Introduction to CyberGIS

Dr. Junjun Yin

CyberGIS Center for Advanced Digital and Spatial Studies
Department of Geography and Geographic Information Science
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign, IL, 61801, USA

January 19, 2016

GEOG 479: Advanced Topics in GIS – CyberGIS

- Course Goals
  - Be familiar with the high-performance computing components in CyberGIS
  - Leverage the CyberGIS architecture and resources to facilitate spatial thinking and knowledge discovery from geospatial Big Data.
  - Interact with geospatial Big Data in various forms (e.g., geo-located Twitter data, Taxi trajectory data and national DEM, etc.)
  - Solve geospatial problems with high-performance computing approaches.
- Course Syllabus
  - https://bitbucket.org/geog479/course/downloads/Syllabus.pdf
- Course materials
  - https://bitbucket.org/geog479/

GEOG 479: Advanced Topics in GIS – CyberGIS

- Course Outline
  - Introduction to CyberGIS, Geospatial Big Data and ROGER
  - High performance computing, cyberinfrastructure and GIS
  - Getting started with ROGER
  - Distributed Spatial Databases
  - Taming Big Geospatial Data with Hadoop
  - Advanced Geospatial Data Analytics
  - Mid-term project
  - Terrain analysis: Parallel MapAlgebra for Spatial Analysis
  - Interactive Geo-visualization of Large-scale Movement Data with Spark
  - Final project

What is CyberGIS?

Cyber + GIS ?

What is GIS?

- Systems
- Science
- Services
- Society
- Synthesis
- Spatial is special

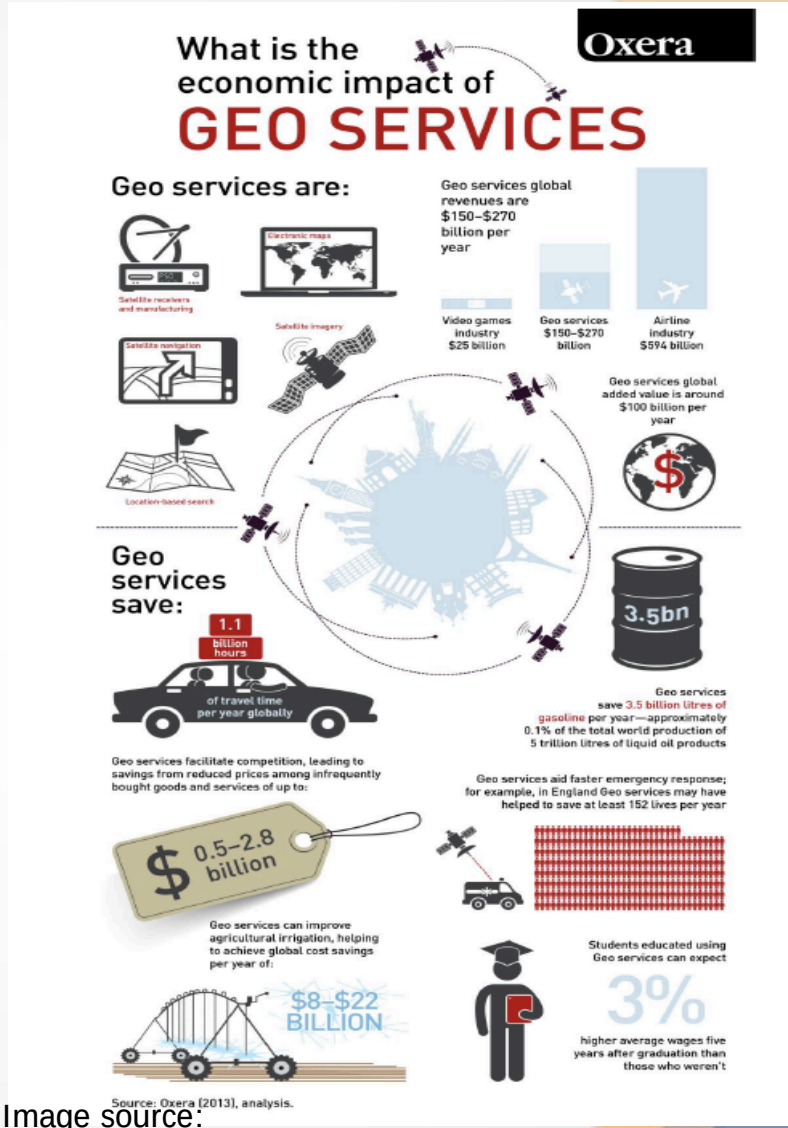


Image source: http://www.oxera.com/Oxera/media/Oxera/downloads/reports/What-is-the-economic-impact-of-Geo-services_1.pdf

What is CyberGIS?

- CyberGIS is the synthesis of cyberinfrastructure, GIS and spatial analysis

- Why CyberGIS?
  - Geospatial discovery and innovation with advanced digital technologies[1]

