Big Data in OpenTopography

Vishu Nandigam

San Diego Supercomputer Center

NSF Big Data in Education Workshop

Jan 28-29 2015

Presentation Overview

- Lidar and OpenTopography
- Data and Workflow
- Cyberinfrastructure
- Data Growth and Challenges
- Data Insights
- Research and Development

LIDAR

- **LI**ght **D**etection **A**nd **R**anging (aka airborne laser swath mapping)
- Billions of of accurate distance measurements with a scanning laser rangefinder + GPS + Inertial Measurement Unit (IMU)

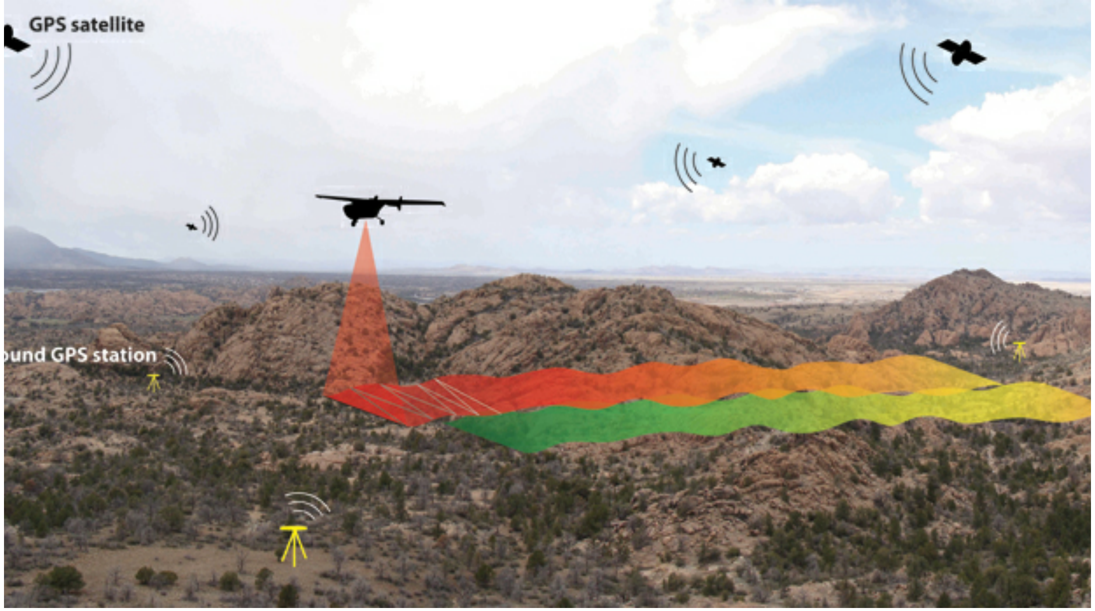
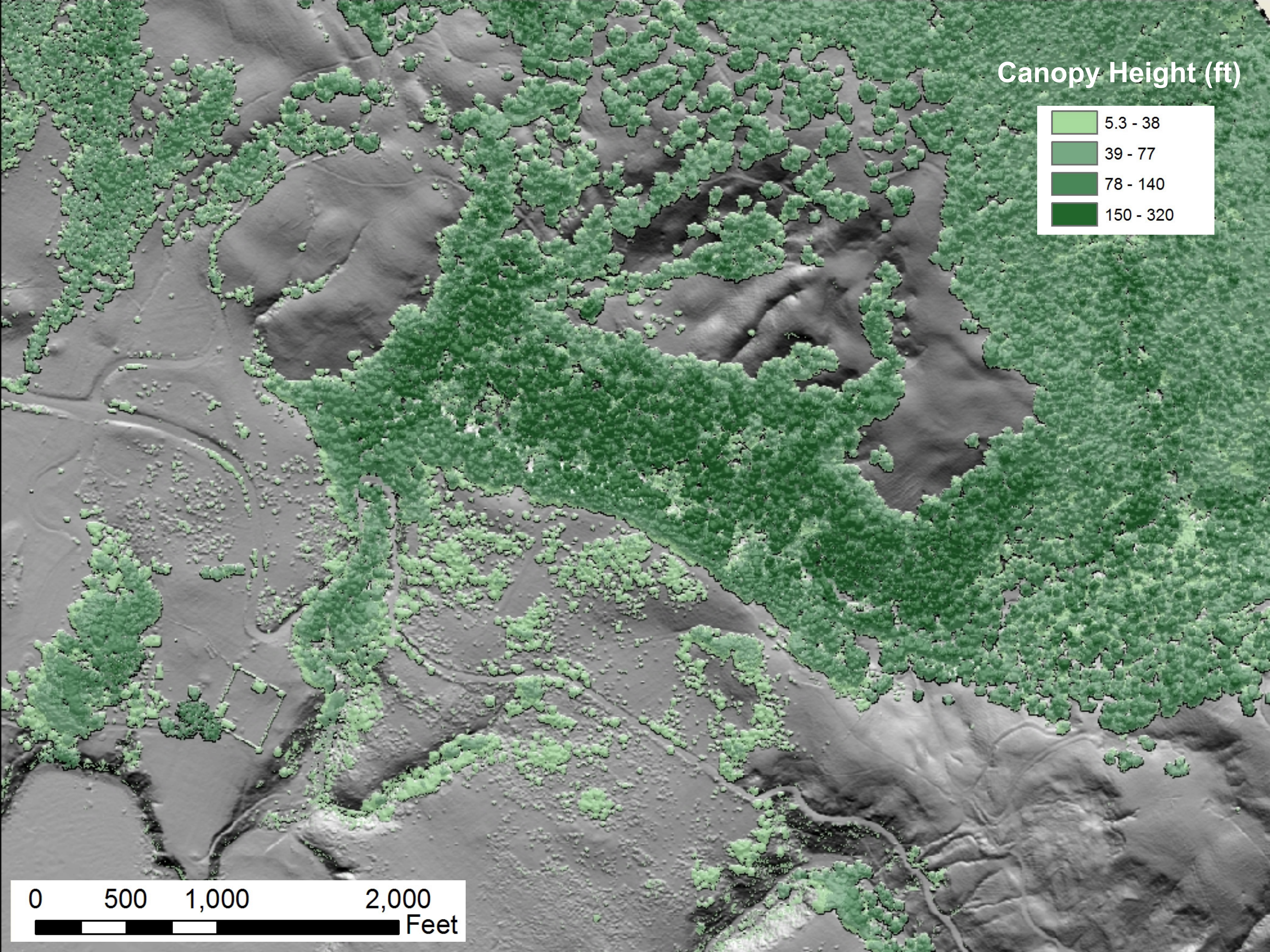


Image: David Haddad, AGS

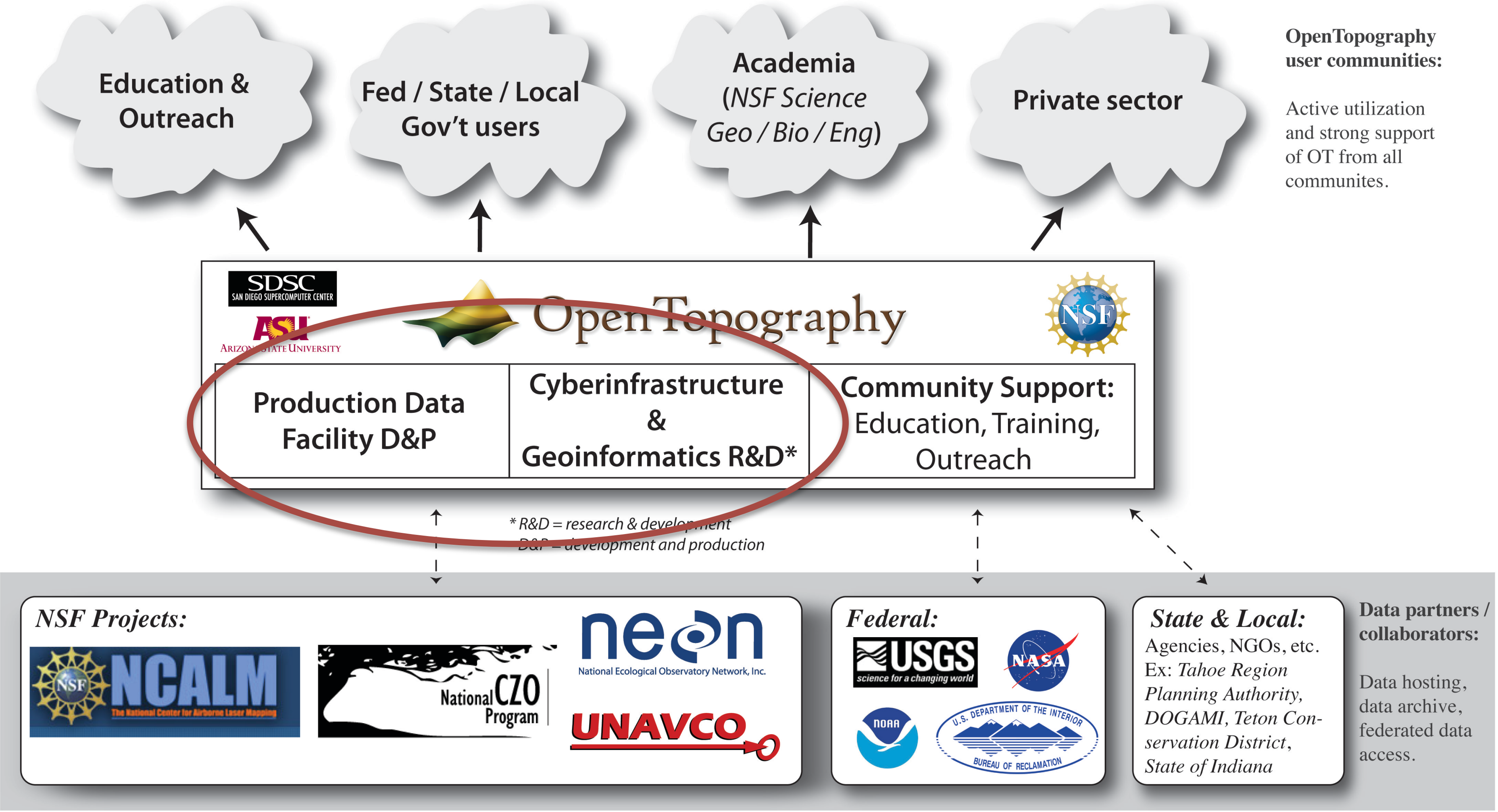
Point cloud (x,y,z coordinates) = fundamental LIDAR data product


