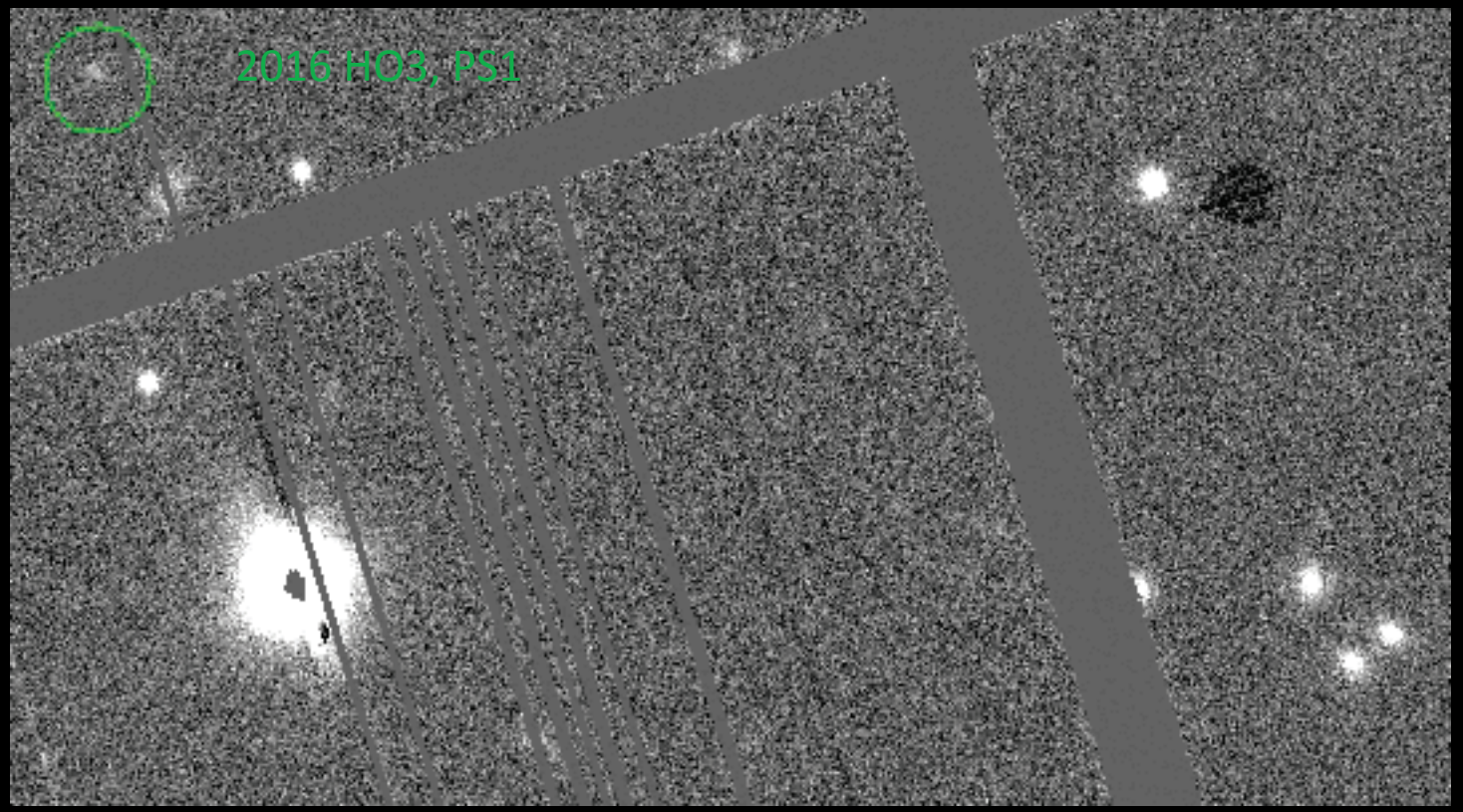
NEO Discovery, Characterization and Data Storage



Gerbs (James) Bauer, UMD, NASA PDS-SBN PI

Introduction: Data Storage is the Memory of any survey

- Data Storage is important to the on-going survey
  - Calibration is based on storage products
  - Basis for inertial source removal and identification
  - Iteration and improvement of the processing pipeline
  - **Necessary for the synthesis of chains of detections**
- Important to the immediate users & receivers of survey data
  - PDCO
  - CNEOs
  - MPC
  - IAWN
- Critical to future users
  - Characterization
  - Pre-covery
  - De-biasing/population studies