1. Wang, S. (2013)

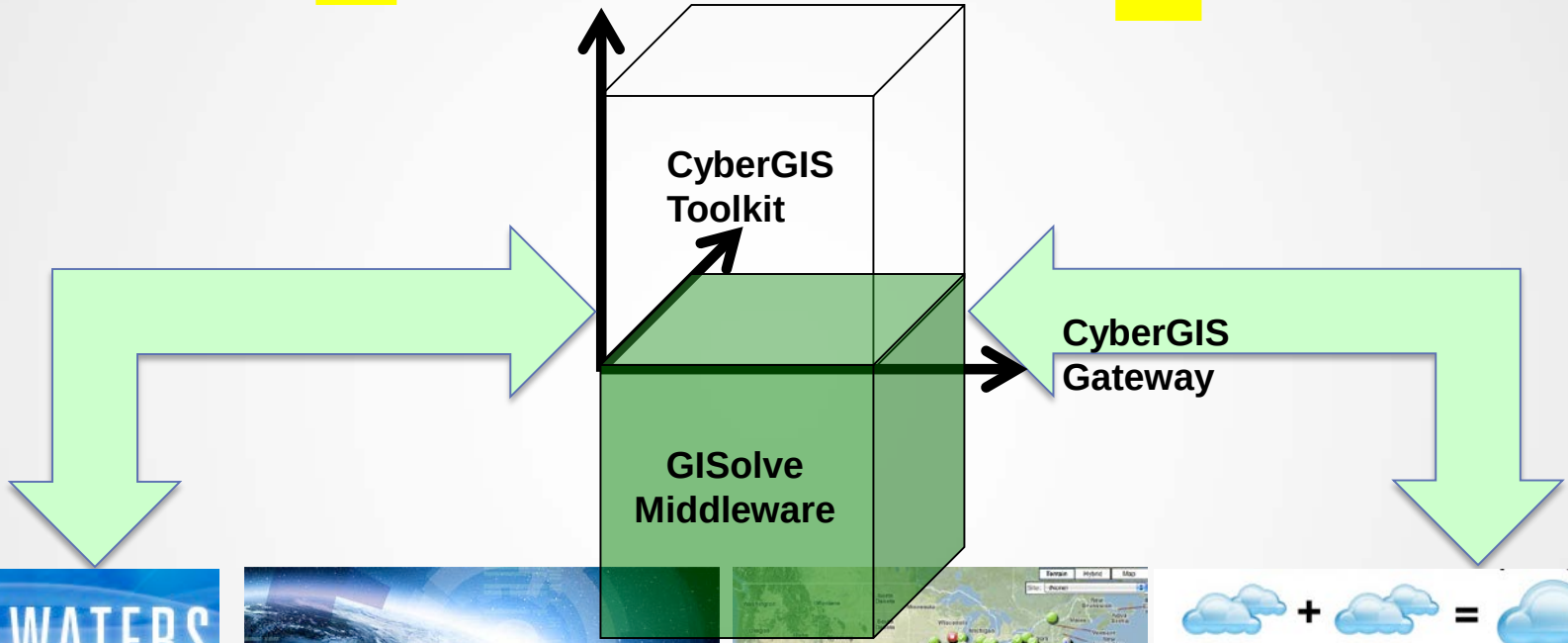
cyberinfrastructure

CyberGIS is Geographic Information Science and Systems based on advanced cyberinfrastructure.

Cyberinfrastructure includes:

- (high performance) computing systems
- data storage systems
- advanced instruments
- data repositories
- visualization environments
- people
- linked by high speed networks

