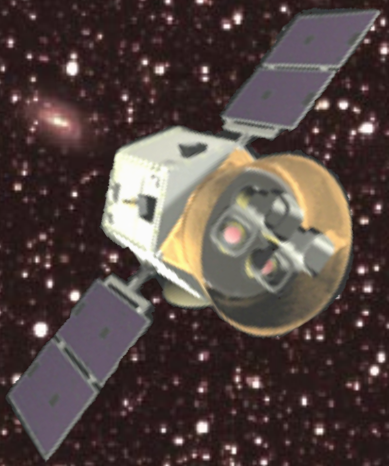
Lessons Learned from Developing and Operating the Kepler Science Pipeline and Building the TESS Science Pipeline

Jon M. Jenkins
NASA Ames Research Center

Monday September 4, 2017



WP11 Exoplanet Science Meeting
University of Warwick
Coventry UK

Overview

- What did it take to build the *Kepler* science pipeline?
- Major modifications to pipeline over lifetime
- High fidelity simulations
- Commissioning, commissioning, commissioning
- High performance computing
- Developing the TESS Science Pipeline
- Communication
- Summary

KEPLER

SCIENCE DATA PROCESSING PIPELINE

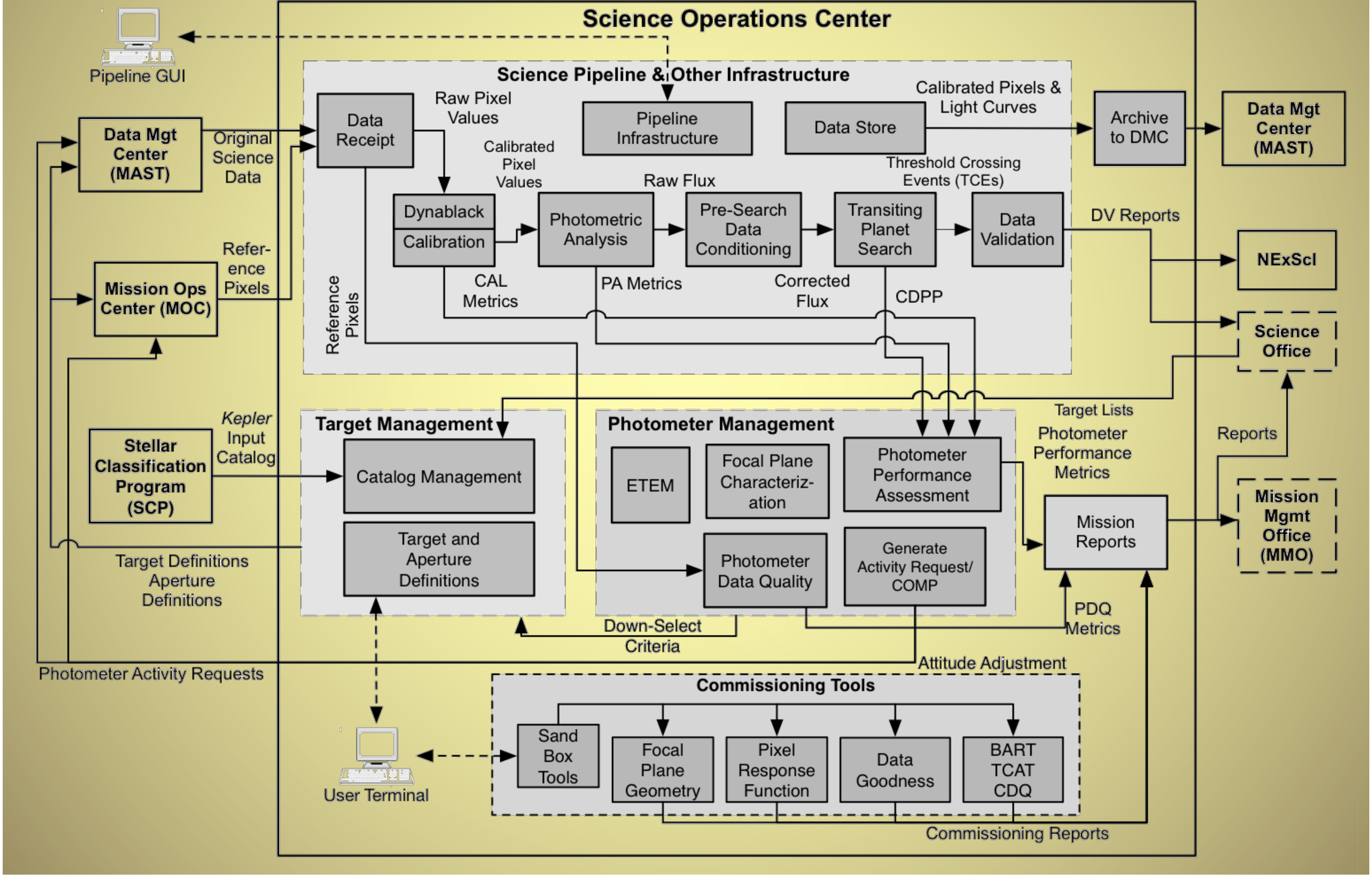
The Science Operations Center: What did it take?

- Design started in earnest in 2004 with launch in March 2009 and operations through May 2013 and reprocessing through 2017
- A total of ~100 person years of effort went into the first complete version of the pipeline (from pixels to planets)
- The staffing was at ~20 individuals per year through 2016, tapering off thereafter (~280 FTEs over project lifetime)
- Build 5.0 was the launch-ready software release
- There were 4 major builds thereafter, with substantive point releases to mitigate issues subsequently identified in flight or full volume re-processing
- Build 9.0, 9.1, 9.2, 9.3 really represented at least two full builds of effort (issues identified in full re-processing and in completeness and reliability processing)
- Unexpected instrumental effects/stellar variability/hardware failures motivated significant software modifications on orbit

Science Operations Center Architecture

Kepler

A Search for Earth-size Planets



Science Operations Center

Pipeline GUI

Science Pipeline & Other Infrastructure

Data Mgt Center (MAST)

Original Science Data

Data Receipt

Raw Pixel Values

Pipeline Infrastructure

Data Store

Calibrated Pixels & Light Curves

Archive to DMC

Data Mgt Center (MAST)

Calibrated Pixel Values

Raw Flux

Threshold Crossing Events (TCEs)

Dynablack

Calibration

Photometric Analysis

Pre-Search Data Conditioning

Transiting Planet Search

Data Validation

DV Reports

NExScI

Refer-ence Pixels

Mission Ops Center (MOC)

Reference Pixels

CAL Metrics

PA Metrics

Corrected Flux

CDPP

Science Office

Target Lists

Target Management

Photometer Management

Stellar Classification Program (SCP)

Kepler Input Catalog

Catalog Management

Target and Aperture Definitions

ETEM

Focal Plane Characteriz-ation

Photometer Performance Assessment

Photometer Performance Metrics

Reports

Mission Reports

Mission Mgmt Office (MMO)

