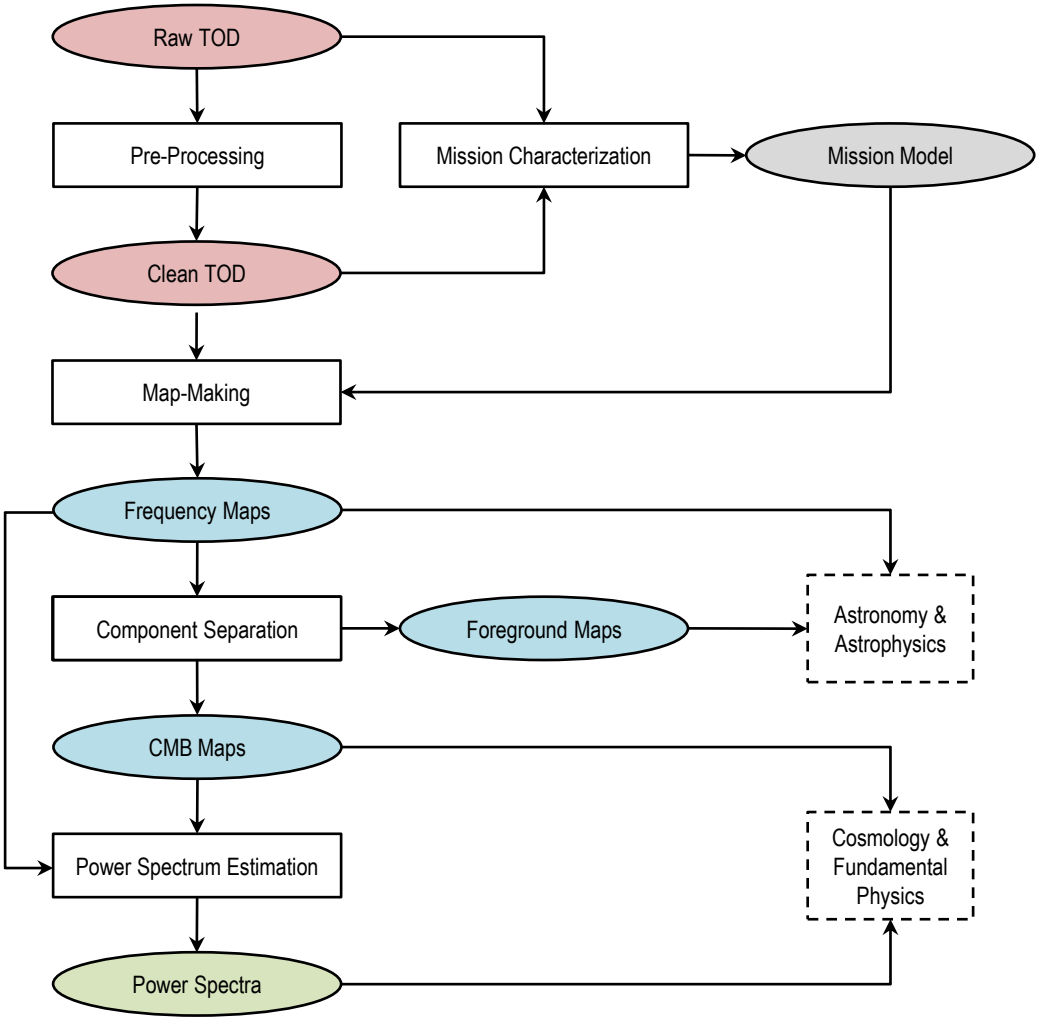
Data Analysis Challenges

Julian Borrill
Computational Cosmology Center, Berkeley Lab
Space Sciences Laboratory, UC Berkeley

Data Analysis

- Sequence of S/N-increasing data compressions via domain transformations:
  - Time
  - Pixels
  - Multipoles
- Each domain exposes different systematics => iterative looping.
- Must propagate both data *and their covariance* for a sufficient statistic.



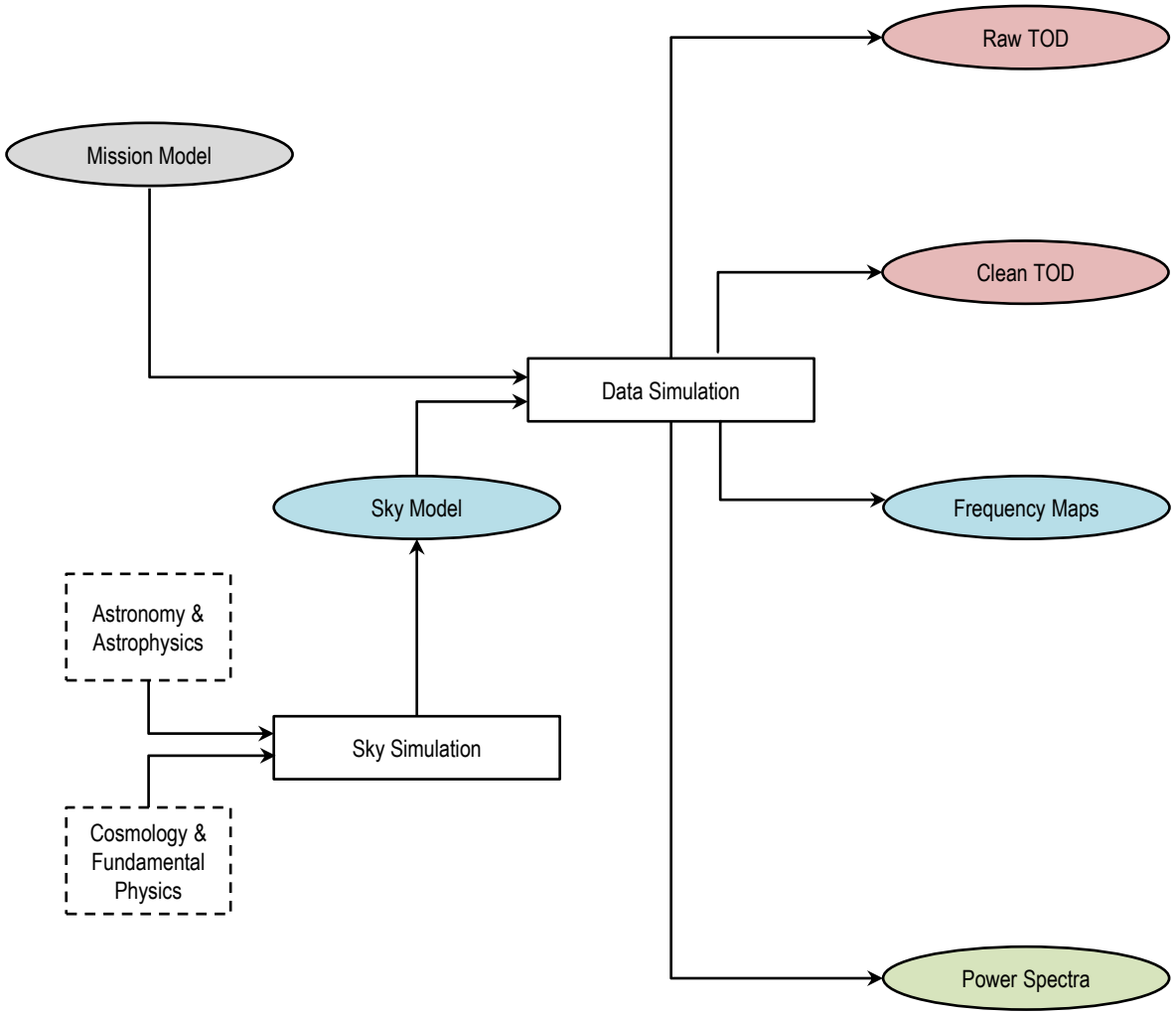
Analysis Methods

- Data volumes:
  - Time domain: $\mathcal{N}_t \sim \Sigma_{det}$ Sampling Rate (Hz) x Observation Time (s)
  - Pixel domain: $\mathcal{N}_p \sim \Sigma_{freq,\ pol}\ 10^9$ x Sky Fraction / [Beam (arcmin)]^2
- Data analysis scaling dominated by:
  - $\mathcal{N}_p^3$ for exact methods with explicit covariance matrices.
  - $\mathcal{N}_{mc}\,\mathcal{N}_t$ for approximate methods with MC uncertainty quantification.
- Computational constraints (1% cycles/year on Top 10 system):
  - 2000 : $\mathcal{N}_p < 10^6$ & $\mathcal{N}_t < 10^{12}$
  - 2015 : $\mathcal{N}_p < 10^7$ & $\mathcal{N}_t < 10^{15}$
  - 2030 : $\mathcal{N}_p < 10^8$ & $\mathcal{N}_t < 10^{18}$

Assumes:
- Moore's Law
- 100% & 1% efficiency

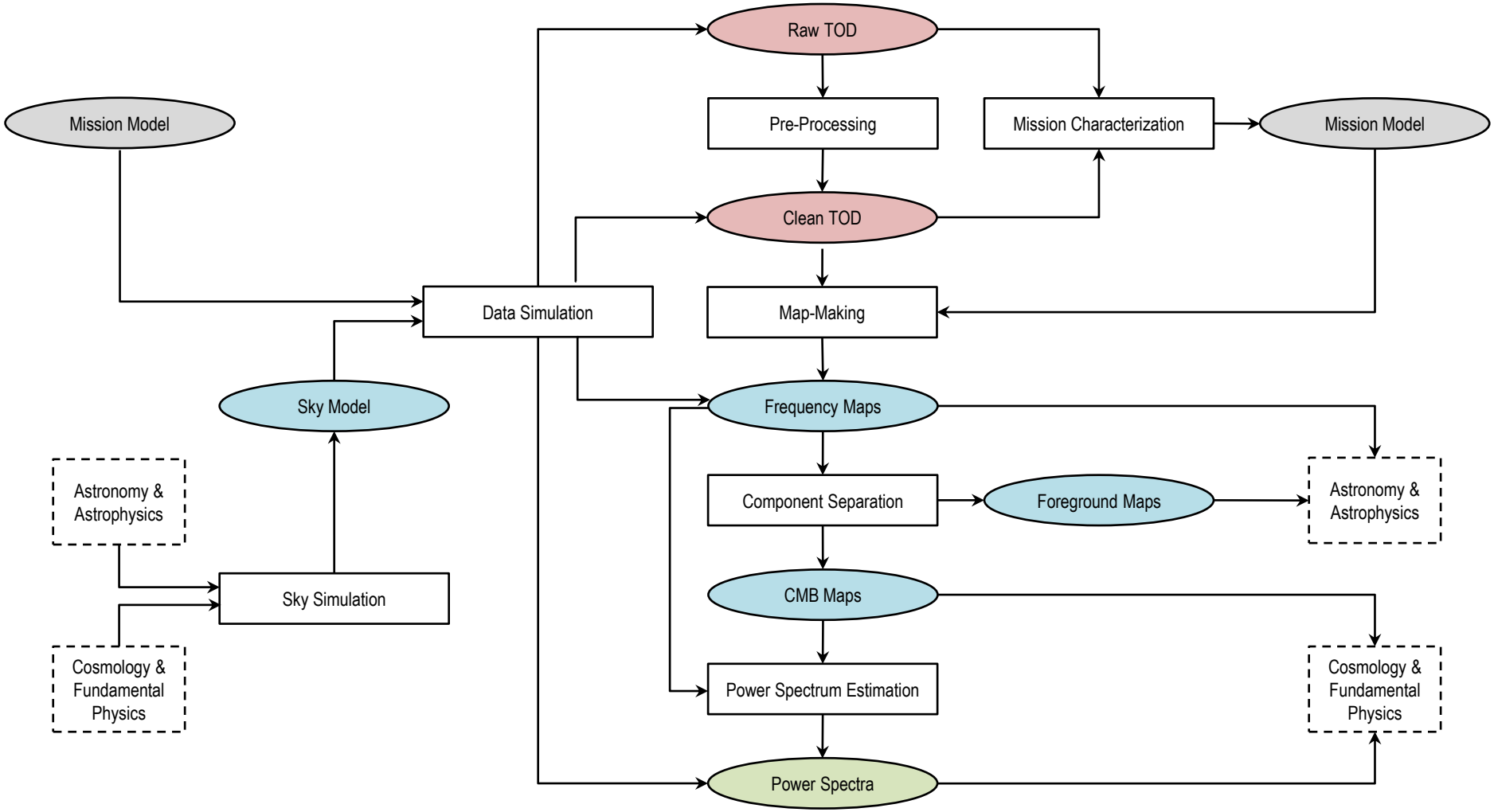
- Except in special cases, exact methods now computationally intractable.

Simulations

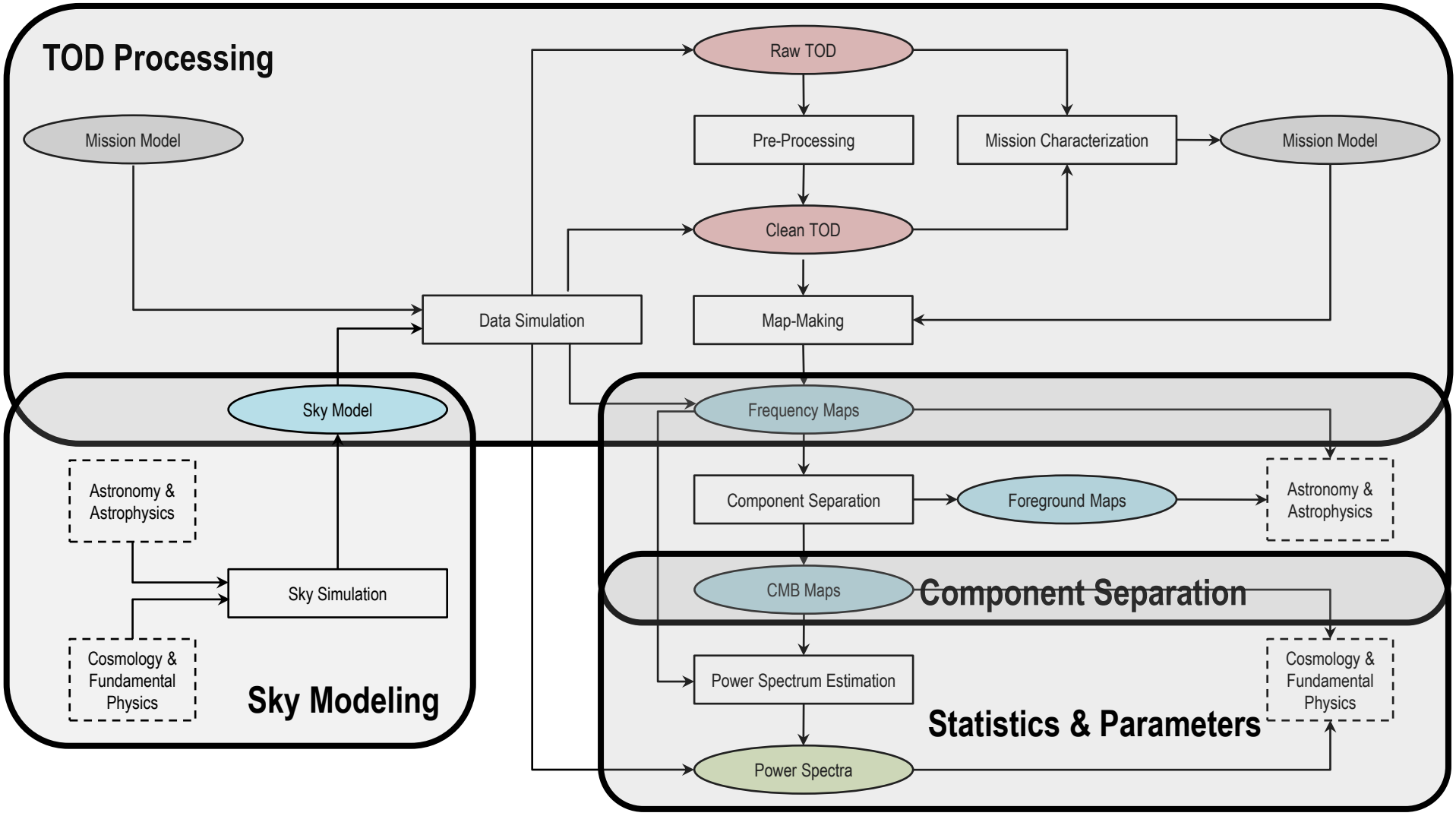


- Needed for:
  - Mission design & development
  - Analysis validation & verification
  - Data uncertainty quantification & debiasing (MC)
- From top to bottom, trade-off between:
  - computational cost
  - realism/reliability

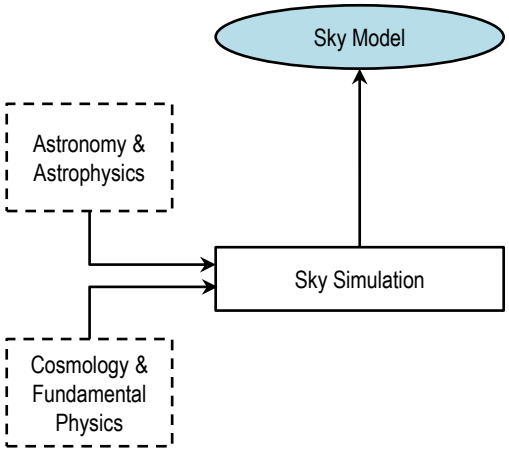
SimDA: Top Down, Wrap Around



SimDA: Domains



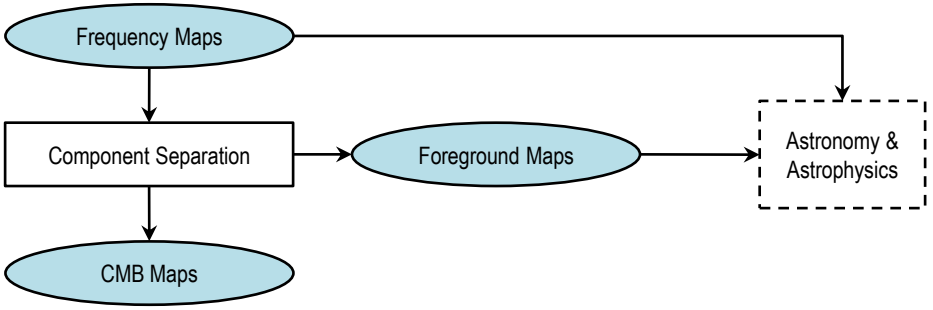
Sky Modeling



- Key Challenges:
  - Reliability: noisy, confused, band-passed, beam-convolved input data, inc. Planck!
  - Self-consistency: eg. CMB secondaries & extra-Galactic foregrounds
  - Mission-independence: decoupling from band-passes
  - Usability: software engineering

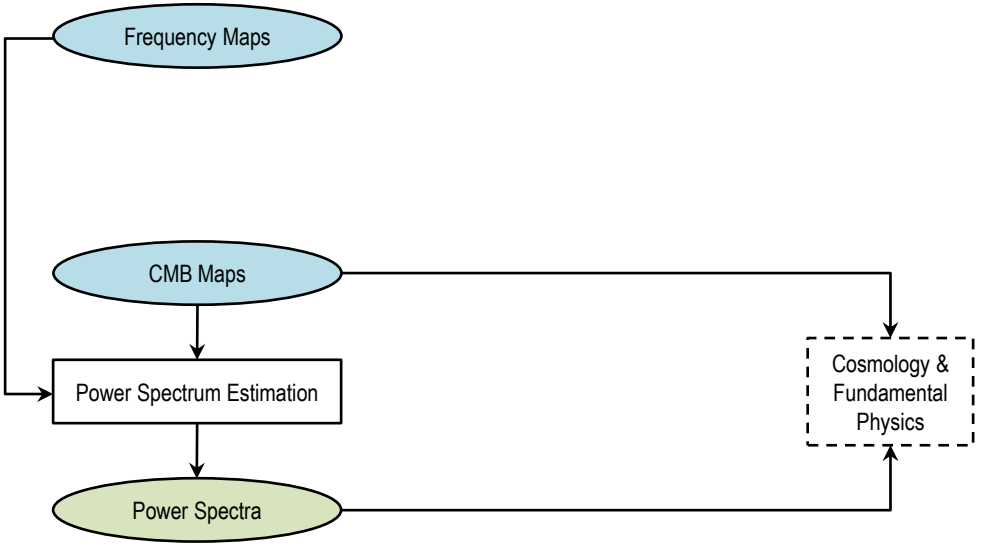
Component Separation

- Key Challenges:
  - Validation: are these the right algorithms for the (as yet unknown) real foregrounds?
  - Verification: are these algorithms right given (as yet flawed) simulated foregrounds?

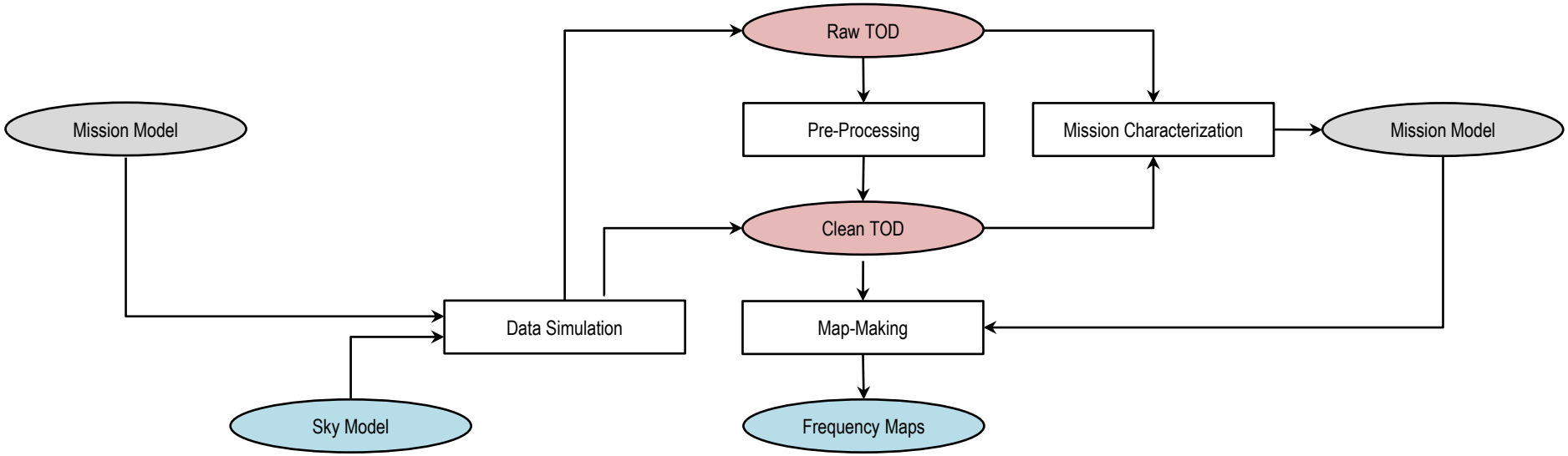


Statistics & Parameters

- Key Challenges:
  - Reliability: sufficiency of real data covariance approximations.
  - Tractability: disk space for many millions of MC maps.

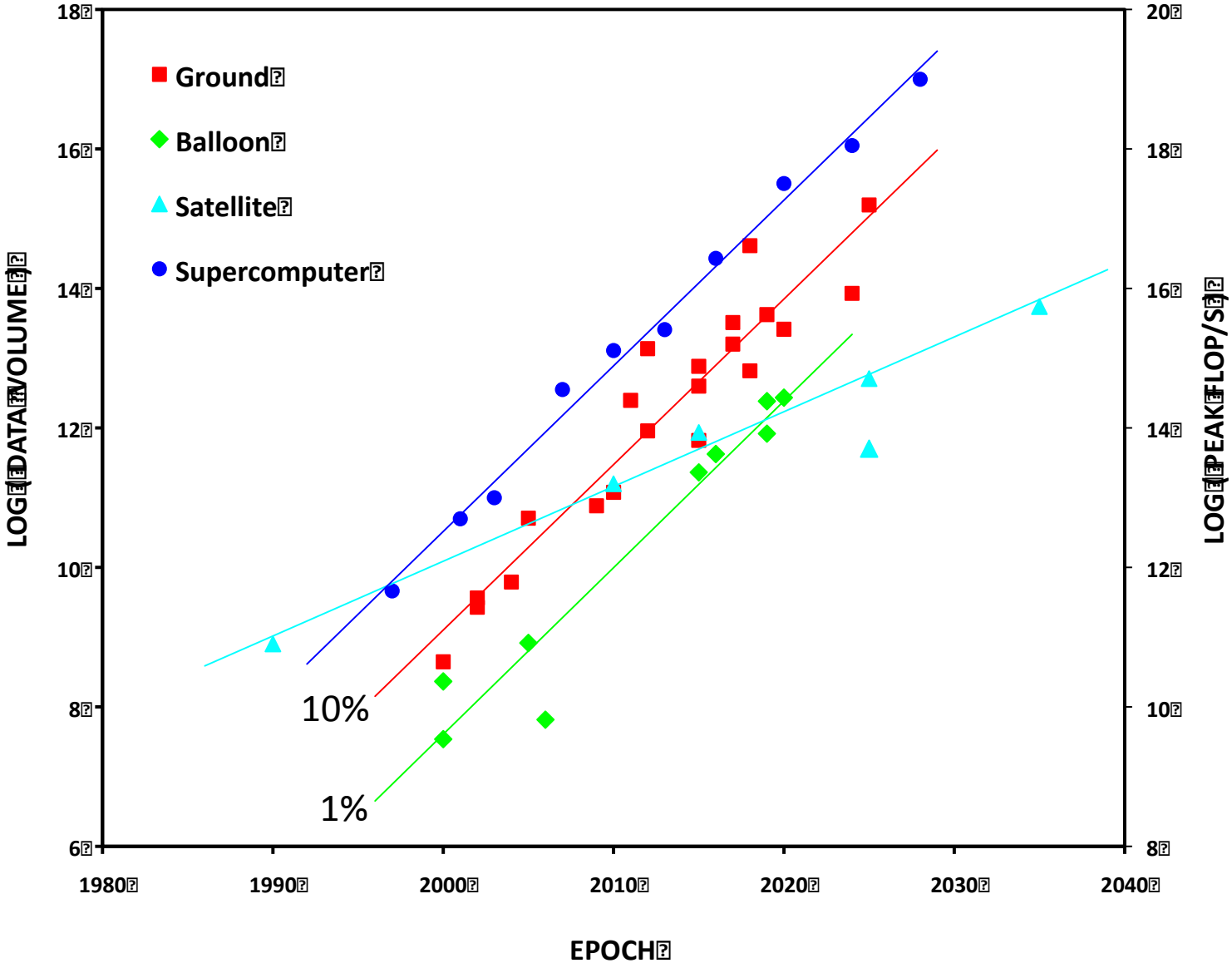


Time-Ordered Data Processing



- Key challenges (primarily) due to data volume

Data & HPC Growth



TOD Challenges

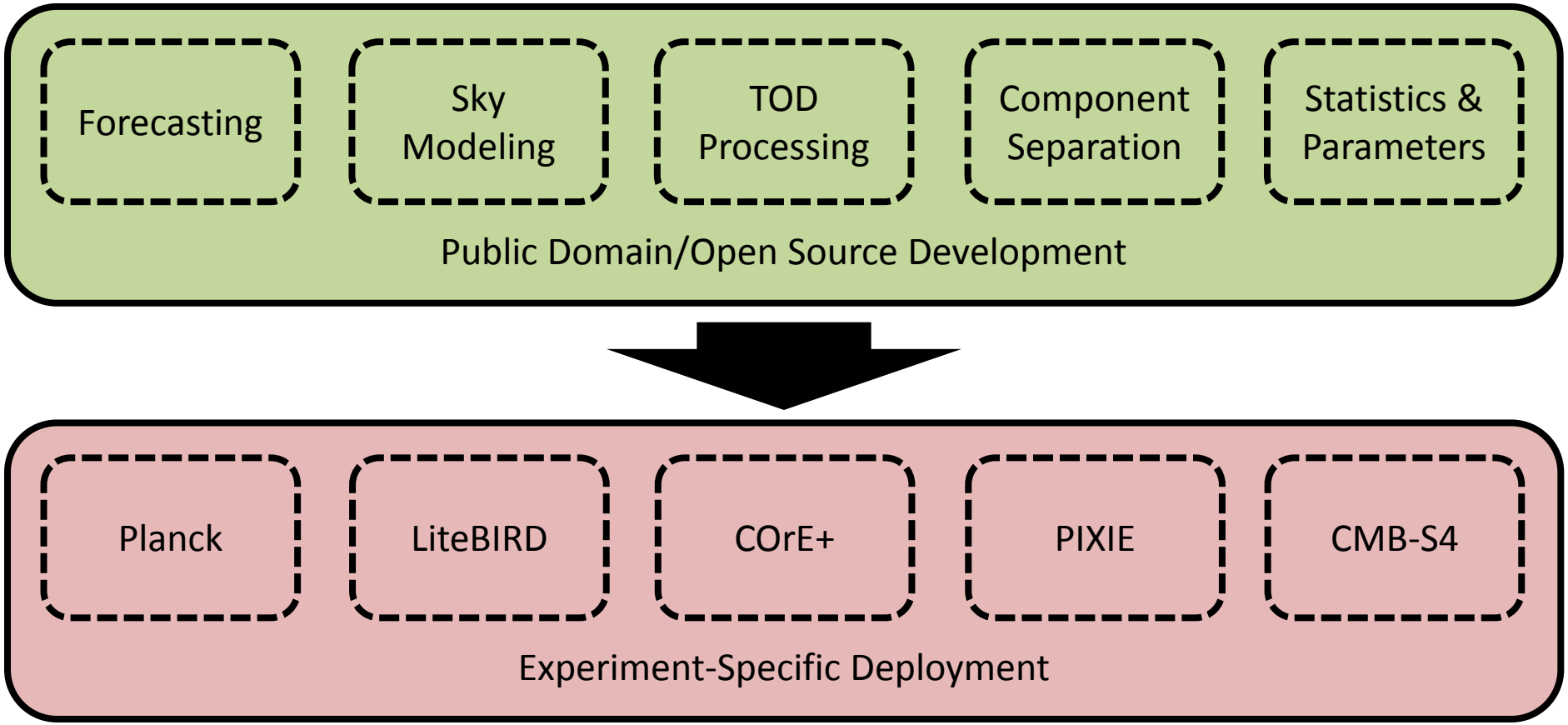
- Monte Carlo uncertainty quantification
  - Requires
    - very many (*statistical*)
    - very representative (*systematic*)

    simulations.
  - Given finite (and challenging!) computational resources, we must balance the competing requirements of speed and realism.

- Pre-processing & mission characterization
  - Require fast/flexible data wrangling
  - Must provide the same MC speed & realism interactively.

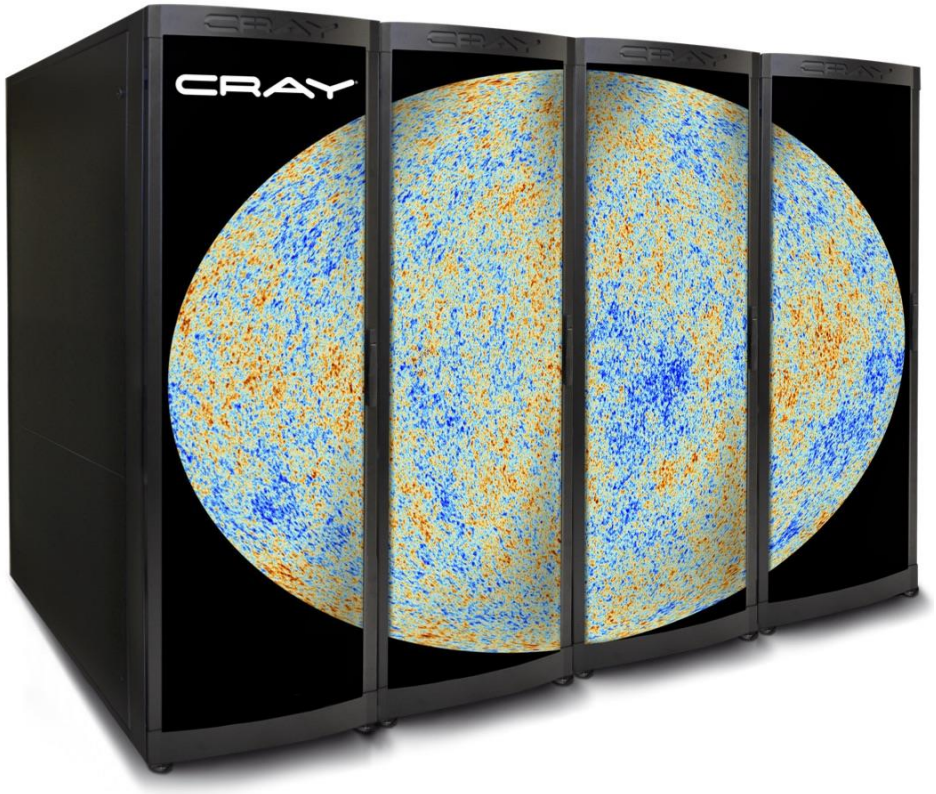
A Modest Proposal

- A two-tier community-wide program:
  - developing common, generic capabilities in the public domain
  - deploying them for specific analyses within each experiment



TOAST

- Generic TOD processing
  - Simulation
  - Map-making
- Experiment-specific extensions
  - Mission
  - Pre-processing
  - …



https://github.com/hpc4cmb/toast

https://github.com/hpc4cmb/toast-core

(& -planck, litebird, cmbs4, …)

Conclusions

- TOD data volumes present tough but predictable computational challenges
  - Computational efficiency must be a key design driver, *and*
  - Moore's Law presents a moving architectural target.

  Efficiency is a journey, not a destination.
- Pure efficiency is sufficient for massive MCs, but must also be made generally useable for pre-processing & mission characterization.
- All next-generation mission-class CMB experiments are facing unique expressions of common simulation and data analysis challenges – the time is ripe for open-source community solutions, uniquely adapted & applied.