

# NSLS-II Data Policy

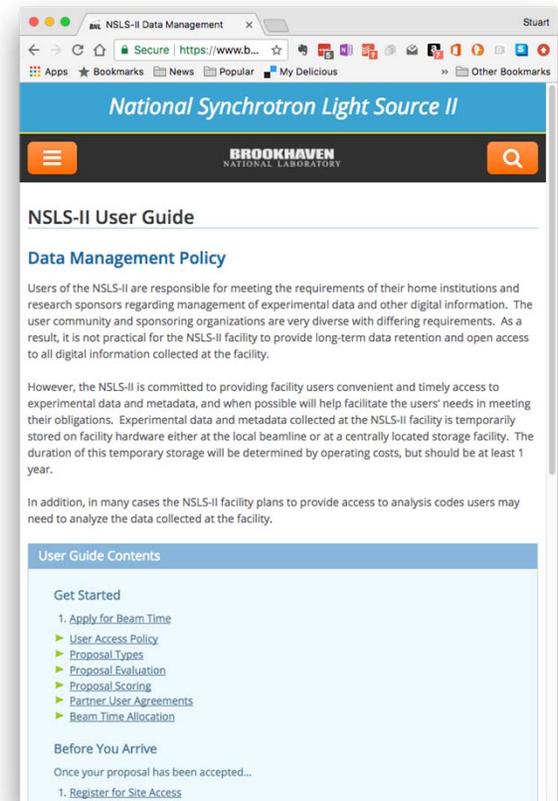
Stuart Campbell  
*5-way Meeting October 9<sup>th</sup> 2017*



# DATA RETENTION POLICY

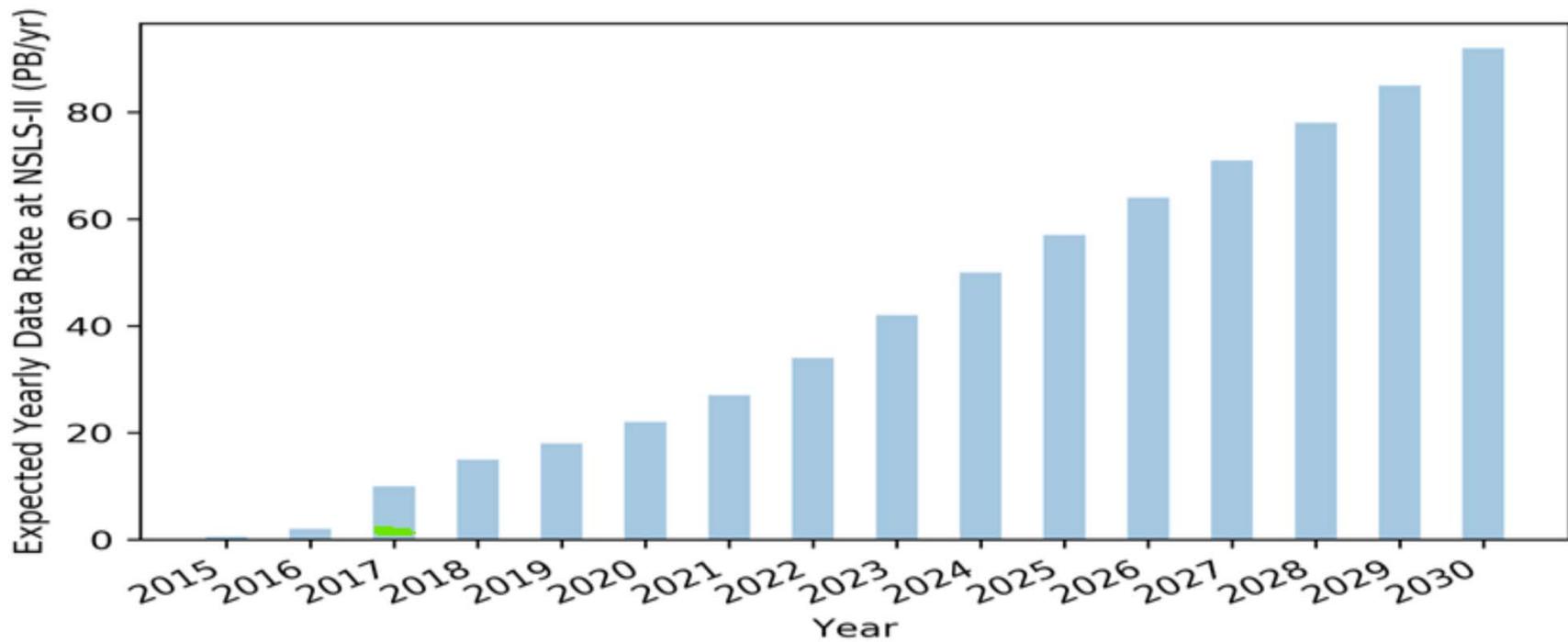
# Data Management Policy (Summary)