Photometer Data Quality

Generate Activity Request/ COMP

Target Definitions Aperture Definitions

Down-Select Criteria

PDQ Metrics

Photometer Activity Requests

Attitude Adjustment

Commissioning Tools

User Terminal

Sand Box Tools

Focal Plane Geometry

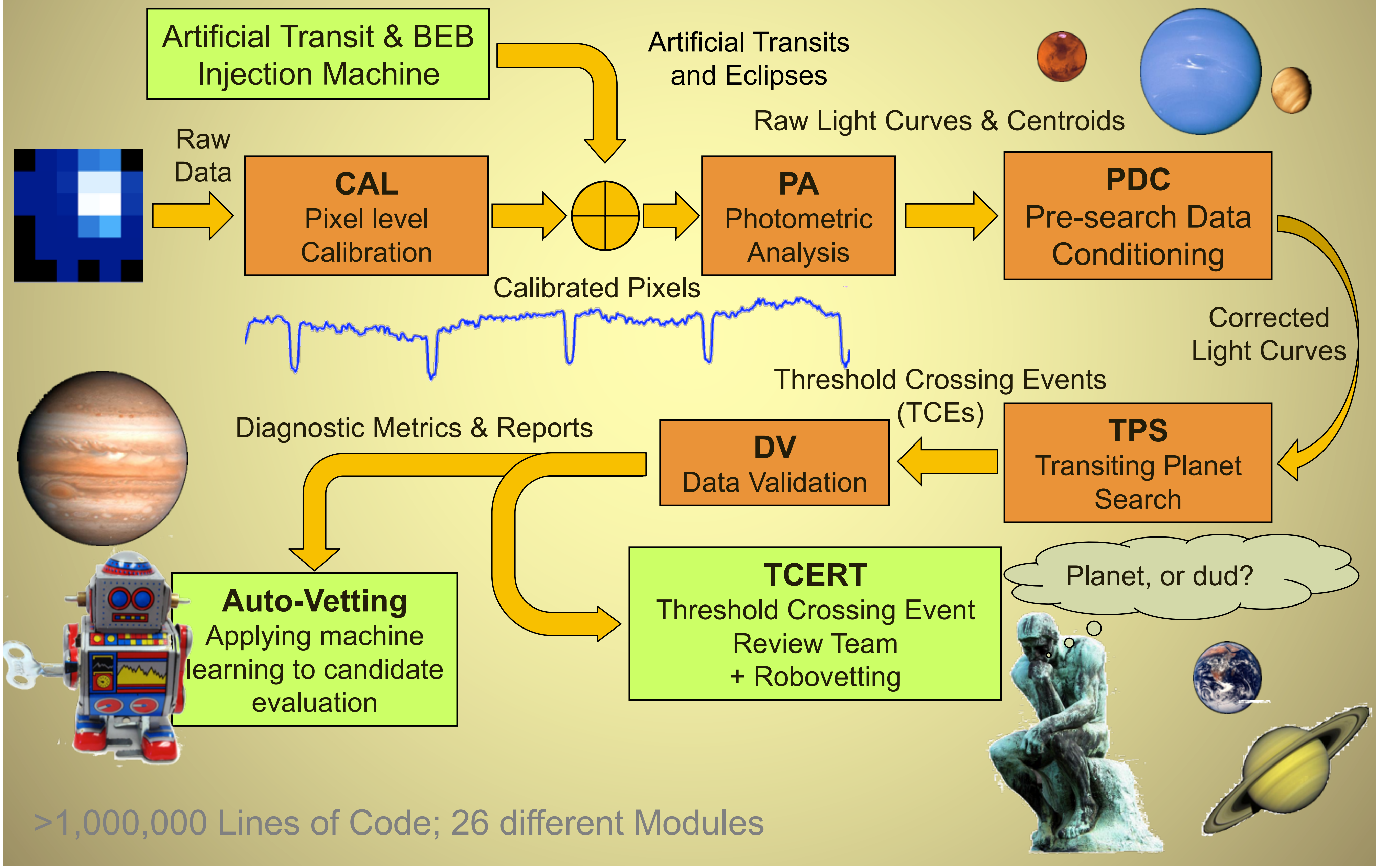
Pixel Response Function

Data Goodness

BART TCAT CDQ

Commissioning Reports

Kepler's Science Pipeline



Artificial Transit & BEB Injection Machine

Artificial Transits and Eclipses

Raw Data

CAL
Pixel level Calibration

Calibrated Pixels

Raw Light Curves & Centroids

PA
Photometric Analysis

PDC
Pre-search Data Conditioning

Corrected Light Curves

TPS
Transiting Planet Search

Threshold Crossing Events (TCEs)

DV
Data Validation

Diagnostic Metrics & Reports

Auto-Vetting
Applying machine learning to candidate evaluation

TCERT
Threshold Crossing Event Review Team
+ Robovetting

Planet, or dud?

>1,000,000 Lines of Code; 26 different Modules

Major Modifications

Every component of the science pipeline saw major evolution over mission

Pixel level calibrations:

- Updates based on actual electronics behavior
- Flagging of electronic image artifacts causing false positives

Identifying optimal apertures

- Use of reconstructed pointing
- Added ability to correct errors in Kepler Input catalog

Photometric analysis

- Major improvements to identifying cosmic rays

Pre-search Data Conditioning

- Development of Maximum a Posteriori approach
- Addition of multi-scale analysis
- Detection of Sudden Pixel Sensitivity Dropouts

Transiting Planet Search

- χ^2 vetoes added

Data Validation

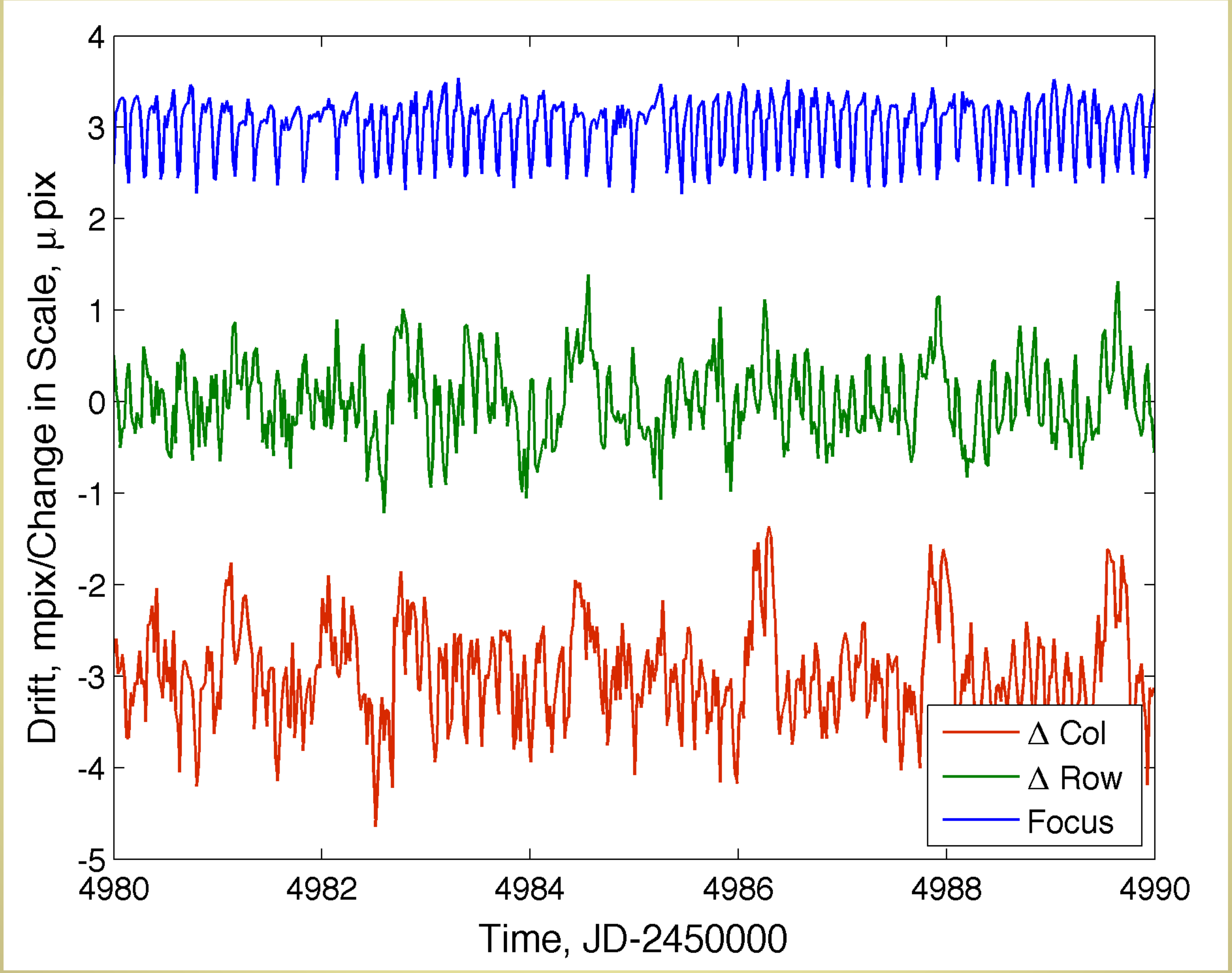
- Difference image analysis
- Ghost Diagnostic + other metrics