Canopy Height (ft)
5.3 - 38
39 - 77
78 - 140
150 - 320
0
500
1,000
2,000
Feet

OpenTopography

- NSF Earth Science Facility: 3 year support in 2009. Renewed in 2012
  (Award No. 1226353 & 1225810 EAR/IF)
- **CI and Science Collaboration**
  - SDSC, ASU & UNAVCO
- Related research efforts
  - NASA ROSES: Extend to Satellite-based LIDAR (waveform data)
  - NSF SI2 CyberGIS: OpenTopo as an exemplar of cyber GIS
  - NSF CluE: Investigate Computer Science issues in big data
- Partnerships with state and local agencies to support data hosting and processing capabilities

OpenTopography Facility Overview



Education & Outreach

Fed / State / Local Gov't users

Academia (*NSF Science Geo / Bio / Eng*)

Private sector

OpenTopography user communities:

Active utilization and strong support of OT from all communites.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

ASU ARIZONA STATE UNIVERSITY

OpenTopography

NSF

Production Data Facility D&P

Cyberinfrastructure & Geoinformatics R&D*

Community Support: Education, Training, Outreach

** R&D = research & development*
D&P = development and production

NSF Projects:

NCALM The National Center for Airborne Laser Mapping

National CZO Program

neon National Ecological Observatory Network, Inc.

UNAVCO

Federal:

USGS science for a changing world

NASA

NOAA

U.S. DEPARTMENT OF THE INTERIOR BUREAU OF RECLAMATION

State & Local:

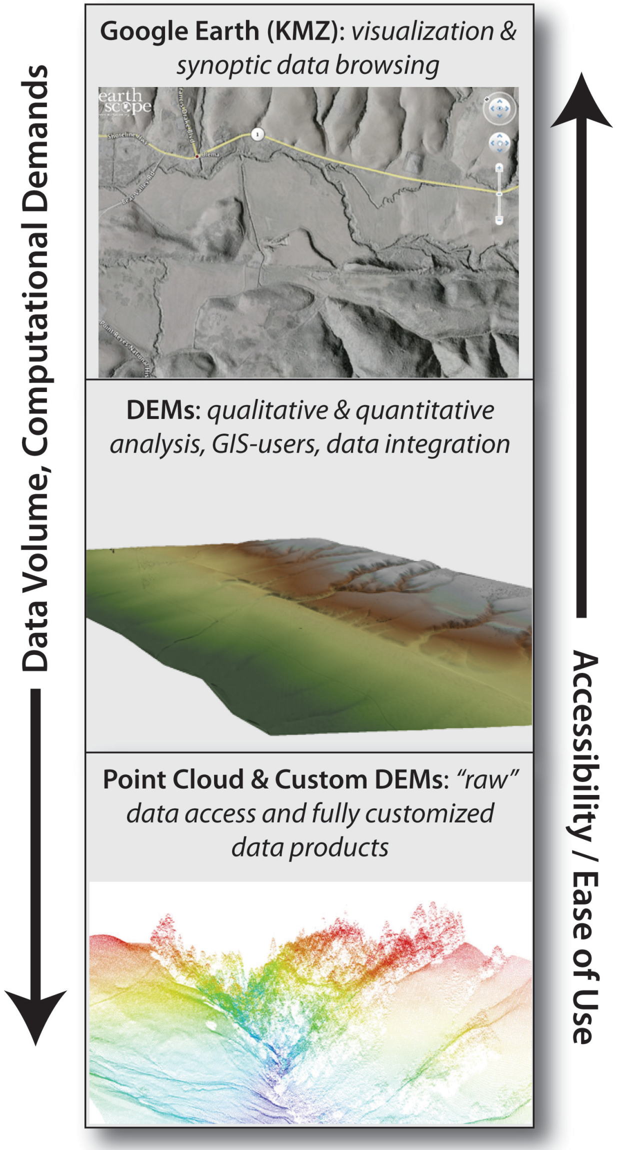
Agencies, NGOs, etc.
Ex: *Tahoe Region Planning Authority, DOGAMI, Teton Conservation District, State of Indiana*

Data partners / collaborators:

Data hosting, data archive, federated data access.

OpenTopography Data

OpenTopography
Multi-Tiered Data Products



Google Earth (KMZ): *visualization & synoptic data browsing*

DEMs: *qualitative & quantitative analysis, GIS-users, data integration*

Point Cloud & Custom DEMs: *"raw" data access and fully customized data products*

Data Volume, Computational Demands

Accessibility / Ease of Use

- Large user community with variable needs and levels of sophistication.
- Goal: maximize access to data to achieve greatest scientific impact.
- Big data
  – treat data as an asset that can be used and reused
  – Co-locate data with on-demand processing

Data Workflow

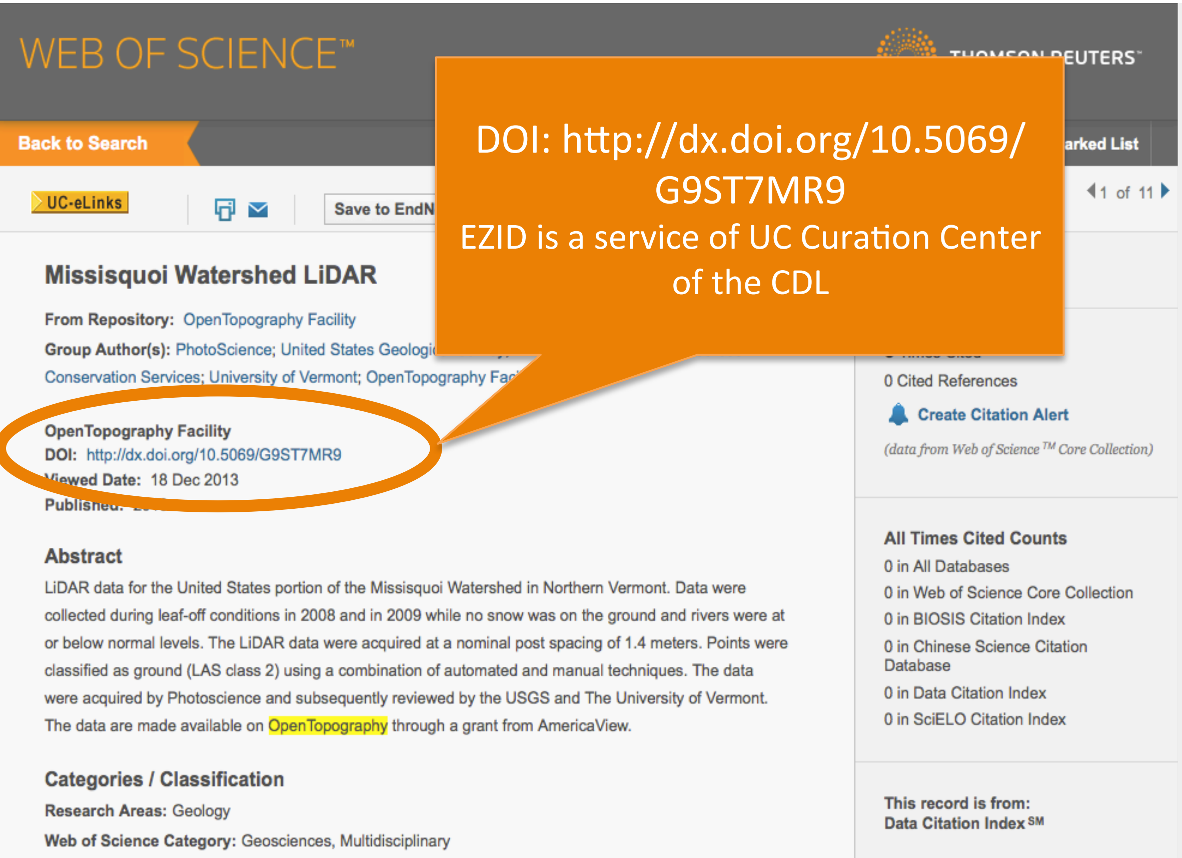
1. Original Source Data from Collector (eg. NCALM)
2. Extract relevant data products
3. Source data is archived in Chronopolis digital preservation network
    1. UCSD Library (Library of Congress)
    2. Three geographically distributed copies of the data
4. Extracted data products go through QA/QC
5. Data transformation and optimization
    1. Error correction
    2. Projection conversion
6. Update Metadata ISO 19115 (Data)
7. Generate additional derived products (e.g. GE hillshades)