Data Storage necessary for any meaningful analysis of the data

Why Worry? Expanding Datasets forthcoming…

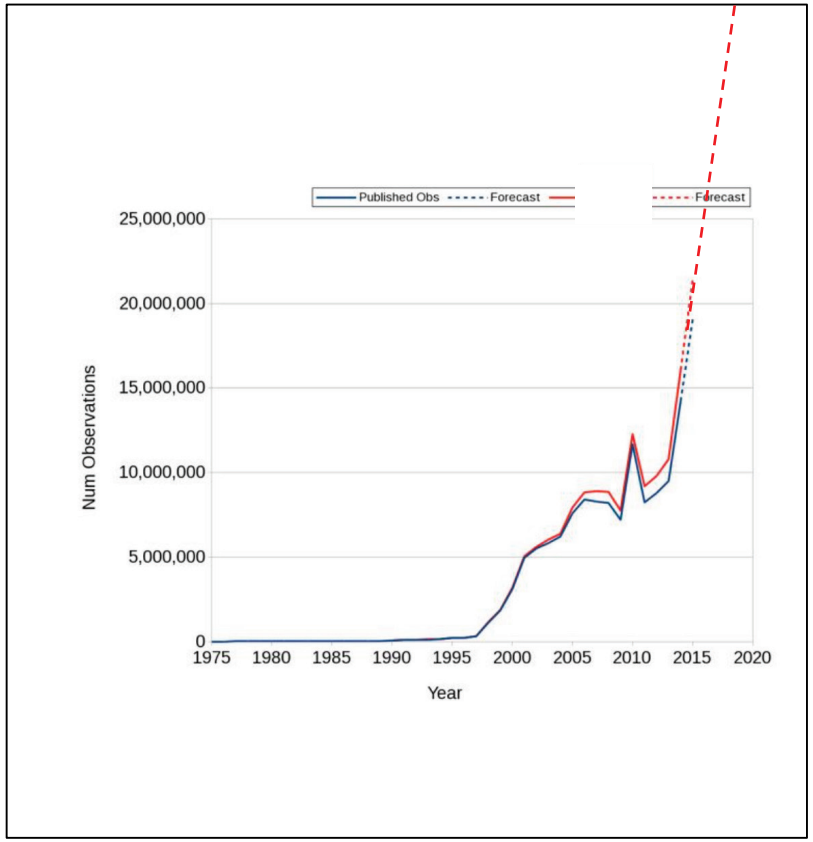
MPC Drivers of Growth

Generalized survey ramp-up

- ~68,000 observations in 1990
- ~10 million in 2007
- ~14 million in 2010
- ~ 25 million observations per year currently

