CyVerse: Cyberinfrastructure for Data Driven Discovery
By Tyson Swetnam

NSF Ecosystem Lightning Talks

March 2, 2022
CI Compass Cyberinfrastructure for NSF Major Facilities Workshop
Making cloud more accessible and inclusive

Tyson L Swetnam, Co-PI
Eroding the digital divide

Problem: The complexity of modern cyberinfrastructure and data fortresses

- Barriers to collaboration
- Limits inclusion and knowledge-transfer
- Limits scientific inquiry and discovery

Solution: Focus on creating open-source platforms which increase accessibility and inclusion and train in their use

- Cloud-Optimized, Analysis Ready Data
- Data-proximate computing
- Dynamic and elastic scaling

Ryan Abernathey’s recent PANGEO talk at OpenOceanCloud: https://vimeo.com/670782104
What is Cyberinfrastructure?

Cyberinfrastructure is a combination of:

- Hardware for compute & storage
- Software, code, & data
- People who provide support & training
# Data Scientists need Training

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

Barone et al. 2017

<table>
<thead>
<tr>
<th>Training need</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training on integration of multiple data types</td>
<td>80%</td>
</tr>
<tr>
<td>Training on data management and metadata</td>
<td>70%</td>
</tr>
<tr>
<td>Training on scaling analysis to cloud/high performance computing</td>
<td>60%</td>
</tr>
<tr>
<td>Multi-data analysis and visualization</td>
<td>50%</td>
</tr>
<tr>
<td>Cloud computing</td>
<td>40%</td>
</tr>
<tr>
<td>Search for data &amp; discover relevant datasets</td>
<td>30%</td>
</tr>
<tr>
<td>Support for bioinformatics and analysis</td>
<td>20%</td>
</tr>
<tr>
<td>Publish data to the community</td>
<td>10%</td>
</tr>
<tr>
<td>Updated analysis software</td>
<td>0%</td>
</tr>
<tr>
<td>Share data with colleagues</td>
<td>0%</td>
</tr>
<tr>
<td>Training on basic computing and scripting</td>
<td>0%</td>
</tr>
<tr>
<td>Sufficient data storage</td>
<td>0%</td>
</tr>
<tr>
<td>High-performance computing</td>
<td>0%</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pcbi.1005755
98,500 users
1500 publications
11 PBs of data
40,000 trained

- Secure Perimeter (VPN)
- HIPAA Compliant Platforms inside of Perimeter
- XNAT
- Training

- Researchers collect data.
- Data is shared and analyzed in the cloud.
- Researchers learn more skills.

- Federation with local and commercial cloud and high-performance computing
- Integration with local user identity management systems
- Security compliance

- ITAR Compliance
- Receive data from multiple sensors
- Policy based data visibility and sharing
- Support for multiple teams and data partitioning
Creating Accessible Platforms
How it works

Interactive Development Environments

Executable Workflows and HPC applications
Building Inclusive Communities
Focus on the problems tomorrow’s data scientists have today

- How do I write a great data management plan to get that next grant?
- How do I finish this analysis and publish?
- How do I scale my analyses to process all of these data?
Which Foundational Open Science Skills are most important?

Findable, Accessible, Interoperable, Reusable

Linked Open Data

OPEN ACCESS

open source initiative

Dev

Ops

code, plan, release, deploy, Operate;
test, monitor
How we do it
DBI-0735191, DBI-1265383, DBI-1743442 and OAC-1664172
Thank You