Advanced Digital and Spatial Studies



BLUE WATERS
SUSTAINED PETASCALE COMPUTING

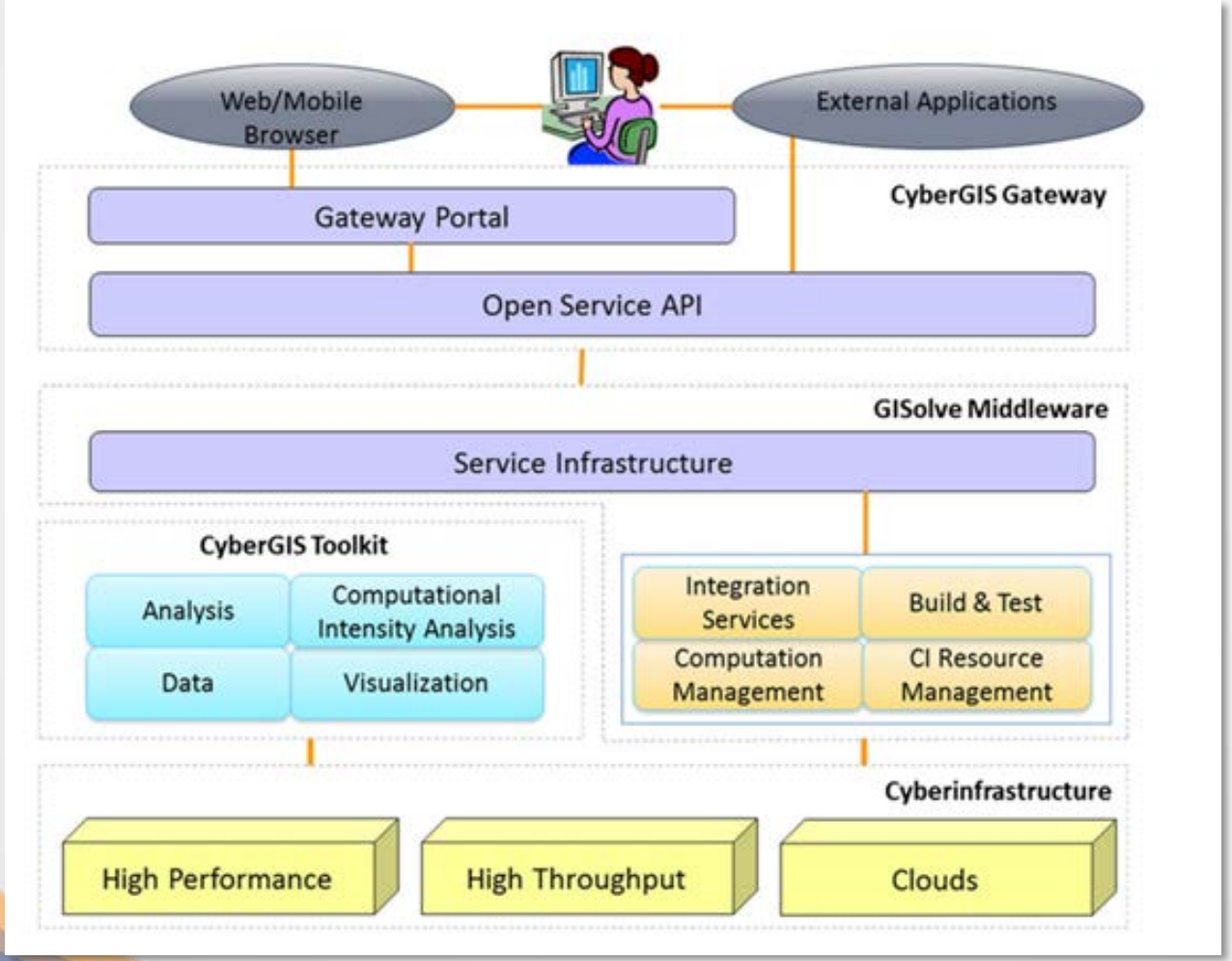
www.ncsa.illinois.edu/BlueWaters/

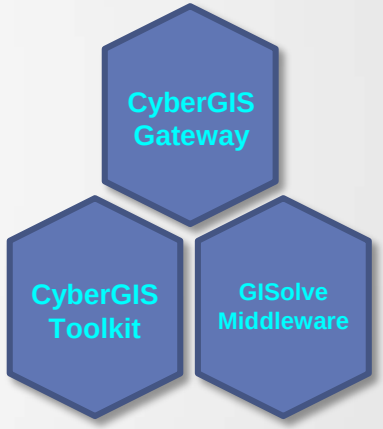
www.xsede.org

www.opensciencegrid.org

http://lakjeewa.blogspot.com/2011/09/what-is-cloud-computing.html

CyberGIS Software Environment





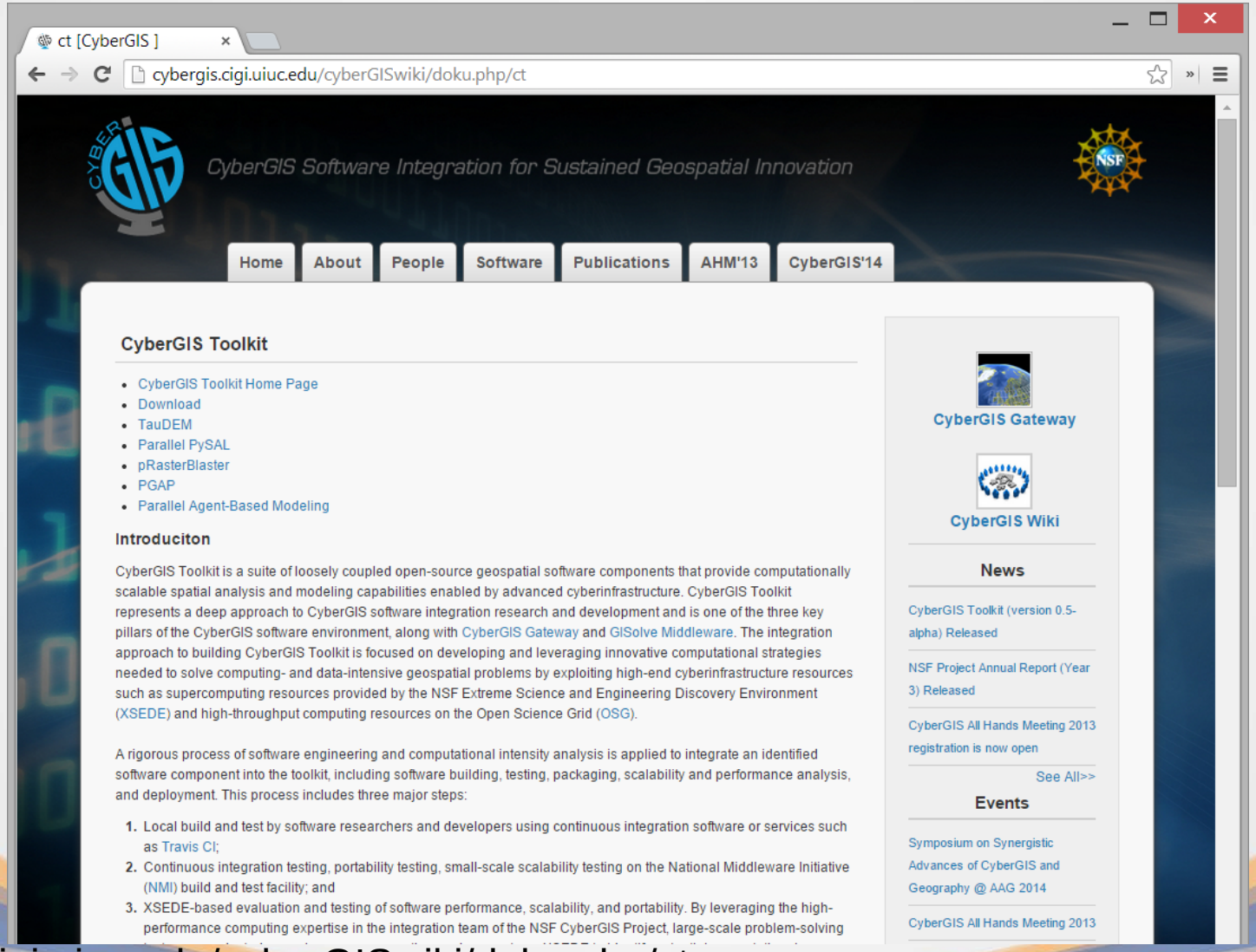
CyberGIS Gateway

GISolve Middleware

GISolve Open Service API

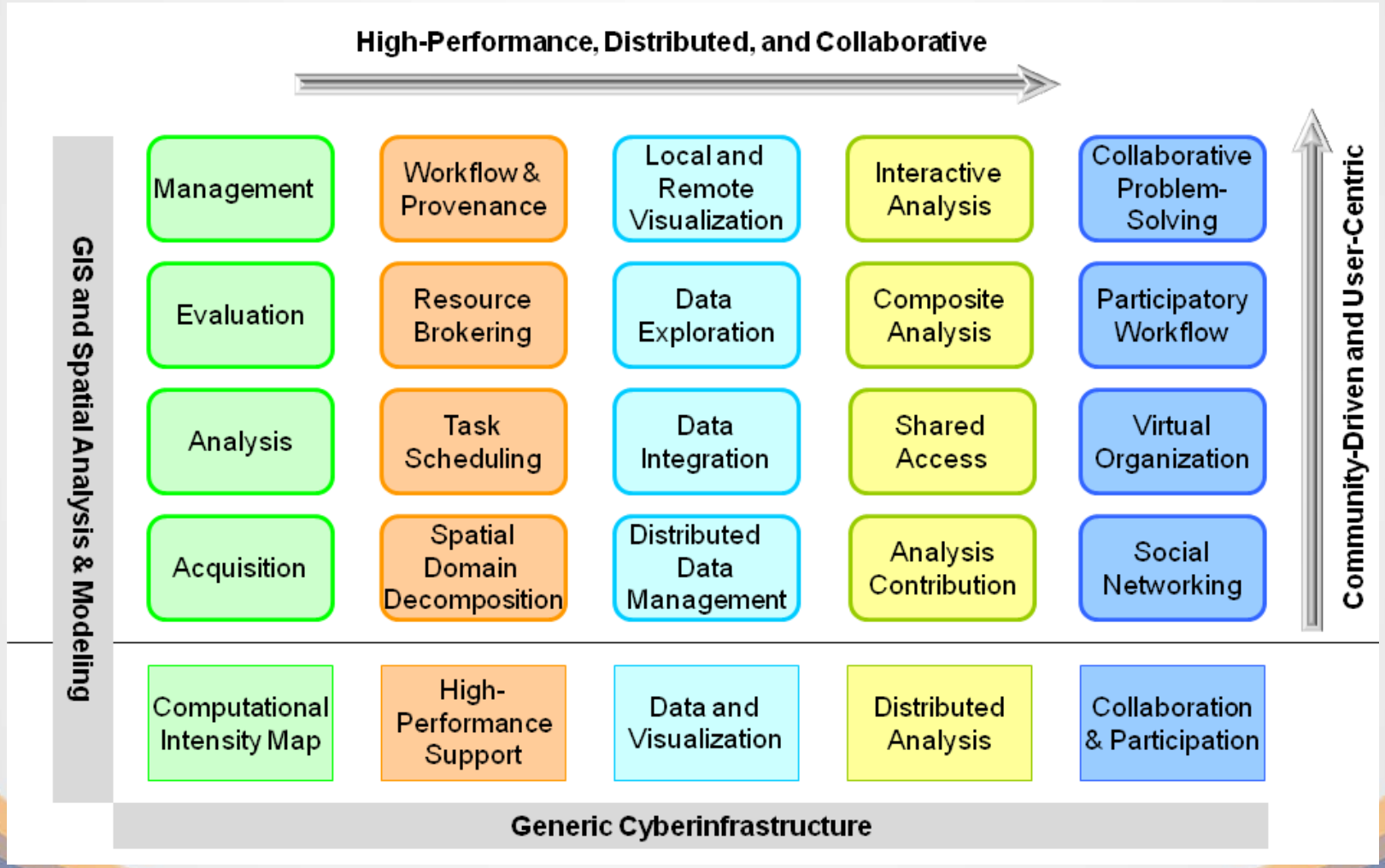
www.cigi.illinois.edu/dokuwiki/doku.php/projects/gisolve/index

