The GISandbox: A Science Gateway For Geospatial Computing

Davide Del Vento, Eric Shook, Andrea Zonca

XSEDE
Extreme Science and Engineering Discovery Environment

Curtis Marean

Curtis.Marean@asu.edu 480-965-7796 166 SHESC Arizona State University

▼ **Institute of Human Origins**

- Bio
- Research
- Teaching
- Public Work

Biography

Foundation Professor, School of Human Evolution and Social Change, College of Liberal Arts and Sciences
Associate Director, Institute of Human Origins

SHESC Themes: Human Origins, Evolution and Diversity; Societies and their Natural Environments

Field Specializations: Archaeology, Modern Human Origins, Paleoanthropology, Paleoecology, Zooarchaeology

Regional Focus: Africa (Southern); Near East

Dr. Marean's research interests focus on the origins of modern humans, the prehistory of Africa, the study of animal bones from archaeological sites and climates and environments of the past. In the area of the origins of modern humans, he is particularly interested in questions about foraging strategies and the evolution of modern human behavior. Dr. Marean has a special interest in human occupation of grassland and coastal ecosystems.

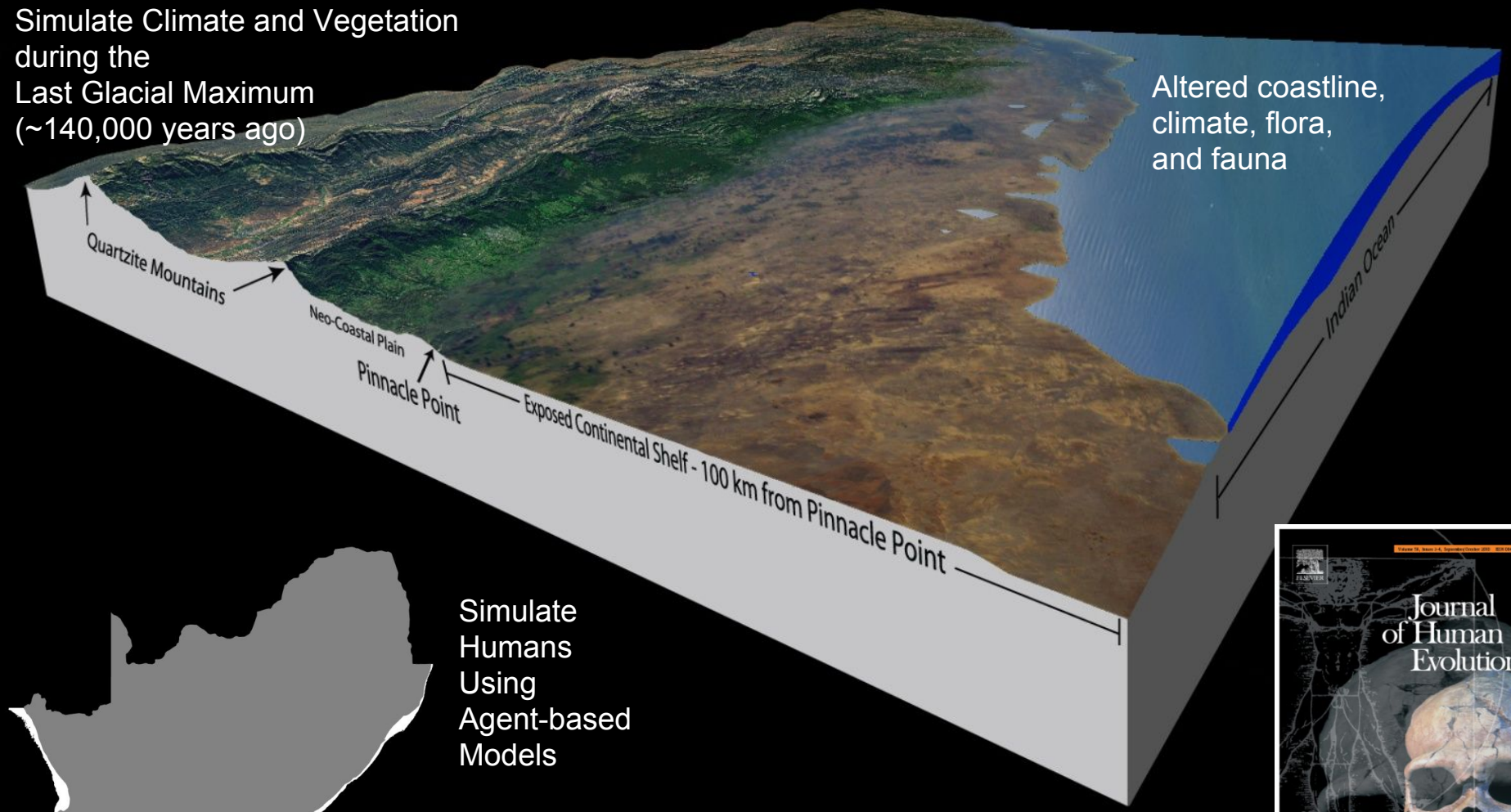
Dr. Marean conducts a variety of studies using zooarchaeology, the study of animal bones, and taphonomy, the study of how bones become fossils. He also is a dedicated field researcher and has conducted fieldwork in Kenya, Tanzania and Somalia, and since 1991 has focused his field efforts in coastal South Africa. He is the principal investigator for the South African Coast Paleoclimate, Paleoenvironment, Paleoecology, Paleoanthropology (SACP4) project based around Mossel Bay in South Africa at the field locality of Pinnacle Point. This large international project, funded by the National Science Foundation and the Hyde Family Foundation, employs a transdisciplinary approach to modern human origins, climate and environment. Under his directorship, Pinnacle Point has become one of the world's most important localities for the study of modern human origins.