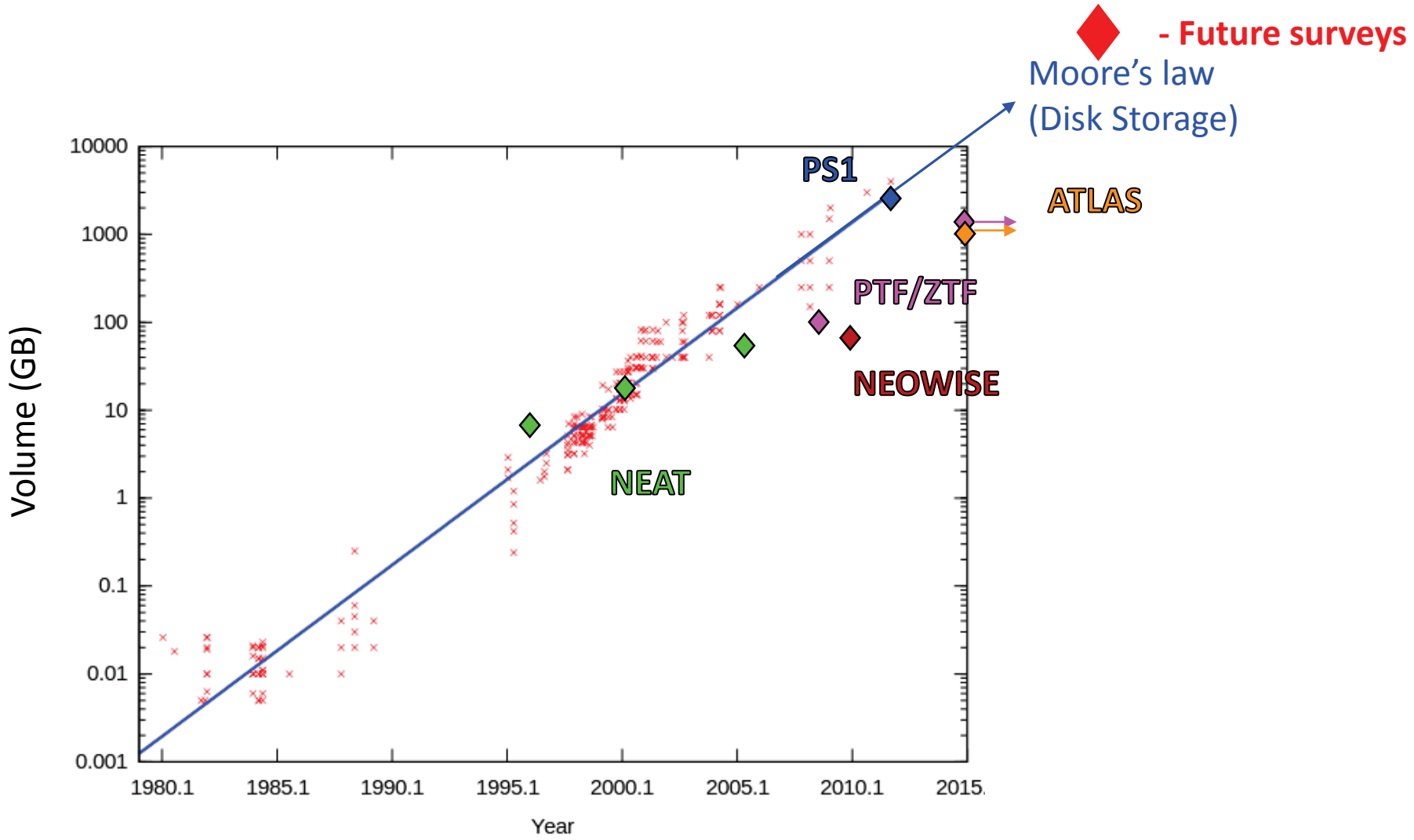
Future Surveys

- Orders of Mag more data



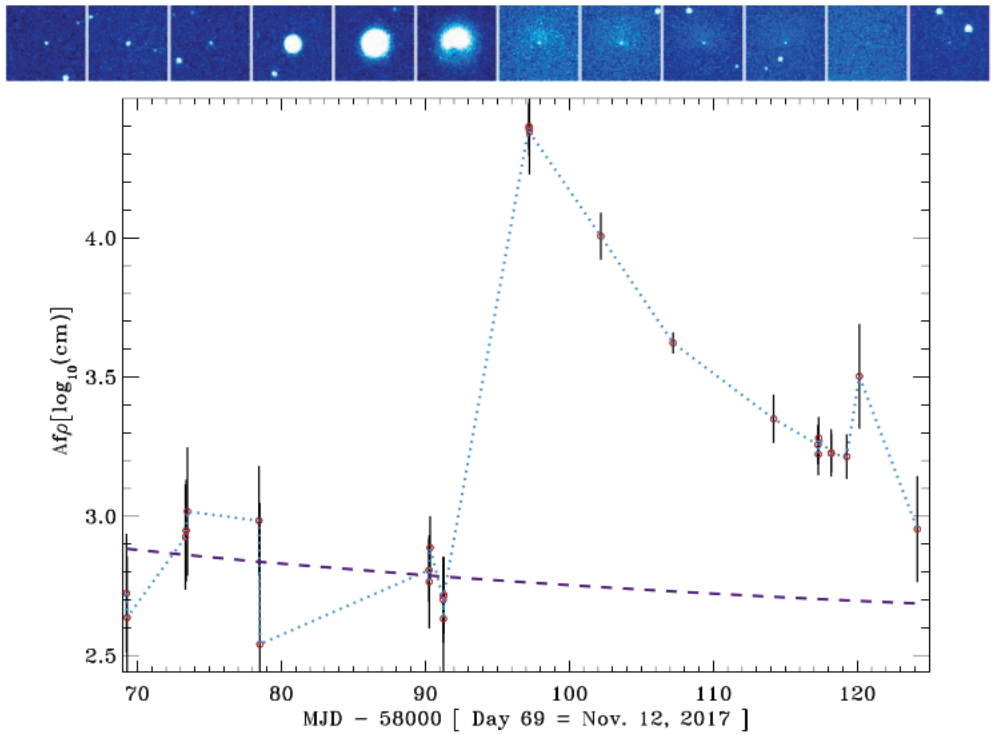
~200 Million Observations/Month (**several million/day**) with NEOCAM, LSST and other surveys

