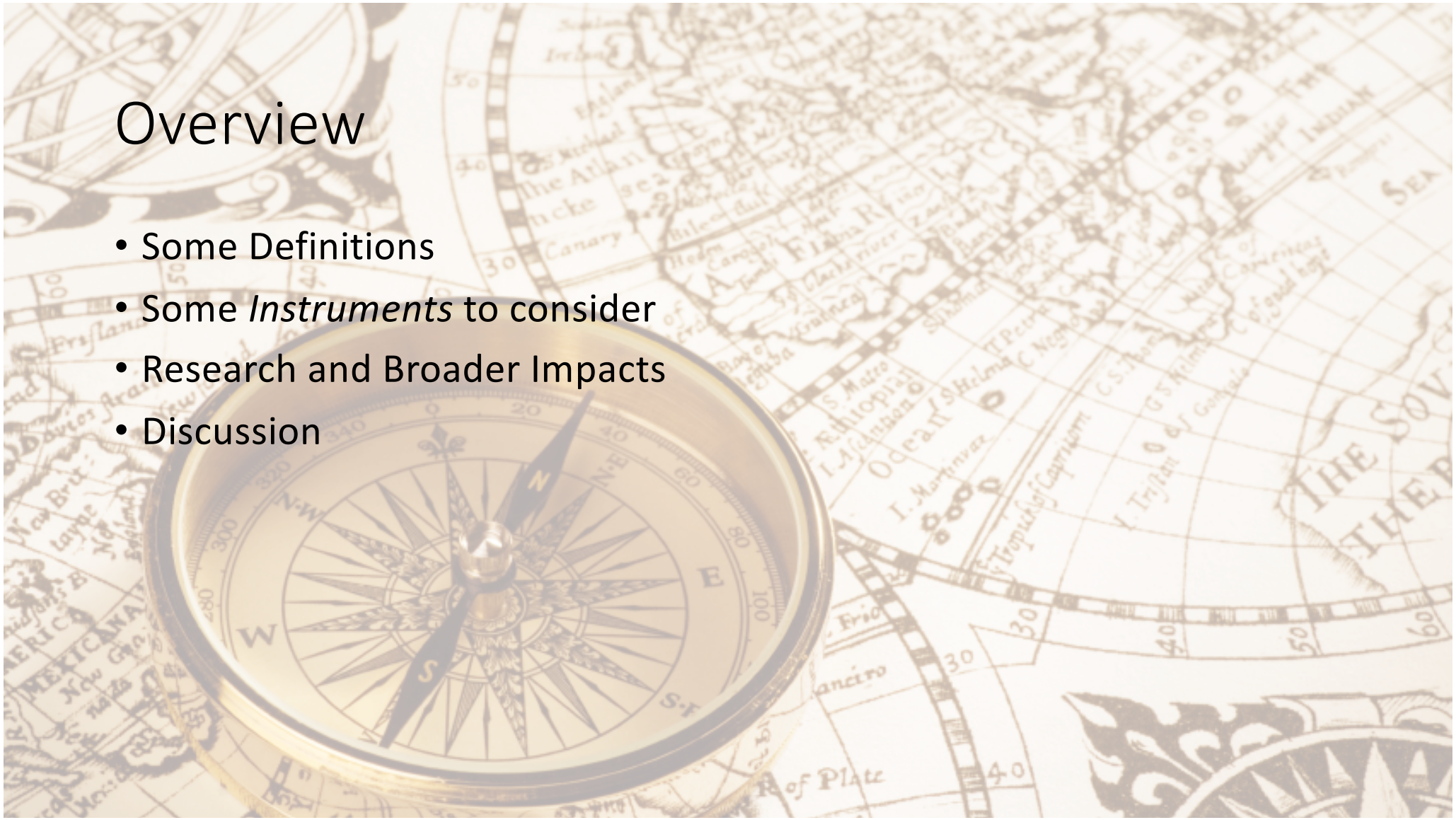
The Network as Instrument

Karl Benedict – kbene@unm.edu
Director of Research Data Services
College of University Libraries & Learning Sciences
University of New Mexico