Expertise Areas

- Archaeology
- Paleoanthropology
- Anthropology

Paleoscape Model and Human Origins

Simulate Climate and Vegetation during the Last Glacial Maximum (~140,000 years ago)

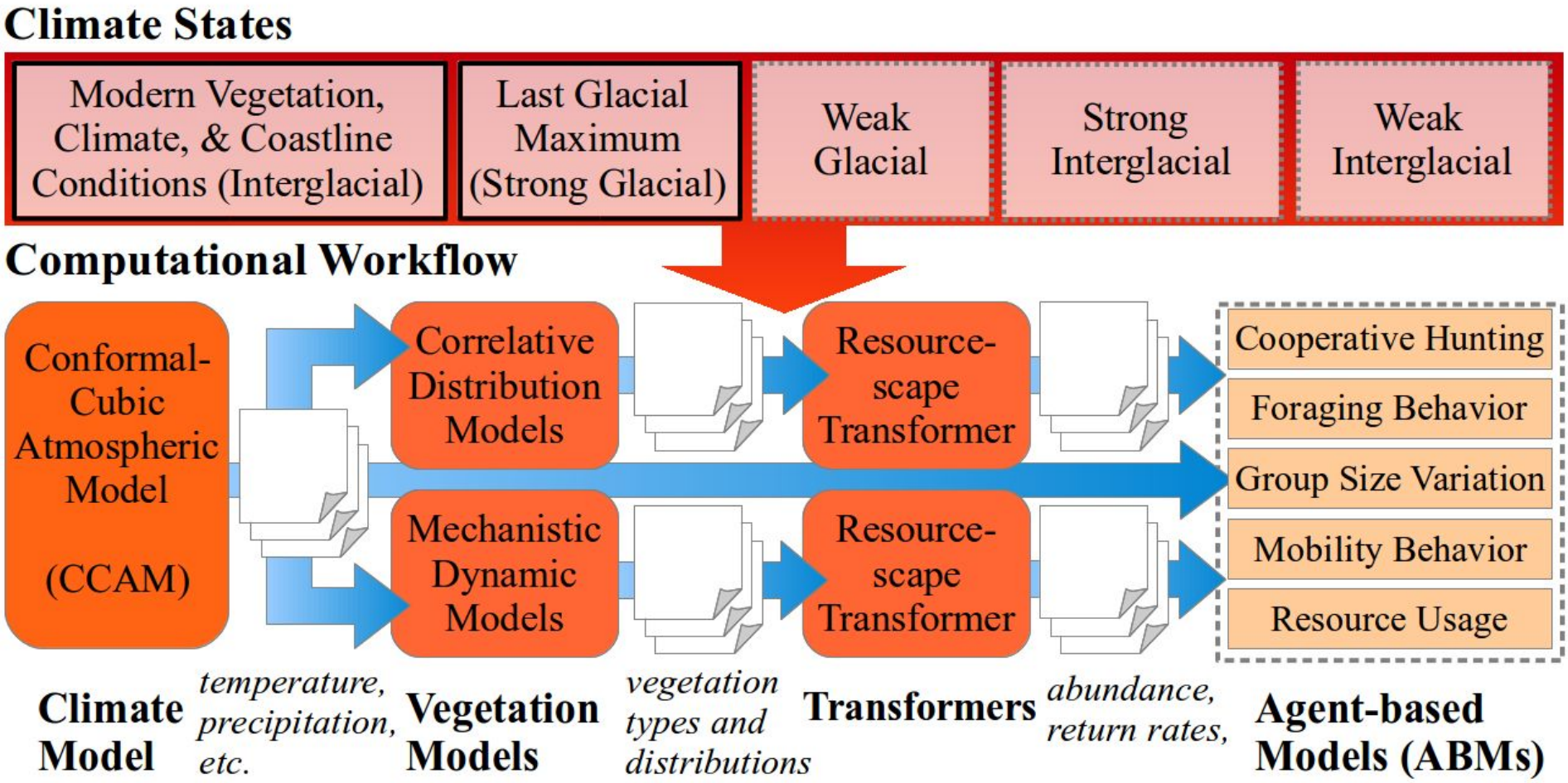


Altered coastline, climate, flora, and fauna

Simulate Humans Using Agent-based Models

Republic of South Africa
Images courtesy of Curtis Marean + Paleoscape Team

Spatio-temporal Data Analytics "Couple" Models for Paleoscape



(Shook, et al. @ XSEDE15)

XSEDE

Extreme Science and Engineering Discovery Environment

- The most advanced, powerful, and robust collection of integrated advanced digital resources and services in the world
- 10-year $220 million NSF project
- Supports 15+ supercomputers and high-end visualization & data analysis resources
- Provides ***free*** access to computing resources and computational experts