NEO Surveys & Data Storage – 1 "Night's" data



NB: demonstrative, not comprehensive

Purposes of Data Storage I



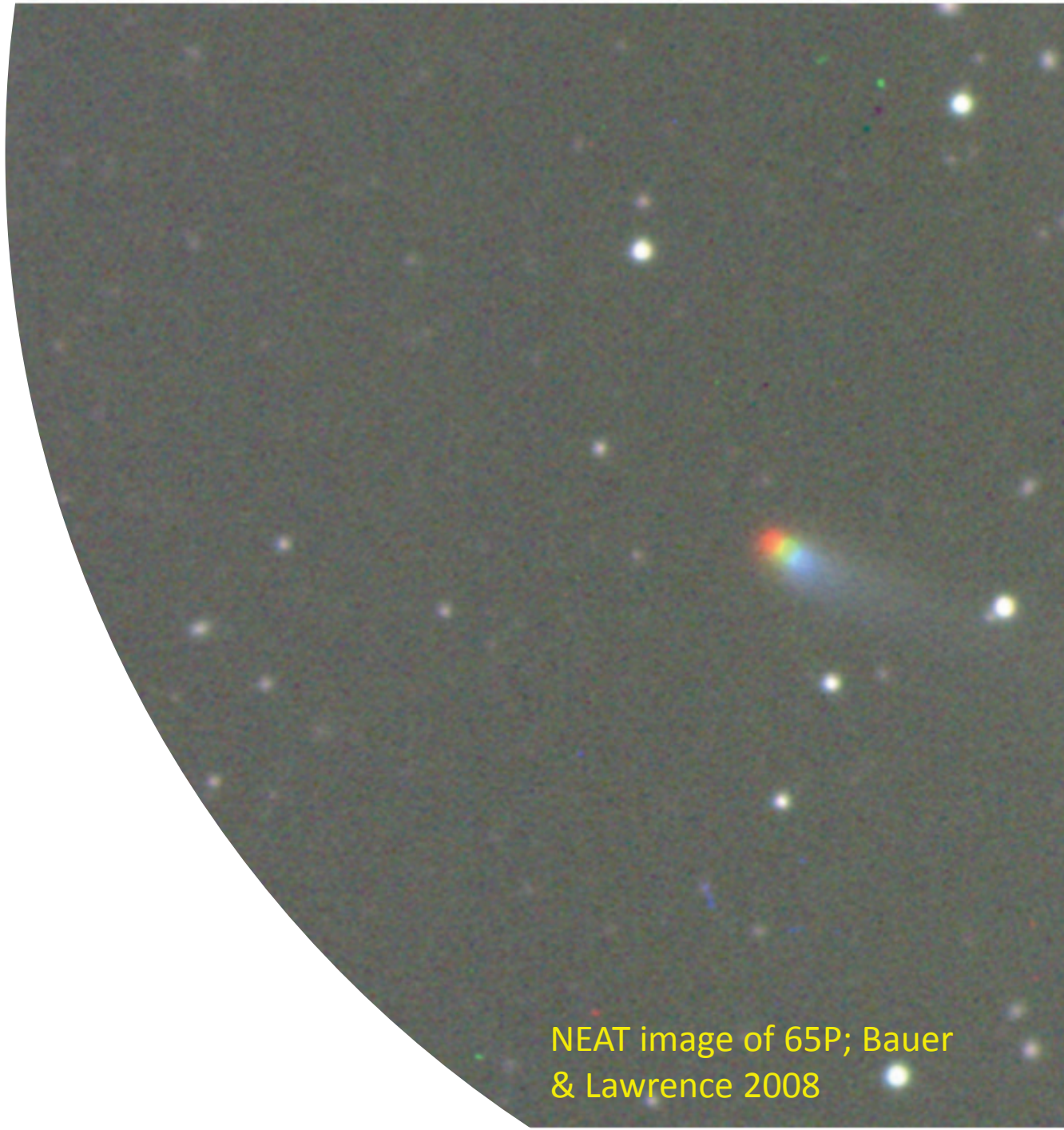
ZTF images and photometry of Echeclus (Bellm et al. 2019)