Data available via OT

Catalog Service for the Web / DOI

(8 &9 of the Data Workflow)

- CSW Catalog – ESRI Geoportal Server
- ISO 19115 (Data)
- CZO, CyberGIS, Thomson Reuters Web of Science



Missisquoi Watershed LiDAR

From Repository: OpenTopography Facility

Group Author(s): PhotoScience; United States Geologi
Conservation Services; University of Vermont; OpenTopography Fac

OpenTopography Facility
DOI: http://dx.doi.org/10.5069/G9ST7MR9
Viewed Date: 18 Dec 2013
Published:

Abstract

LiDAR data for the United States portion of the Missisquoi Watershed in Northern Vermont. Data were collected during leaf-off conditions in 2008 and in 2009 while no snow was on the ground and rivers were at or below normal levels. The LiDAR data were acquired at a nominal post spacing of 1.4 meters. Points were classified as ground (LAS class 2) using a combination of automated and manual techniques. The data were acquired by Photoscience and subsequently reviewed by the USGS and The University of Vermont. The data are made available on OpenTopography through a grant from AmericaView.

Categories / Classification

Research Areas: Geology

Web of Science Category: Geosciences, Multidisciplinary

0 Cited References

Create Citation Alert

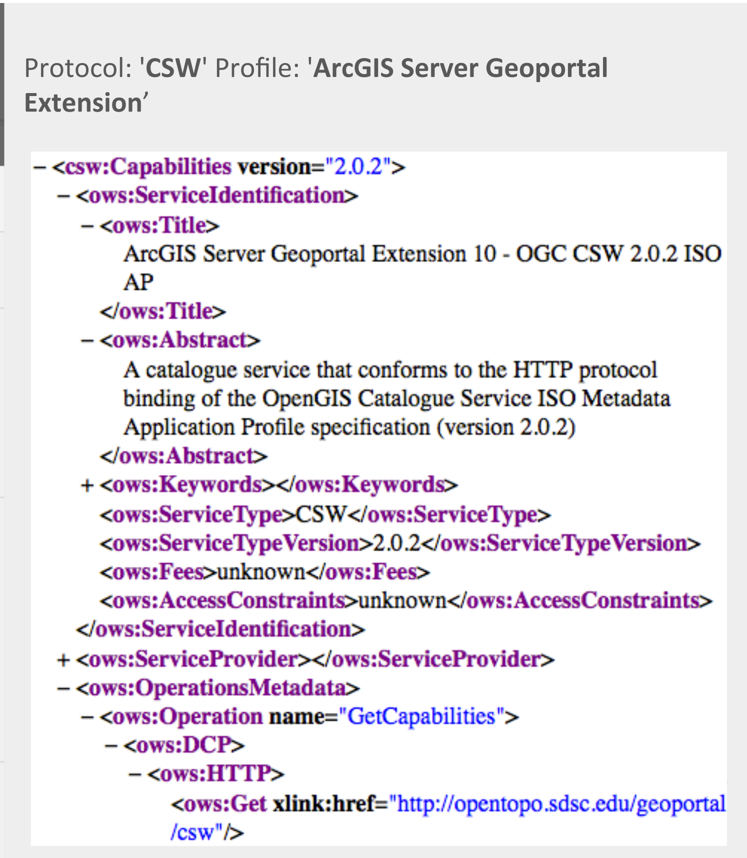
(data from Web of Science ™ Core Collection)

All Times Cited Counts

0 in All Databases
0 in Web of Science Core Collection
0 in BIOSIS Citation Index
0 in Chinese Science Citation Database
0 in Data Citation Index
0 in SciELO Citation Index

This record is from:
Data Citation Index SM

DOI: http://dx.doi.org/10.5069/G9ST7MR9
EZID is a service of UC Curation Center of the CDL



Protocol: '**CSW**' Profile: '**ArcGIS Server Geoportal Extension**'

```
– <csw:Capabilities version="2.0.2">
  – <ows:ServiceIdentification>
    – <ows:Title>
        ArcGIS Server Geoportal Extension 10 - OGC CSW 2.0.2 ISO
        AP
      </ows:Title>
    – <ows:Abstract>
        A catalogue service that conforms to the HTTP protocol
        binding of the OpenGIS Catalogue Service ISO Metadata
        Application Profile specification (version 2.0.2)
      </ows:Abstract>
    + <ows:Keywords></ows:Keywords>
      <ows:ServiceType>CSW</ows:ServiceType>
      <ows:ServiceTypeVersion>2.0.2</ows:ServiceTypeVersion>
      <ows:Fees>unknown</ows:Fees>
      <ows:AccessConstraints>unknown</ows:AccessConstraints>
    </ows:ServiceIdentification>
  + <ows:ServiceProvider></ows:ServiceProvider>
  – <ows:OperationsMetadata>
    – <ows:Operation name="GetCapabilities">
      – <ows:DCP>
        – <ows:HTTP>
            <ows:Get xlink:href="http://opentopo.sdsc.edu/geoportal
            /csw"/>
```

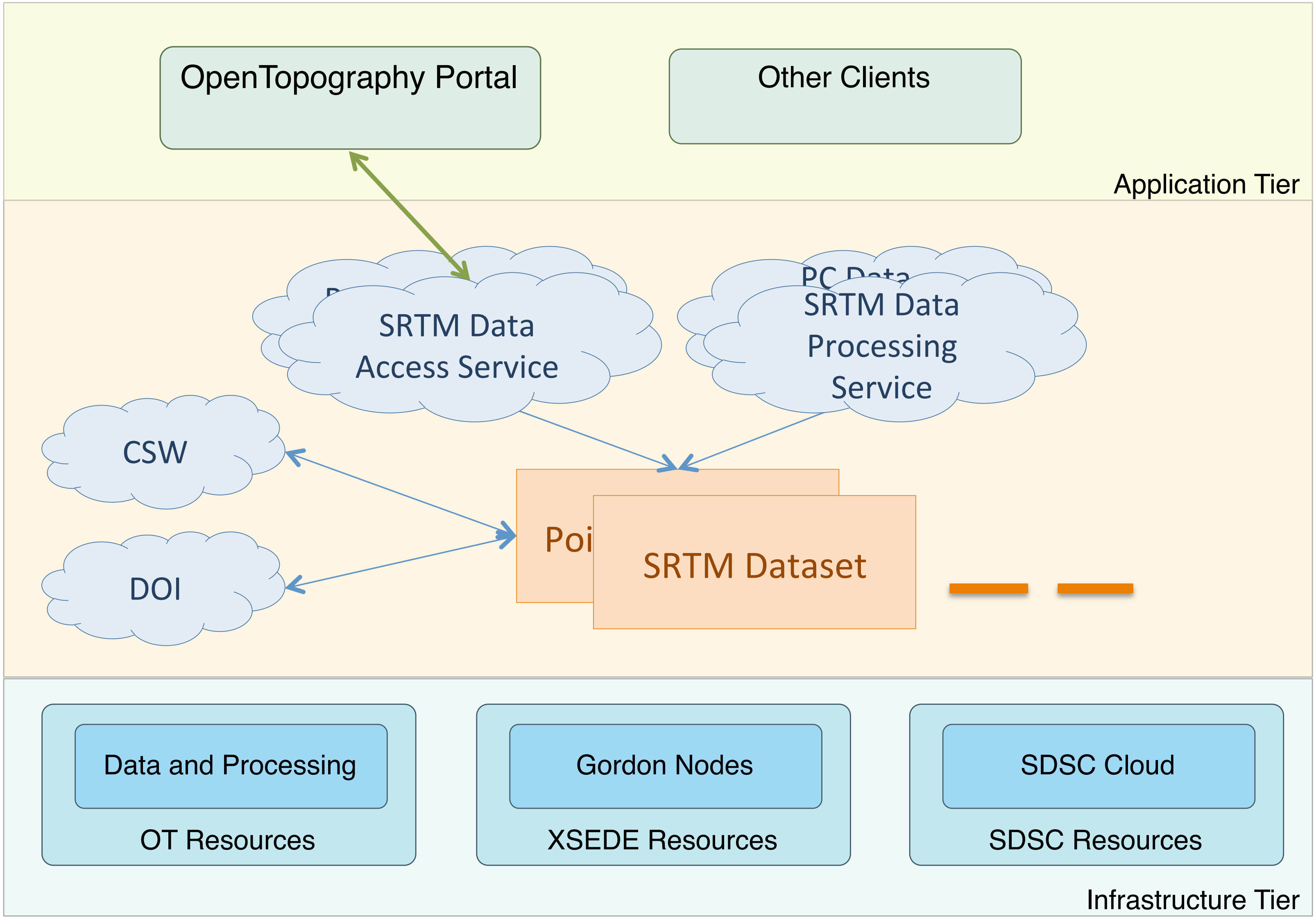
Current Data Holdings

- Lidar Point Cloud Datasets
  - 770 Billion+ lidar returns.
  - Each return has additional attributes
  - On-demand processing capabilities
- SRTM, Raster (multiple layers)

  e.g. Sonoma - several intensity products, canopy height, bare earth, hydro-enforced bare earth, canopy top, etc.
- 30+ TB of on-demand online data