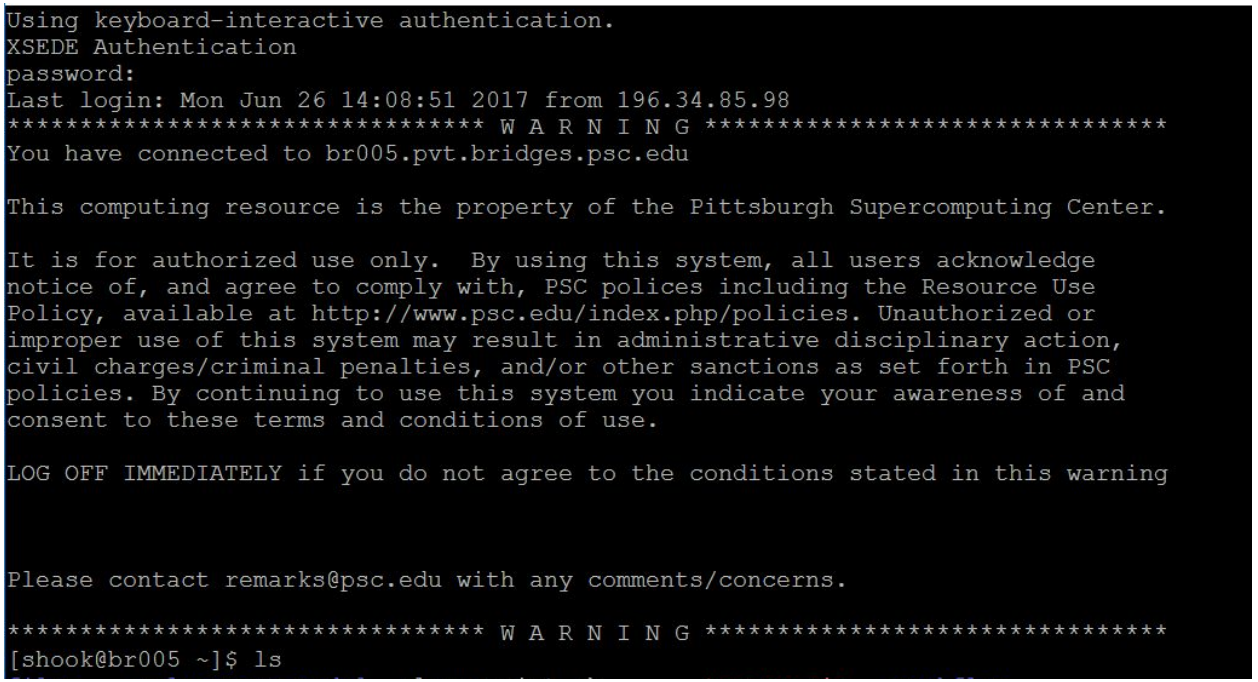
http://xsede.org

- **Startup Allocation**
  - ~1/2 page proposal discussing what you aim to do
  - Allocations up to 50,000 hours for a single resource
  - Quick turnaround time (~1 week evaluation)
  - Considered the “first step” for most researchers
- **Research (or Education) Allocation**
  - 10 page proposal + computational details
  - Allocations up to 10 million hours per year
  - Reviewed quarterly

** Both free for research and education*

A Challenge for Many Users

- Command-line Interface
- Batch Queuing System
- Split architecture:
  - Head Nodes,
  - Compute Nodes,
  - Viz Nodes,
  - Data Sharing Nodes
- Complex Software Stack
  - environment