Short Term:

- Verification/validation of detections and associated discoveries.
- Quantify Uncertainties and systematic sources (trailing, seeing, scattered light).
- Rapid Analysis of features (H, Δmag).
- Immediate Identification of unique and classifying features (coma, companions).

Purposes of Data Storage II

- Long Term:
  - Re-visit most of Previous…
  - Linking detections across epochs or oppositions
  - Pre-covery (Different!! – stacking, lower SNR thresholds, iterative regional searches).
  - Behavior over time(LC variation, Phase response, activity at different epochs).



NEAT image of 65P; Bauer & Lawrence 2008

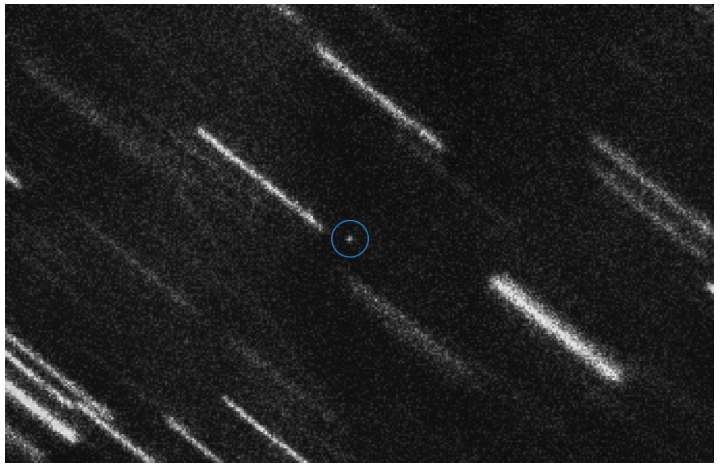
Data Types (NEO Surveys)

- Detections -> MPC/Archives/Brokers

Example XML file structure

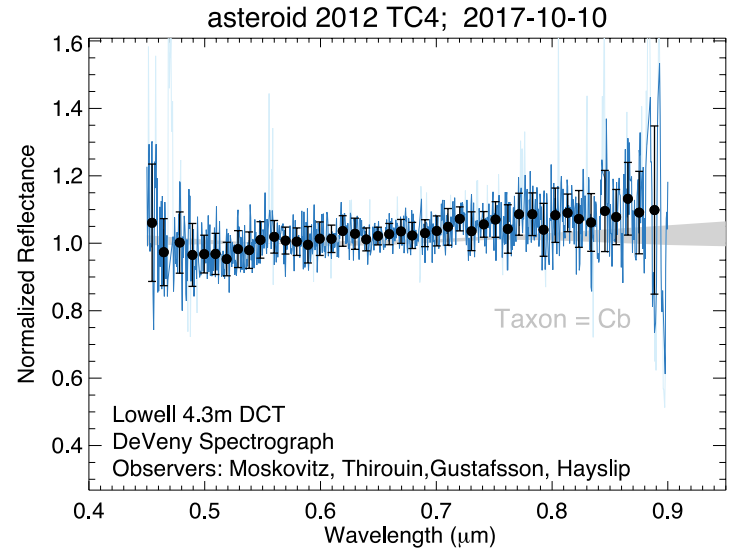
```
<observationBatch>
<observationContext>
 <observation>
  <count>12</count>
 </observation>
 <observatory>
   <mpcCode>F51</mpcCode>
   <name>Pan-STARRS 1</name>
 </observatory>
 <contact>
   <name>P. Villa</name>
   <ackMessage>Sample header</ackMessage>
   <ackEmail>pancho.villa@gmail.com</ackEmail>
 </contact>
 <telescope>
   <aperture>1.5</aperture>
   <design>Reflector</design>
   <detector>CCD</detector>
 </telescope>
 <observers>
   <name>P. Villa</name>
   <name>F. Madero</name>
 </observers>
 <measurers>
   <name>P. Villa</name>
   <name>F. Madero</name>
 </measurers>
</observationContext>
<observations>
 <optical>
  <provID>2015 HD76</provID>
  <trkSub>aa</trkSub>
  <mode>CCD</mode>
  <notes>Klmn</notes>
  <stn>F51</stn>
  <prg>3l</prg>
  <obsTime>2015-04-25T12:32:34.5</obsTime>
  <ra>215.65605</ra>
  <dec>-13.547872</dec>
  <astCat>2MA</astCat>
  <rmsRA>0.15</rmsRA>
  <rmsDec>0.13</rmsDec>
  <rmsCorr>0.21</rmsCorr>
  <mag>21.9</mag>
  <photCat>2MA</photCat>
  <band>w</band>
  <rmsMag>0.25</rmsMag>
```