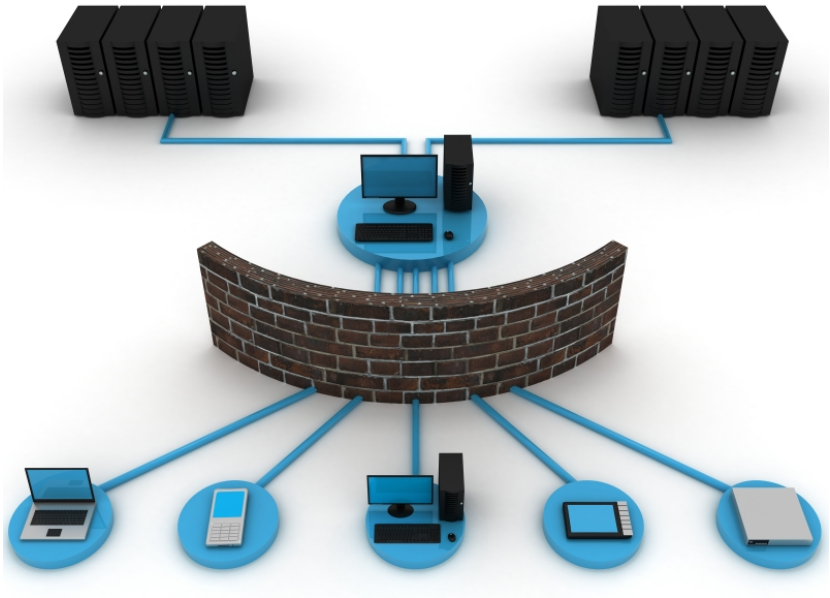


Overview

- Some Definitions
- Some *Instruments* to consider
- Research and Broader Impacts
- Discussion

Some Definitions

Network



Computer Network:

A computer network is a group of computer systems and other computing hardware devices that are linked together through communication channels to facilitate communication and resource-sharing among a wide range of users.

- https://www.techopedia.com/definition/25597/computer-network

Instrument

Scientific Instrument:

"A scientific instrument is an instrument used for scientific purposes. Most are measurement instruments. They may be specifically designed, constructed and refined for the purpose. Over time, instruments have become more accurate and precise."

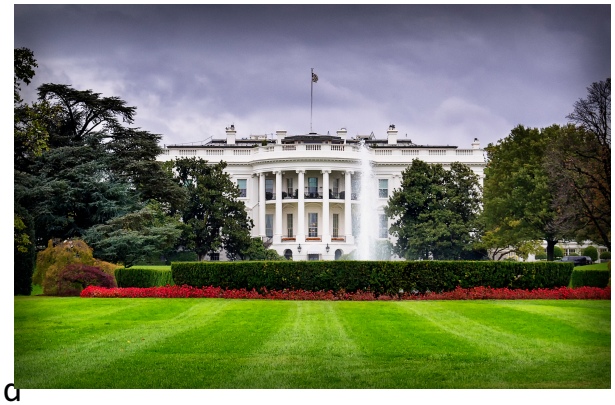
- Increasingly – instruments are network connected and integrated into Laboratory Information Systems

- https://en.wikipedia.org/wiki/Scientific_instrument


Large Hadron Collider


Very Large Array

Data



“Data: For the purposes of this Memorandum, the term "data" refers to all structured information, unless otherwise noted. “ – OMB Memorandum M-13-13

“All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science.” – Science General Information for Authors – Data and Materials Availability Statement

“Data (/ˈdeɪtə/ day-tə, /ˈdætə/ da-tə, or /ˈdɑːtə/ dah-tə) is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.” - https://en.wikipedia.org/wiki/Data

Science

AAAS

WIKIPEDIA
The Free Encyclopedia

Computation

Computational science (also scientific computing or scientific computation) is a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problems. Computational science fuses three distinct elements:

