transgender Dignity in Islam : Related Articles | OOUZ  
[www.oouyz.com/geturl?aid=11737352](http://www.oouyz.com/geturl?aid=11737352) ▾  
May 23, 2016 - London Mayor Sadiq Khan offers Trump tour - Business Insider. -Business Insider ...  
California man berates Muslim women at ice cream store. -NY Daily News.

*No results on google regarding event yet*

## Topic 3

[('trump', 15797), ('donald', 3496), ('clinton', 1321), ('hillary', 1265), ('supporters', 1119), ('live', 1104), ('wont', 1046), ('let', 1006), ('trying', 897), ('ice', 864), ('cream', 860), ('eat', 824)]



# UCIPT In Action: Early Events / News Item Prediction

Cal @halbur · 2m

**palestina**  
@itsdatnunu

When you just t  
won't let you liv

**palestina**  
@itsdatnunu

Shoutout to Andrew  
sensitive and suppo  
10:54 PM - 23 May 20

1,238


**This Guy Was Kicked Out Of An Ice Cream Parlor After Telling Two Muslim Women “I Don’t Want Them Near My Country”**

“When you just trying to eat your ice cream but trump supporters won’t let you live.”

*posted on May 24, 2016, at 2:14 p.m.*

**Tasneem Nashrulla**  
BuzzFeed News Reporter

**Nura Takkish, a 22-year-old woman from California, tweeted a video Monday showing a man at an ice cream parlor telling her and her friend — they were both wearing hijabs — “I don’t want them near my country.”**



Google

Donald Trump surrogate is quietly courting Muslims to downplay ...  
/.../Trump-surrogate-quietly-courting-Muslims-downpla... Daily Mail ▾  
p called for a temporary ban on non-American Muslims entering the United ....  
k yoga pants as she feasts on ice cream in Saint Tropez ...

closes on Frankfort Ave. - The Courier-Journal  
com/story/news/.../ice-cream.../84777110/ ▾ The Courier-Journal ▾  
emade Ice Cream kitchen on Frankfort Avenue has closed; new store to open

Jerry Ice Cream at CVS!!! - The Killeen Daily Herald  
cream.../article\_0866331e-2132-11e6-ab0b-a7... Killeen Daily Herald ▾  
& Jerry Ice Cream is on sale for \$3.99 this week at CVS! (originally \$5.99). If you  
im inserts, there's a manufacturer coupon for a \$1 off ...

suicide of top Clinton aide: 'Very fishy' | TheHill  
280999-trump-on-suicide-of-top-clinton-aide-very-fishy ▾ The Hill ▾  
ad intimate knowledge of what was going on," Trump said of Foster's  
 Clintons. "He knew ... The Rubes will lap it up like it's chocolate ice cream.  
ered me to scrub records of Muslims with terror ties.

gnity in Islam : Related Articles | OOOYUZ  
url?aid=11737352 ▾  
on Mayor Sadiq Khan offers Trump tour - Business Insider. -Business Insider ...  
as Muslim women at ice cream store. -NY Daily News.

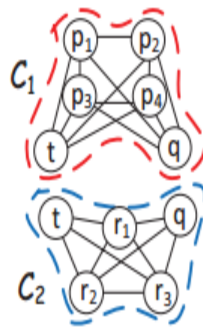
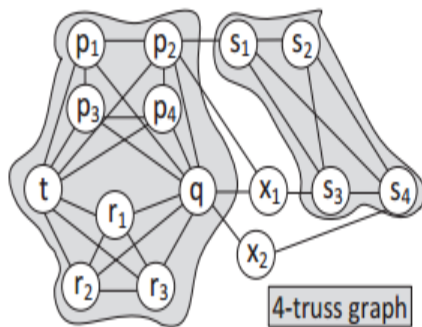
rters', 1119), ('live', 1104), ('wont',