- Users are responsible for meeting home institution & sponsors data management requirements
- NSLS-II will not provide long term storage
- Temporary storage for a period determined by operating costs (should be at least 1 year)
- provide access to analysis codes users may need to analyze the data collected at the facility.



## What we are currently doing

- Able to keep all data so far...



- Estimated for the 28 NSLS-II beamlines under development
- Fortunately, we are not generating huge amount just yet – but it's coming...

# External Data Processing and Access (Plans)

- Central 'processing/analysis' facility
  - Accessible from offsite
  - Front end login servers reside at NSLS-II
  - Seam-less use of compute and storage from NSLS-II and BNL Central Computing
- Different access for different usage requirements
  - Command line access
  - Complete Graphical Desktop
  - Python Notebooks
  - Custom web applications
- Data Transfer off site
  - Globus, Cloud (Dropbox, Google, Amazon, etc..)

# EXTERNAL RESOURCES

# NERSC

- Startup Allocation for NERSC
  - OK for prototyping and gaining experience
  - Allocated 50,000 CPU hours for the next 18 months
- Prototype pipeline for submitting jobs
  - We've had some initial discussions with NERSC over collaborating regarding some of their efforts deploying JupyterHub (and Dask). This would allow users to make use of NERSC resources from a user friendly interface at BNL.

## NERSC Data Storage ?

- Planning test data transfer rates from NSLS-II → NERSC
  - Networking Issues need to be resolved
  - GlobusOnline (should this be a standard)
  - SFTP/SCP
  - What standard data transfer mechanisms should we support?
- Are we staging data for compute only ?
- Is longer term storage feasible ?
  - If so, who pays ?

## Brookhaven Resources

- Computational Science Initiative (CSI)
  - Initial testing and access to cluster
  - Many common R&D projects
  - Offers a potential storage solution  
\$8.33+overhead per TB/month
- Working towards a single user authentication across BNL

# WORKING TOGETHER ?

## Current Activities

- SLAC
  - Data Acquisition: bluesky and ophyd
  - Graphical interfaces – looking at PyDM
  - Overall excellent start to collaboration
- APS
  - Software R&D working group formed
  - Developed Plan
  - Many common projects, e.g. bluesky, tomopy, ...

# LEAPS



- “League of European Accelerator-based Photon Sources ”
- Started to Engage with WG3 (Lead: Mark Heron)
- Learn about best practices
- More potential collaborations

# LEAPS WG3 (Information Technology)

## Address Challenges in:

- IT governance, IT security, IT procurement strategies
- Harmonised open data policy, data format, meta-data capture, e-log-book, federated identity management, data archival and curation
- High-speed data acquisition, visualisation, and reduction
- Utilisation of European Open Science Cloud Services for data analysis and storage.
- Data analysis software developments and software optimisation
- Training of data scientists and IT specialists
- Tutoring services for data analysis
- Networking with e-infrastructure providers and stake-holders, EU and world-wide
- Networking with other scientific communities
- Outreach