- Images ->Archives



VLT 2012 TC4, 2017.08.06

- Other Characterization -> Archives/Publications



Detections

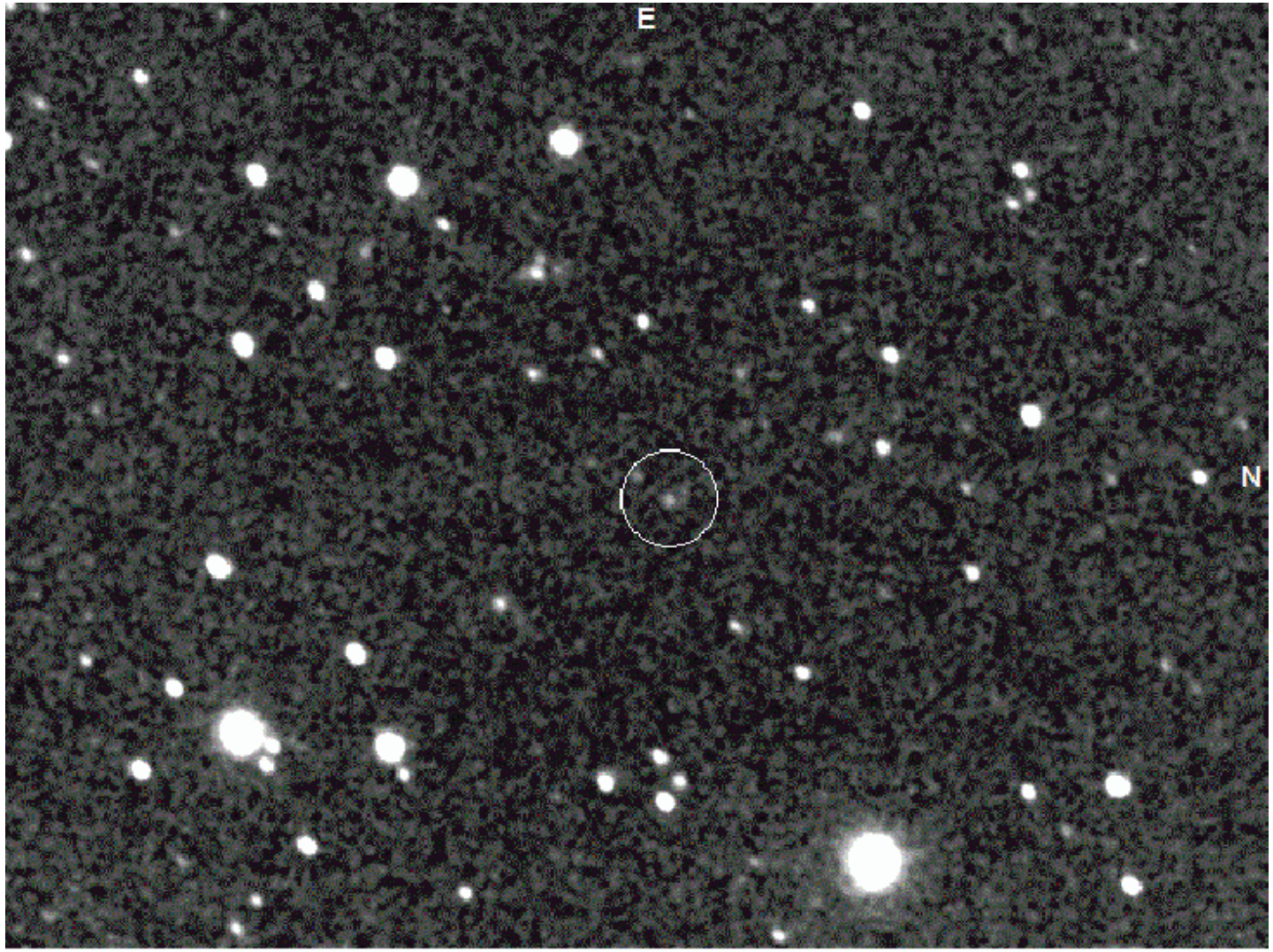
Courtesy T. Grav, PSI

Cyan – Earth
Blue – Mercury, Venus & Mars
Grey – MBAs
Green – NEOs
Yellow - Comet

- Astrometry
  - Uncertainty
  - Position Reference Grid & Epoch
  - Temporal accuracy and time (UTC? Ephemeris Time? Leap Seconds?)
- Photometry
  - Uncertainty
  - Reference Catalog
  - Bandpass
  - Temporal Variations
- Special Cases
  - Radar
  - Infrasound
  - Etc.
- Specific information
- Less Volume
- Less context
- Potentially less cross-disciplinary science
  - Depends on Selection Criteria and extracting information…
  - Can serve to guide sub-frame extraction, but still need the …

Images

- Artifact ID, background and noise source identification
- Background/foreground object subtraction
- Extended objects
  - Gradients and SBP
  - Morphological analysis
- Faint source extraction and Stacking
  - Known sources
- Re-calibtration
  - Improved processing
  - Improved reference grids
- **Largest Data Volume**



361P/Spacewatch, m=18.6, 2017/12/10 0252-0322UT, 3x600s GIF, FOV 13'x10'
11" SCT f/6.4 STF-8300M 1.24"/pixel Mike Olason, Denver Colorado

Other Characterization Data

- Radar
  - Ranging
  - Shape
  - Rotation
- Polarization
- Spectroscopy
  - Composition
  - Surface properties

2017 YE5

- Data volumes small, and quantities are relatively small. Sample size considerably smaller; typically targeted observations.

Storage Vs. Archive

