Data Management

Plenary session

- Slides
- JC: *Simulations validate design, but how do we validate simulations*? A lot of this will be part of the technical-to-measurement session.

Parallel session

Participants: Eli Dart, Ted Kisner, Don Petravick, Duncan Watts, Julian Borrill, François Bouchet, Anna Ho, Ragnhild Auerlien, Sara Simon, Colin Bischoff, Debbie Bard, Ankur Dev, Tom Crawford, Cosmin Deaconu, Katie Harrington, Andrea Zonca, Brian Koopman, Kolen Cheung, Marius Millea, Christos Giannakopoulos, Nigel Sharp, Christian Reichardt, Jesse Treu, Nathan Whitehorn, Vyoma Muralidhara, CArlo Baccigalupi, David Rapetti, Reijo Keskitalo, Marcelo Alvarez, Sasha Rahlin, Kimmy Wu, Allen Foster

- Data Movement
    - DP: *It is said that transients can use IRIDIUM for the transport to alerts.*
    - *JB: What kind of access to timestream data should be provided to the collaboration?*
        - *SR: Some kind of gating is probably necessary.*
        - *TK: facilitate lots of small scale exploration, it will inform larger processing efforts.*
        - *CB: Make data available in human readable format.*
        - *TC: Do we anticipate a lot of people outside of DM wanting to re-analyze the TOD? I don't feel like there were many people clamoring to do that for WMAP or Planck.*
        - *FB: Yes, on Planck there were not that many people ready and capable of toi processing. I mean full toi processing. But we had of course the capability to see the data on screen when acquired*
        - *NW: We've benefited a lot from that in SPT, I think*
        - *KH: I've done a lot of instrument characterization with TOD analysis that wasn't headed directly toward making maps*
        - *JT: There are not that many people working on pre-cleaning raw tods in SO or in ACT. Which WBS section will work on that for CMB-S4 ?*
        - *TC: This one.*
        - *TC: Katie - Yes, I agree with both you and Colin on this, and it is probably up to DM to supply those tools. But I'm more responding to Julian's*

question about how much computing we need to be prepared to supply to people who want to analyze large fractions of the TOD into maps.
    - *NW: And I think we really want to ensure that more people can engage here, which sets some requirements on tooling*
    - *TC: I.e., the instrument characterization analyses should not require large allocations of supercomputer time.*
    - *JB: I think Nathan's point on facilitating engagement is critical.*
    - *JT: Tool development shouldn't require huge amounts of computer time; the testing might though, although testing really ought to be able to be done with low levels of processing time.*
- Software Infrastructure
    - Question on DM/SI interface on simulated data registration/archiving/distribution
    - *NW: What is the development plan for software infrastructure? How to avoid parallel development of software for lab use? TK: Interfaces will be established for DC1.*
- Data Simulation
    - JB: *Extragalactic foreground models (websky) for different cosmologies? AZ: Will have to ask Marcelo. Carlo mentions Euclid work. Pan-Expt Extragalactic Group. Issue for ECC too.*
    - *Trade-offs between systematics modeling & mitigation and residuals modeling.*
- Data Reduction
    - What can DQ say about the best use of SP bandwidth?
    - ML for DQ seems like an obvious area of research (cf. SO group)
    - DR/DM interface on data viewers
    - Expt. characterization will be critical for commissioning.
    - What is the best map (characterization, cost, etc) for each science case?
- Transients
    - *CB: Daily maps are motivated by transient science, so if the science would improve with more frequent maps then we should do that.*
    - *JB: how does that vary between Chile & South Pole … maybe 2 visits/day in Chile; many per day at Pole.*
    - *Time series analysis!*
    - *RK: CHLAT will visit a given site at 5-min intervals for 15-20 minutes at a time, once per day.*
    - *NW: Issuing alerts for short time scale events has not been part of the design yet. It could be.*
- Site Hardware
    - How do we want to capture SP bandwidth issues to maintain pressure on agencies to improve it?
    - *CB: What is the redundancy for SP storage? TC: SPT3G hasn't had excessive HW failures but have a full year of redundancy. NW: about to disk failures per year. TC: Horrible failure rate with tapes. DP: Budget for disaster recovery. JB: Let's budget for this.*

Report back

Plenary Session 3/10/21

From Zoom chat:

- Keith Thompson: Which WBS is involved in the actual near-term data quality efforts -- human-hours, timescale within days of the data being taken? Clearly the software used would be included in this data management category, but the actual operations involved in monitoring the data from this many receivers is significant -- probably at least half a dozen full-time equivalent, needs to be planned for (if it hasn't already).
- Ted Kisner: It does fall under Data Reduction. Obviously with interaction with Software Infrastructure and Data Movement. And yes, ramping up FTE people is and will be a challenge.
- Keith Thompson: Thanks.
- Julian (responding verbally): High on DM's risk register addressing FTE issues.
- Jesse Treu: If you're talking about the so-called SO machine learning group… the truth is that there are just a couple of us working on ACT data using neural networks to beef up the current "cuts" pipeline which has plenty of human intervention involved. As of now, there are some promising results with neural networks, but the training requires "supervised learning", which requires human intervention. Next steps include graduating into "unsupervised learning" but we have not yet made that step.
- Don Petravick: Note that the Transients concerns also might call for a review at Science Requirements level. not just work within DM.

  Julian [response] - given how rapidly transient science is evolving, important for DM to respond to the science case from the wG, not to lead that effort.

Phillip Lubin: Local foregrounds - satellite clusters. Should this be seen as a threat?
Julian - Yes certainly on our radar. Discussion Friday RFI session. Emerging issue and understanding what sort of mitigation is necessary is definitely important to keep in mind.

Grant Teply - any plan to respond to external alerts [of transient/variable sources]?
Julian - no, not within the survey strategy. Our primary strategy focuses on repeated scans of the sky. Certainly the analysis WG could respond by pulling out data from the appropriate time/space. But plan is not to go after targets of opportunity.