MAGIC Scientific Data Life Cycle Summary- Spring 2019

Rich Carlson (DOE/SC)
Vipin Chaudhary (NSF)
Joyce Lee (NCO)

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Life Cycle Speaker Series

- Purpose of series:
  - Cover all aspects of the Data Life Cycle
  - Identify changes and updates found in the past few years
  - Identify agency programs that address these changes
- Monthly Meeting topics:
  - February: Data Sources and Topics in 2019
  - March: Data growth in science and society
  - April: Data Analysis and processing
  - May: Data management and storage
  - June: Interactive analysis tools and services
  - July: Data Triage and discovery
  - August: Interim Summary

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Sources and Topics in 2019

- Deb Agarwal
  - Position: Department Head and Senior Scientist, Lawrence Berkeley National Laboratory
  - Title: *Deduce (Distributed Dynamic Data Analysis Infrastructure for Collaborative Environments*
- Suhas Somnath
  - Position: Computer Scientist, Oak Ridge National Laboratory
  - Title: *Scaling Data-Driven Scientific Discovery – A Data Lifecycle View*
- Alexander Hexemer
  - Senior Staff Scientist, Lawrence Berkeley National Laboratory
  - Title: *Addressing the Data Challenges at the Advance Light Source*

Data Sources and Topics in 2019

- Data Life Cycle includes:
  - Planning and expectations
  - Data/Metadata generation and collection
  - Analysis (triage, in-depth, (re-)calibration, re-processing, reuse)
  - Curation and provenance generation and checking
  - Storage (initial to archival)
  - Publication (DOI, papers, data, data subsets)
- There are multiple loops in this cycle and multiple iterations through loops and subloops

Data Sources and Topics in 2019

- Scientific data comes from many sources
  - Earth Science (AmeriFlux, FLUXNET Carbon Flux, NGEE Tropics)
  - Experimental (Microscopy, Astronomy, Typography)
  - Simulations (Climate, Combustion, Materials)
- Significant Data attributes
  - Experimental data is rapidly increasing as detectors and sensor capabilities improve
  - Simulation data is rapidly increasing as processing power improves
  - Data may be reprocessed multiple times, changing old values
  - Multiple file formats are in common use

Data Sources and Topics in 2019
Challenges and Issues

- Data changes as it is processed, analyzed, and reprocessed
  - What data changed
  - What impact does this have on previous papers and/or conclusions
  - Who used old data and how are they notified when changes are found
- Multiple, possibly disparate, sources of data
  - Many contributors and/or sources
  - Data may have been originally collected for different purpose
  - Original meaning or format of data may have been lost

Data Sources and Topics in 2019
Challenges and Issues

- Data generation
  - Commercial instruments may use proprietary data formats
  - Commercial instruments may require dedicated computer to operate
  - Multi-stage processing pipeline becoming necessary as data volumes grow
- Analysis tools and services
  - Proprietary software is close-source, antiquated, and expensive to maintain
  - Simple to use high performance programming/analysis environments are needed
  - Routinely moving data and/or code at high rates over variable distances is hard
  - Need to take advantage of advances in ML/AI along with more traditional analysis tools

Data Sources and Topics in 2019

- Challenges and Issues
  - Data curation, provenance, and metadata
    - Tracking how data was generated, changed, updated, or moved is important
    - File formats like HDF5 build metadata into the file, but not every format has that function
    - Distributed workflows and data analysis tools need to add to this curation process
  - Make science discoveries as easy as possible
    - Better visual and interactive analysis/discovery systems
    - Automate as much of the routine process as possible
    - Provide recognition to everyone involved in the discovery process

Data Sources and Topics in 2019

- Conclusions and Findings
  - Overall Challenge: Moving from the current situation where individuals work in isolation to ecosystems that automate as much of the discovery process as possible
  - Need to train current and future scientists to work in this environment
  - Data and software needs to be open source allowing its free exchange with sufficient documentation to understand/use
  - Cultural Change: Moving from individual data sets to an ecosystem where communities have major infrastructures and individuals receive credit and recognition for their contributions to their science domain