CyberGIS Toolkit



cybergis.cigi.uiuc.edu/cyberGISwiki/doku.php/ct

CyberGIS Toolkit



CyberGIS architecture
(source: Wang et al., 2013)

Terminologies

- Cloud computing
- High performance computing
- High-throughput computing
- Distributed computing
  - Hadoop and Spark
  - Non-relational database (NoSQL)
  - Distributed database
- Parallel computing

- Computational geography
- WebGIS

CyberGIS Center for Advanced Digital and Spatial Studies

Empower advanced digital and spatial studies through innovation of CyberGIS technologies and applications

Research Application Focus Areas:

Agriculture and Food
Disaster and Emergency
Earth and Environment
Energy and Water Resources
Health and Wellness

CyberGIS Gateway and ArcGIS Online Integration

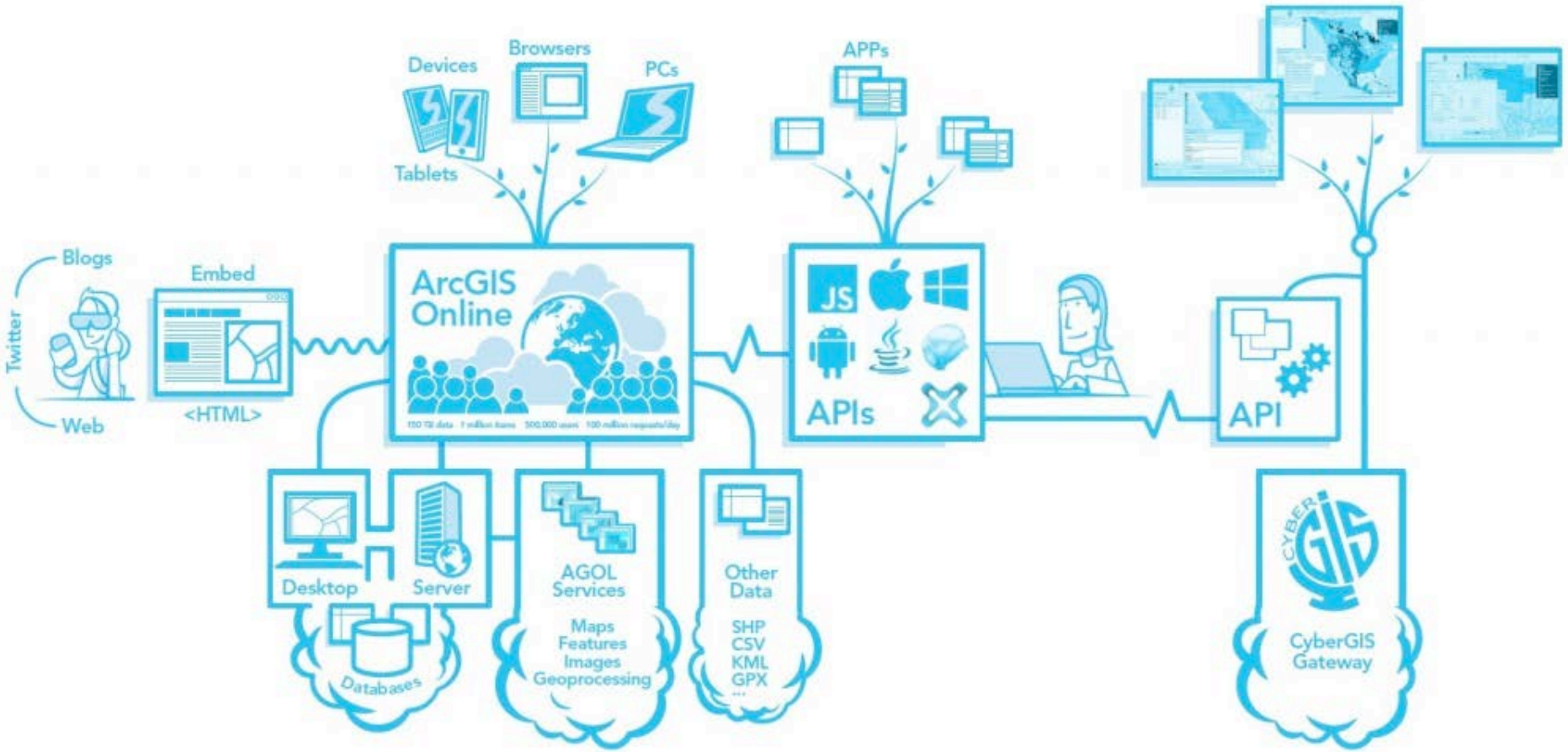
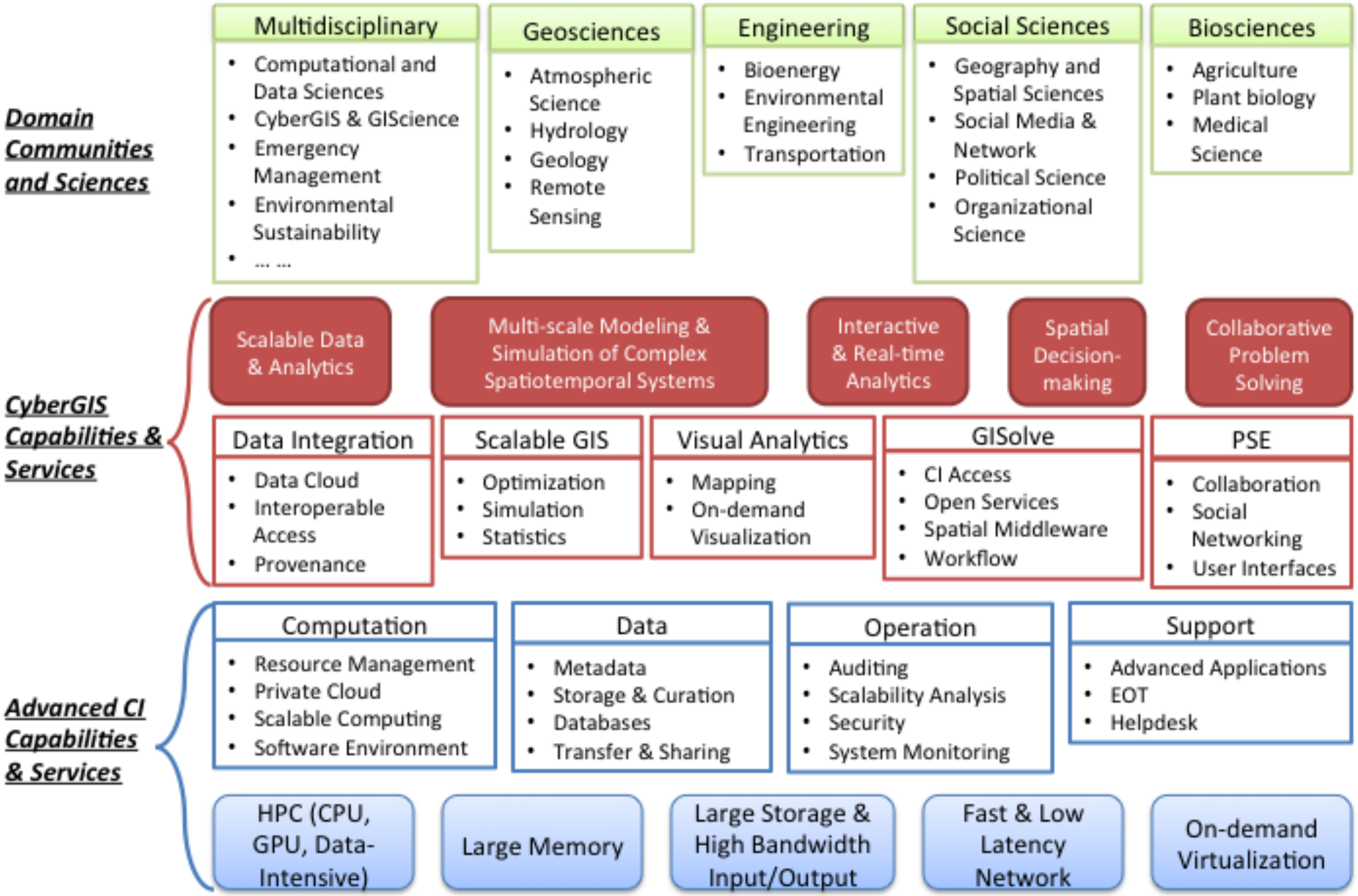


Image source: http://blogs.esri.com/esri/arcgis/2013/10/01/what-is-cybergis/



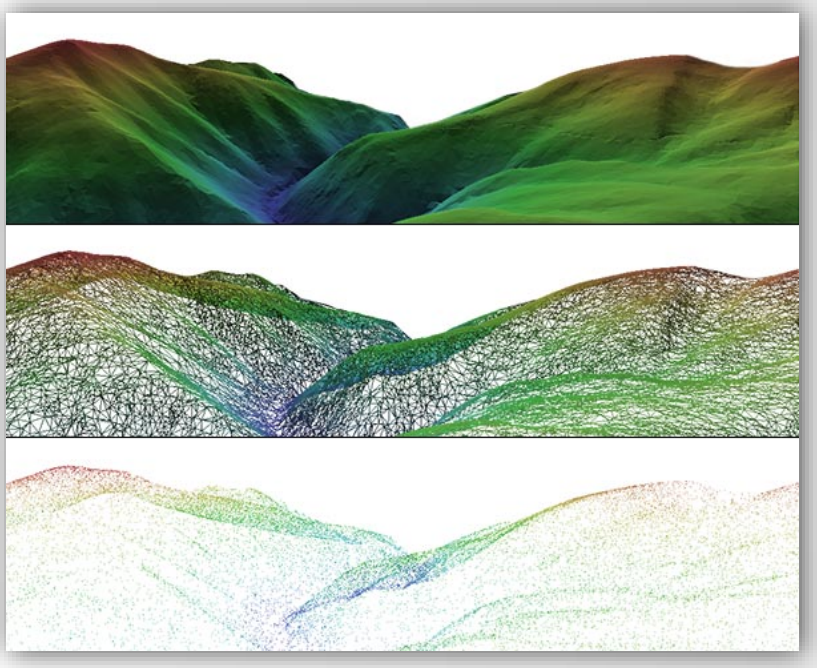
* EOT – Education, Outreach, and Training; HPC – High Performance Computing; PSE – Problem-Solving Environment

Education, Outreach, and Training Activities

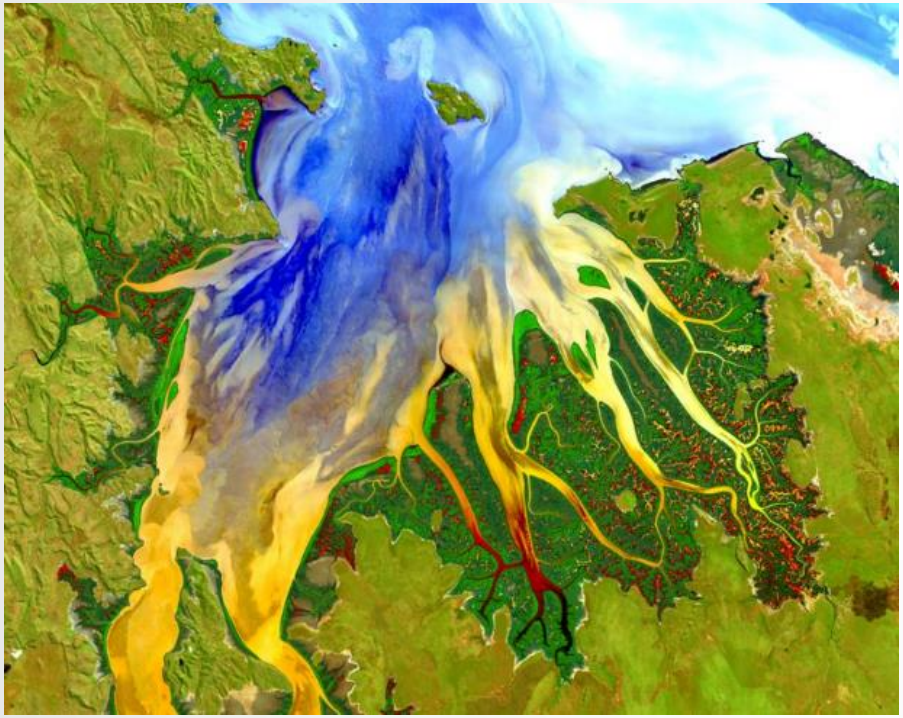
- CyberGIS Commons
  - For hands-on learning and training
- CyberGIS Collaboration Space
  - For groups to develop collaborative research
- CyberGIS Helpdesk
  - For in-depth technical consulting
- CyberGIS Brown Bag Seminars
  - Once every two to three weeks on average
- CyberGIS Distinguished Speaker Series
  - Once every two months on average
- CyberGIS MOOCs
- CyberGIS Webinars

Geospatial Big Data

- Volume
  - USGS NED 30m(175GB), 10m(500GB)
- Variety
  - Forms and formats (Raster and Vector)
  - Shapefile, GeoTiff, GeoJSON, WMS, WFS, NetCDF, HDF5
- Velocity
  - All Geocoded Tweets
    - 2.5kb/tweet
    - 6,000 average tweet per second
    - 3% of tweets are geo-located
    - 38GB per day
  - Landsat 8
    - 650GB per day
- Veracity
  - Uncertainty



Geospatial Big Data



Landsat 8

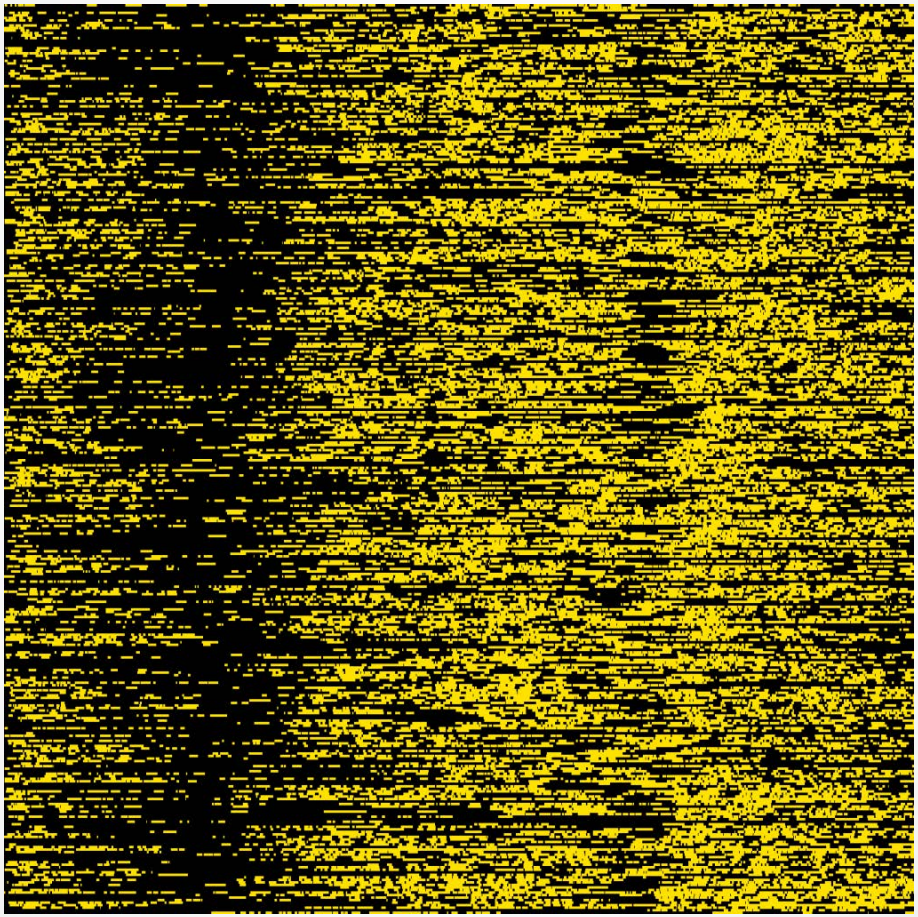
- ~650 images per day
- ~1GB each (compressed)
- ~650gb per day

Twitter Firehose

All Geocoded Tweets

- ~2.5KB / tweet
- ~6,000 average tweet per second
- ~3% of tweets are geolocated
- ~38GB per day

Geospatial Big Data



Visualization of New York taxi data
The black pixel means the taxi did not have fare and yellows means it did
Source: http://chriswhong.com/open-data/my-first-data-art-nyc-taxi-defrag/

CyberInfrastructure Resources

NSF Blue Waters @ UofI
13,300 TFlop/s

NSF XSEDE: SDSC Comet
2,000 TFlop/s

UofI Campus Cluster

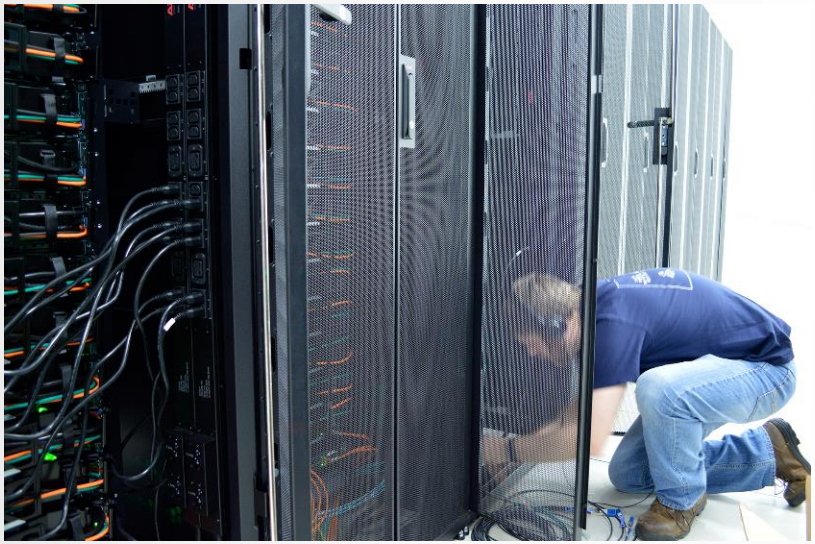
CyberGIS ROGER

CyberGIS ROGER
Resourcing Open Geospatial Education and Research
After Roger Tomlinson
goo.gl/8MpYC9
~60 TFlop/s

Why ROGER?

- Both a research project and a research resource
- Configured to best support geospatial data, with emphases on local memory and shared storage size and speed
- Supports multiple paradigms: traditional batch HPC, Hadoop, and Cloud (OpenStack).
- Integrate the three paradigms and leverage their strengths
- Inform the design of future geospatial supercomputers

ROGER: Overview



- Provides the CyberGIS Center with HPC, Hadoop and OpenStack functionality all on one system
- Managed by Systems Group

  With support from our storage, network, security and services groups

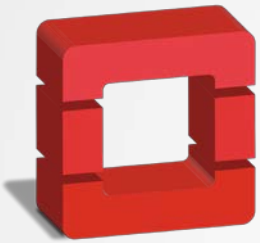
ROGER: File system

- IBM General Parallel File System (GPFS)
- Parallel file system means all of the nodes get fast access to hard drive storage.

ROGER for this class

- Will create ROGER accounts for every student in this class
- Interact with various components in ROGER
  - Job submission
  - Handling and processing massive data with Hadoop, Spark, R, and Python, etc.
  - Getting to know OpenStack
  - Working with Geospatial libraries

ROGER for this class

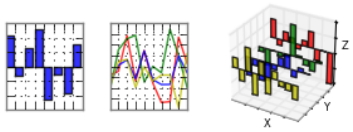


openstack
CLOUD SOFTWARE

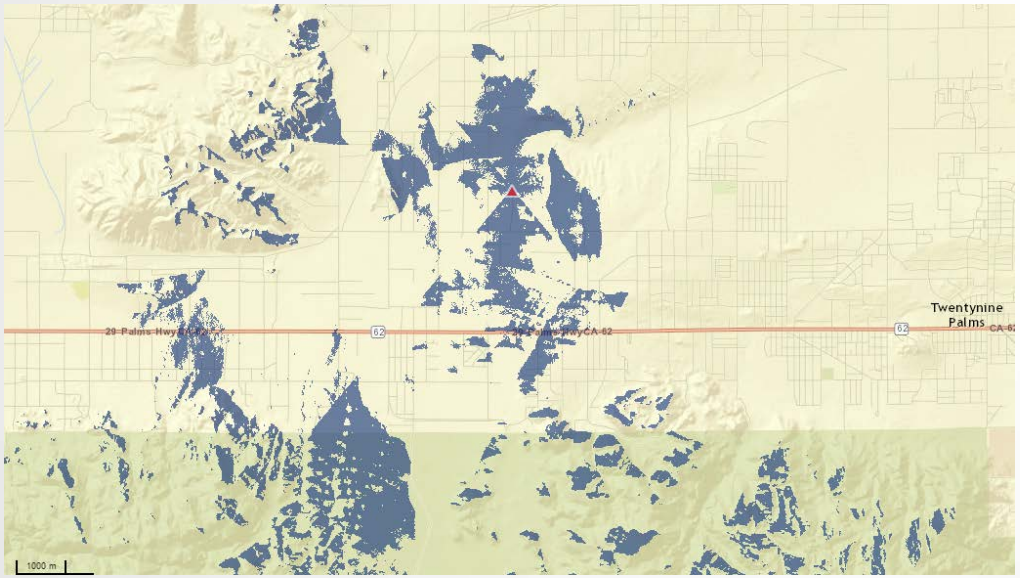
mongoDB

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



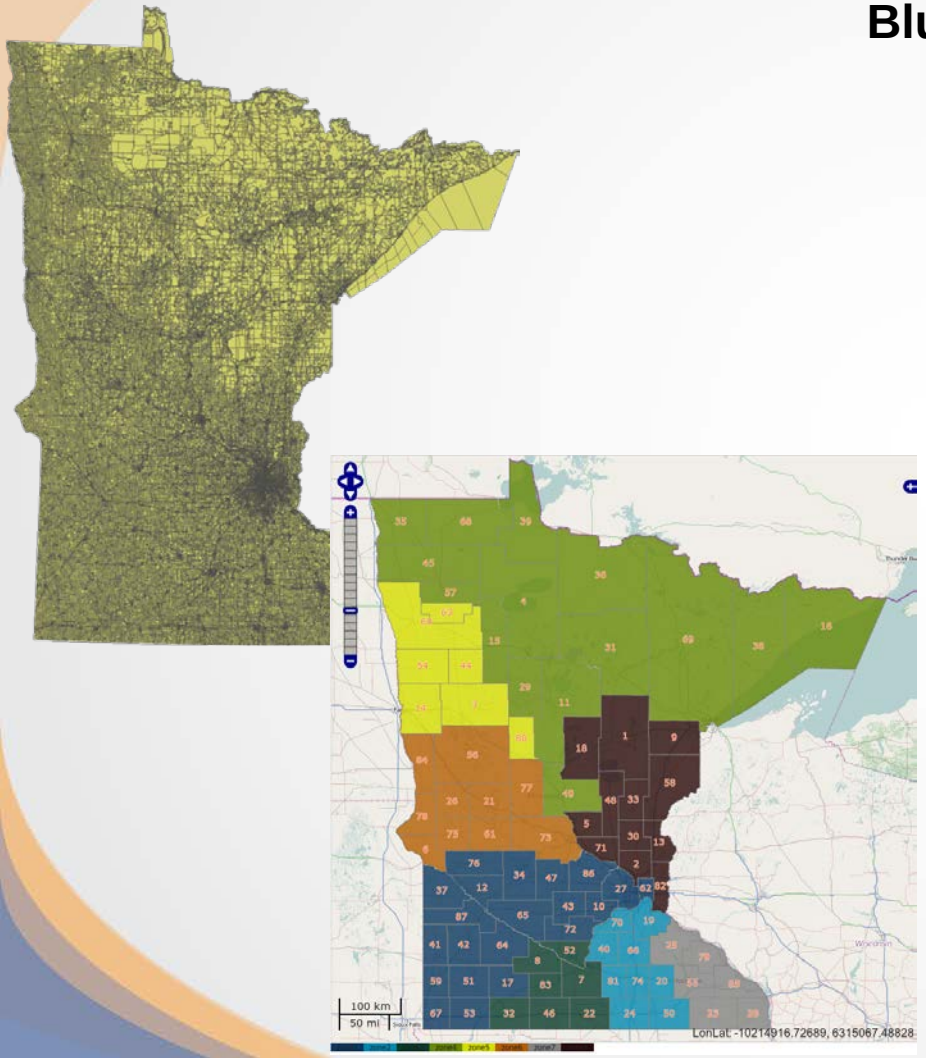
CyberGIS Viewshed App



Viewshed

Uofl Cluster

Parallel Genetic Algorithm for Redistricting Optimization on Blue Waters

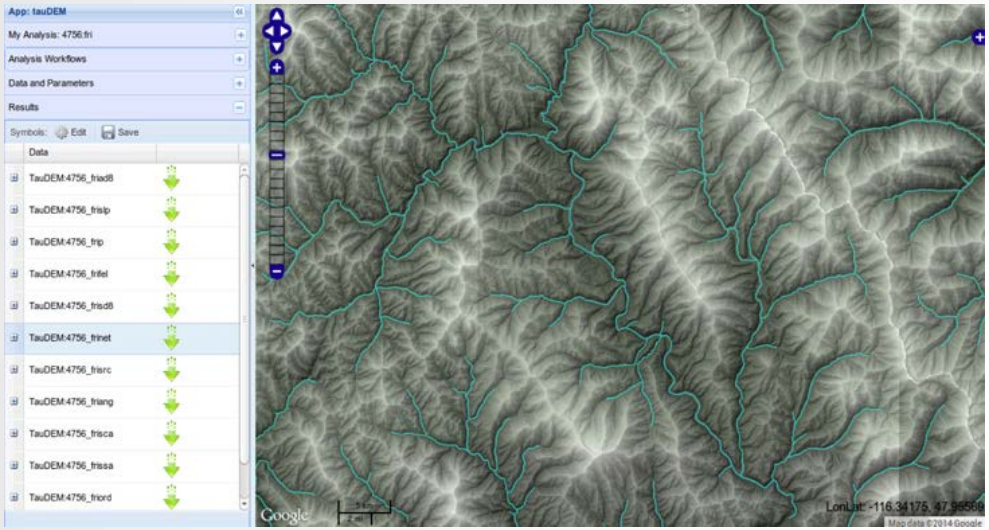


NSF Blue Waters @ UofI

13,300 TFlop/s

Collaborative Work with Wendy K. Tam Cho @ UIUC

CyberGIS TauDEM App



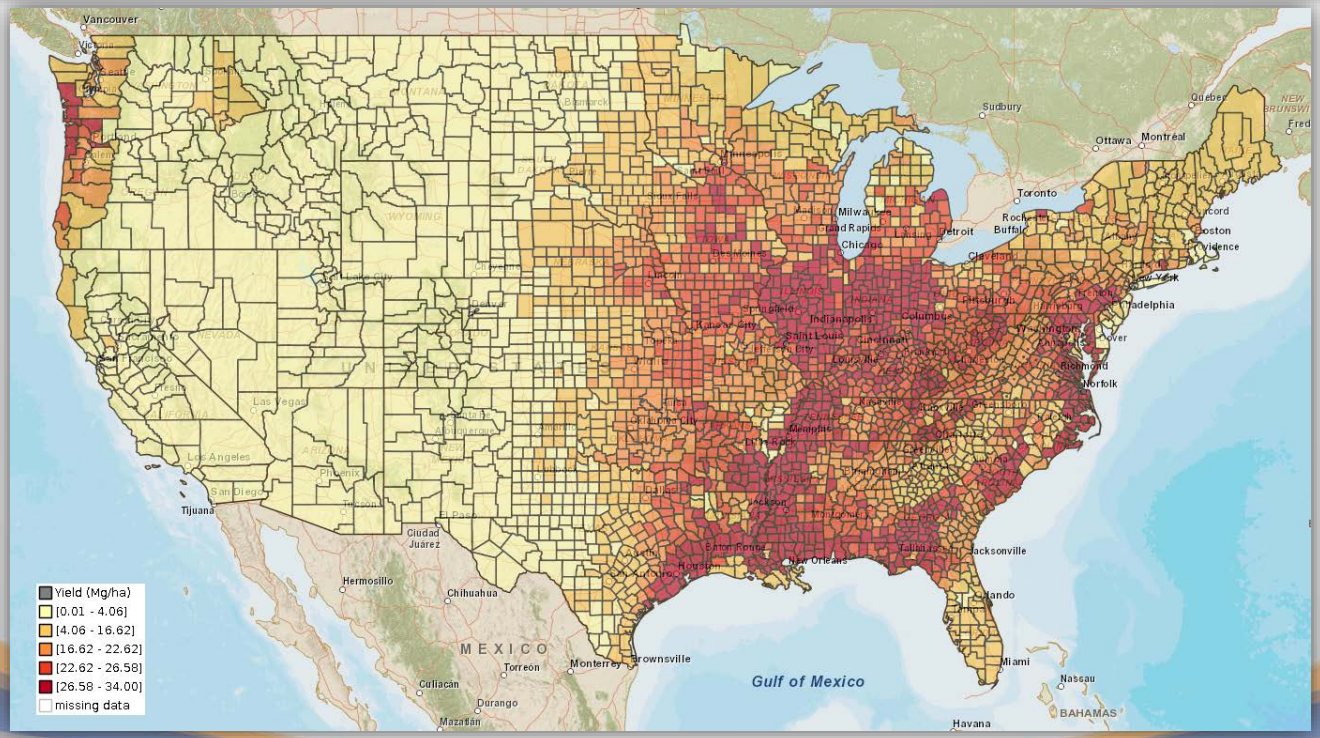
TauDEM

NSF XSEDE: SDSC Trestles

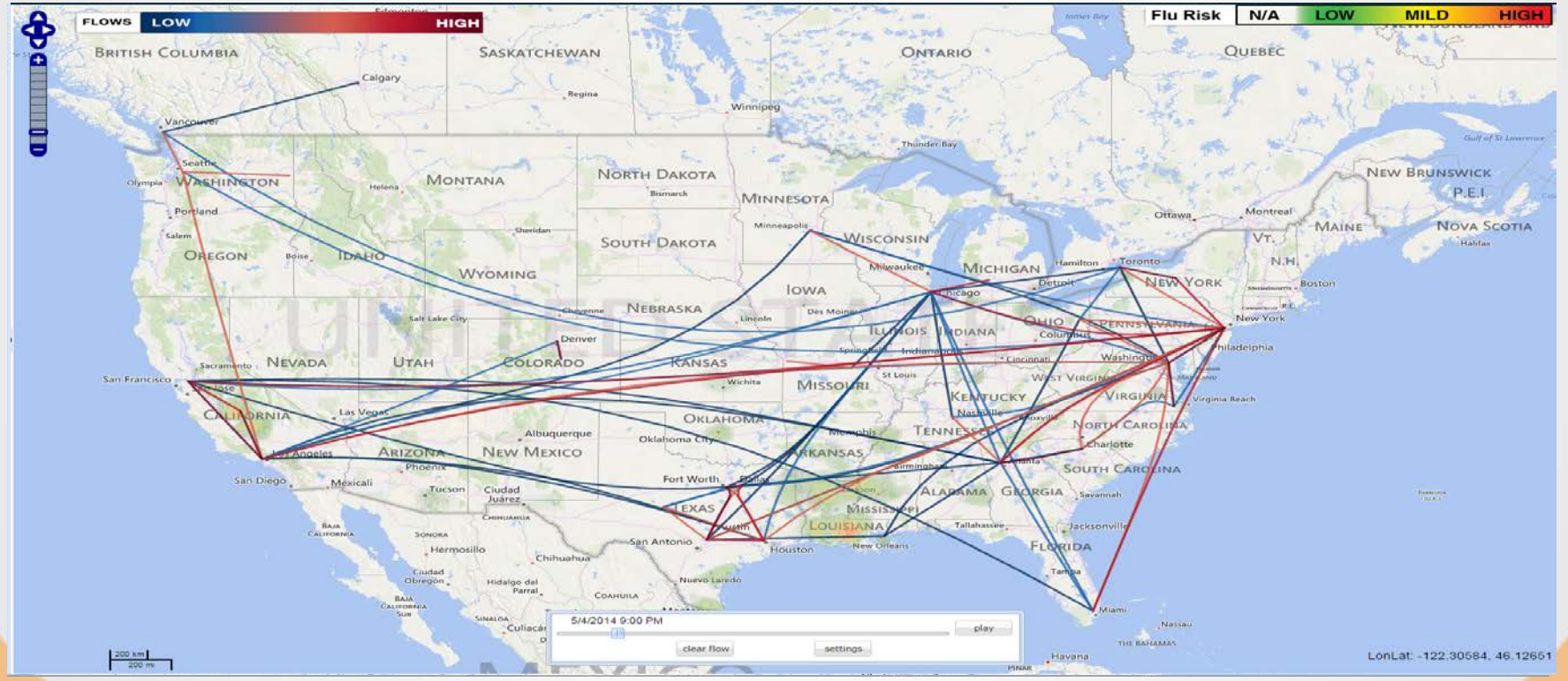
100 TFlop/s

Collaborative Work with David Tarboton @ Utah State University

CyberGIS BioScope App



CyberGIS FluMapper App



Lab

1. FluMapper
   - Social media, point data, streaming data, Hadoop, routing, raster and vector visualization
2. BioScope
   - Agriculture, supply chain optimization, chart generation, reference base maps
3. TauDEM
   - Hydrologic data analysis, digital elevation models, code reuse

Reference

- Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., and Nyerges, T. L. (2013). Cybergis software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science*, 27(11), 2122–2145.
- Wang, S. (2013) "CyberGIS: Blueprint for Integrated and Scalable Geospatial Software Ecosystems." *International Journal of Geographical Information Science*, 27 (11): 2119-2121