~80% of the science code was re-written

Kepler
A Search for Earth-size Planets

Short Timescale Instrumental Errors

Signature of a heater cycling on the reaction wheels 3/4

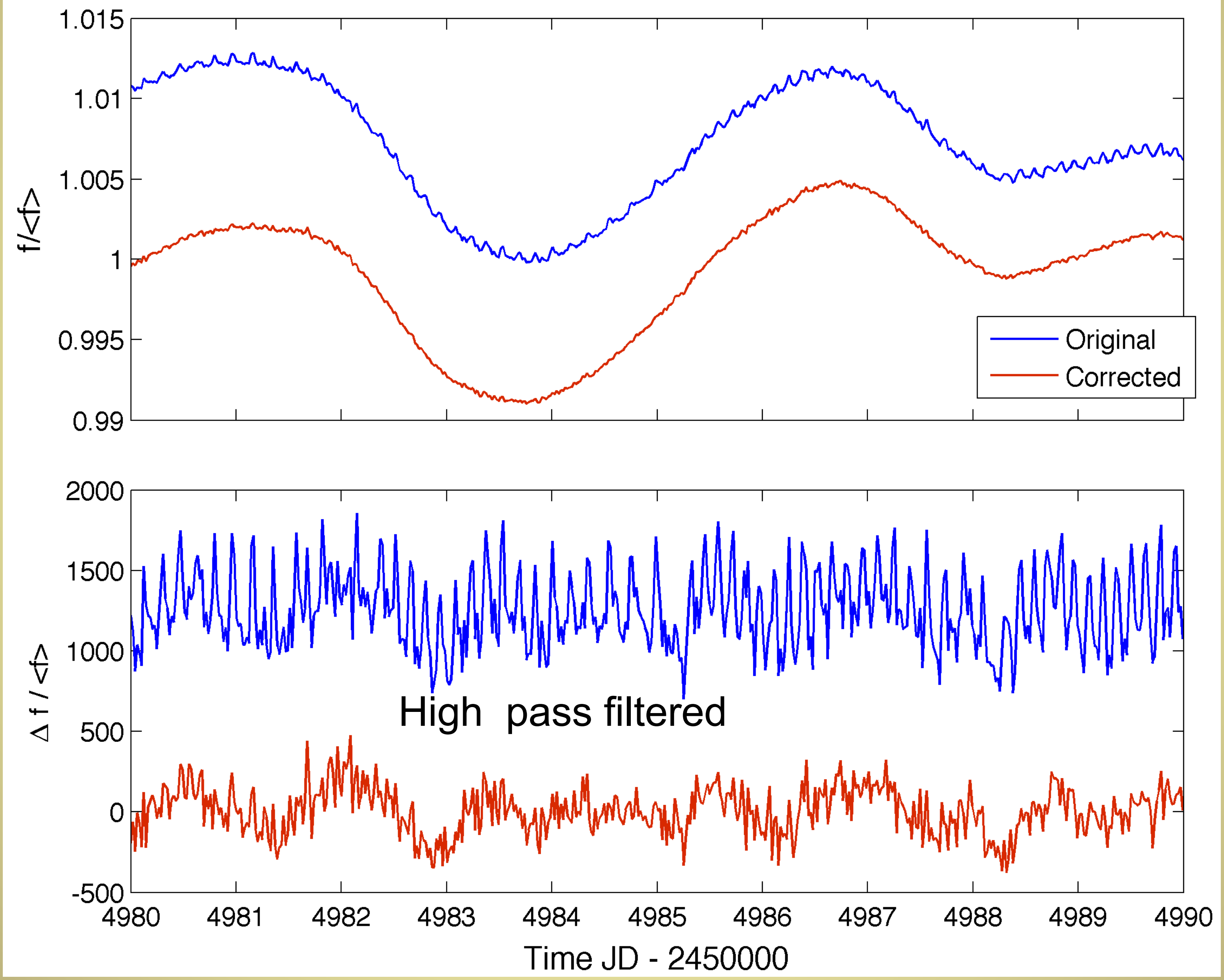


Kepler is sensitive to its thermal environment

Instrumental Effects in Photometry

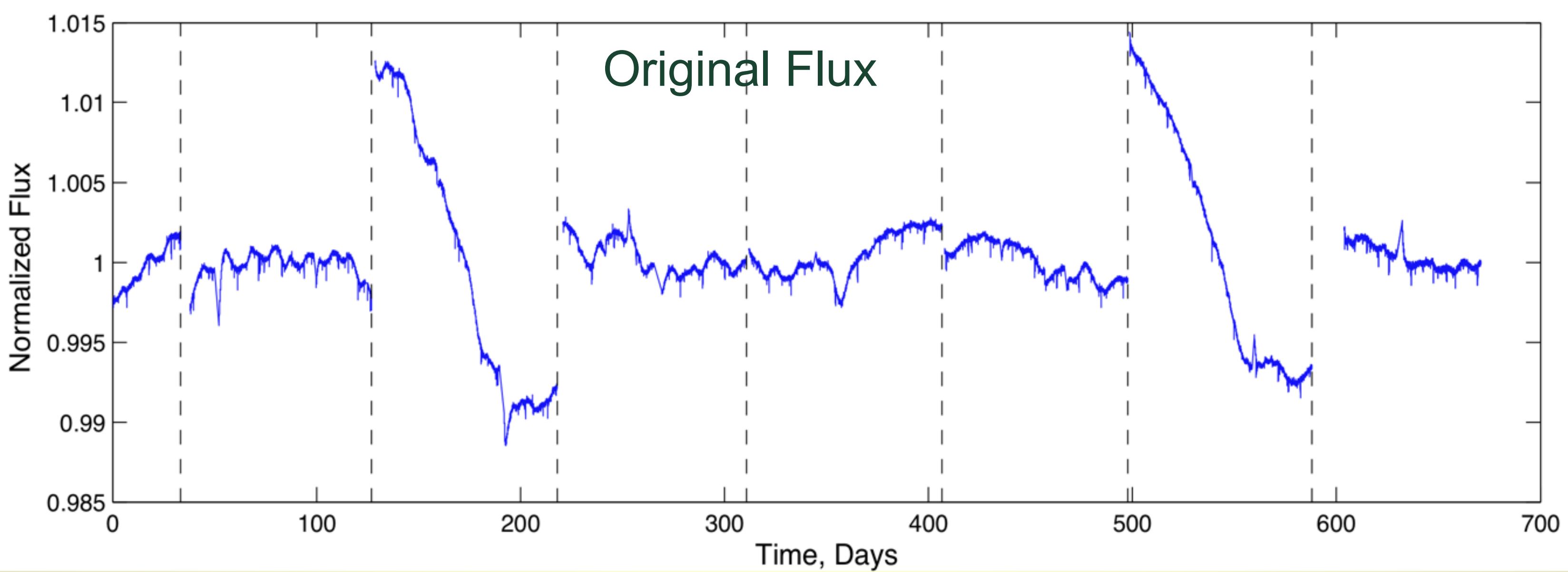
Kepler

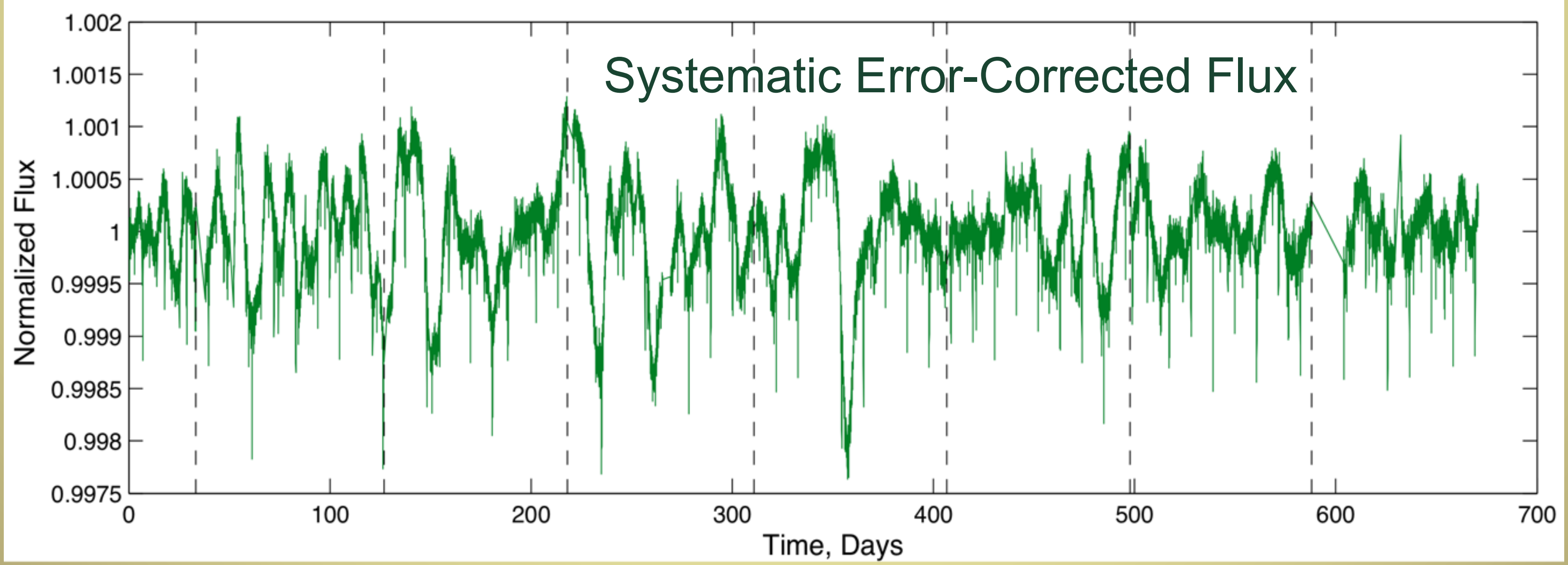
A Search for Earth-size Planets



Correcting Systematic Errors

A Search for Earth-size Planets





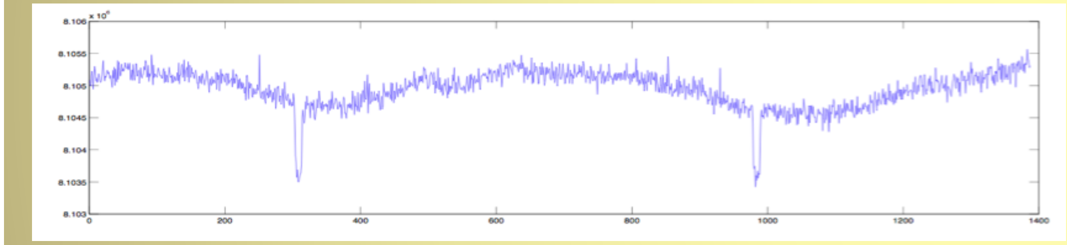
We apply a Maximum A Posteriori approach as per Stumpe et al. 2014