Data Sources and Topics in 2019 Discussion Summary

- Top 2 challenges to realizing your vision?
  - Overall challenge: having all needed tools to go from end-to-end. Moving from current situation to where it is reasonably routine. Trying to put together a data pipeline from instrument to user service to analytics, ML, etc., is very challenging and takes much expertise.
  - Who will pay for holding visitors' data in data catalogue
  - How to train domain scientists to view each other's data sets
  - Motivating younger scientists to have more exposure to computing; perhaps curriculum changes are needed
  - Commercial proprietary data formats in software. Force vendors to provide more open data formats; would benefit scientists nationwide. Need support of from funding agencies and journals to require community standard
  - Who will pay for data storage? Note: Future increase of compute requirements
  - Funding agencies need to mandate that awardees have a data plan, not just in hard drives. Also have data stored in archive and funding necessary components. Helps shift community
- Cultural Change
  - Cultural change from individual data sets to moving data to compute facility and build data into major infrastructures (working in the community). Need to start prepping now for future change (e.g., more light sources). Also ensure data credit, data publishing work well when no longer in complete control of data. How would it positively impact scholarly reputation.
  - ALS: How to correctly throw away raw data.

Data Growth in Science and Society

- Charlie Catlett
  - Position: Senior Computer Scientist, Argonne National Laboratory
  - Title: *Using Computation and New Sources of Data to Understand Cities*
- Peter Nugent
  - Position: Senior Staff Scientist, Lawrence Berkeley National Laboratory
  - Title: *Multi-Messenger Astronomy and the Discovery of a Neutron Star – Neutron Star Merger*
- Shawn McKee
  - Position: Research Scientist, University of Michigan
  - Title: *Scientific Data Lifecycle: Perspectives from an LHC Physicist*

Data Growth in Science and Society

- Scientific data comes from many sources
  - Societal and Environmental science
  - Observational Astronomy
  - Particle Physics
- Significant Data attributes
  - Experimental data is rapidly increasing as detectors and sensor capabilities improve
  - Data from multiple sources needs to be combined/fused to provide greater understanding
  - Some data may contain personalizes or private information that need to be protected

Data Growth in Science and Society Challenges and Issues

- Data Triage – Fast Analysis of important data first
  - Observed events have a finite lifetime and important information is enclosed at the start of the event
  - Wireless network bandwidth limits data transmission rates, requiring fast in situ processing
  - Detector data volume far exceeds data transfer rates even on multi-gigabit/sec networks
- Data Fusion – Scientific insights require data from multiple sources
  - Different sensors provide different views of the same event
  - Multiple co-located data sensors provide a wide view of different events
  - Both simulations and observations are required to gain a full understand of an event

Data Growth in Science and Society
Challenges and Issues

- Follow-up activities
  - Observed events have a finite lifetime and important information is enclosed at the start of the event
  - Wireless network bandwidth limits data transmission rates, requiring fast in situ processing
  - Detector data volume far exceeds data transfer rates even on multi-gigabit/sec networks
- Long term storage and archiving
  - Data curation artifacts and metadata need to be stored along with raw/processed data to allow follow-on studies
  - Data volumes and global science communities require data storage occur at multiple geographically separate locations
  - New storage architectures needed to handle all aspects of data movement and storage

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Growth in Science and Society
Challenges and Issues

- Public access and policy guidance
  - Use and reuse of data by the public is important to maintain interest and trust – requires public data archives with usable APIs and adequate metadata/documentation
  - Science policy – drives new instrument design and deployment, create roadmaps for future investigation campaigns, support research funding
  - Public policy – drives political and social policy by providing reliable and understandable data
- Make science discoveries as easy as possible
  - Better visual and interactive analysis/discovery systems
  - Automate as much of the routine process as possible
  - Provide recognition to everyone involved in the discovery process

Data Growth in Science and Society Conclusions and Findings

- Overall Challenge: Virtuous cycle has emerged where more data (simulation and observational) leads to deeper insights which leads to new questions which requires more data
- Need to collect data over months or years to draw scientifically accurate insights, yet need fast analysis to confirm when to begin collecting data
- The existence of data at certain sites/sensors can be just as important as the absence of data at those sites/sensors
- Cultural Change: Publicly available scientific and environmental data can drive public interest, which supports general knowledge and funding priorities