**SDSC** SAN  
SUP

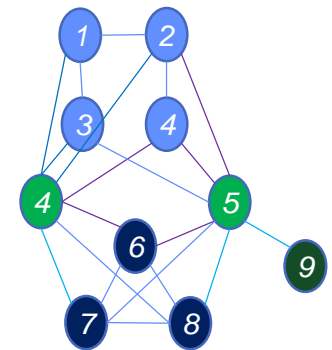
UNIVERSITY  
OF  
CALIFORNIA

# Community Evolution

- How are the important entities (e.g., topics, people, ...) in a networked system connected? How do these connections change over time and with external events?
- How to define a community?
  - $k$ -truss subgraph
    - A subgraph for which each edge has at least  $k - 2$  triangles



- Nodes
  - Hashtags
  - Users
  - Tweets
- Edges
  - Mentions
  - Corefs
  - Tweeted



# The UCIPT Project

- **NIH R01 Grant Awarded**
  - UCLA
    - Lead, Predictive Analytics
  - UCI
    - AsterixDB enhancements, Visualization
  - UC Riverside
    - Spatiotemporal Analysis
  - UC San Diego Preventive Medicine
    - Data set provider
  - AWESOME
    - Data Layer
    - Heterogeneous Features for Prediction
      - Continuous topic discovery
    - Solution Builder

## **HIV Risk on Twitter: The Ethical Dimension of Social Media Evidence-based Prevention for Vulnerable Populations**

Nadir Weibel, Purvi Desai, Lawrence Saul, Amarnath Gupta, Susan Little  
*University of California, San Diego,*  
9500 Gillman Dr., La Jolla, CA 92093, USA  
{weibel, pdesai, lsaul, a1gupta, little}@ucsd.edu

*Funding provided by NIH #20162575*

# NSF West Big Data Innovation Hub (WBDIH)



Funding provided by NSF CISE (NSF 15-562)

**Scale our successes  
Spark sustainable partnerships**

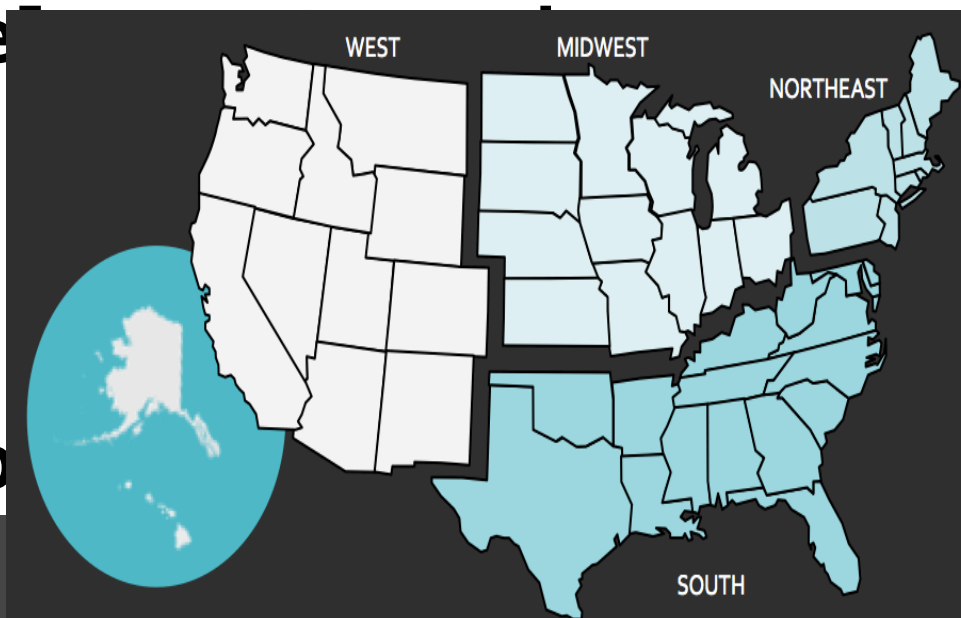


# Regional Big Data Innovation Hubs

- Build and strengthen partnerships across industry, academia, nonprofits, government to
- Address societal + scientific challenges,
- Spur economic development
- Foster a national ecosystem

P.S.