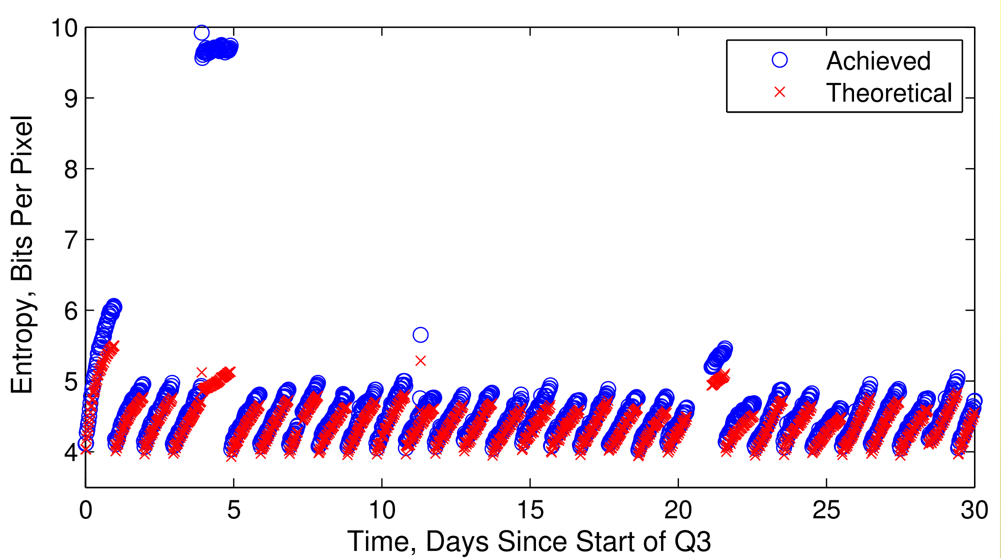
A Search for Earth-size Planets

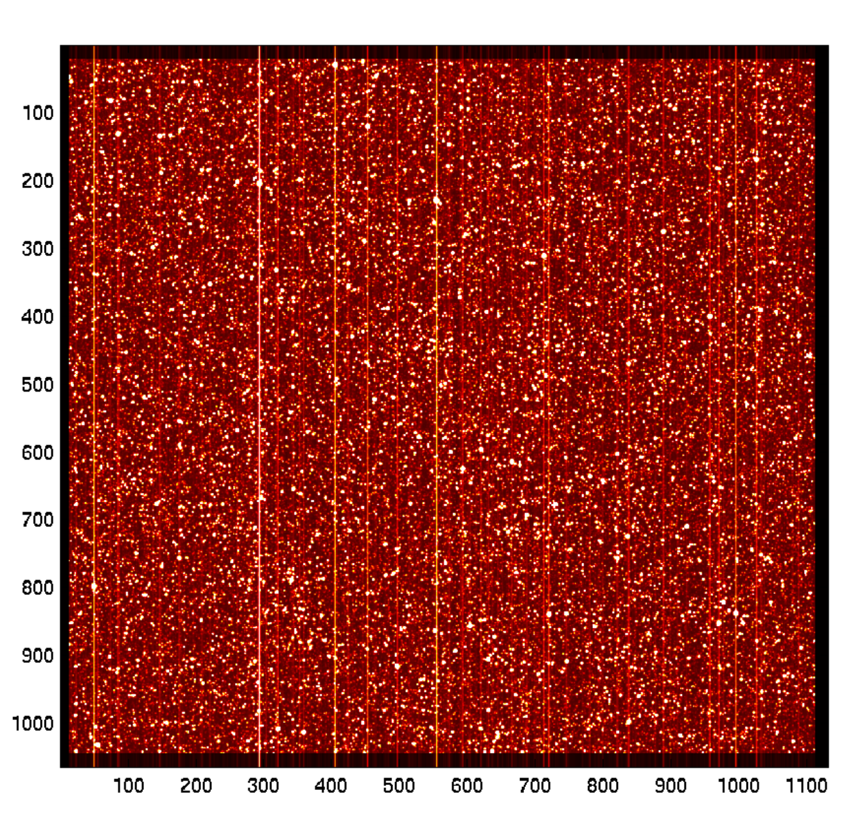
High Fidelity Simulations are Indispensable

End-End Model (ETEM) drove design of SOC and testing of entire ground segment

Simulated data were so good that we didn't need to update the compression tables after launch (the achieved compression (~4.5-5 bits per pixel) was within 0.1 bits of ideal performance







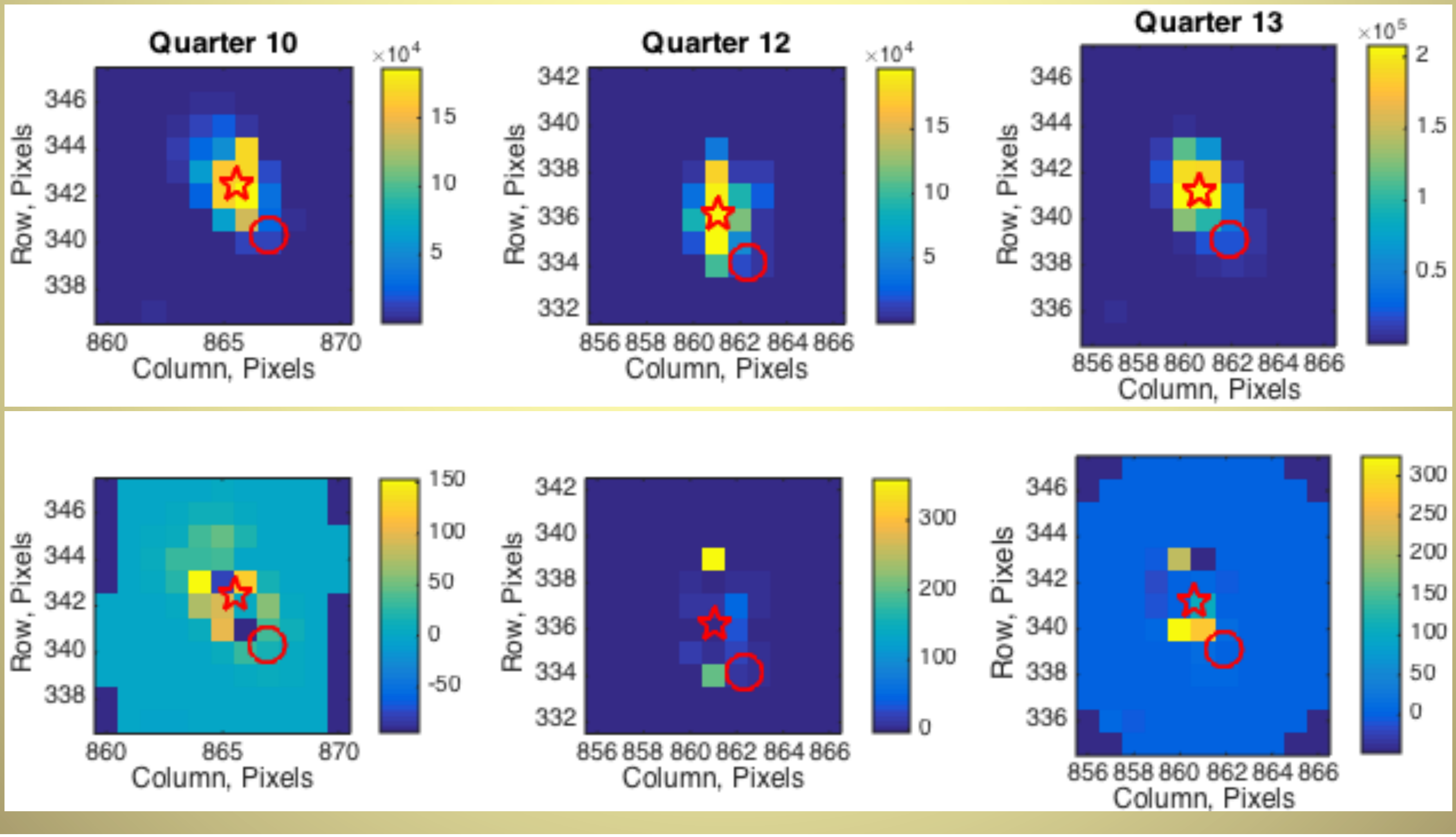
Difference Image Analysis

A Search for Earth-size Planets

Difference image analysis was key for Kepler for excluding false positives from background eclipsing binaries