- Complete Capture of Data
- Ready Data Formats
- Access to Data by Science Team
- Useable (with some effort)
- Multi-Level Data Products
  - Including ancillary products for debugging

- Long-Term Stability
- Simple Data Formats
- Easy Access to Data
- Searchable
- Higher Level Data Products
  - Calibrated, cleaned products… these will get the most use.

Archives: Accessibility/Discoverability

- Previous Emphasis was on Storage Longevity and Future Usability
  - Not necessarily on **EASE** of use…
- New Emphasis on Data Discovery
  - More extensive and descriptive meta-data
  - More sophisticated search tools
  - More automated delivery to data distribution nodes
- Cross-Discipline Usage
  - Format conversions
  - Complete descriptions of data and tools

Archives: F.A.I.R.

- Findability,
- Accessibility,
- Interoperability, and
- Reusability

Provide a more consistent mapping between data products, analysis, and publications.

New Data *Archive* Developments: DOIs

Data Object Identifiers

- Collections of data used in generating analysis products
- Proper referencing and tracking of archived data usage
- Attributes for authorship and refereed status (if data is reviewed)
- Immediate notice by data distribution nodes – common coin of exchange
- MAST and PDS issuing DOIs for their data collections

Backup and Volume

Currently:

- Primary
- Secondary (on site)
  - Disk failures, etc.
  - Immediate access
- Remote-site backup
  - Often NSSDCA or like
  - Week or more to restore
  - **No cost**, takes all data