- Azure credits
- Open Storage Network



# West Hub Priority Areas



## METRO / URBAN

Smart Cities,  
Transportation,  
Housing, Police Data,  
Economic Development



## PRECISION MEDICINE

Diagnostics,  
Treatment, Genomics,  
Environment/ Exposome



## NATURAL RESOURCES & HAZARDS

Water, energy,  
agriculture, disaster  
response, sustainability



## SCIENTIFIC DISCOVERY & LEARNING

Open science &  
reproducibility,  
education, workforce

Convene  
Curate  
Communicate

## CROSS-CUTTING EFFORTS:

- Cloud Computing Task Force
- Public Policy, Ethics, Privacy, Security
- Data Sharing, Infrastructure
- Data-Driven Storytelling / Data Literacy
- Data Hackathons Community of Practice



# The Science of Data-Driven Storytelling

Translating our work to generate societal impact

#datascistories  
@datascienceinc @westbigdatahub

# Metro Science for Public Good

THINK  
GLOBAL  
ACT  
LOCAL

- **Cities are Dynamic Environments**
  - Many sensed and organizational open datasets
  - Potential to improve public safety and quality of life



SDSC CDSO:  
*Ilkay Altintas*

AS  
SEEN  
ON

**coursera**





# Knowledge Discovery and Real-Time Interventions from Sensory Data Flows in Urban Spaces

- California Energy Commission Electricity and Gas Consumption
- US Energy Information Administration Dataset
- San Diego Power Outage Data



**Energy Utility**

**Traffic**



- San Diego and Carlsbad Traffic Flow Data
- San Diego Adaptive Traffic Light Data
- Airport Traffic Delay Reporting Service
- GeoLife GPS Trajectories
- WAZE Connected Communities

- **Residential:** AMPds, Dataport, ECO, iAWE, PLAID, REDD, UK-DALE
- **Commercial:** COMBED



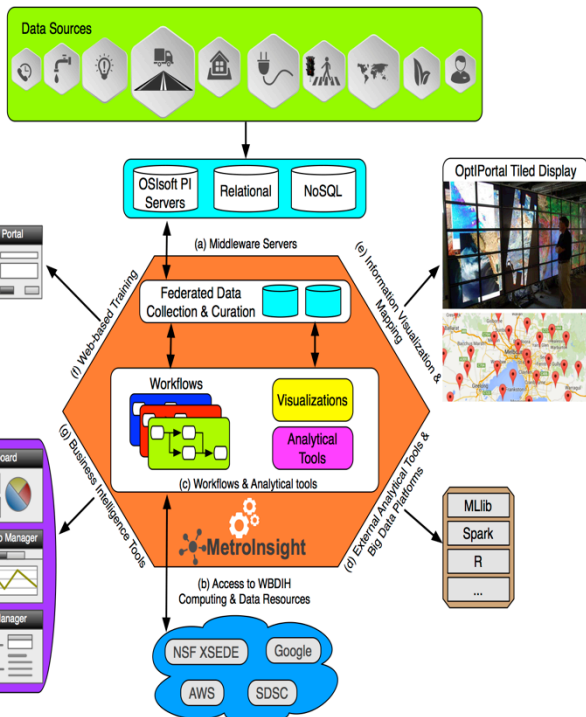
**Energy Consumption**

**Weather and Hazards**



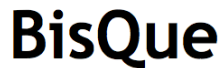
- Earthquake Notification Service
- Weather Forecast
- WIFIRE Wildfire Archive
- DEMROES
- HPWREN Sensor and Camera Data

**MetrolInsight Data Sources**



# Data Platform & Tools Ecosystem

- Hasn't this been done already?



