Data Management Technology Development

Julian Borrill, Joy Didier & Laura Newburgh
for the
“Cat's Been On My Keyboard” Working Group
(CMBS4DMTDWG)
especially
Yuji Chinone, Tom Crawford, Ken Ganga,
Christian Reichardt, Graca Rocha & Nathan Whitehorn

CMB-S4 Collaboration Meeting – Argonne National Laboratory – March 2018

Starting Point

The CDT report identified key DM challenges in five specific areas:

1. Instrument Data: acquisition, transmission, storage
2. Time-Domain: live-monitoring, pre-processing, map-making
3. Analyses: foregrounds, science goals, feedback
4. Simulation: modeling inputs, data synthesis, coordination
5. Publication: data products, software products, archiving

and one overarching area:

6. Computational Resources: bandwidth, storage, cycles

Filling In The Details

The CMBS4DMTDWG went through the CDT list adding “Level 4” detail.

One example:

DM.2 – Time-Domain Processing

DM.2.3 – Map-Making

DM.2.3.1 – Implementation

DM.2.3.2 – Standardization

DM.2.3.3 – Matching Maps To Science Goals

DM.2.3.4 – Characterization

Prioritization Principles

We then adopted 2 principles for prioritizing technology development

1. The importance of the technology to the current project definition phase, and to the various milestones within it:

   Decadal Survey, CD-0, CD-1, MREFC, CD-2

2. The likelihood that the required technology will not have been developed for other (S3, satellite) experiments by the time S4 needs it.

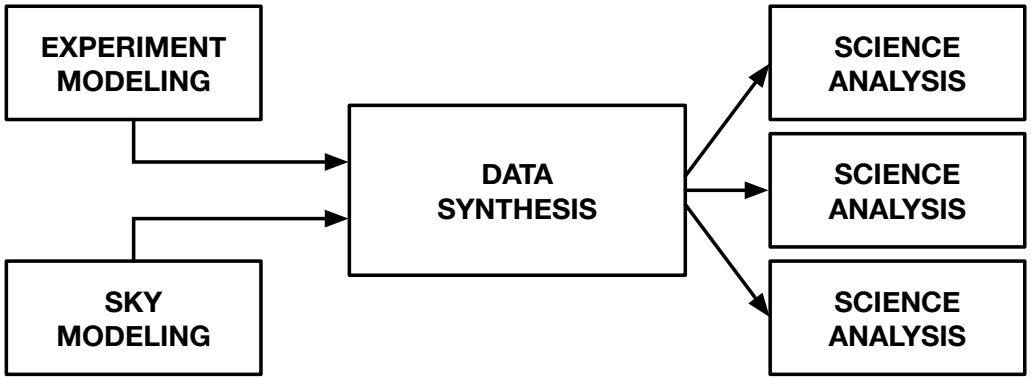
Evaluation Metrics

Using the cost-schedule-statistics-systematics metric, current DM risks are mostly concentrated in the statistics and systematics domains.

- Getting the design wrong won't necessarily increase (and could actually decrease) the construction cost and/or schedule.
- Getting the design wrong will increase the statistical and/or systematic uncertainties, requiring longer & costlier operations.
- One cost-risk is failure to secure sufficient free (National Facility) compute resources, by under-allocation or code inefficiency.

Project Definition Phase

- During the project definition phase we will progressively:
  - Refine the details of the experiment configuration **(P3.2)**
  - Extend the breadth and depth of the science analyses **(P2.3)**
- Current DM priorities are the technologies needed to support this



- Other elements are important, but none have pre-CD2 lead times.

Experiment & Science Inputs

- To make informed configuration choices we must be able to forecast the full spectrum of science outcomes from the various options.
- Greater configuration detail corresponds to more degrees of freedom in the inputs to the data synthesis:
  - Instrument (optics & electronics)
  - Observation (sites & scanning strategies)
  - Environment (atmosphere & ground pickup)
  - Foregrounds (galactic & extragalactic)

Data Synthesis

- Currently S4 data are synthesized in the map domain.
- Many of the degrees of freedom we want to incorporate will ultimately require time-domain data synthesis.
- Time-domain data synthesis implies reduction to maps for analysis
  - We have a wealth of S3 approaches to compare & contrast.
- Time-domain synthesis & reduction requires a tightly-coupled, computationally efficient, HPC/HTC framework
  - Need to be able to respond to architecture evolution.

Science Analyses

- Analyses *per se* aren't the provenance of DM, but supporting them is.
- Each analysis requires maps that
  - Include the necessary science signals
    - Scalar/Tensor/NG CMB; Extragalactic FG & Lensing; …
  - Are well-matched to the analysis algorithm
    - Full frequency, split frequency, foreground-cleaned, …
  - Are sufficiently well-characterized
    - Full covariance, diagonal covariance, Monte Carlo, …

Prioritization - I

"Importance to project definition phase milestones"

- All of the elements identified will be needed at some point during the project definition phase.
- The relative prioritization of sub-elements requires a schedule of the required level of detail (configuration, science) at each milestone.
- The overall framework that couples the elements, including professionalizing our coding, is a priority on any schedule.

Prioritization - II

"Likelihood S4 will need it first"

Instrument Modeling	Progressively extended from S3
Observation Modeling	Uniquely multi-site, multi-survey
Environment Modeling	Progressively extended from S3
Sky Modeling	In progress for S3, S4 & satellites
Time-Domain Synthesis	Unique + extended inputs, data volume
Time-Domain Reduction	Variety of S3 approaches, data volume
Map-Domain Synthesis	In progress for S3, S4 & satellites
Science Analyses	In progress for S3, S4 & satellites
Framework – Scale & Quality	