- Algorithms (numerical and non-numerical) and modeling and simulation software developed to solve science (e.g., biological, physical, and social), engineering, and humanities problems
- Computer and information science that develops and optimizes the advanced system hardware, software, networking, and data management components needed to solve computationally demanding problems
- The computing infrastructure that supports both the science and engineering problem solving and the developmental computer and information science

- https://en.wikipedia.org/wiki/Computational_science

Collaboration

Research Collaboration Currently has a number of dimensions:

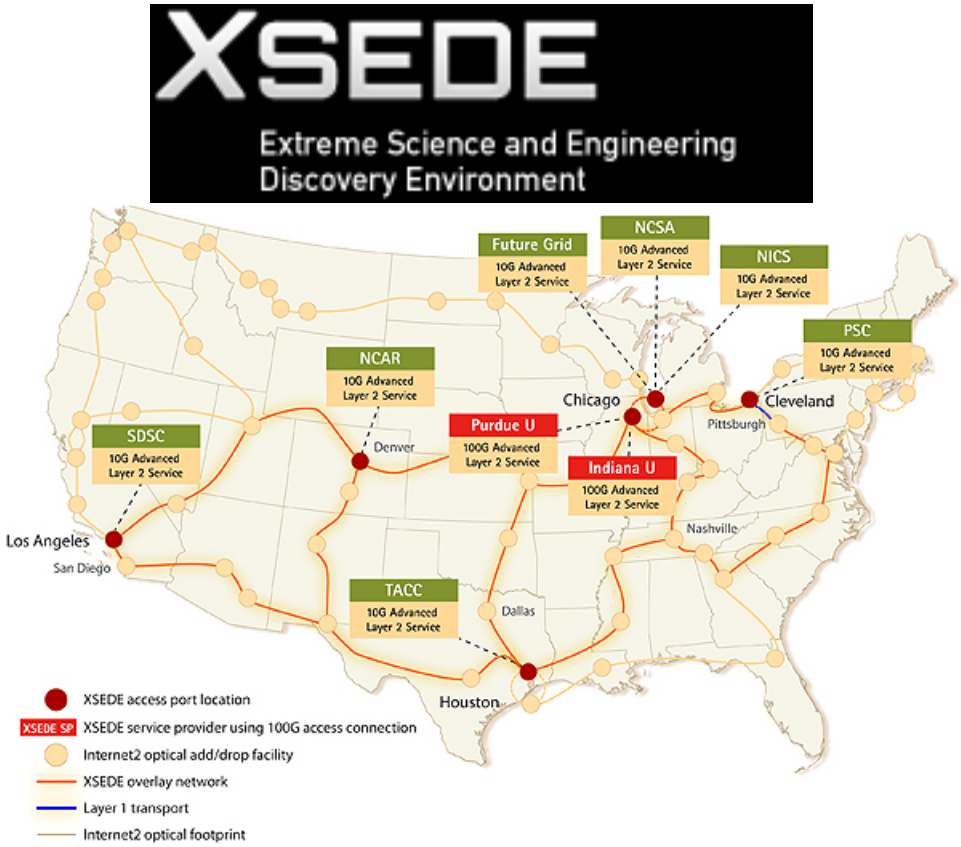
- Communication – web and video conferencing
- Data management and sharing
- Document development
- Resource sharing
- All potentially across large distances


Some *Instruments* to Consider

Computation – High Performance Computing

The XSEDE network is an example of a collection of network connected computational resources

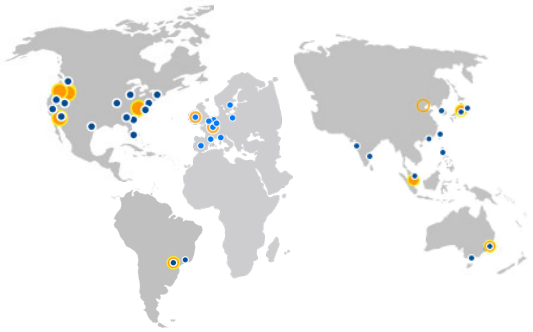
- Compute Resources
- Visualization
- Storage
- High-throughput computing
- Other services