Data Growth in Science and Society Discussion Summary

- Privacy and public access
  - Privacy of individuals captured by video sensors can be mitigated by processing data in situ using ML/AI algorithms which identify categories (person, car, bus, bicycle, dog, etc.), counting those items, and discarding the raw image (i.e., counting the number of people in a crosswalk or the number of cars waiting at a red light)
- ML/AI algorithm validation
  - Save sampled video data on restricted access server to continuously update algorithm training
- Public access
  - Public has access to processed data and raw data that does not contain privacy information
  - Strict access controls for all other data

Data Analysis and Processing

- Margaret Johnson
  - Position: Assistant Director, National Center of Supercomputing Applications
  - Title: *Continuous Learning About Data: Experience from the Dark Energy Survey and NCSA*
- Dan Petravick
  - Position: Senior Project Manager, National Center of Supercomputing Applications
    Title: *Continuous Learning About Data: Experience from the Dark Energy Survey and NCSA*
- Yong Chen
  - Associate Professor of Computer Science, Texas Tech University
  - Title: *Empowering Data-driven Discovery with a Lightweight Provenance Service for High Performance Computing*

Data Analysis and Processing

- Focus on Data Provenance
  - Document history of an object with particular value in authentication
  - In computer science, lineage of data including processes that act on data and agents responsible for those processes
  - Good governance if documented in detail sufficient for reproducibility
  - Complex scientific workflows make it hard provide efficient provenance
  - Gaining lot of attention
    - Provenance Markup Language
    - Research Data Alliance has Research Data Provenance Interest Group
    - Open Provenance Model
    - W3C Provenance Working Group

Data Analysis and Processing Challenges and Issues

- Performance requirements
  - HPC users are performance sensitive
  - Low overhead of processing and memory required
- Coverage requirements
  - Provenance generated from multiple physical locations
  - Provenance could have various granularities
- Transparency requirements
  - Users should not change or recompile code for provenance
  - Users cannot disable it
  - Non-repudiation

Data Analysis and Processing Challenges and Issues

- Framework able to run community codes while maintaining single provenance (and metadata) framework
- Provenance levels (granularity) needs to be same for all levels of products
- Lot of data is mutable
  - Lots of files are mutable
  - No mechanism to add characterization as more was learned about the data
- Scientific workflows are complex
  - Types of computing and analysis (hardware)
  - Geographical distribution of resources

Data Analysis and Processing Conclusions and Findings

- Overall Challenge: Expensive process with much variation in granularity and non-repudiation has not been adequately addressed
- Using ontologies to relate labels
- Using graphs for provenance – property graph model

Data Analysis and Processing Discussion Summary

- Provenance – how to prevent changes being made to what has already been done
  - DES- database ingested in production system; then goes to read-only database table to users.
  - Secure campus: off-campus user must use VPN or secure gateway to access system. Always issue of whether users on HPC systems have privacy protections vs. users running jobs and information should be accessible. Consider privacy from job angle.
- NCSA collect data at certain time of day or 24 hours? Is part of analysis how data is changing from one night to the next? Reduce amount of data stored? Also in context of scientific analysis.
  - Cosmological survey run by optical telescope is only run at night. Further constrained by 2 factors: 1) need to observe when object is directly overhead (avoid galactic cap) 2) moon, weather
  - Took data in northern hemisphere fall/winter to avoid galactic cap. Irregular processing/workload.
  - Astronomy has efficient domain-specific data compression to reduce data. 1TB or less/night was compressed data.

Data Analysis and Processing Discussion Summary

- Next 2 things needed for your system?
  - Clowder able to work on missing features; notes work with geospatial communities.

  - Yong Chen:
    - Working on prototype; representing provenance in property graph model. Assembling and processing pieces of code and get feedback from community.
    - Provenance survey – from HPC system and system administrator’s perspective – how to leverage provenance collected here and how to leverage and optimize valuable resources.