  - usually hard to tailor for special needs
- and more ...

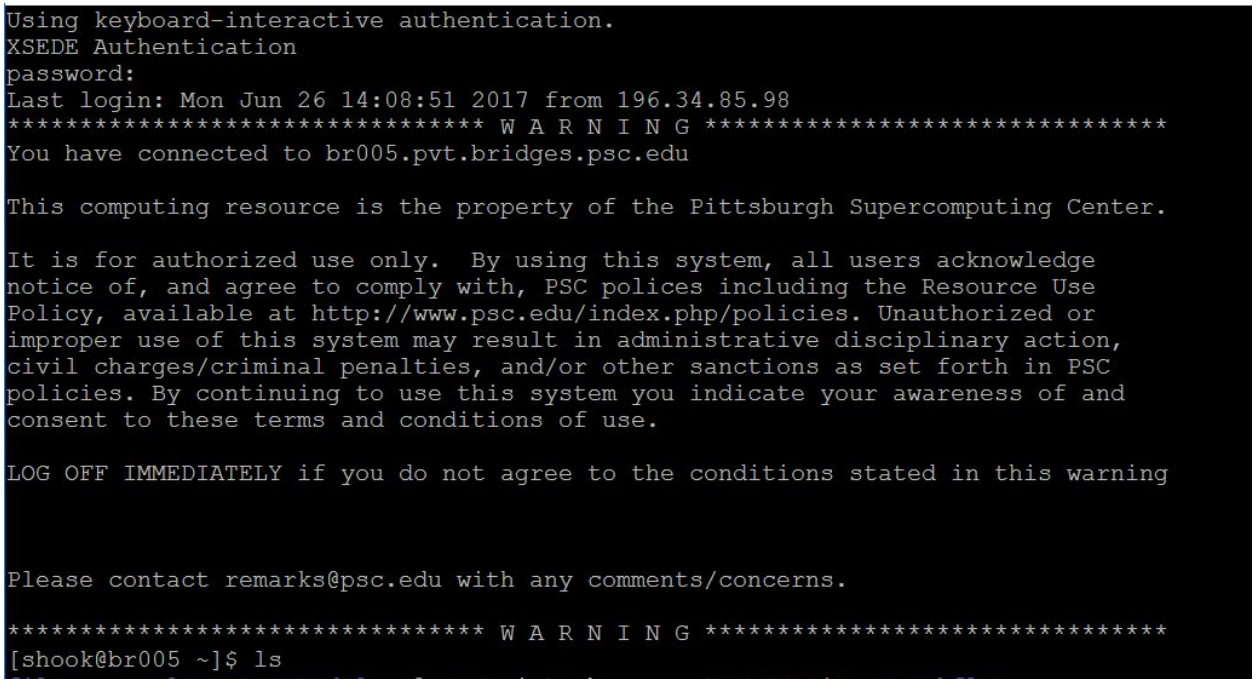


Bridges supercomputer at the Pittsburgh Supercomputing Center

A Challenge for Many Users

- Command-line Interface
- Batch Queuing System
- Split architecture:
  - Head Nodes,
  - Compute Nodes,
  - Viz Nodes,
  - Data Sharing Nodes
- Complex Software Stack
  - environment
  - usually hard to tailor for special needs
- and more ...

- ***High learning curve***
- ***Hard to adapt online instructions (e.g. for containers, installs, etc)***



Bridges supercomputer at the Pittsburgh Supercomputing Center

Science Gateways

Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines.