Especially important for bright, saturated (bleeding) targets

KIC 3542116



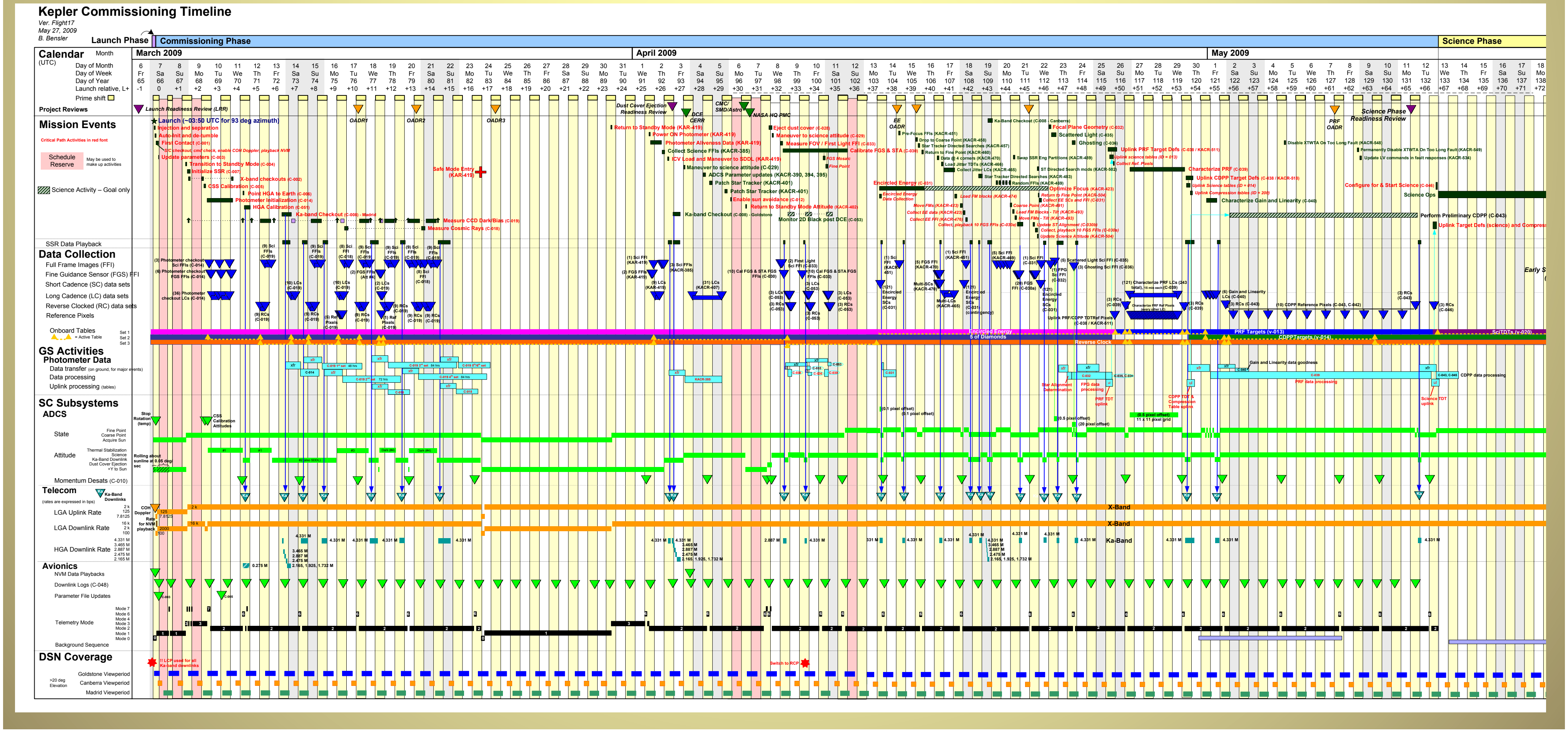
Rappaport et al.2017, arxiv1708.06069

Commissioning, Commissioning, Commissioning!

Kepler

A Search for Earth-size Planets

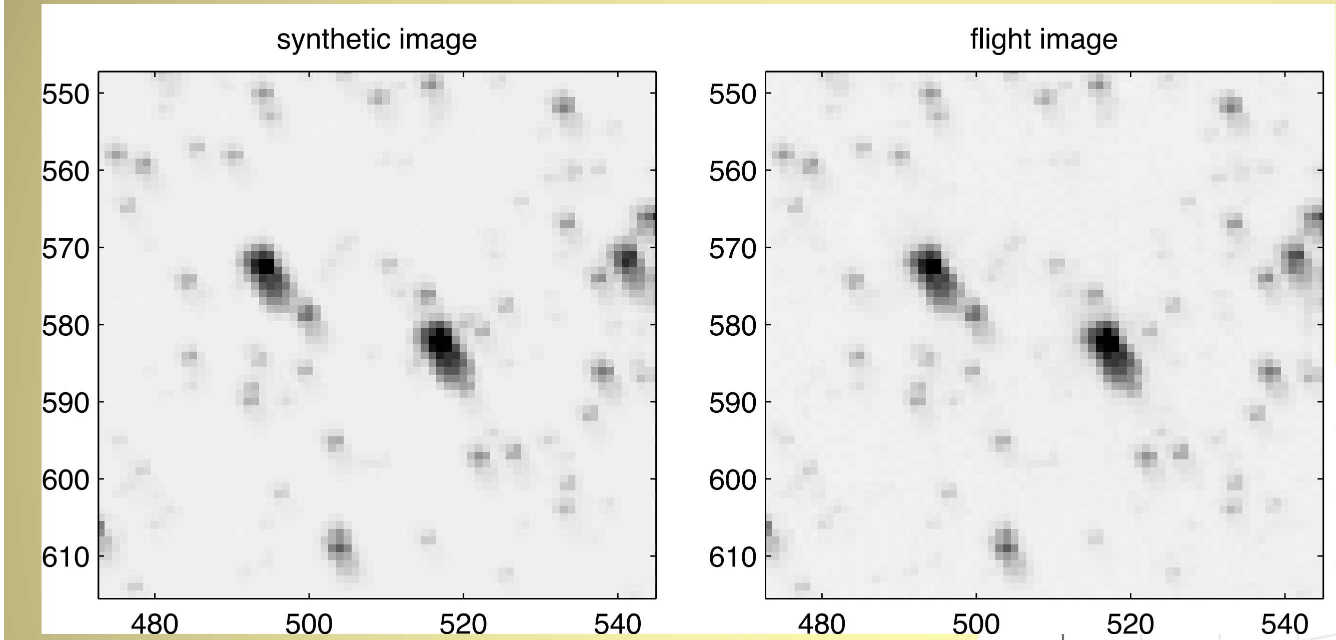
- Commissioning tools require special attention and data sets
- Effort for commissioning tools may be as great as that for major science pipeline modules
- Don't leave commissioning tool development to the last