  (Excluding big custom job runs, original archived data, pre processing data and user generated products)

OT CyberInfrastructure



OpenTopography Portal

Other Clients

Application Tier

PC Data

SRTM Data Access Service

SRTM Data Processing Service

CSW

DOI

Poi

SRTM Dataset

Data and Processing

OT Resources

Gordon Nodes

XSEDE Resources

SDSC Cloud

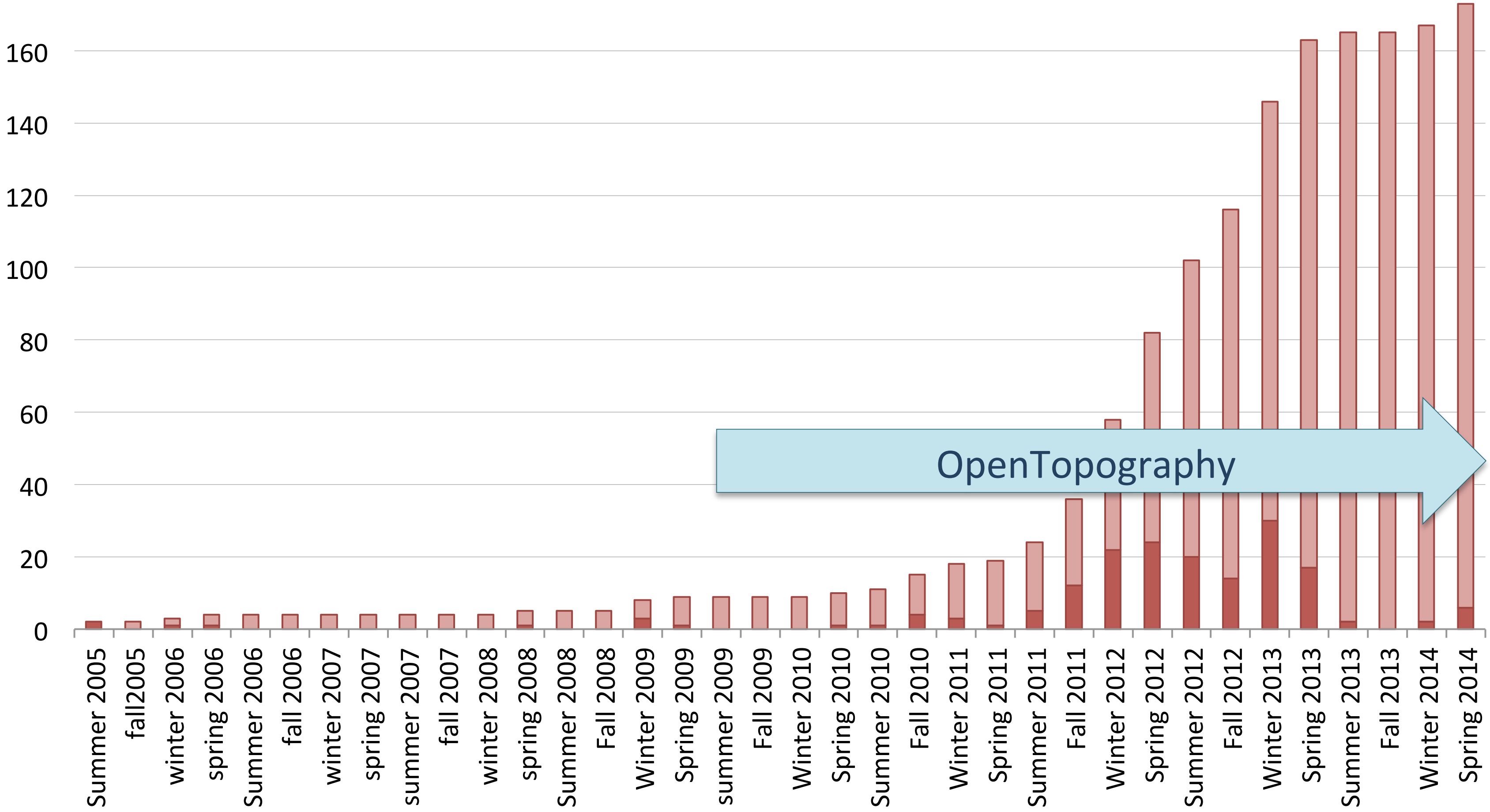
SDSC Resources

Infrastructure Tier

OT Challenges

(Keeping pace with data and user growth and advancing science!)

LIDAR Point Cloud Data Growth



Sensor Hardware Technology

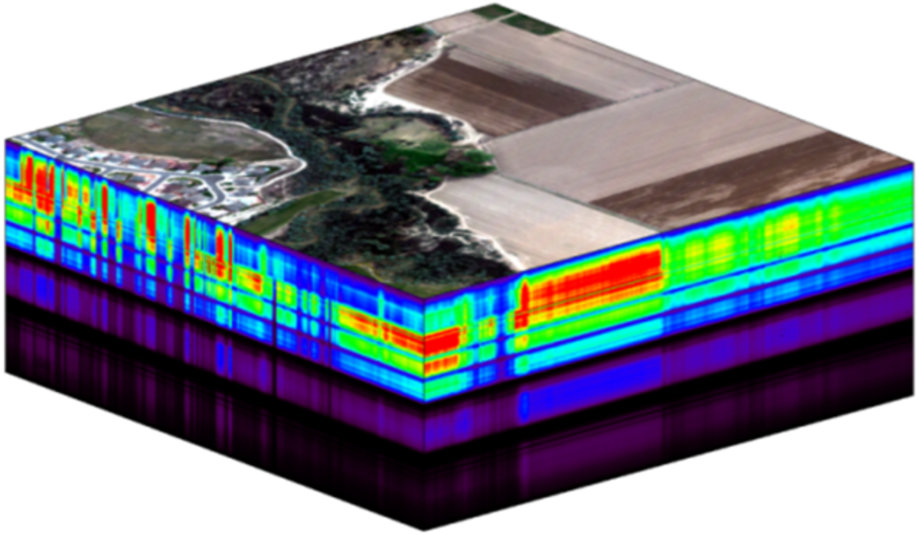
- Rapid Evolution of Laser Scanner Technology
- Data is being collected on multiple channels (different wavelengths) and capturing full waveform data.
- Early datasets collected with scanners operating at less than 33Khz. Current systems collect data at ~900Khz

Greater Resolution =
Larger Data Volumes

Image: RIEGL USA

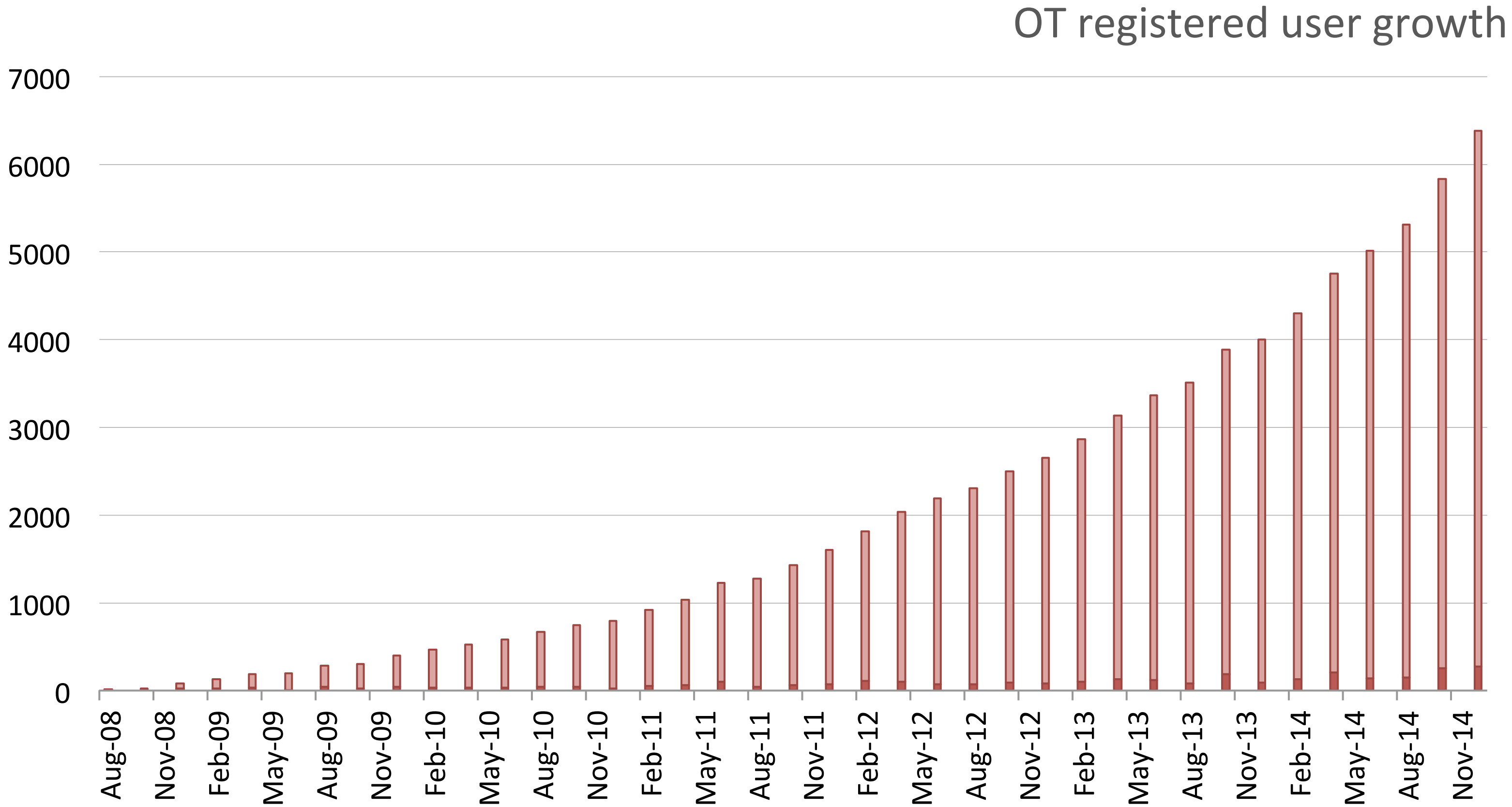
Diverse and complex datasets support

- Discrete return lidar
- Full waveform lidar
- Optical imagery (R,G,B orthophotography)
- Hyperspectral imagery