From Science Gateway Community Institute the lead provider of resources, services, experts, and ideas for creating and sustaining science gateways

SGCI Science Gateways Community Institute

XSEDE Science Gateways: https://www.xsede.org/gateways-listing .
Science Gateway Community Institute: https://sciencegateways.org .

Science Gateways cont'd

- Lower the barrier to entry for science disciplines
- Common platform for collaborative science
- Scientists use the allocation for the science gateway so there is no need to write an allocation proposal

SGCI Science Gateways Community Institute

XSEDE Science Gateways: https://www.xsede.org/gateways-listing .
Science Gateway Community Institute: https://sciencegateways.org .

GISandbox Vision

Interactivity

- User-friendly GUI
- Unified interface for different resources

Scalability

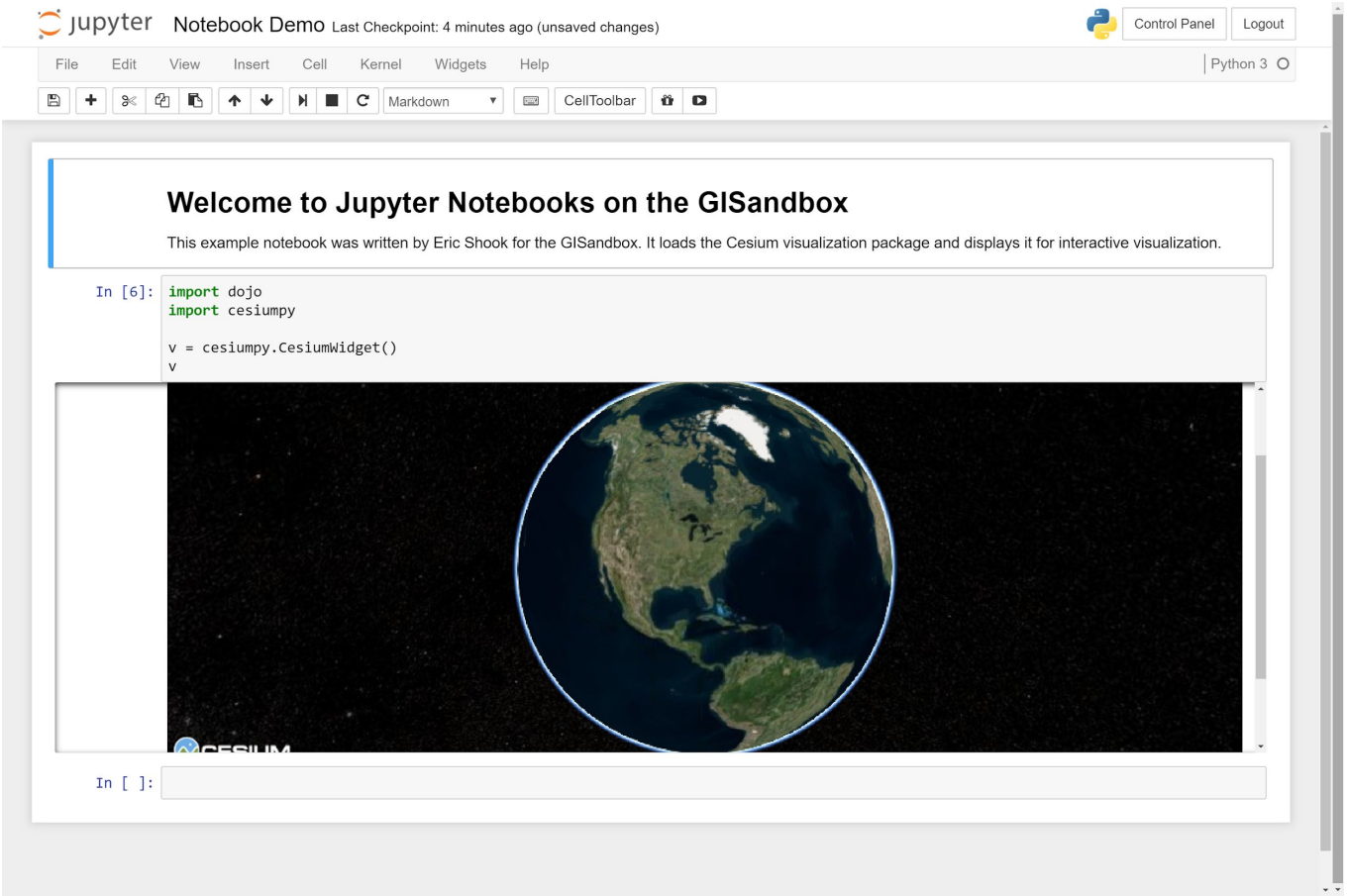
- Datasets much larger than what a laptop or workstation can handle
- Large number of repeated computations with different parameters