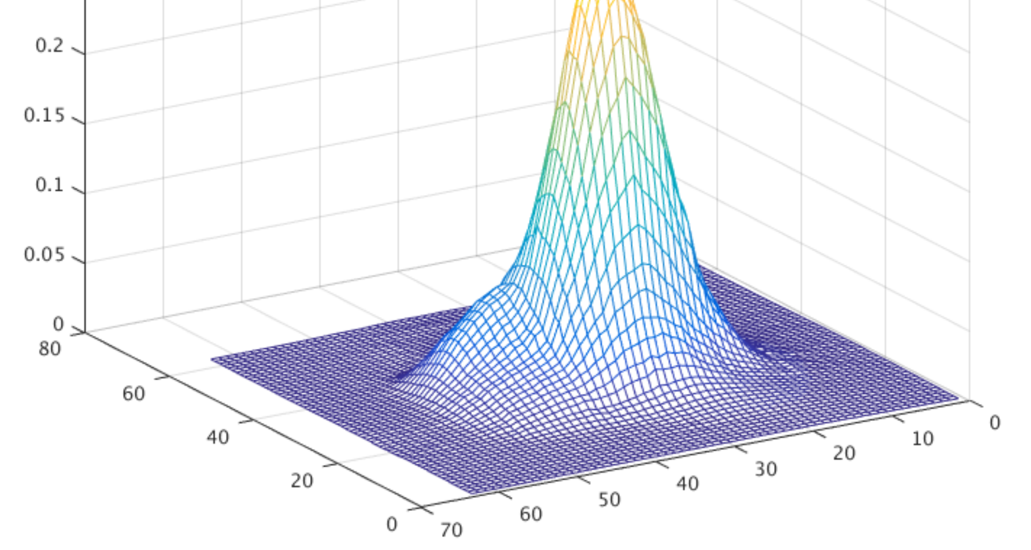
A Search for Earth-size Planets

Pixel Response Function Characterization

Kepler PRF



TESS PRF



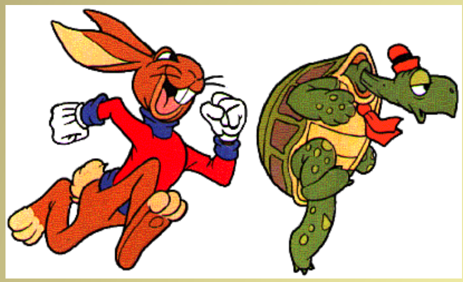
Keeping Up with the Data

Kepler

A Search for Earth-size Planets

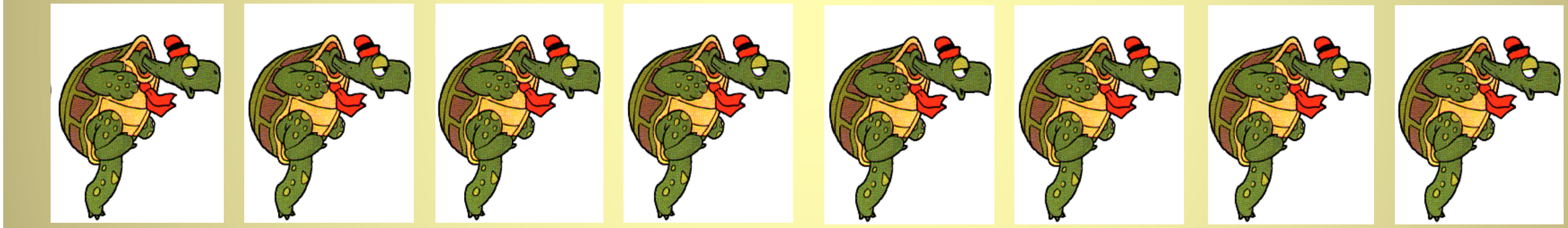
Improving the Throughput

A Search for Earth-size Planets

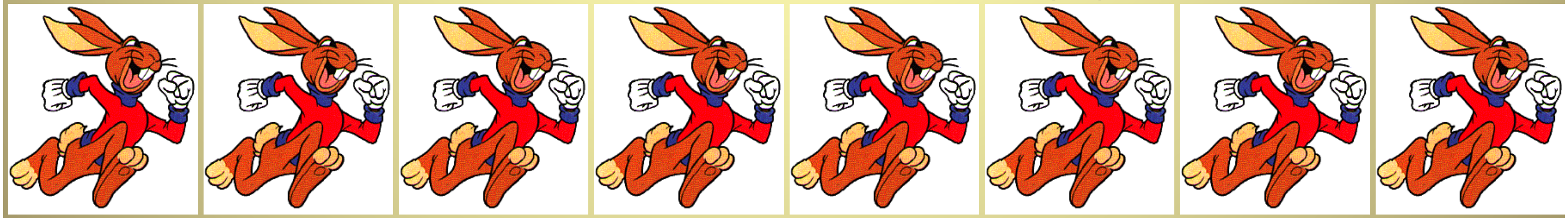


Some fast code; Some slow code

Step 1: Parallelize all code



Step 2: Make slow code fast(er)



Hardware Architecture: *Kepler* Science Operations Center

64 hosts, 712 CPUs,

3.7 TB of RAM,

148 TB of raw disk storage

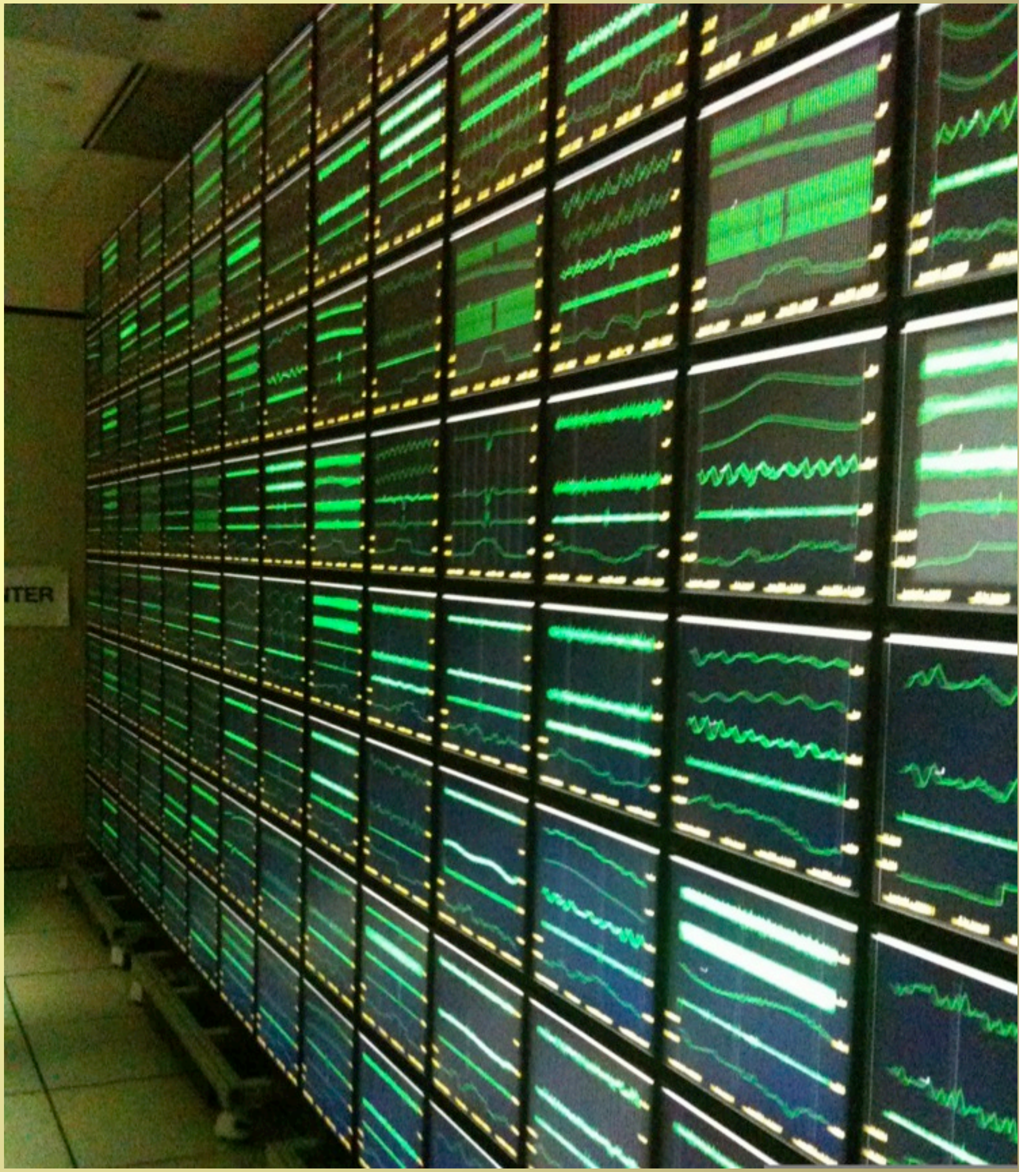
Hardware Architecture: NAS Pleiades Supercomputer

5.34 Pflop/s peak cluster

211,872 cores

724 TB of memory

15 PB of storage



η_{earth}: Mapping Completeness and Reliability

A Search for Earth-size Planets

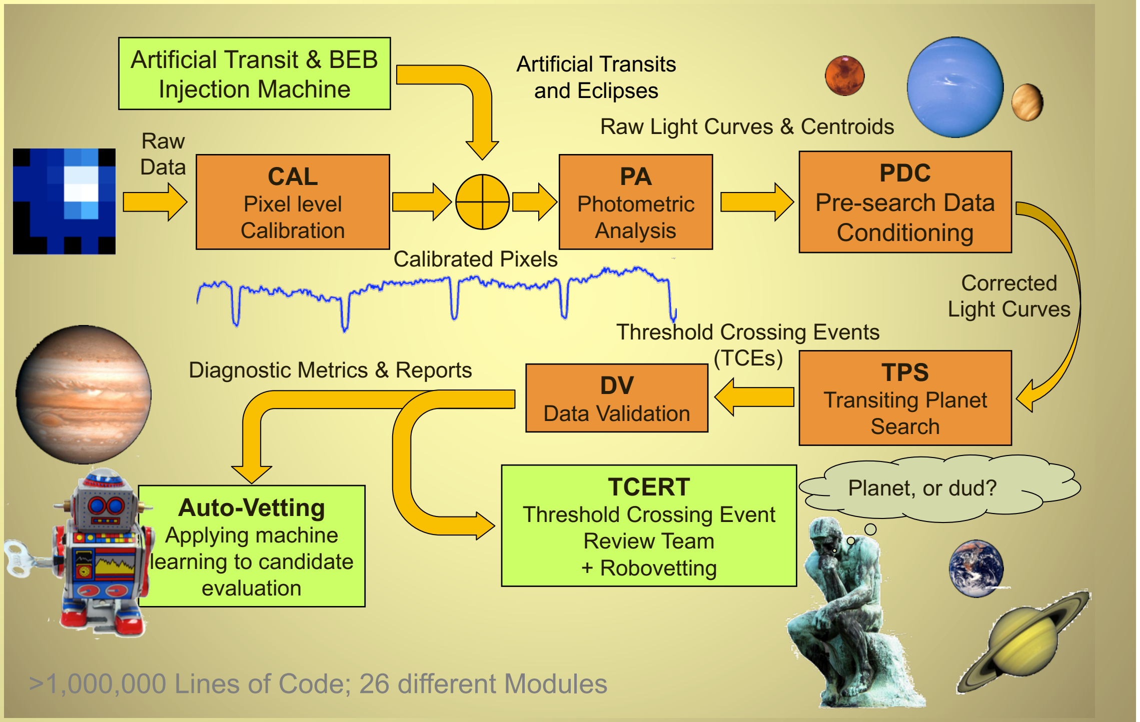
Characterizing completeness and reliability of software/people pipelines is extremely resource intensive

Kepler shipped the final light curve products in April 2015

We've spent the remainder of the time until present adding artificial transits, BEBs, scrambling the data temporally, inverting the light curves etc., etc.