RESEARCH DATA ALLIANCE



PLATFORM FOR SCIENTIFIC COLLABORATION

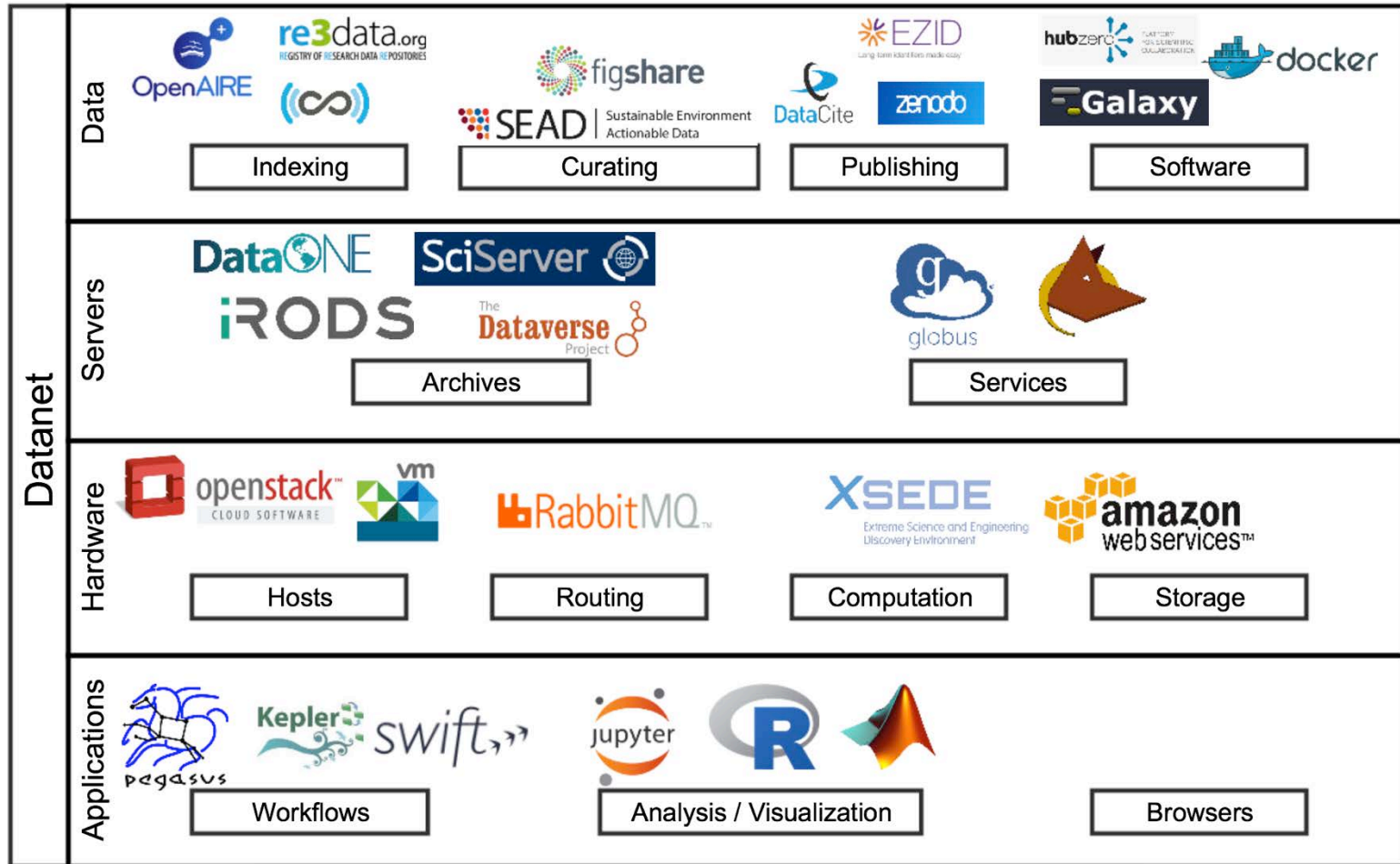


SAN DIEGO SUPERCOMPUTER CEN



UNIVERSITY OF CALIFORNIA

# Where We Are



# What is the National Data Service?

National effort to bring together infrastructure supporting the *publication, discovery, and reuse* of data

## 1. Large-scale Data Service Interoperability

- Distributed cloud and compute
- Innovation in the gaps: services, software, integration

## 2. Incubator of Data Projects & Pilots

- Quick start sandbox
- Choose services based on features (not time to install)

## 3. Training, course, hackathon support





# Who We Are: NDS Consortium

- **Communities**

- Astronomy, Biology, Engineering, Geoscience, Information Science, Material Science, Medicine, Social Science

- **Universities, Libraries, Archives, and Publishers**

- CU Boulder, Harvard, Indiana, Johns Hopkins, Notre Dame, Purdue, UC San Diego, UIC, UIUC, U Michigan, ICPSR ...
- Nature, Science, APS, IEEE, PLOS, Elsevier, ...