Reproducibility

- Sounds obvious for science, but not always obvious in computing

GISandbox

Play place for researchers and educators to learn about, experiment with, and advance geographic information systems and science (gisandbox.org)



GISandbox User Interface (Jupyter Notebooks)

Jetstream

- Jetstream is an NSF-funded (NSF-1445604), user-friendly *cloud environment*
- Led by the Indiana University Pervasive Technology Institute, in collaboration with TACC
- Allocated via XSEDE

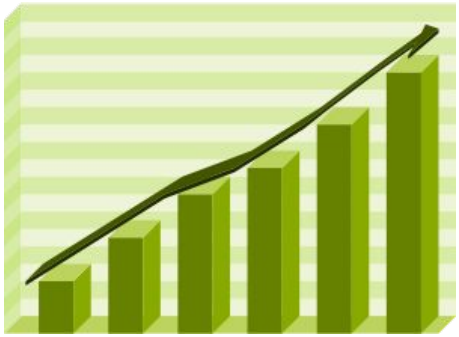
GISandbox Interactivity

- Achieved via Jupyter Notebooks
  - interactive browser-based documents, containing both code, rich text and figures

- Augmented by Jupyter-based access to HPC resources
  - that we have specifically developed for this project

GISandbox Scalability

- Leverage XSEDE HPC resources
- Using container tech to maintain identical environment on different HPC resources
- Data will be already hosted there
- Dramatically increase the number of cores, memory, disk space for scientific applications



GISandbox fostering Reproducibility

- Jupyter Notebooks are at the same time production code and draft journal paper

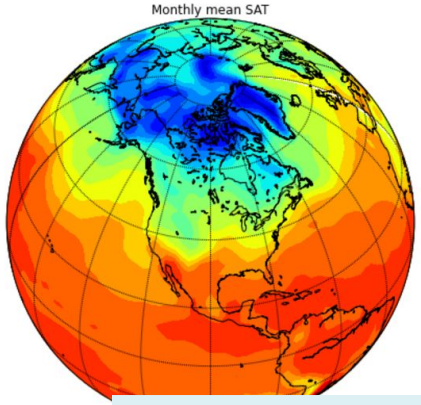
- GISandbox will use container technology to provide a consistent environment across machines and time

- GISandbox itself is reproducible, because it uses ansible to ensure consistency of installation and deployment

GISandbox Architecture (10,000 foot view)

While the rest of the code might be the same:

```
In [22]: fig = plt.figure(figsize=(15,7))
         #m.fillcontinents(color='gray',lake_color='gray')
         m.drawcoastlines()
         m.drawparallels(np.arange(-80.,81.,20.))
         m.drawmeridians(np.arange(-180.,181.,20.))
         m.drawmapboundary(fill_color='white')
         cs = m.contourf(x,y,air_c[0,:,:],20)
         plt.title('Monthly mean SAT')
Out[22]: <matplotlib.text.Text at 0x1116bd210>
```