Mapping completeness and reliability and characterizing the candidate vetting process is difficult

Recommendation: Pursue machine learning for conducting or modeling the candidate vetting process



Artificial Transit & BEB Injection Machine

Artificial Transits and Eclipses

Raw Light Curves & Centroids

Raw Data

CAL Pixel level Calibration

PA Photometric Analysis

PDC Pre-search Data Conditioning

Calibrated Pixels

Corrected Light Curves

Threshold Crossing Events (TCEs)

Diagnostic Metrics & Reports

DV Data Validation

TPS Transiting Planet Search

Auto-Vetting Applying machine learning to candidate evaluation

TCERT Threshold Crossing Event Review Team + Robovetting

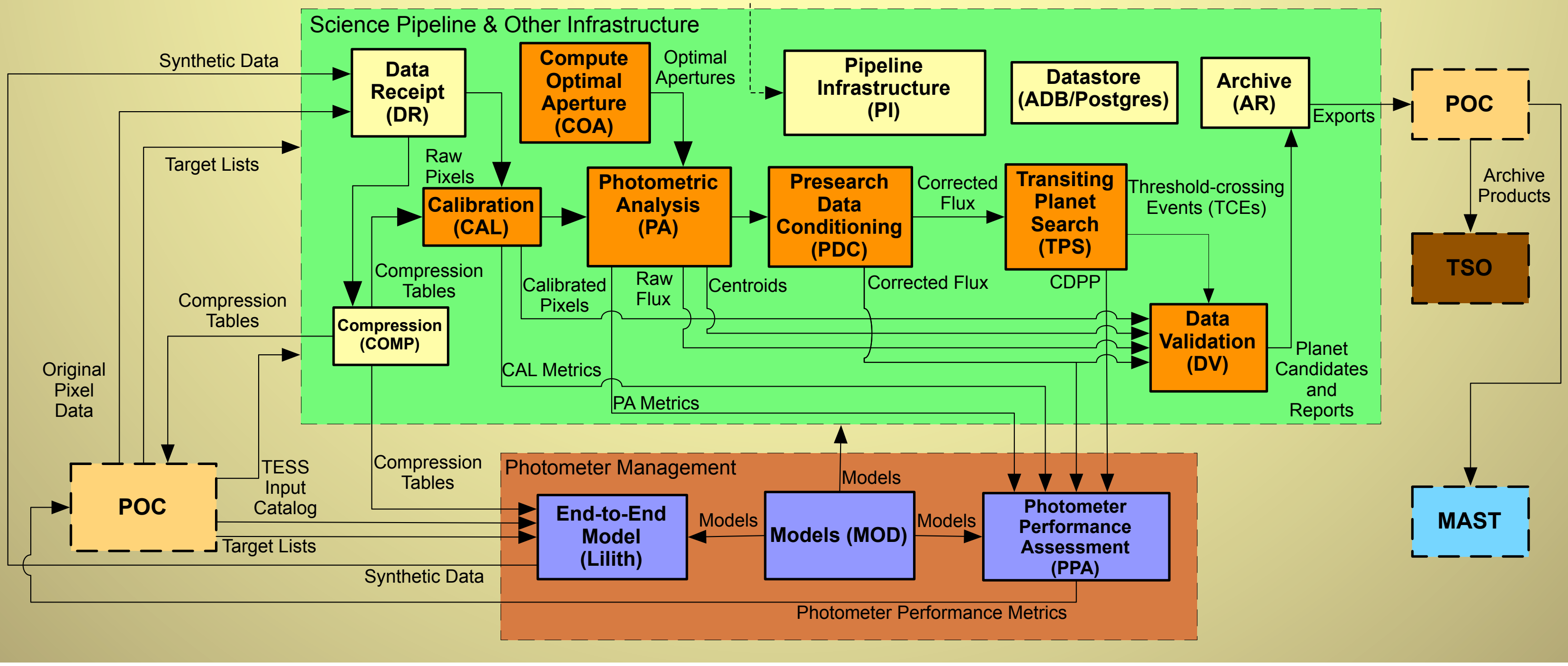
Planet, or dud?

>1,000,000 Lines of Code; 26 different Modules

Developing the TESS Pipeline

Kepler
A Search for Earth-size Planets

- ~13X pixel data rate over Kepler
- Leveraged heritage from Kepler pipeline
- Significantly lower cost (~46 FTEs over project lifetime)
- Significant speed improvements:
  - Colocated servers and storage with NAS Pleiades supercomputer
  - Moved pixel-level calibrations to C++
  - Sped up Presearch Data Conditioning by 10X
  - Originally projected 20+ days to process one sector
  - Complete pipeline requires ~5 days to process one sector



Science Pipeline & Other Infrastructure

Data Receipt (DR)

Compute Optimal Aperture (COA)

Optimal Apertures

Pipeline Infrastructure (PI)

Datastore (ADB/Postgres)

Archive (AR)

Exports

POC

Archive Products

TSO

MAST

Synthetic Data

Target Lists

Raw Pixels

Calibration (CAL)

Photometric Analysis (PA)

Presearch Data Conditioning (PDC)

Corrected Flux

Transiting Planet Search (TPS)

Threshold-crossing Events (TCEs)

Compression Tables

Calibrated Pixels

Raw Flux

Centroids

Corrected Flux

CDPP

Data Validation (DV)

Planet Candidates and Reports

Compression Tables

Compression (COMP)

CAL Metrics

PA Metrics

Original Pixel Data

POC

TESS Input Catalog

Target Lists

Synthetic Data

Compression Tables

Photometer Management

End-to-End Model (Lilith)

Models

Models (MOD)

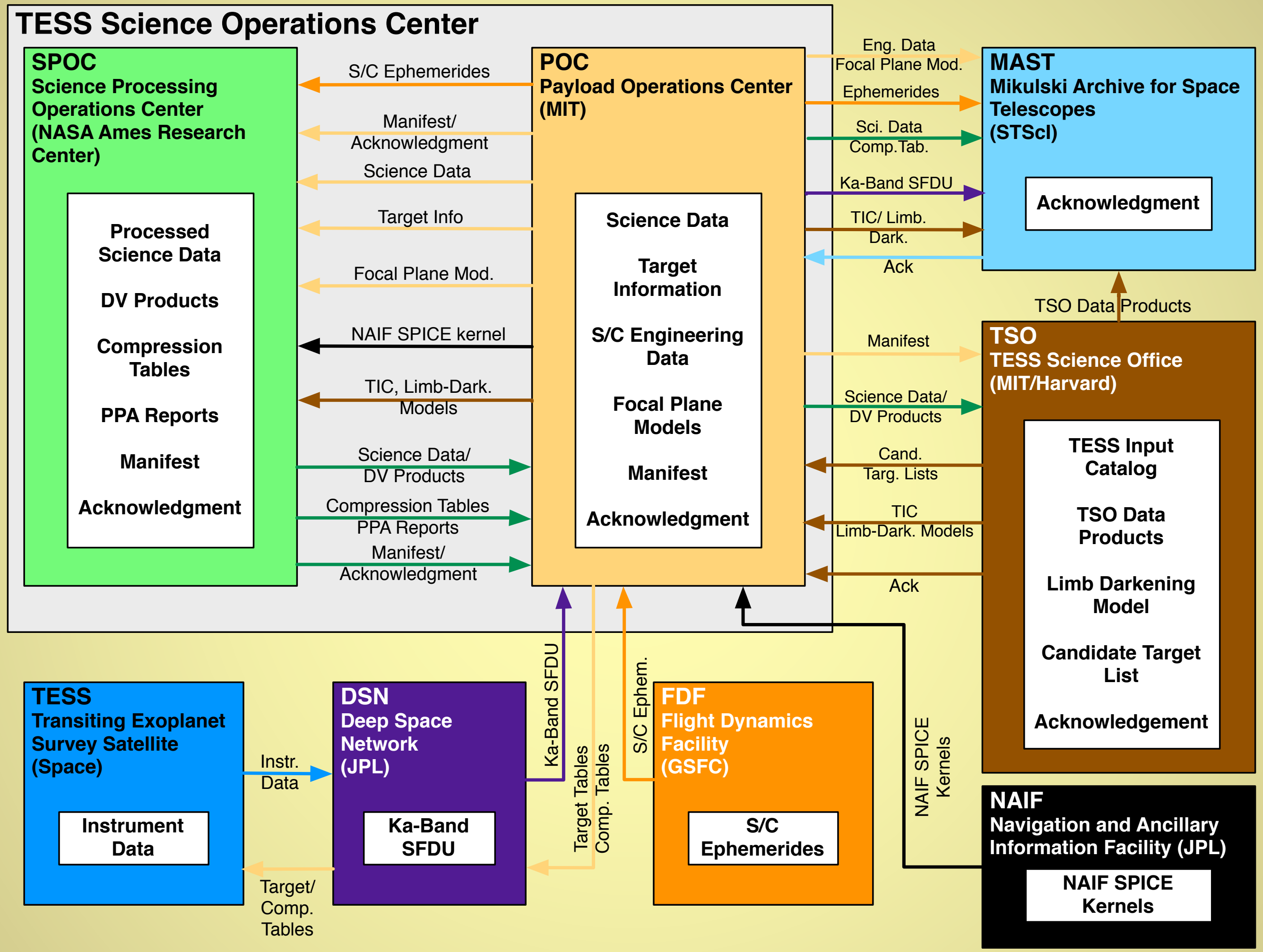
Models

Models

Photometer Performance Assessment (PPA)