Computation – On-demand Computing (i.e. cloud computing)

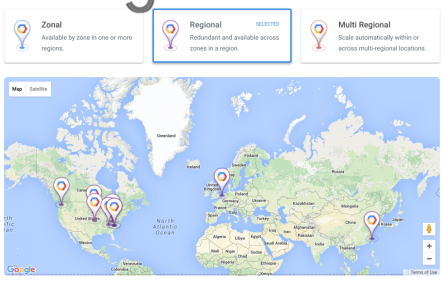
amazon web services

Amazon EC2



Microsoft Azure

Google Cloud Platform



openstack

Both public and private cloud computing may be used to effectively scale up non-HPC computational resources on an as needed, “pay-as-you-go”

- Tasks easily distributed to many individual computers
- Development and testing platforms
- Platform/Infrastructure As a Service

Computation - Distributed

Computation can also be distributed throughout ad-hoc networks of contributing computers. Examples include:

- The HTCondor project from the University of Winconsin
- Berkeley Open Infrastructure for Network Computing (BOINC) project from UC Berkeley

Data – Collaborative Data Sharing

A variety of public and private cloud solutions enable collaborative data sharing

- Dropbox
- Google Drive
- Amazon S3
- OwnCloud

Google Drive

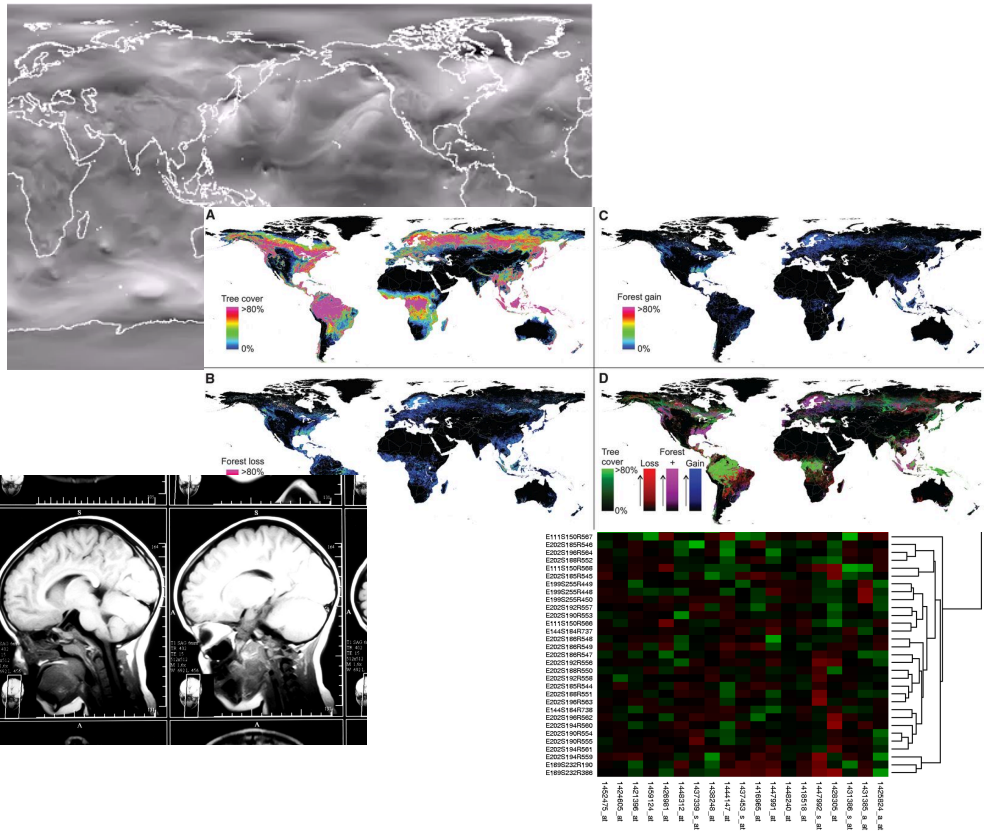
Amazon S3

Data – Streaming Data

Data may be generated in a continuous fashion and made available as continuously updated streams or frequenty updated snapshots