- Doubles size of Necessary Disk Capacity
- ***No cost to restore or access***

Backup and Volume

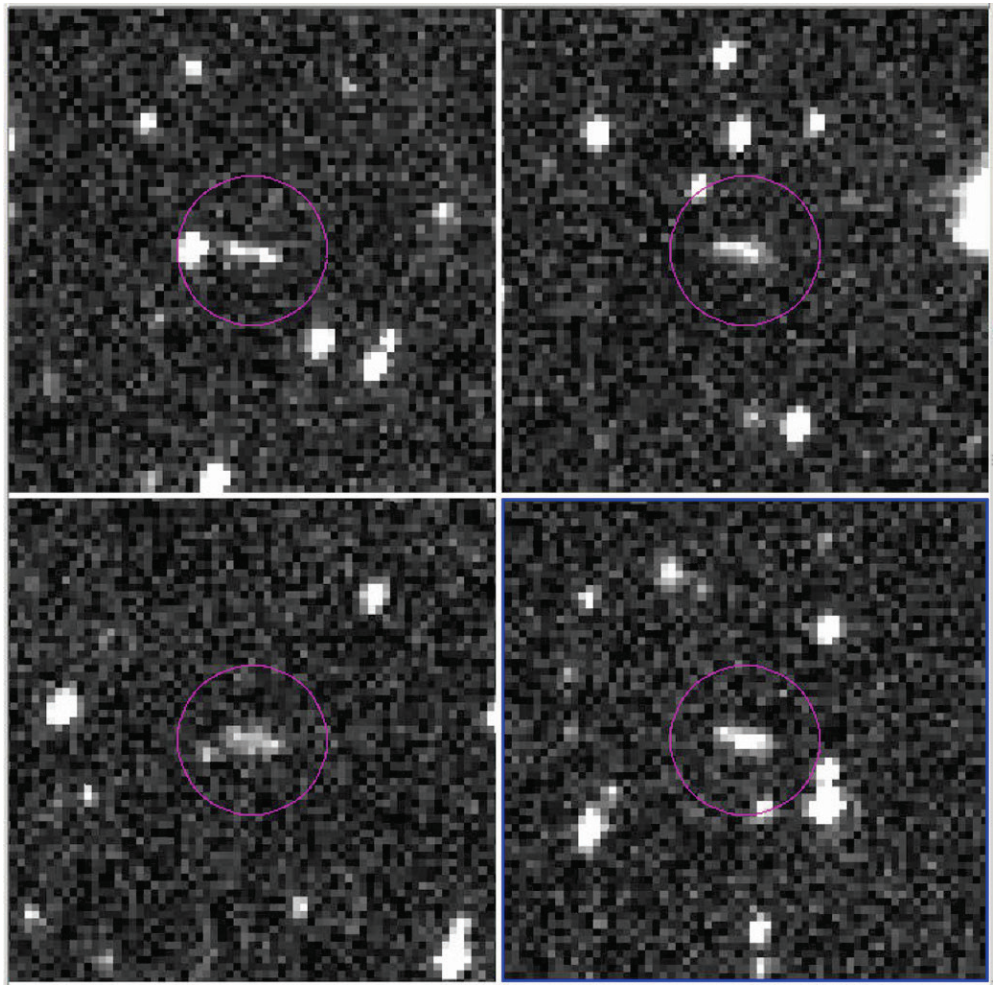
Near-future…

The move to cloud-based storage:

- Basic problem with costing – pay to download from cloud, and its per-file and volume-scaled.
- First as backup is feasible
  - Costing is not straight-forward, but reasonably competitive with on-site storage units & maintenance costs
  - Restoration costs (and pacing), but can be reserved
  - If volumes are ~PB, may become too costly
- As Primary…
  - Much more costly (who pays?)
  - Access to web-based computing tools

New Data Storage Developments: The MPC – planned data distribution enhancements

- Database driven processing (more automated)
- Distribution of Live copy of MPC database
- APIs to access live DB and queries
- Can be used to enhance storage products in near-real-time



New Data Storage Developments: The MPC

Generalized MPChecker

Goal

- Statistically robust attribution of detections / tracklets to known orbits.

Requirements

- Accurate integration of orbits, incorporating multiple non-gravitational forces
- Robust generation of covariance statistics for orbital fits
- Rapid propagation of orbits (& uncertainties) to generate statistically robust uncertainty regions on the sky
- Rigorous criteria for the association of candidate detections / tracklets with propagated catalog orbits
- Parallel (test) deployment: May - July 2019
- **Can use to populate attributes for known objects in frame**

Conclusions

- Data Storage will grow, but not at such an alarming rate…
- Surveys are getting deeper and more sensitive– growth in **detections**.
- Still need to save the images(!), and MORE metadata and products for discoverability and F.A.I.R.
- Alternative storage vehicles (e.g. the cloud) and backup schemes are progressing, but presently have some limitations (potentially unconstrained costs – will likely be solved soon).
- More tools are coming online to enhance data storage products in real-time or near-time processing.