- **Computing and Data Centers/Cyberinfrastructure**

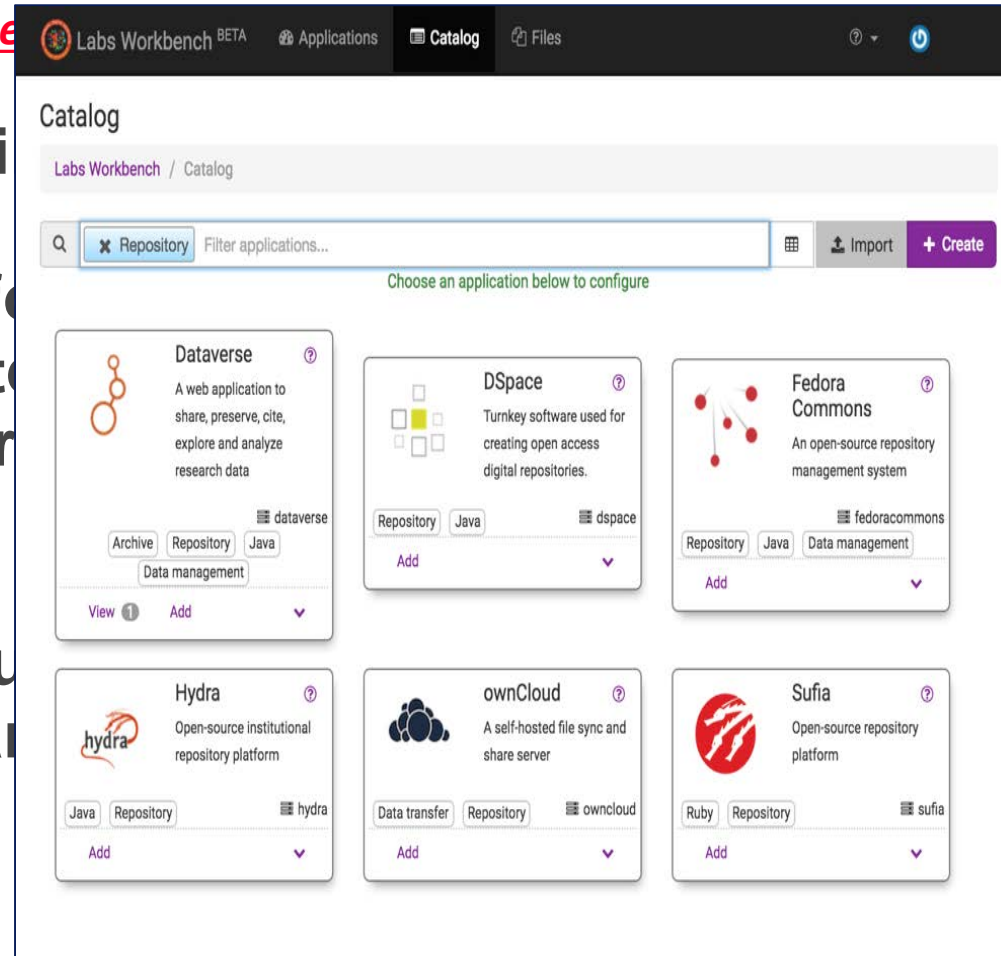
- ANL, NCSA, PSC, SDSC, TACC
- Brown Dog, Data Excacell, DataONE, DFC, CyVerse, GABBs, IN-CORE, iRODS, Globus, SciServer, SEAD, Terra Populus, TERRA REF, Whole Tale, LSST, LIGO, ...



