Data Management Report-Back

For Eli Dart & Sasha Rahlin; Debbie Bard & Ted Kisner; Sara Simon & Andrea Zonca; Colin Biscoff & Reijo Keskitalo; Don Petravick & Nathan Whitehorn; Tom Crawford

CMB-S4

Data Movement (Eli & Sasha)

- What does it mean to provide access to the raw data to the collaboration?

  - Data exploration? Yes, facilitated by Data Reduction data quality data products and visualization tools

  - Experiment characterization? Yes, facilitated by Data Reduction instrument/observation characterization tools

  - Algorithm design? Yes, facilitated by Software Infrastructure and Data Simulation tools

  - Small-scale re-processing? Yes, facilitated by Data Reduction tools

  - Bulk processing/re-processing? No, we don't have the cycles

CMB-S4

Software Infrastructure (Debbie & Ted)

- What is the interface between Software Infrastructure and Data Movement for pipeline data/metadata indexing & provenance tracking?

- How does the development plan avoid parallel efforts for short- and long-term use-cases (cf. DAQ)?

  - Data challenges exercise Software Infrastructure as part of the whole DM subsystem throughout the construction project.

CMB-S4

Data Simulation (Sara & Andrea)

- Can we obtain extragalactic foreground simulations consistent with multiple cosmologies (eg. null & threshold parameters to test for false positives & negatives)?

    - Need to work with simulations teams (eg. Websky)

    - Need a pan-experiment extragalactic foregrounds group!

- Need to evaluate the trade-offs between systematics simulation and mitigation vs simply simulating residuals.

Data Reduction (Colin & Reijo)

- Can data quality inform the optimal use of South Pole bandwidth?

  - Yes, and we need to evaluate this - where would we most benefit from more time to understand the data?

- Does machine-learning have a role in data quality?

  - Yes! For many reasons (data volume to funding opportunities).

  - Liaise with SO ML DQ group.

- Experiment characterization tools are particularly critical for commissioning.

- Given finite resources, what is the best "well-characterized single-frequency map" for each science case?

  - Evaluation of alternatives at CD-1; selection by CD-2

CMB-S4

Transients (Don & Nathan)

- Does the transients element change with the event rates implied by SPT (~500 alerts/year)?

- Since most SPT/ACT transients are short, do we need to increase the cadence and/or reduce the alert time? What is possible at each site?

- With so many alerts, do we need a hierarchy of interest?

- What does "an alert" consist of? Does it have multiple stages? If so, who is responsible for each?

- Are we tracking variability as well as transients?

This is an evolving science case, possibly implying evolving requirements - how does the DM subsystem respond to that?

Site Hardware (Tom)

- What is the right degree of redundancy for South Pole site storage?
  - Current plan is 2 copies x 12 months of data, but this is not enough:
    - Time between shipments could be as much as 16 months
    - Overhead of RAID storage
    - Overhead of disaster recovery
    - Very low cost compared to obtaining the data!