NEON Hyperspectral Imagery collected light reflected across the electromagnetic spectrum for a total of 426 bands of information. Image: Nathan Leisso, NEON AOP

User Growth



CyberGIS, NASA/UNAVCO communities

Service Level Access

Pluggable Services Infrastructure

- Methods for scientific data processing are evolving
- Users demanding more processing services
- Pluggable services infrastructure
  - OT development sandbox
  - Assist researchers with their code
  - Deployment of the algorithm as a service
  - Update processing workflow and UI

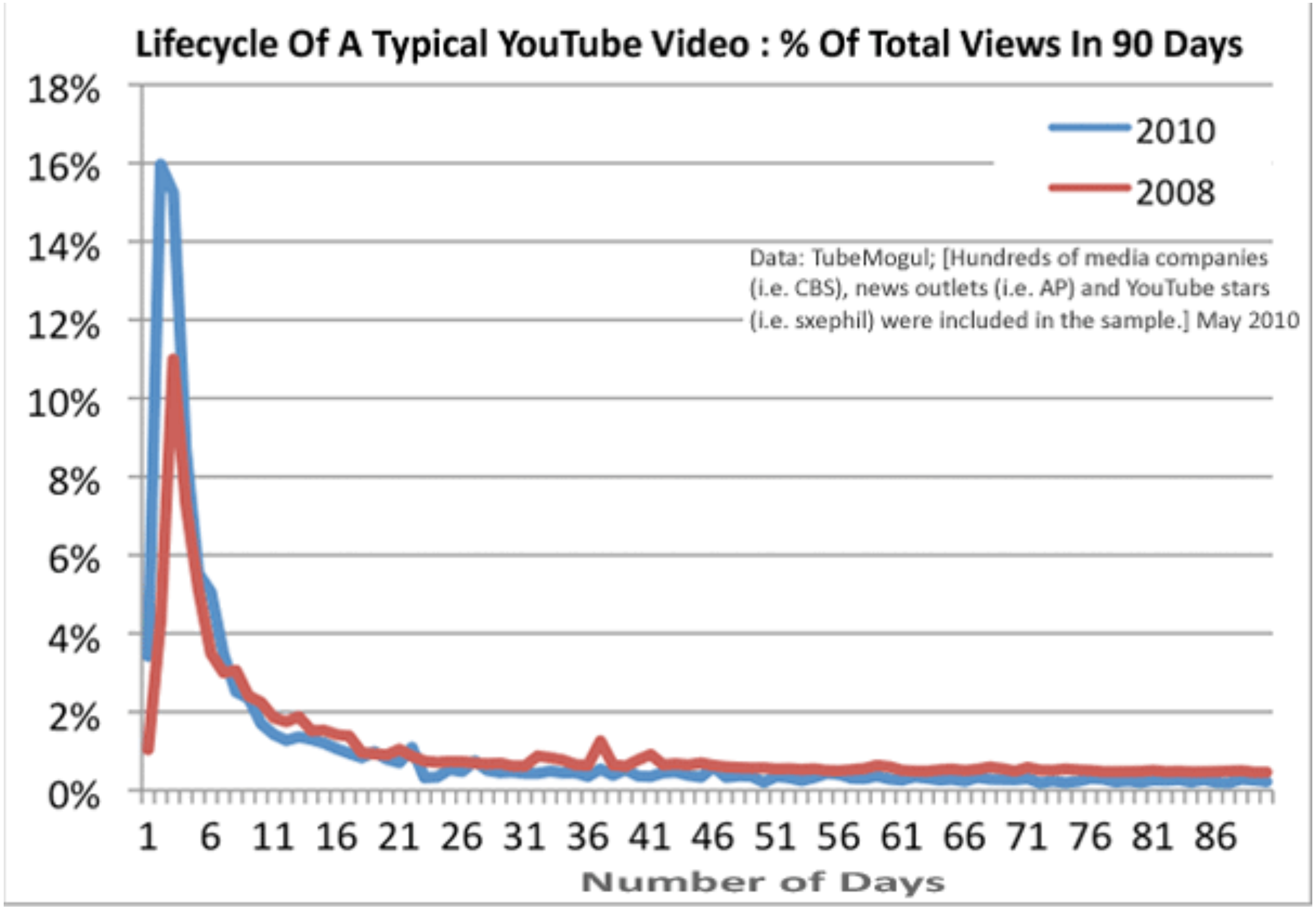
Increase in user Generated Derived products!

Data Insights

What can we learn from:

40,000+ custom PC jobs
1.2 trillion lidar points processed
15,000+ custom raster jobs (past year)

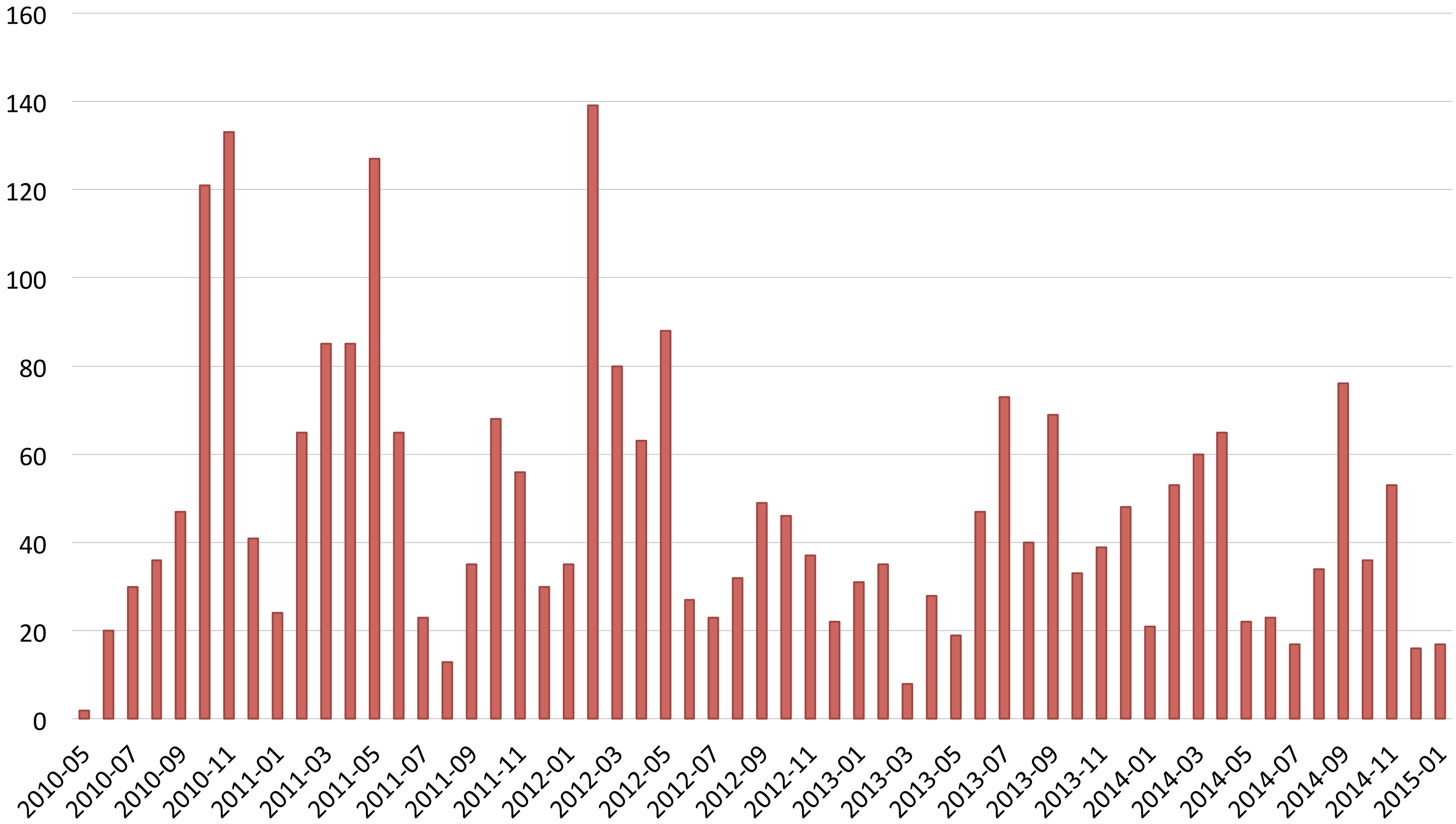
Data Access Patterns



Source: http://www.businessinsider.com/chart-of-the-day-the-lifecycle-of-a-youtube-video-2010-5