Jetstream Cloud Resource

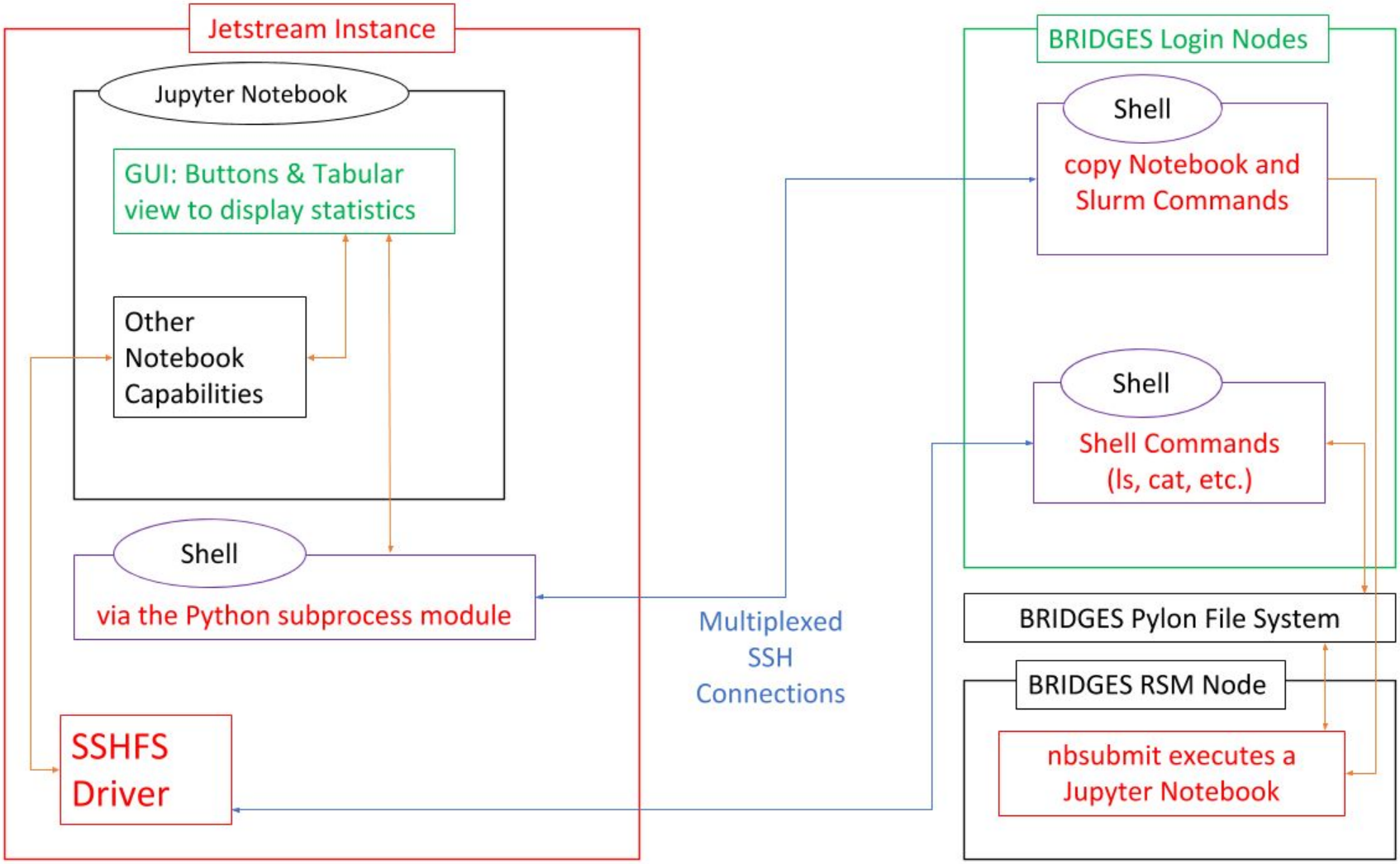
Jupyter "magic" command to run code cells on Comet or Bridges supercomputrs

Comet supercomputer at the San Diego Supercomputing Center
Source: http://ucsdnews.ucsd.edu

Bridges supercomputer at the Pittsburgh Supercomputing Center
Source: http://insidehpc.com

Credits: https://github.com/koldunovn/python_for_geosciences

GISandbox Architecture & Implementation



GISandbox Architecture & Implementation

Under the hood

- JupyterHub for access
- CILogon authentication via XSEDE credentials
- Jupyter Notebook running on Jetstream
- Remote mounting of supercomputer filesystem
- Remote launch of jobs to the supercomputer
- Remote monitor of jobs in the queue

GISandbox Architecture & Implementation

Dashboard

- Authentication via XSEDE credentials
- Unified Jupyter Notebook interface for:
  - small datasets & jobs (jetstream)
  - large datasets & jobs (supercomputer)

- Many Python and R libs GIS researchers need:
  - rgdal, raster, dismo, sp sdm, ….
  - gdal, rasterio, scipy, pandas, fiona, cesiumpy, ...
- Other not-notebook software with hooks to use it
  - NetLogo (agent-based modeling environment)
  - Esri ArcGIS Enterprise *(* Still in testing and piloting mode)*