# NDS Labs Workbench

<https://www.workbench.nationaldataset.org>

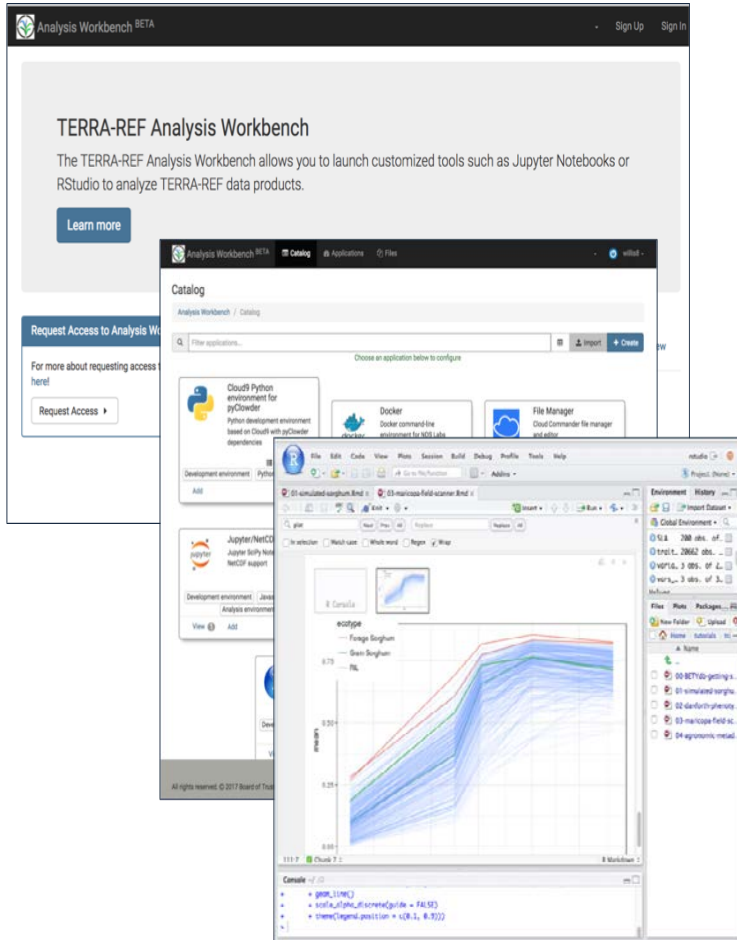
- NDSC initiative started in January 2016
- Community-driven platform to share, discover, evaluate, develop, and test research data management and analysis tools
- Open platform -- community members recommend and contribute tools



# Use case: TERRA-REF Analysis Workbench

<http://www.terraref.ndslabs.org>

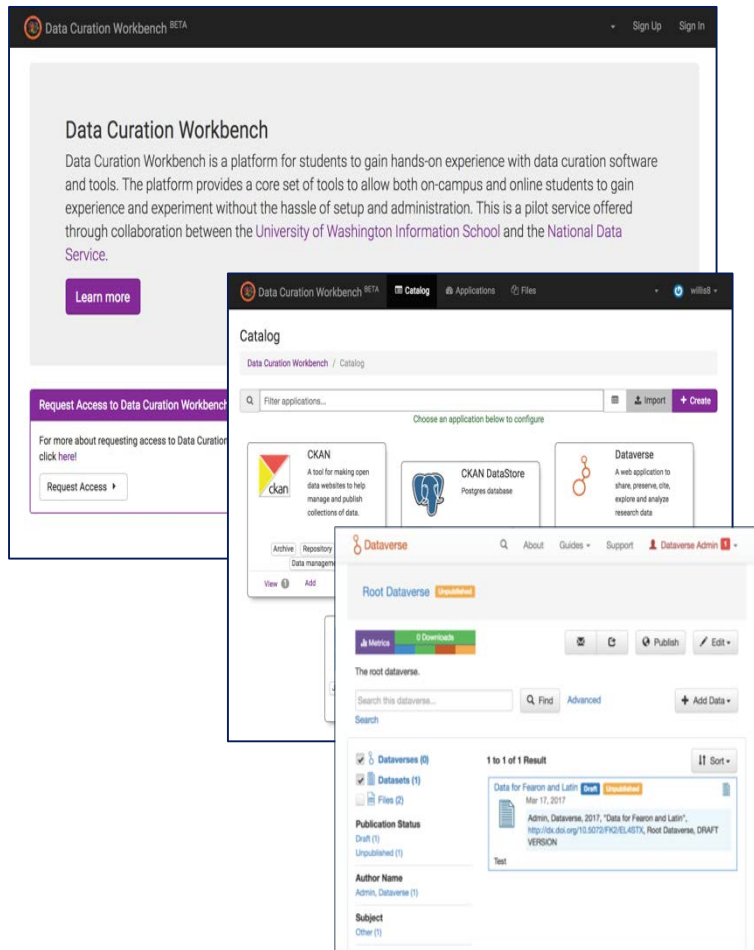
- Initially used for workshop tutorial to provide consistent environments, scaling to support ~50 participants.
- Customized Labs Workbench instance hosted as part of TERRA-REF infrastructure at NCSA.
- Custom analysis and development environments with direct access to TERRA-REF data. Used by collaborators and alpha users.



# Use case: Data Curation Education Workbench

<http://www.ischool.ndslabs.org>

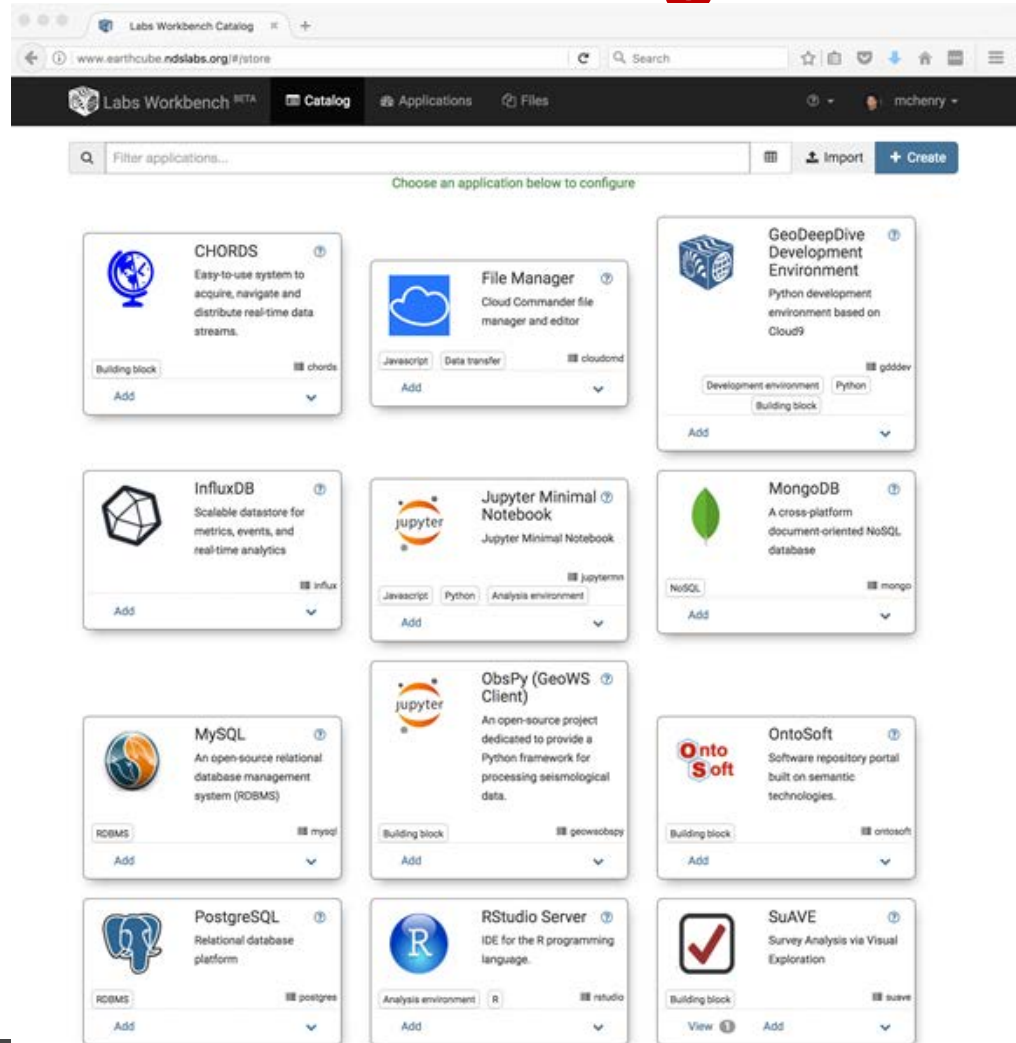
- NDS pilot project initiated during NDSC6, led by Carole Palmer at the University of Washington iSchool
- Customized instance of Labs Workbench hosted on SDSC Cloud
- Sandbox for students to get hands-on experience with data curation systems and tools including CKAN and Dataverse.





# Use Case: EarthCube Building Blocks

## Custom catalog for Geoscientists












# Beyond Traditional Publications

- **Equal Partners:  
Pubs, Data, Analysis**
- **Transparency and  
access to methods**
- **Not feasible to re-run  
computation.**
- **Analysis, computed  
data**



# NDS Share: DataDNS

## Registered Data Sets

Dataset	Publications	Location	Launch Notebook	Show Metrics
<a href="#">Renaissance Simulations</a> O'Shea, Brian (oshea@msu.edu); Wise, John; Xu, Hao; Norman, Michael <a href="#">Cite Dataset</a>	<ul style="list-style-type: none"><li>O'Shea, B. W., Wise, J. H., Xu, H., &amp; Norman, M. L. (2015). PROBING THE ULTRAVIOLET LUMINOSITY FUNCTION OF THE EARLIEST GALAXIES WITH THE RENAISSANCE SIMULATIONS. <i>The Astrophysical Journal</i>, 807(1), L12. doi:10.1088/2041-8205/807/1/L12</li><li>Ahn, K., Xu, H., Norman, M. L., Alvarez, M. A., &amp; Wise, J. H. (2015). SPATIALLY EXTENDED 21 cm SIGNAL FROM STRONGLY CLUSTERED UV AND X-RAY SOURCES IN THE EARLY UNIVERSE. <i>The Astrophysical Journal</i>, 802(1), 8. doi:10.1088/0004-637x/802/1/8</li></ul> <a href="#">More</a>			
<a href="#">Dark Sky Simulations</a> Warren, Michael; Friedland, Alexander; Holz, Daniel; Skillman, Samuel; Sutter, Paul; Turk, Matthew (mjturk@illinois.edu); Wechsler, Risa <a href="#">Cite Dataset</a>	<ul style="list-style-type: none"><li>S. W. Skillman, M. S. Warren, M. J. Turk, R. H. Wechsler, D. E. Holz, P. M. Sutter. <i>Dark Sky Simulations: Early Data Release</i>.</li><li>Warren, M. S., Friedland, A., Holz, D. E., Skillman, S. W., Sutter, P. M., Turk, M. J., &amp; Wechsler, R. H. (2014). Dark Sky Simulations Collaboration. ZENODO. <a href="https://doi.org/10.5281/zenodo.10777">https://doi.org/10.5281/zenodo.10777</a></li></ul>			
<a href="#">Magnetohydrodynamic Turbulence Simulations</a> Mösta, Philipp (pmoesta@berkeley.edu) <a href="#">Cite Dataset</a>	<ul style="list-style-type: none"><li>Mösta, P., Ott, C. D., Radice, D., Roberts, L. F., Schnetter, E., &amp; Haas, R. (2015). A large-scale dynamo and magnetoturbulence in rapidly rotating core-collapse supernovae. <i>Nature</i>, 528(7582), 376–379. doi:10.1038/nature15755</li></ul>			

**“...to make the fruits of research and scholarship better and available to all who need or want them.” Berman et al.  
christine@sdsc.edu**

**Christine Kirkpatrick**

**Division Director, IT Systems & Services, SDSC**

**Deputy Director, West Big Data Innovation Hub**

**Executive Director, National Data Service**