Data Management and Storage

- Ilkay Altintas
  - Position: Chief Data Science Officer, San Diego Supercomputing Center
  - Title: *Storage Challenges and Opportunities for Data Science in the Continuum*
- Ben Blaiszik
  - Position: Research Scientist, Globus, Univ of Chicago and Argonne National Laboratory
  - Title: *A Data Ecosystem to Support Machine Learning in Materials Research*
- Glenn Lockwood
  - Position: Storage Architect, Advanced Technologies Group, National Energy Research Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory
  - Title: Simplifying data management through storage system design
  - Note recent talk on *Understanding and Tuning I/O Performance ATPESC 2019* *Glenn K. Lockwood National Energy Research Scientific Computing Center Lawrence Berkeley National Laboratory* regarding the importance of developing new tools to understand I/O performance

Data Management and Storage

- Problems are increasingly data driven, need collaboration and are very heterogeneous.
  - Edge, cloud, HPC
- Data science requires data
  - Different modalities and life cycle of what goes into storage
  - Storage needs to be coupled with data management services and
  - Need dynamic storage optimization through intelligent workflows that can take advantage of underlying intelligent infrastructure
  - Access to datasets, models and codes
- Storage architecture is getting complicated
  - Not cost effective to deploy fastest possible for all

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Management and Storage Challenges and Issues

- Scalable data science requires effective storage management involving a number of storage modalities
  - Active vs. passive, hot vs. cold, various filesystems, object storage, …
- Storage needs to be coupled with data systems and data management services to enable the full data lifecycle - Findability, interoperability,
- Dynamic storage optimization using workflows
  - possible if we have intelligent storage systems
- Dynamic composition requires dynamic network, compute and storage resources combined with intelligent software for steering applications.

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Management and Storage Challenges and Issues

- Number of publications across domains is growing rapidly and access to datasets is improving but still a challenge
- Access to models and codes is a particular challenge
- Curating datasets
  - Asynchronous process whose requirements defined by each organization
  - Specific metadata fields may be required or reused

Data Management and Storage

- Challenges and Issues
  - Storage has gotten complicated
    - Deletion/retention policy and practice
    - Cost, duplication, hot/cold storage
    - Types of storage

  - Dealing with deeper storage pyramid
    - Burst buffer Fastest storage (Flash-based storage)
    - Scratch Tier: Large capacity, disk based storage system
    - Project tier/Community file system tier: higher capacity tier with more data resilience; not purged
    - Archive Tier: tape-based, cheap and high capacity, slow storage with tiers in between for users to move and store data in most economical location in relation to performance requirements (e.g., longer wait to access data from archive)

Data Management and Storage Conclusions and Findings

- Dynamic storage composability matters
- Storage systems are more useful if they are integrated as a part of application optimization workflow
- Data management needs to be coupled with storage
- Need to build cohesive infrastructure for AI-driven science
  - Data, compute, models, workflows
- Empower researchers to publish data, regardless of size, type and location
  - Automate data and metadata extraction
  - Enable unified search and discovery across disparate data sources
  - Deploy with APIs to simplify connection to other data efforts and enable automation

https://www.nitrd.gov/news/2019-MAGIC-Data-Life-Cycle-Series.aspx

Data Management and Storage Conclusions and Findings

- DLHub – a data and learning hub for science
  - Collect, publish, categorize models and processing
  - Serve models via DLHub operated service to simplify sharing, consumption and access
  - Mint persistent identifiers for all artifacts
  - Enable new science through re-use, real-time integration and synthesis of existing models
- Simplify data management through innovations in system hardware/software
- Manage complexity to simplify user interfaces

Data Management and Storage Conclusions and Findings

- Develop better middleware and tools: libraries and methods that applications can interact with, so user can write single source code and run at different facilities, using middleware tools/abstractions to hide these differences/characteristics from application
- Optimization by delivering better/smarter system software- make storage system more agile/responsive to users' intent (e.g., APIs for users to tag their metadata)
- Simpler storage system hardware (e.g., eliminating tiers in storage hierarchy to extent possible)
- Old way: storage expert for monitoring, not sustainable as storage systems become more complex.

Data Management and Storage Discussion Summary

- Experience with debugging pathological behavior of user activity in storage? When something acting poorly in storage system, TOKIO gives insight of everything else that is going wrong in system. If part of system is dead, key is better resilience.
- Security concerns with storage
- DL Hub- centralized service where calls come in, compute distributed among execution sites. Implementing model to pull down container and run locally. Sustainability (e.g., set up other sites)
- Rapid growth in volume of data from simulations and experimental facilities: how to aid science community in discarding data from data streams to enable effective storage?
  - NERSC: trying to provide foundational tools to enable communities to make their decisions – not qualified to do it for the communities (data management policy is to delete data after 30 days)
  - MDF: provide automated tools to index data to help inform users; decision
  - SDSC: Data triage is issue. One approach: allocate imagery and leave to users to decide