## What can we do together?

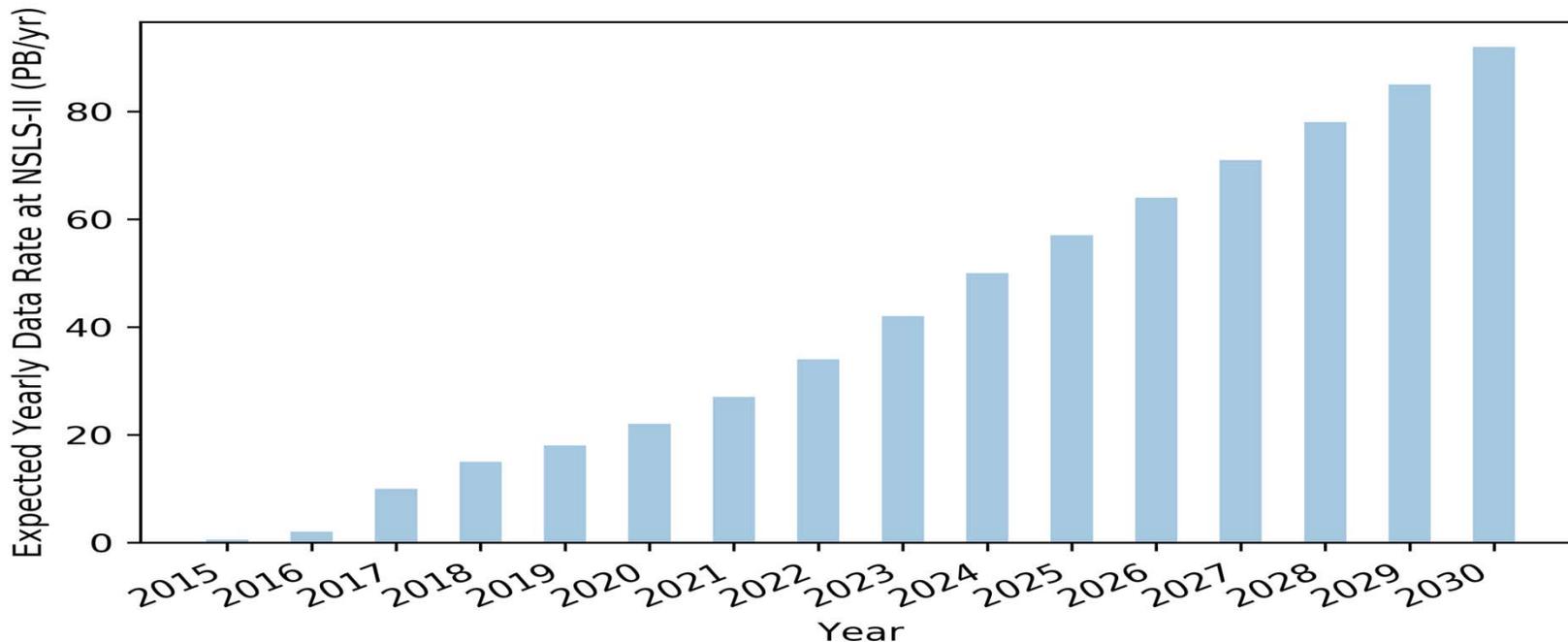
- Consistent data management policy
  - Life time of data
  - Embargo period
- Long Term Storage
  - Is it possible to get core funding for NERSC
- Development of tools that bridge the gap between facility users and computational resources
- Develop Metadata standards



# THANK YOU

# ADDITIONAL SLIDES

# Estimated NSLS-II Storage Needs - ?



- Estimated for the 28 NSLS-II beamlines under development
- Was based on current information on operating beamlines
- Takes into account estimated increases in both detector and beamline operational reliability.

## Data Management Policy

- Users of the NSLS-II are responsible for meeting the requirements of their home institutions and research sponsors regarding management of experimental data and other digital information. The user community and sponsoring organizations are very diverse with differing requirements. As a result, it is not practical for the NSLS-II facility to provide long-term data retention and open access to all digital information collected at the facility.

## Data Management Policy

- However, the NSLS-II is committed to providing facility users convenient and timely access to experimental data and metadata, and when possible will help facilitate the users' needs in meeting their obligations. Experimental data and metadata collected at the NSLS-II facility is temporarily stored on facility hardware either at the local beamline or at a centrally located storage facility. The duration of this temporary storage will be determined by operating costs, but should be at least 1 year.

## Data Management Policy

- In addition, in many cases the NSLS-II facility plans to provide access to analysis codes users may need to analyze the data collected at the facility.

# Interactive Data Browser

Search by any metadata field, including custom ones invented at experiment time.

The screenshot displays the Interactive Data Browser interface. On the left, a list of search results is shown, with a red arrow pointing to the search input field above it. The results include fields like 'scan', 'count', and 'relative\_scan'. The middle section, titled 'View Header (metadata):', shows a tree view of metadata for a specific scan, including details like 'detectors', 'plan\_args', and 'stop'. The right section, titled 'Interactive figure', shows a line plot of 'noisy\_det' versus 'sequence #' for scan\_id 337. The plot shows a sharp drop in the noisy detector signal at sequence 3.0. Below the plot is a toolbar with navigation and analysis icons.

**Search results**

- scan ['b307a...']
- scan ['c30c81']
- count ['d03544']
- count ['0377ef']
- count ['fdbf13']
- scan ['4a188b']
- count ['33f002']
- scan ['ae831b']
- relative\_scan ['39a98e']
- scan ['8b4c9a']
- count ['74a26f']
- scan ['1da333']
- scan ['0e0cc0']
- count ['20bc5f']

**Metadata**

View Header (metadata):

- 1
- descriptors
  - start
    - a: 1
    - detectors
      - 'noisy\_det'
      - num\_steps: 4
      - operator: 'Ken'
    - plan\_args
      - detectors
        - 'SynGauss(name=noisy\_det)'
        - num: 4
        - plan\_name: 'count'
        - plan\_type: 'generator'
        - proposal\_id: 3
        - sample: 'B'
        - scan\_id: 337
        - some\_calibration\_number: 1.6
        - time: [2017-01-21 21:47:48] 1485053268.77...
        - uid: '0377efb9-b21d-4786-bc59-e9282fd3c...
  - stop
    - exit\_status: 'success'
    - run\_start: '0377efb9-b21d-4786-bc59-e928...
    - time: [2017-01-21 21:47:48] 1485053268.83...
    - uid: 'b15a395d-0bd3-4dfe-a676-39fe7f786...

Export Events (data):

CSV Excel Copy UID to Clipboard

**Interactive figure**

Allow overplotting

noisy\_det

sequence #

scan\_id 337

Data acquisition system (Bluesky) captures some metadata automatically and provides many ways for the user to provide more. This makes searching, provenance, and analysis easier.

# Development Roadmap

- Version 1.0 release (end of FY17) Highlights include:
  - Step scans (N-dimensional, complex spiral trajectories, etc)
  - Fly scans
  - Reciprocal space scans
  - Adaptive scans (e.g. feed measurements back into plan)
  - Sophisticated streaming analysis (e.g. live tomographic reconstruction)
  - Manual and automatic suspend/resume (e.g. in response to beam dump)
  - Simulation mode
  - All known hardware on operational beamlines is supported
  - Automated end-of-run data export to TIFF, CSV, HDF5, etc.
  - Sample management inventory database.
  - Provide basic feature for commenting on or tagging datasets
- Currently feature complete
  - Code re-structuring, more documentation, more testing

# Development Roadmap

- Version 1.1/2.0 release (FY18/19) Highlights include:
  - Improved graphical tools for experiment setup
  - Domain specific graphical tools for collection
  - Remote access to streaming data
  - Role based security
  - Improved graphical tools for data discovery
  - Improved provenance handling
  - Integration of cloud services such as Dropbox, Google Drive and Amazon Web Services, etc.
  - Integration of Globus Security and Transfer
  - Support for more data export formats
  - Integration with site wide proposal system

## Develop Meta-Data Standards

- Development of a standard meta-data dictionary to allow data sharing
  - Between beamlines
  - Between facilities
  - Enables Multi-modal experiments
- Engage with beamlines, scientific areas, other facilities and organizations
  - S. Campbell is now a member of NeXus International Committee
  - Excellent initial progress (agreed 16 core set of meta data items with APS/ALS/ORNL)