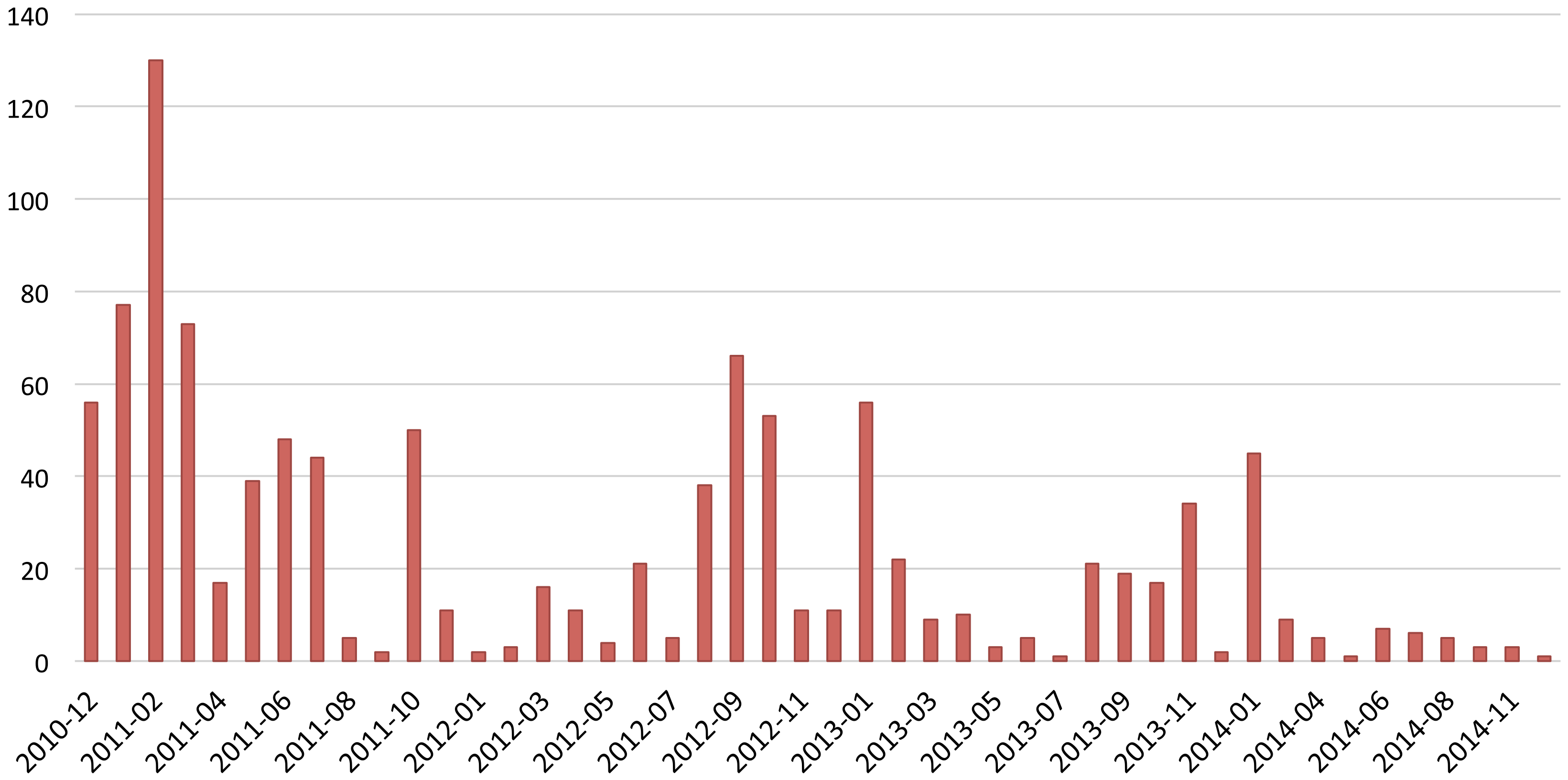
Access Patterns in LIDAR Scientific Datasets

B4 – San Andreas Fault



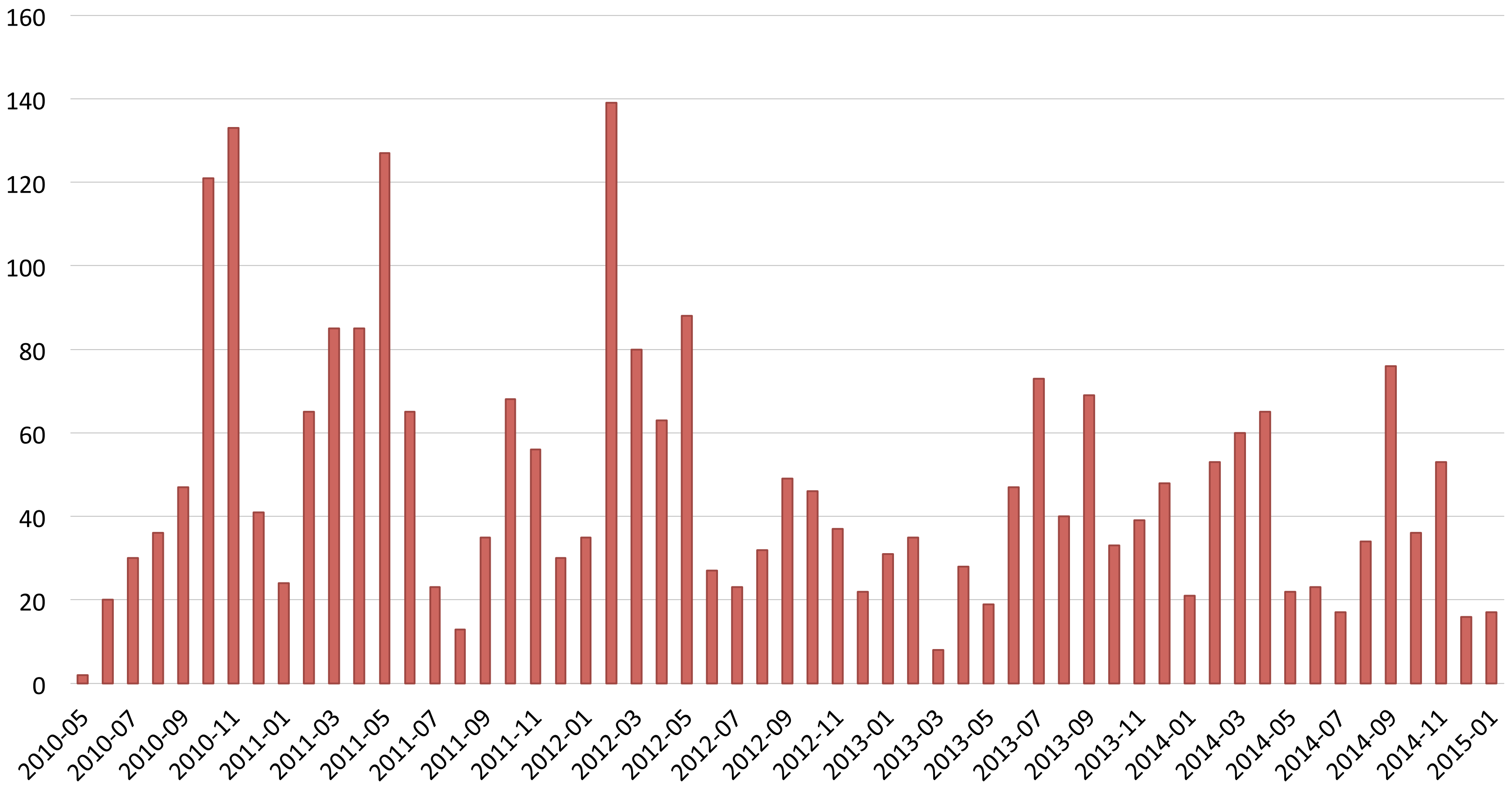
Access Patterns in LIDAR Scientific Datasets

Event based datasets - El Mayor-Cucapah Earthquake



Access Patterns in LIDAR Scientific Datasets

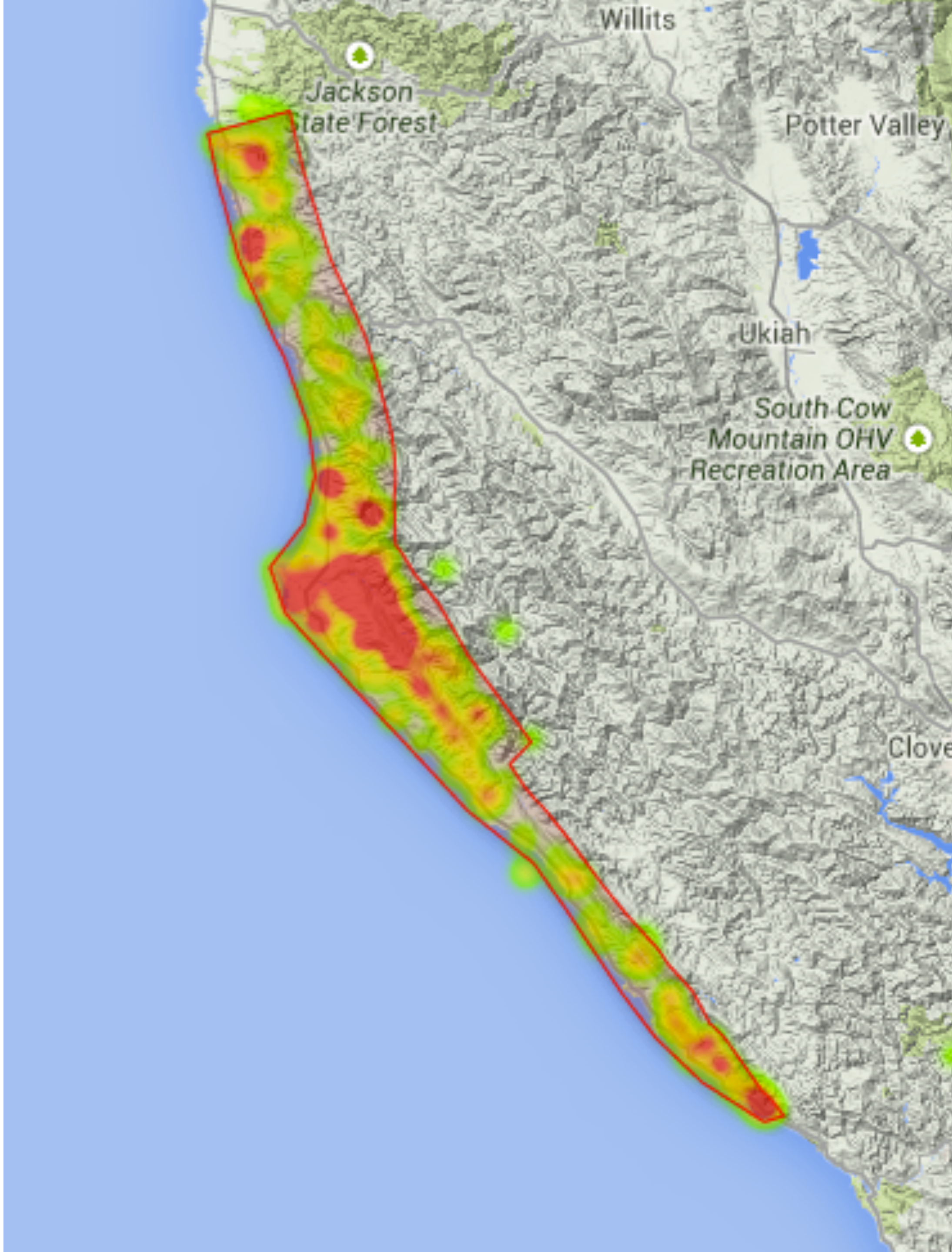
Event based datasets – Haiti Earthquake



Data Usage Analytics

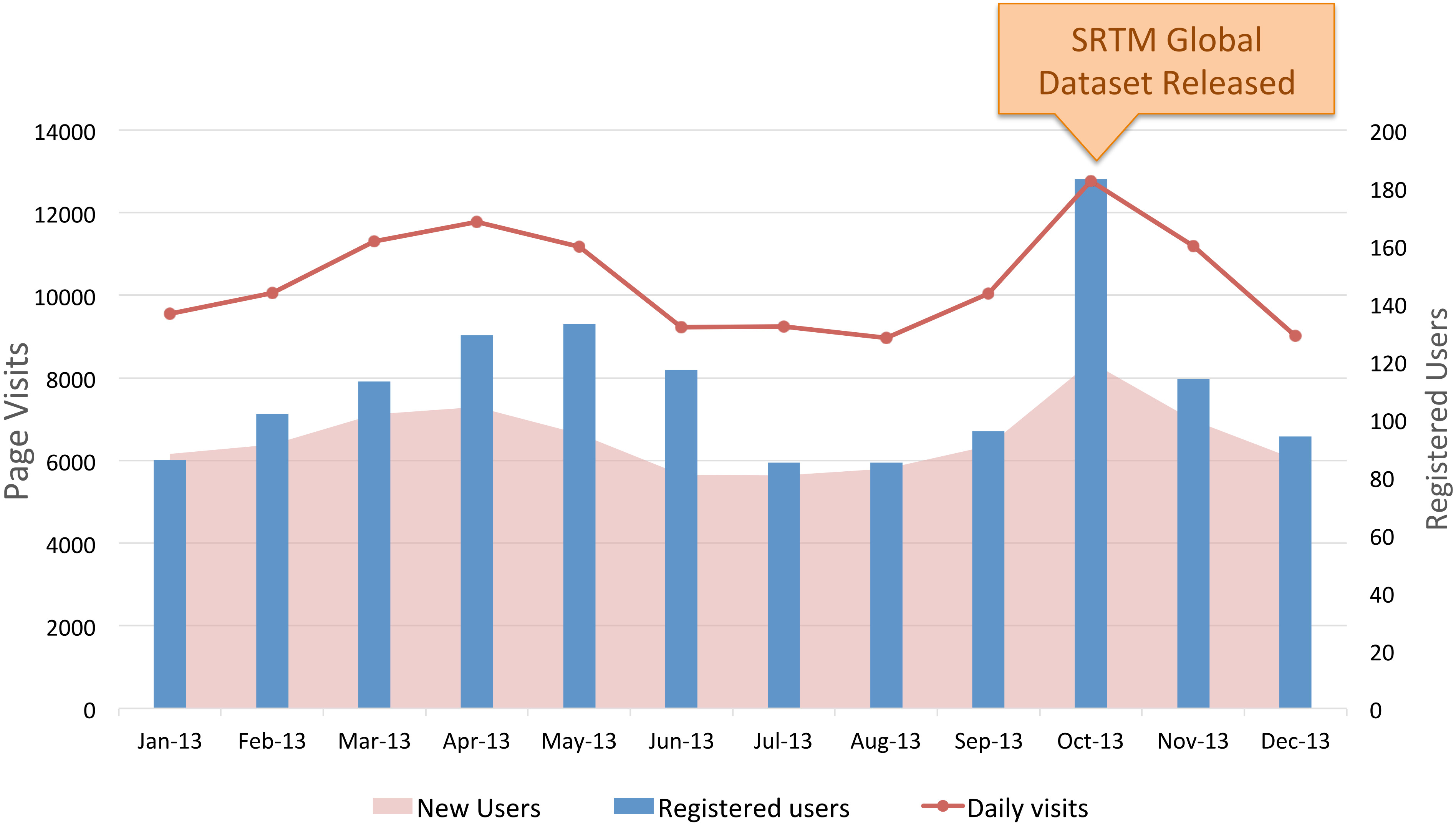
Northern San Andreas Fault

Social networking with data
Recommendation system



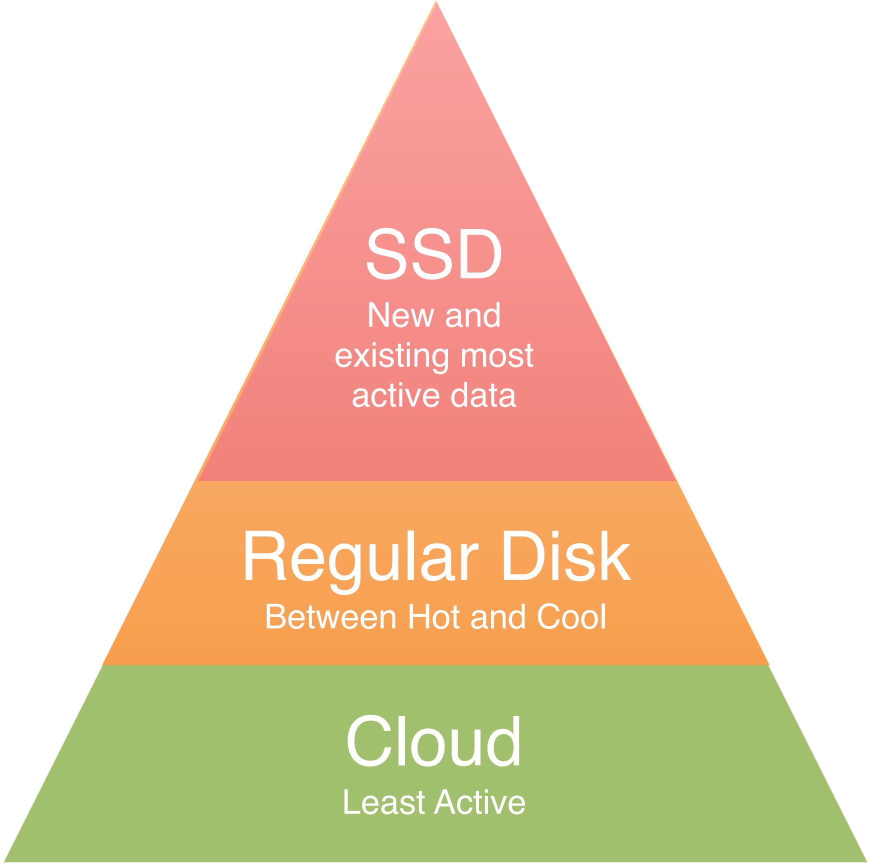
Understanding Traffic Patterns

Impact of dataset releases and workshops on traffic



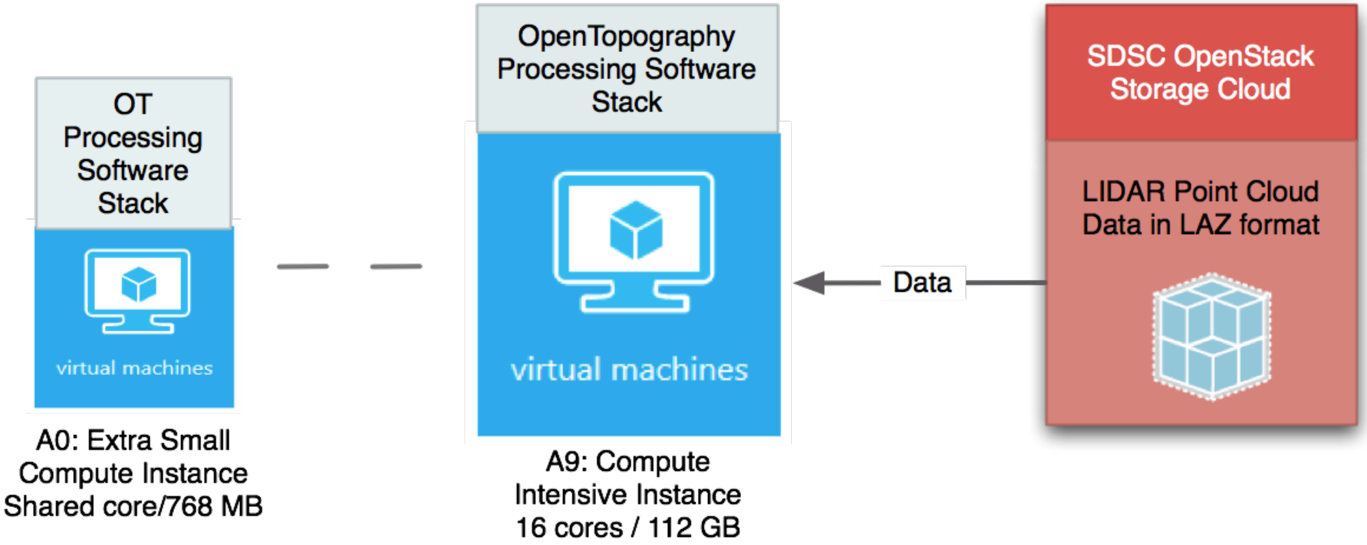
Tiered storage based on Data Access

- Activity based data ranking and tiered cloud integrated storage



Cloud Computing

- Cost effectiveness & feasibility of data science facilities on the cloud
- Microsoft Azure for Research Award Integration of cloud based on-demand geospatial processing services into community earth science data facilities.



OT Processing Software Stack

virtual machines

A0: Extra Small Compute Instance Shared core/768 MB

OpenTopography Processing Software Stack

virtual machines

A9: Compute Intensive Instance 16 cores / 112 GB

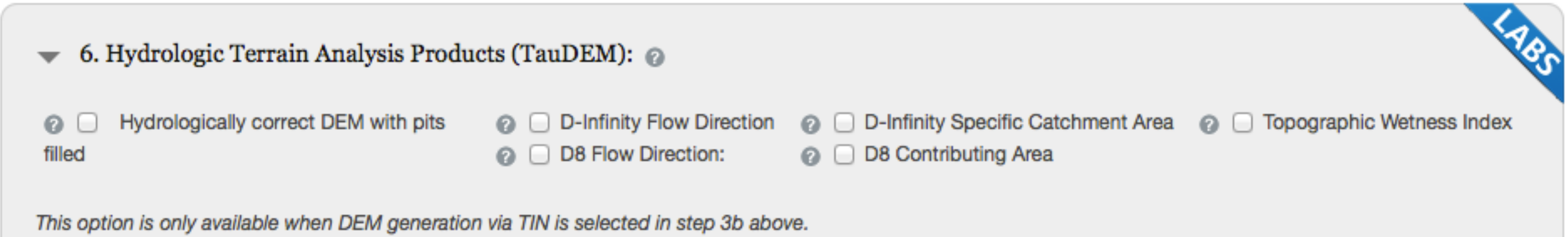
Data

SDSC OpenStack Storage Cloud