- Data from instruments
- Data from user interactions with systems

Data – High Volume Observational or Model Data



Some data sources produce huge volumes of data that must be transferred from one location to another, for example

- Satellite Imagery
- Climate and weather models
- Medical imaging
- Genomics

Data – Replication, Backup & Preservation

Large volumes of data may need to be transferred between facilities to meet data security, access and preservation needs

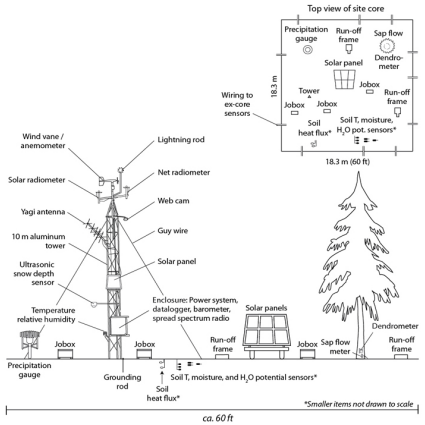
- Replication for efficient local access
- Backup and disaster planning
- Long-term preservation of access

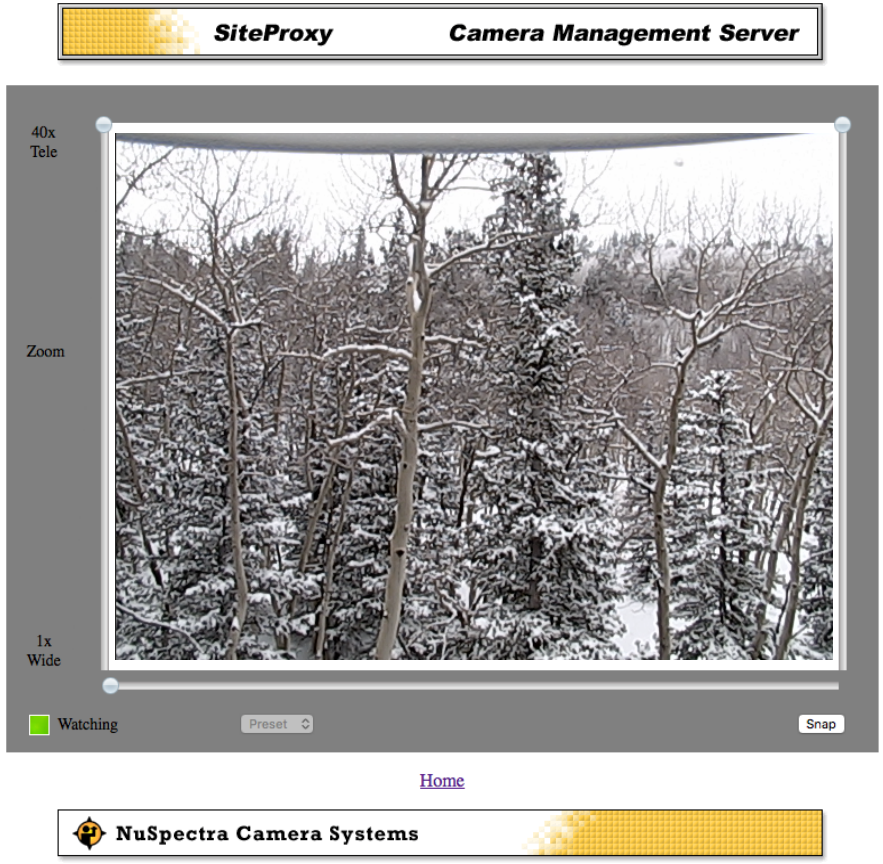
UNM | LoboVault

DPN
THE DIGITAL PRESERVATION NETWORK

Data – Remote Control

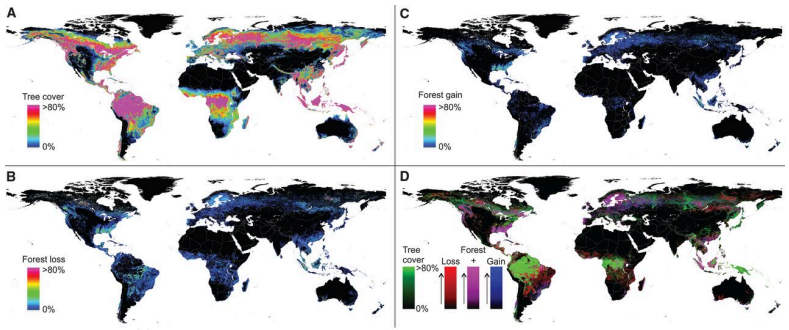
Network connectivity with sensor systems can enable remote control capabilities for those sensors





Data & Computation – Co-location

An increasingly important model for data-intensive science is to co-locate data and computational capabilities to maximize analytic performance



High-Resolution Global Maps of 21st-Century Forest Cover Change

M. C. Hansen,[1]* P. V. Potapov,[1] R. Moore,[2] M. Hancher,[2] S. A. Turubanova,[1] A. Tyukavina,[1] D. Thau,[2] S. V. Stehman,[3] S. J. Goetz,[4] T. R. Loveland,[5] A. Kommareddy,[6] A. Egorov,[6] L. Chini,[1] C. O. Justice,[1] J. R. G. Townshend[1]

Quantification of global forest change has been lacking despite the recognized importance of forest ecosystem services. In this study, Earth observation satellite data were used to map global forest loss (2.3 million square kilometers) and gain (0.8 million square kilometers) from 2000 to 2012 at a spatial resolution of 30 meters. The tropics were the only climate domain to exhibit a trend, with forest loss increasing by 2101 square kilometers per year. Brazil's well-documented reduction in deforestation was offset by increasing forest loss in Indonesia, Malaysia, Paraguay, Bolivia, Zambia, Angola, and elsewhere. Intensive forestry practiced within subtropical forests resulted in the highest rates of forest change globally. Boreal forest loss due largely to fire and forestry was second to that in the tropics in absolute and proportional terms. These results depict a globally consistent and locally relevant record of forest change.

Google Earth Engine

128.8 M km2
143 Billion 30m pixels
654,178 Landsat 7 ETM+ Scenes

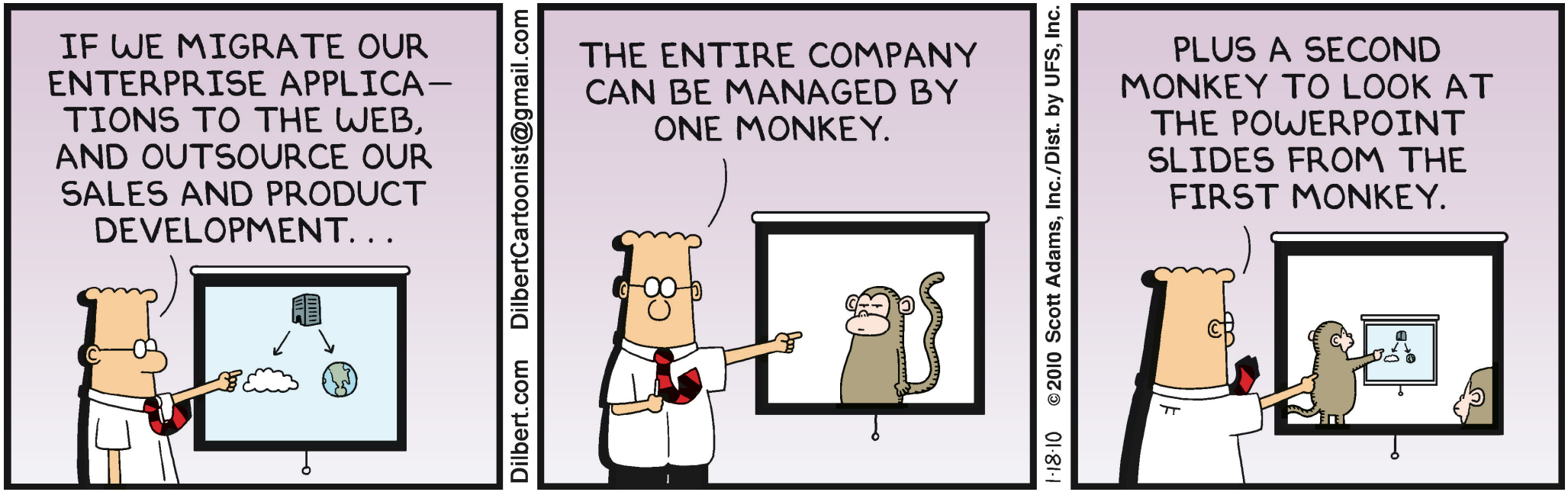
"… one million CPU-core hours on 10,000 computers …"

Collaboration

The increasing focus on collaborative interdisciplinary research yields a demand for effective tools for collaboration and communication

- Audio and video conferencing
- Desktop sharing
- Shared access to data and documents
- Shared document production

Research and Broader Impacts



Discoverability

Network accessible data require machine readable documentation (metadata) to allow for discovery at scale through a variety of systems

METADATA
Dublin Core® Metadata Initiative
INNOVATION

FederalGeographicDataCommittee fgdc

ISO

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

LTER Network Data Portal

UNM | LoboVault

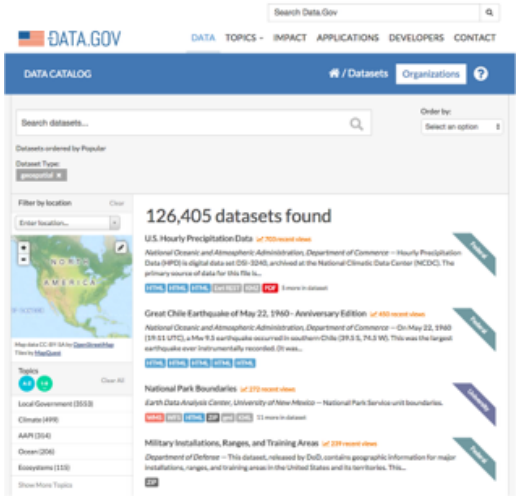
figshare

GEO GROUP ON EARTH OBSERVATIONS | GEOSS Portal
Discover, Access, Contribute
Earth Observations, Information and Services