Interactive Analysis Tools and Services

- Alan Chalker
  - Position: Director of Strategic Programs, Ohio Supercomputer Center
  - Title: *Open OnDemand Overview*

- Shreyas Cholia
  - Position: Group Leader, Useable Software Systems, Lawrence Berkeley National Laboratory
  - Title: *Jupyter - An Interactive Platform for Scientific Computing*

Interactive Analysis Tools and Services
Open on Demand

- Finished OnDemand 1.x;
- Developing OnDemand 2.x (Jan 2019-Dec 2023)
- NSF-funded open-source HPC portal enabling interactive applications and cluster access.
- Goal: to provide an easy way for system administrators to provide web access to their HPC resources.
- Benefit: Lowers barrier to entry; faster for science

Interactive Analysis Tools and Services
Open OnDemand 2.0 Project Overview

- 4 Main Areas:
  - Visibility: Enhance resource utilization visibility by integrating existing Open XDMoD platform
  - Scalability: Support more types of computing resources
  - Accessibility: appeal to more scientists from different domains
  - Community engagement: building community and making it a community-driven project
- Evaluating Open OnDemand: will provide account if interested

Interactive Analysis Tools and Services
Jupyter

- Jupyter Notebook provides interface for interactive computing for science
- Jupyter enables web-based notebook interfaces, which can combine backend computing language kernels (Python, R, Julia etc.), documentation elements, and interactive frontend visualizations and widgets
- Widely used across scientific domains
- Jupyter platform enables scientific data analysis and exploratory computing

Interactive Analysis Tools and Services
Jupyter

- New developments in the Jupyter ecosystem:
  - Reproducible Research (mybinder.org: shareable reproducibility from public git repository)
    - Binder provisions reproducible environments for Jupyter
    - BinderHub: Jupyter Hub + Binder provides complete software environment
      - Built on Kubernetes, cloud-agnostic, scalable, community driven and deployable by anyone
      - Jupyter Hub provides multiuser support for Jupyter – centralized deployment, access to big data sets and enable shared workflow and results
  - JupyterLab
    - Common backplane that provides an integrated environment for Jupyter based tools.
    - Goes beyond notebooks to bring in other applications

Interactive Analysis Tools and Services Challenges & Issues

- Enable Jupyter at scale on big systems and large datasets
- How do we connect and capture distributed facilities and workflows in Jupyter?
- How do we enable reproducible science at scale? Binder for HPC?
- Seeking more input from community (e.g., reproducible science at scale, Binder HPC)
- Jupyter R&D at LBL:
  - Superfacility- How Jupyter can act as the primary Superfacility interface
  - Usable Data Abstractions – Enabling access to very remote datasets through Jupyter and Jupyterlab
  - Jupyterhub – Software development to enable integration of Jupyterhub with NERSC systems in a scalable, secure and intuitive manner

Interactive Analysis Tools and Services Discussion Summary

- Fairly broad adoption across the world
- Data integrity and security- taken seriously, talking to Thomas Mendoza* (Lawrence Livermore) who is working on secure, end-to-end installation of Jupyter, from kernel to end user. Mendoza will be speaking on Jupyter Security at Trusted CI Webinar, NSF Cybersecurity Center of Excellence on 9/23/19
- Binder integration with Google – on paid basis. Many projects to integrate with HPC, Binder?
- Binder uses Kubernetes to spin up containerized environment.
  - Roadblock: Kubernetes not running on Cori system.
- Assume Binder uses Kubernetes, but can't do it yet. Can implement Binder on own Kubernetes infrastructure.
- Reach out to Shreyas with feedback. Can pass along specific use cases to Jupyter team. Jupyter can do a lot more in the HPC, large scientific workflows
- Rick Wagner working on integration of Globus Auth

Data Triage and Discovery

- Dr. Francine Berman
  - Position: Hamilton Distinguished Professor of Computer Science, RPI, Co-founder, Research Data Alliance
  - Title: *Organizational challenges to promoting data sharing, stewardship and preservation*
- Hubertus van Dam
  - Position: Application Architect, Brookhaven National Laboratory,
  - Title: *Online data analysis of molecular dynamics simulations for exascale computing platform*

