

December 18, 2014

Big Data: Where can EPSCoR states use big data and what tools do EPSCoR states need to thrive

Craig Stewart – XSEDE Campus Bridging manager, Jetstream Principal Investigator

XSEDE

Extreme Science and Engineering
Discovery Environment



Initial assertions

- What EPSCoR states typically have in abundance:
 - Excellent faculty
 - Interested and energetic students
- What EPSCoR states sometimes have too little of:
 - Support staff
 - Lab equipment
 - Cyberinfrastructure resources
 - Release time for faculty to do curriculum development

What is cyberinfrastructure, anyway?

- “Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, **and people**, all linked by high speed networks to make possible scholarly innovation and discoveries not otherwise possible.”



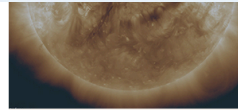
The best friend of researchers, research educators, and students interested in research in EPSCoR states is...

- The gap between the US and international research communities' ability to produce research-quality data and ability to analyze these data



Just a few examples

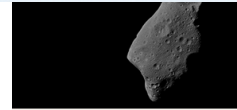
- NCBI
- Zooniverse
- VAO
- Many, many national databases of DNA and RNA sequences, weather data, etc.



Sorting out Sunspots

Help us organize sunspot images in order of complexity to better understand the Sun's magnetic activity.

SUNSPOTTER



Help us discover near-Earth asteroids

We need your help to protect the Earth, seek out potential future resources, and better understand the form...

ASTEROID ZOO

Climate



Model Earth's climate using historic ship logs

Help scientists recover Arctic and worldwide weather observations made by US Navy and Coast Guard ships.

Weather



Classify over 30 years of tropical cyclone data.

Scientists at NOAA's National Climatic Data Center need your help.

CycloneCenter

Humanities



Study the lives of ancient Greeks

The data gathered by Ancient Lives helps scholars study the Oxyrhynchus collection.

ANCIENT LIVES



Explore soldiers' diaries from the First World War

Annotate and tag diaries from the First World War.

OPERATION WAR DIARY

Nature



Hear Whales communicate

You can help marine researchers understand what whales are saying



Help explore the ocean floor

The HabCam team and the Woods Hole Oceanographic



You're hot on the trail of bats!

Help scientists characterise bat calls recorded by citizen



Go wild in the Serengeti!

We need your help to classify all the different animals caught in millions of camera

XSEDE (xsede.org) is a national source of cyberinfrastructure resources

- Allocated
 - Cycles
 - Data storage
 - Support
 - Get help the first time you apply - help@xsede.org and/or your local campus champion
- Available to all (without allocations)
 - Globus Transfer
 - Training & curriculum materials
 - Campus Bridging



The logo features the word "Jetstream" in a bold, italicized, red sans-serif font. A light blue swoosh underline starts from the left edge of the frame, passes behind the letters, and tapers off to the right. Below the main text, the tagline "A national science & engineering cloud" is written in a smaller, white, sans-serif font.

Jetstream

A national science & engineering cloud

**funded by the National Science Foundation
Award #ACI-1445604**

What is Jetstream?

- NSF's first cloud for science and engineering research across all areas of activity supported by the NSF
- Jetstream will be a user-friendly cloud environment designed to give researchers and research students access to interactive computing and data analysis resources “on demand.”
- It will provide a user-selectable library of virtual machines that users can select from to do their research.
- Software creators and researchers will also be able to create their own customized virtual machines -or- their own “private computing system” within Jetstream.
- It will enable countless discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.
- Two especially important biology platforms will be supported - iPlant and Galaxy.



XSEDE

What does the name mean? And is it really a cloud?

- Name
 - In the atmosphere the Jetstream lies at the border of two different air masses
 - The Jetstream system stands at the border of the existing NSF-funded XD program and advanced cyberinfrastructure resources and users who have not previously used such NSF funded infrastructure before.
- Yep, it's really a cloud, or at least a cloud environment (one could quibble over the definition of cloud vis-à-vis expansibility). Software layers:
 - Atmosphere interface
 - KVM
 - OpenStack
 - CentOS Linux



Dashboard

Images

Favorites

My Images

Projects

Cloud Providers

Quotas

Settings

Search Images

Search by App Images, Tag, OS, and more

Popular Searches: [R](#) [Bisque](#) [NGS](#) [Community: Astrophysics](#)


Quick Sort: Popularity Recency Rating

[Advanced Search Options](#)

Quick Filter:

View as:

Popular Images from All Communities




Math Kernel Library

[blas](#) [fft](#) [fortran](#) [lapack](#)

Community: Mathematics

52 likes, 0 dislikes, 7 comments



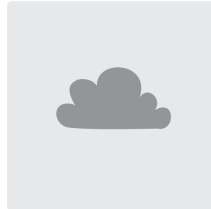
RNASeq Analysis Tools

[bowtie2](#) [blast](#) [blat](#) [edgeR](#)

[R](#) [rnaseq](#) [tophat2](#)

Community: Biology

30 likes, 2 dislikes, 4 comments




Atmospheric Dispersion Modeling

[aermod](#) [aermet](#) [aermap](#)

Community: Atmospheric Sciences

20 likes, 0 dislikes, 0 comments





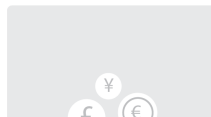
MrBayes with TreeMix

[bayesian inference](#) [mrbayes](#)

[treemix](#)

Community: Phylogenetics

25 likes, 1 dislike, 10 comments



Science Domains and Users

- Biology
- Earth Science/Polar Science
- Field Station Research
- Geographical Information Systems
- Network Science
- Observational Astronomy
- Social Sciences
- Jetstream will be particularly focused on researchers working in the “long tail” of science with born digital data
- Enabling analysis of field-collected empirical data on the impact and effects of global climate change will be one of the specific foci of Jetstream
- Whatever *you* do



XSEDE

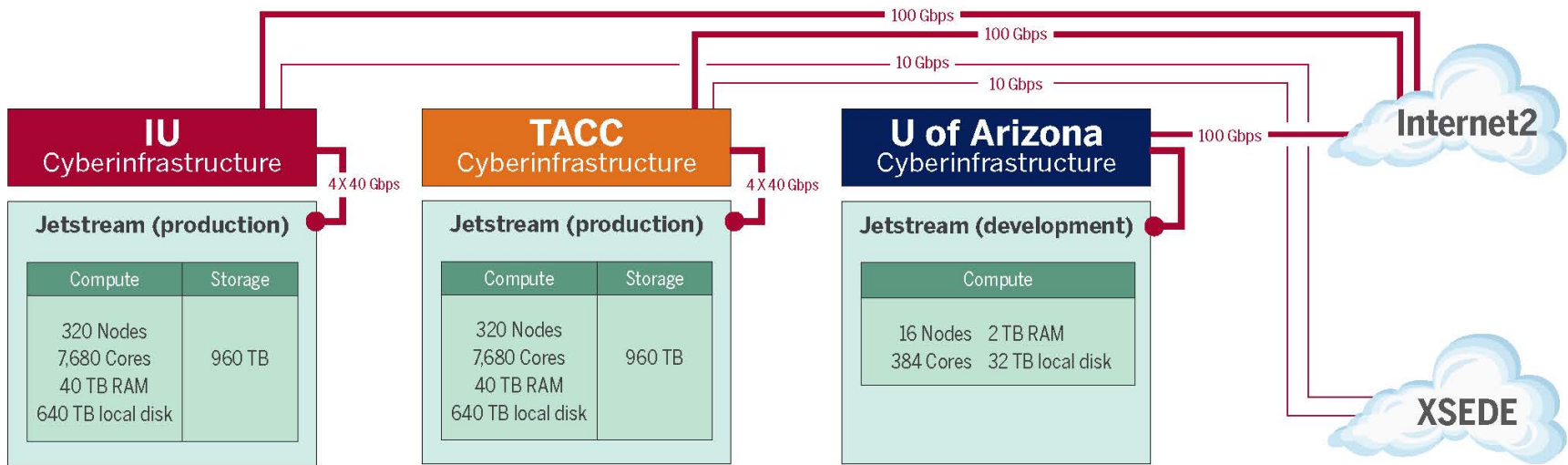
21st century workforce development

- Jetstream will include virtual Linux desktops and applications specifically aimed to enable research and research education at small colleges and universities including HBCUs (Historically Black Colleges and Universities), MSIs (Minority Serving Institutions), Tribal colleges, and higher ed institutions in EPSCoR States
- Jetstream will also support deployment of user-friendly Science Gateways



XSEDE

Jetstream System Diagram



Jetstream Collaborators

- University of Chicago - Globus
- Arizona University – iPlant
- Johns Hopkins University and Penn State University
- Cornell University –Ms. Susan Mehringer, Lead. Cornell® Virtual Workshops about Jetstream and applications running on jetstream.
- University of Arkansas at Pine Bluff – Dr. Jesse Walker, lead. cybersecurity education,, Minority Serving Education outreach
- University of Hawaii – Dr. Gwen Jacobs, lead. EPSCoR early adopter/user. Jacobs will chair Science Advisory Board
- National Snow and Ice Data Center (NSIDC) – Dr. Ron Weaver, lead. Data retrieval from NSIDC, application integration with ice sheet analysis applications
- University of North Carolina, Odum Center –Dr. Thomas Carsey , lead. Data retrieval from Dataverse Network
- National Center for Genome Analysis at Indiana University – providing genome analysis software. Includes TACC, PSC, and SDSC as partners



XSEDE

EPSCoR states can thrive in the big data era by linking

- Your local talent
- Publically available, research quality data
- National CI resources
- *To create new, meaningful, and important discoveries*

Your thoughts and questions?

- Now or...
- send follow up questions to stewart@iu.edu




Please cite as: Stewart, C.A. 2014. “Big Data: Where can EPSCoR states use big data and what tools do EPSCoR states need to thrive?” Presentation before the EPSCoR / IDEA Board of Directors. (Arlington, VA, December 3, 2014). Available at: <http://hdl.handle.net/2022/19210>

Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.

Except where otherwise noted, the contents of this presentation are copyright 2014 by the Trustees of Indiana University. This content is released under the Creative Commons Attribution 3.0 Unported license (creativecommons.org/licenses/by/3.0). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.





Our reach will forever
exceed our grasp, but,
in stretching our horizon,
we forever improve our world.

XSEDE

Extreme Science and Engineering
Discovery Environment