Challenges

- Local Notebook performance on remote data
  - XSEDE resources connected with hi bw and low latency
  - may need to educate users about this
- Myriad of tools and dataset
  - to address the varied needs of different researchers
- Consistent deployment and maintenance
  - using ansible to fully automate the process
- Long term sustainability, exploring options
  - long term system and security administration
  - long term allocation on various resources

Future work (in progress)

- Complete the security audit & implement its recommendations
- Complete the containerization of remote jobs
- Complete the integration with NetLogo
- Complete legal and technical details for ArcGIS
- Work with partners to use and test the system
- Some partners are building curriculum material for using GISandbox in class and lab projects
- Improve usability and utility of the system
- Transition from prototype to production

Acknowledgements

- XSEDE for the allocation of resources
  - on Jetstream, Comet, Bridges
  - NSF ACI-1548562

- The Science Gateway Community Institute
  - Especially Mark Krenz for conducting a cybersecurity audit about the design and implementation of GISandbox

- Sergiu Sanielevici and Nancy Wilkins-Diehr
  - for their guidance during the initial formulation of the GISandbox

Thank You

Questions?

Davide Del Vento <ddvento@ucar.edu>

Additional slides for discussion

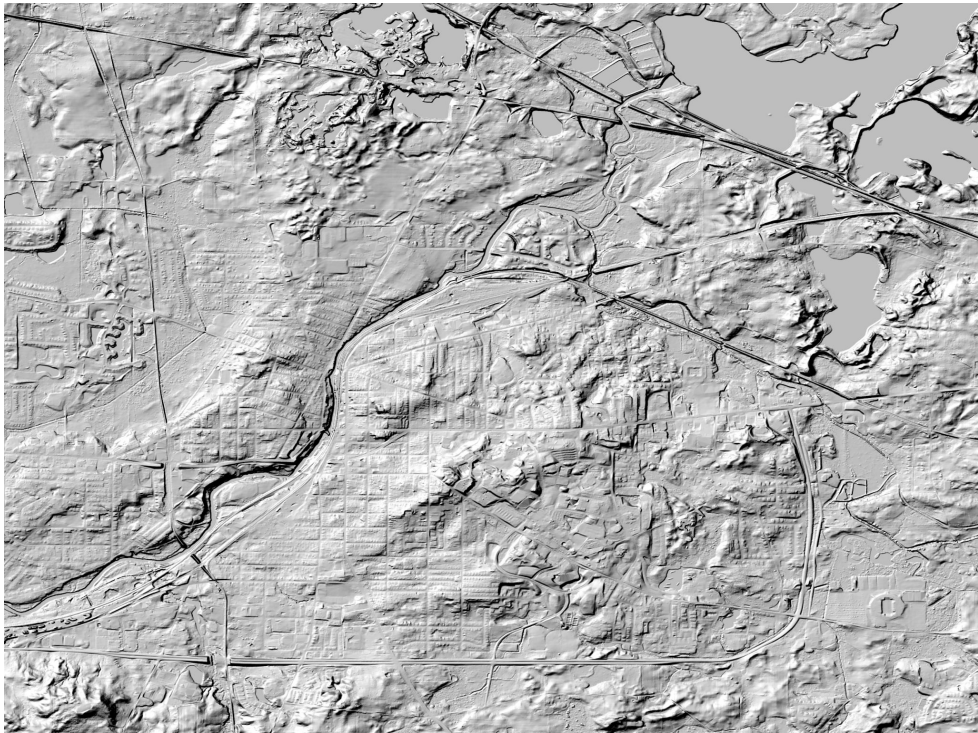
Ansible

- Ansible is software that automates software provisioning, configuration management, and application deployment
- Think of it as a way to “script” everything that has ever been done on a server (e.g. installing sw, adding users, changing configuration files, etc)
- But with powerful features such as error management, idempotentence, declarative syntax, etc

Geospatial Computing and Geographic Information Science and Systems (GIS)

Geographic Information Science (GIScience) studies the science and theory behind geographic information systems, which visualize, analyze, and manipulate geographic information

Geospatial Computing is situated at the intersection of GIScience and computational science



HillShade operation applied to Digital Elevation Model

Agent-based Models (ABMs)

“ABMs are artificial societies: every single person (or ‘agent’) is represented as a distinct software individual. The computer model tracks each agent, ‘her’ contacts and her health status as she moves about virtual space — travelling to and from work, for instance... Agents can be made to behave something like real people: prone to error, bias, fear.”

(Epstein, 2009) (page 687)