LIDAR Point Cloud Data in LAZ format

Leveraging HPC

- Dedicated Gordon nodes via XSEDE (democratization of supercomputing resources)



6. Hydrologic Terrain Analysis Products (TauDEM):

□ Hydrologically correct DEM with pits filled

□ D-Infinity Flow Direction

□ D8 Flow Direction:

□ D-Infinity Specific Catchment Area

□ D8 Contributing Area

□ Topographic Wetness Index

This option is only available when DEM generation via TIN is selected in step 3b above.

GEO Data and Cyberinfrastructure Imperative: Harness the Power of Computing and Computational Infrastructure.

- GEO Priorities and Frontiers: 2015-2020

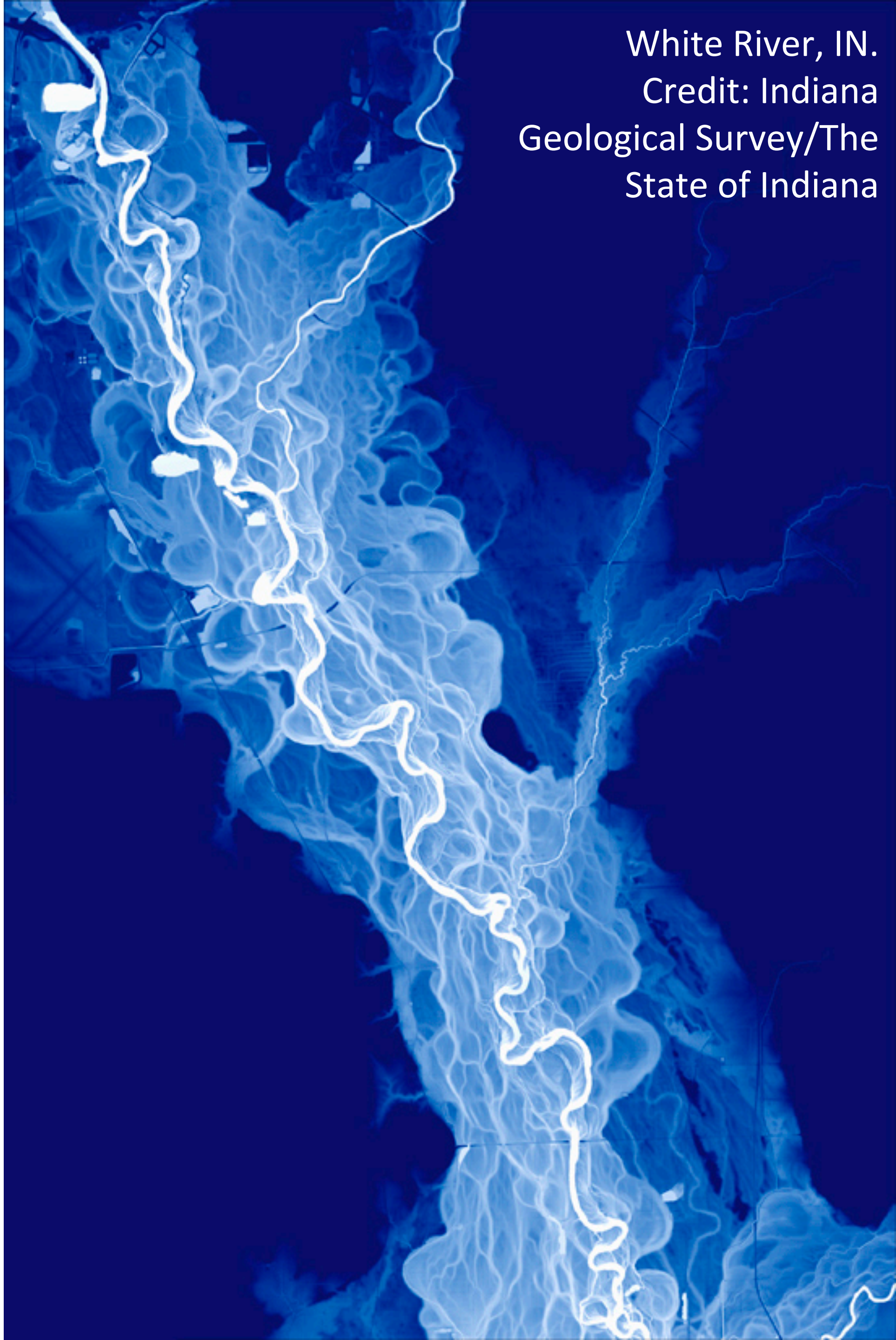
Summary

- OpenTopography is an modern agile data facility
- Cyberinfrastructure driven by science use cases – CI and science collaboration
- Big Data needs to be usable - Community not only wants access to data but also wants tools for processing these data.
- Concept of OT can be used as a template for other large data facilities

Thank You!

viswanat@sdsc.edu
info@opentopography.og

@OpenTopography



White River, IN.
Credit: Indiana Geological Survey/The State of Indiana