Photometer Performance Metrics

Communication is Key



TESS Science Operations Center

SPOC
Science Processing Operations Center (NASA Ames Research Center)

Processed Science Data
DV Products
Compression Tables
PPA Reports
Manifest
Acknowledgment

POC
Payload Operations Center (MIT)

Science Data
Target Information
S/C Engineering Data
Focal Plane Models
Manifest
Acknowledgment

S/C Ephemerides
Manifest/ Acknowledgment
Science Data
Target Info
Focal Plane Mod.
NAIF SPICE kernel
TIC, Limb-Dark. Models
Science Data/ DV Products
Compression Tables PPA Reports
Manifest/ Acknowledgment

MAST
Mikulski Archive for Space Telescopes (STScI)

Acknowledgment

Eng. Data Focal Plane Mod.
Ephemerides
Sci. Data Comp.Tab.
Ka-Band SFDU
TIC/ Limb. Dark.
Ack

TSO Data Products

TSO
TESS Science Office (MIT/Harvard)

TESS Input Catalog
TSO Data Products
Limb Darkening Model
Candidate Target List
Acknowledgement

Manifest
Science Data/ DV Products
Cand. Targ. Lists
TIC Limb-Dark. Models
Ack

NAIF
Navigation and Ancillary Information Facility (JPL)

NAIF SPICE Kernels

NAIF SPICE Kernels

TESS
Transiting Exoplanet Survey Satellite (Space)

Instrument Data

Instr. Data
Target/ Comp. Tables

DSN
Deep Space Network (JPL)

Ka-Band SFDU

Ka-Band SFDU
Target Tables Comp. Tables
S/C Ephem.

FDF
Flight Dynamics Facility (GSFC)

S/C Ephemerides

The TESS Project is distributed geographically with the Science Pipeline separated by a continent from the Science Office'

Resolving data issues requires good communication between the Payload Operations Center, the Science Processing Operations Center and the Science Office

New Ideas for Every Step Will Emerge

New ideas for improving photometry/astrometry will emerge, both within the team and without

- "Halo" photometry on K2 data on Pleiades (White et al. 2017, MNRAS 471)
- "Everest" K2 photometry (Luger et al. arXiv:1702.05488)
- Machine learning/Deep learning neural networks

Preserving ability to re-process the pixel data with better algorithms and tuned parameters is a really good idea

Take advantage of the compressibility of your data

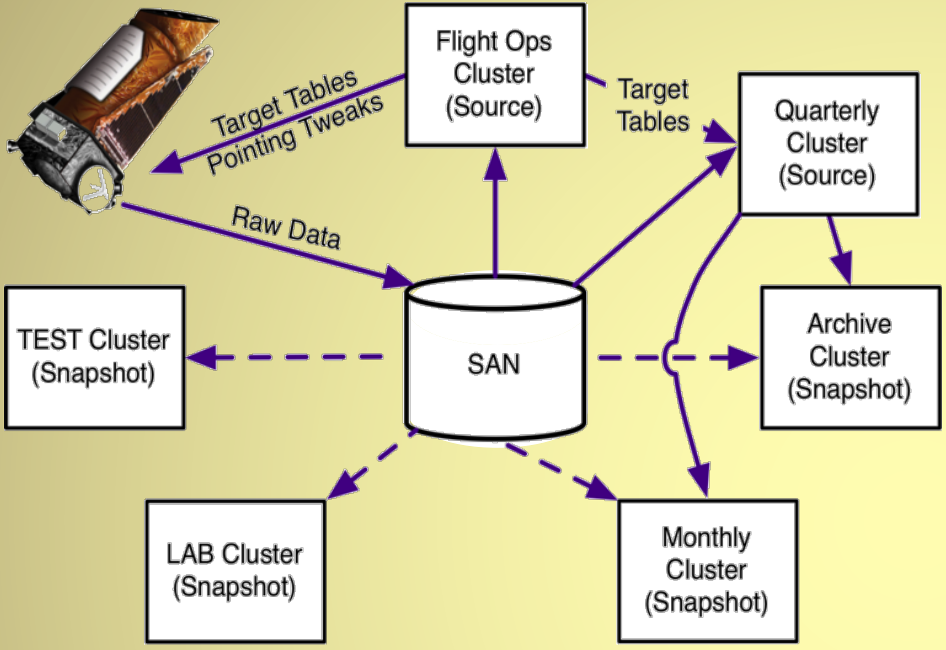
- *Kepler* achieved compression rates of 4.5 bits per pixel
- TESS should achieve compression rates of ~3 bits per pixel for 2 minute data and ~4 bits per pixel for 30 min FFIs

Summary

- Science pipelines require significant planning and effort
- Previous pipelines can be leveraged to reduce development time (but this does not reduce time required for V&V testing)
- Plan to rewrite the majority of the science code in light of unexpected in-flight characteristics/behavior/hardware changes
- High fidelity simulations are indispensable
- Determining η_{earth} is computationally intensive and huge effort
- Give adequate attention to developing commissioning scenarios and associated tools
- Take advantage of data compression to increase the amount of pixel data downlinked from PLATO

SOC Cluster Architecture

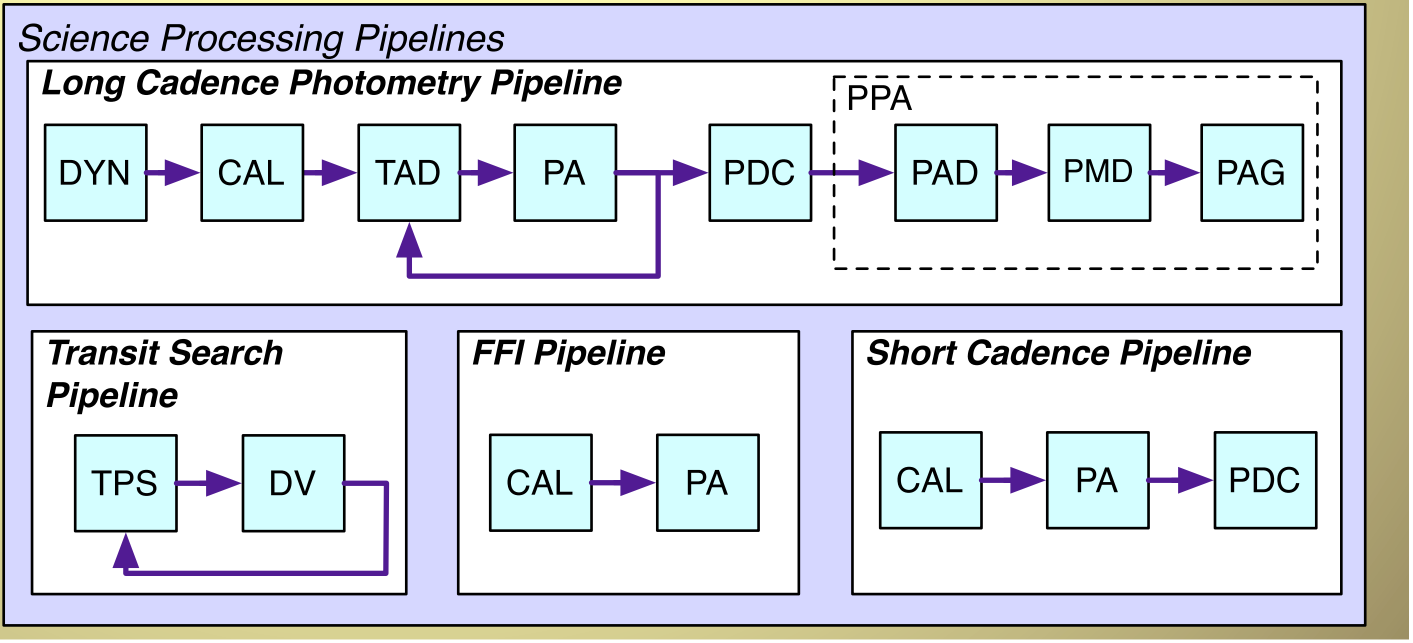
A Search for Earth-size Planets


Flight Ops Cluster (Source)
Target Tables
Pointing Tweaks
Target Tables
Quarterly Cluster (Source)
Raw Data
SAN
TEST Cluster (Snapshot)
Archive Cluster (Snapshot)
LAB Cluster (Snapshot)
Monthly Cluster (Snapshot)

6 Clusters:

4 Operations Clusters:
Flight Ops, Quarterly, Monthly & Archive)

2 Test Clusters:
LAB & TEST


Science Processing Pipelines

Long Cadence Photometry Pipeline

DYN → CAL → TAD → PA → PDC → PPA (PAD → PMD → PAG)

Transit Search Pipeline

TPS → DV

FFI Pipeline

CAL → PA

Short Cadence Pipeline

CAL → PA → PDC