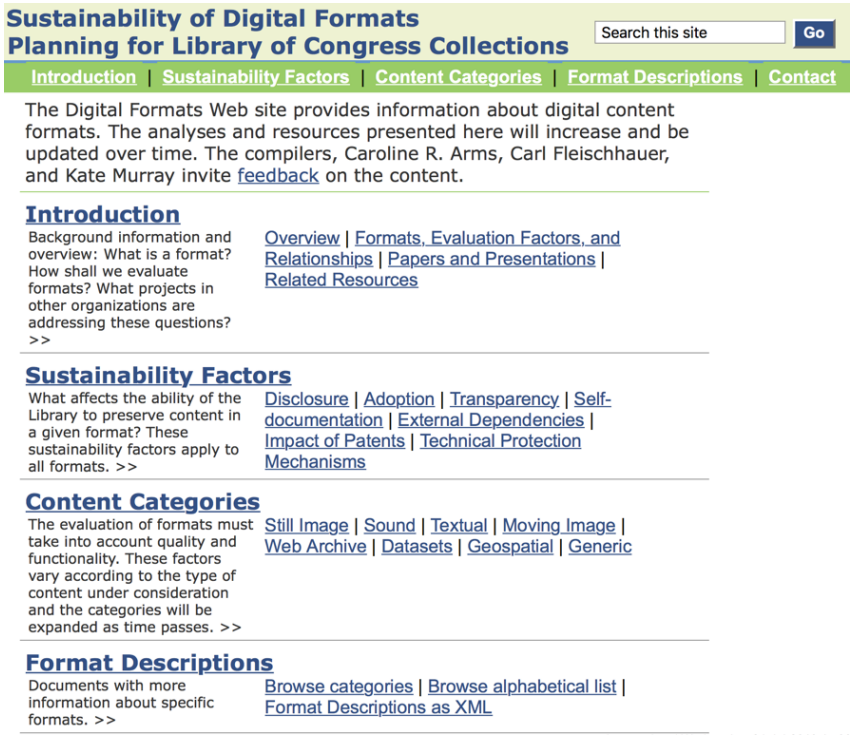
zenodo

DRYAD

DataONE
Data Observation Network for Earth



Access - Download



http://www.digitalpreservation.gov/formats/index.shtml

To maximize impact data must be accessible in formats that are broadly usable, easlily preserved and well documented for understanding and use

- Open standards
- Well documented
- Self-documenting
- Robust

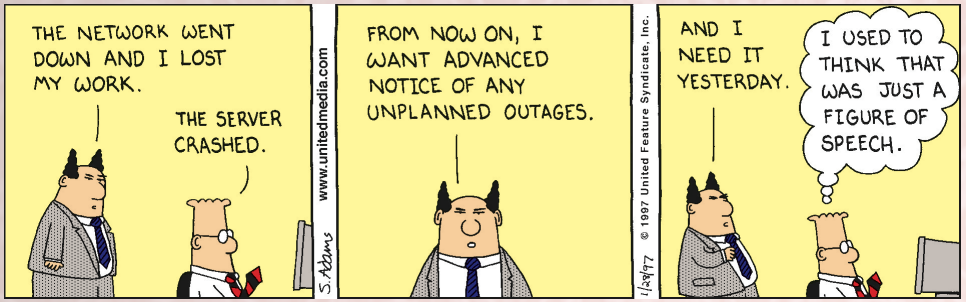
Access – Data Access, Analysis & Visualization Services

Data access, analysis and visualization services enable "smart" data access and use – making otherwise difficult to use data usable and set the stage for apps through an API (Application Programming Interface).

OGC®
Making location count.

REST –
Representational State Transfer

SOAP –
Simple Object Access Protocol



Questions/Discussion

Acknowledgements

- Dilbert cartoons are used under license from United Media
- Large Hadron Collider Image Credit: *alpenthread* - https://flic.kr/p/4fdCYX
- Very Large Array Image Credit: *Hajor* - https://commons.wikimedia.org/wiki/File:USA.NM.VeryLargeArray.02.jpg
- Stack of Hard Drives Image Credit: *Kenny Louie* - https://flic.kr/p/5Ut2pe
- XSEDE Network Map - https://www.xsede.org/networking
- Global Forest Cover Change Map – *Science*, 15 November 2013. Vol. 342 no. 6160. pp. 850-853. DOI: 10.1126/science.1244693
- DNA Microarray Image - https://en.wikipedia.org/wiki/DNA_microarray#/media/File:Heatmap.png
- Haiti Earthquake Image Credit – *United Nations Development Programme* - https://flic.kr/p/7vJBLd
- All other images licensed from iStockPhoto or Shutterstock
- This presentation is presented as part “Connecting Science, Education, & Economic Development” An NSF Workshop Series for Capacity and Community Planning funded by the NSF Campus Cyberinfrastructure Data, Networking, and Innovation Program (CC*DNI) - NSF Award Number: 1541340