Data Triage
Challenges and Issues

- I/O on HPC not keeping up with peak performance
  - Compute-Data gap presents challenge for the exascale
- Center for Online Data Analysis and Reduction (CODAR) Collaboration with NWChemEx project. Examined 2 applications:
  - Simulate behavior of molecules and calculate trajectories as evolve over time
  - Developing codes for exascale computers and want to analyze codes' performance; worked on extracting data from these applications as applications are running
- Chimbuko: Performance analysis tool to examine the performance of scientific codes as the data comes off simulation (e.g., anomaly detection)

Data Triage
Findings and Conclusions

- Solutions:
  - De/compress data (use substitute compute cycles to do IO) or
  - Online data analysis (no data storage)
- Change to the practice of data analysis is needed. Need model able to compute-and-analyze on-the-fly.
  - Applied to molecular dynamics simulations with NWChemEx for scientific data as well as performance data

Data Discovery

- Data drives discovery
- Data is an effective and critical tool to address problems
- Stewardship, preservation, infrastructure and tools are needed to support data sharing and data-driven discovery
  - Need to be able to find data and know what it means
  - Ecosystem of tools and structure make data useful for analysis
  - Data needs to be preserved for results to be reproducible
  - For effective data stewardship and sharing - all levels, from the user level on up, must work together

Data Discovery
Challenges and Issues

- Stewardship in government, academia, industry
  - Stewardship, preservation and data use more problematic in academia, where there is very little stakeholder alignment, than industry
- Stable organizational support for stewardship and preservation a harder sell in academia:
  - Fear of economic commitment - Costly maintenance and economic sustainability issue
  - Lack of newsworthiness- not “moonshot” research
  - Misalignment with conventional incentive structure - no recognition for infrastructure and difficult to measure success/impact of effective data infrastructure

Data Discovery
Findings and Conclusions

- Stewardship, preservation and use of the data that drives modern discovery require an ecosystem of researchers, beneficiaries, stakeholders and organizations to operate effectively
- Move towards a solution to the lack of adequate stewardship and preservation infrastructure
  - Increase importance of infrastructure – make it sufficiently important
  - Identify effective research/infrastructure ratio
  - Focus on value - Everything should have a sustainability model that makes business sense

Data Triage and Discovery Discussion Summary

- Impact of ML and IoT
  - Research Data Alliance community volunteering to build infrastructure together in a way that promotes effectiveness and usefulness
  - ML calculation under the covers; hard to figure out. Will require sea change in thinking - meaning, retention, who does it, etc. No good answer but requires a different approach
- NASA leads in handling unique data sets. Lessons from NASA experience?
  - Many stakeholders align better with NASA (have instruments, data storage; recognize importance). Thus, NASA closer to commercial entities, which helps community stability.
  - NASA very engaged with community. NASA and partners have been thoughtful about standards, which makes a difference in data usefulness/ tools used to ascertain meaning of data.
  - From researcher's point of view: get most longevity out of data due to being part of infrastructure, stakeholder alignment and more applicable standards
  - Data collected for one mission is sometimes recycled decades later for different purpose (note Data Management Challenges – reuse data sites for other disciplines; findability problem; unpredictability experienced in data world not accepted in other areas, international data set challenges)
  - Need narrative to support stability in data management and preservation

Data Triage and Discovery Discussion Summary

- Success stories:
  - Databurst and Internet archives
  - See incremental path - more realistic but, difficult to stay the course
- Trends in how institutions view their responsibilities
  - Concept of data management plan helped shape thinking of data as a “first class object” in research world
  - While agencies experimenting more with funding technical and social infrastructure around data, universities experiencing difficulties in being run as businesses. 1% universities have more discretionary funds. More awareness at national level that someone needs to be responsible, but universities can’t pick up slack
  - Seeking economies of scale; different approaches. If get CIOs, libraries together, VP of Research on same page, seems to work.
  - Continue this discussion in Coalition for Academic Scientific Computation (CASC) that has been discussing policy, technology overlaps for more ideas

Data Triage and Discovery Discussion Summary

- Infrastructure inadequate for science proposals
  - Certain proposals not offered because infeasible or would be declared such. Researchers pitch “safe” proposals within a range of reasonableness

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov