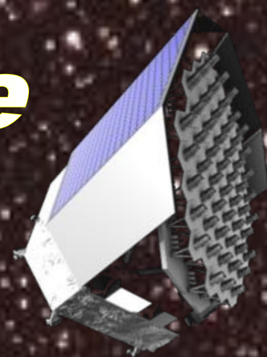
Lessons Learned from Developing and Operating the Kepler Science Pipeline and Building the TESS Science Pipeline



Jon M. Jenkins

NASA Ames Research Center

Monday September 10, 2018

Institute for Planetary Science

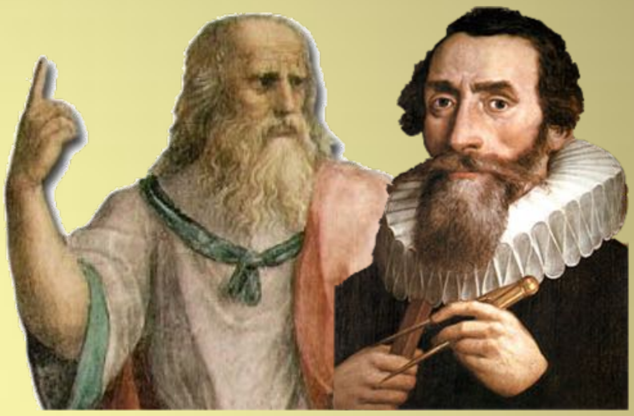
DLR German Aerospace Center

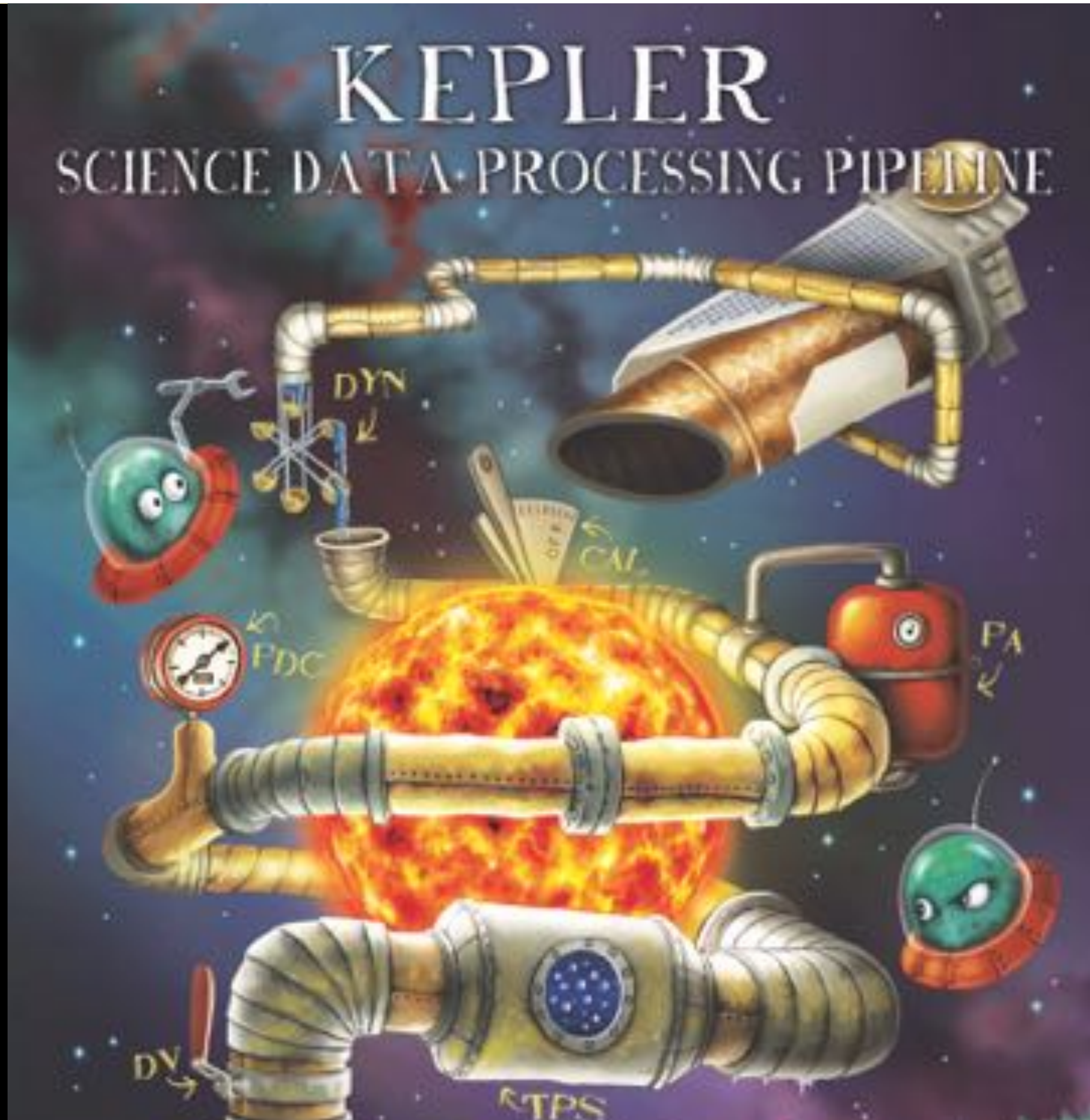
Berlin, Germany

Overview

A Search for Earth-size Planets

- 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
DYN
CAL
PA
PDC
DV
TPS

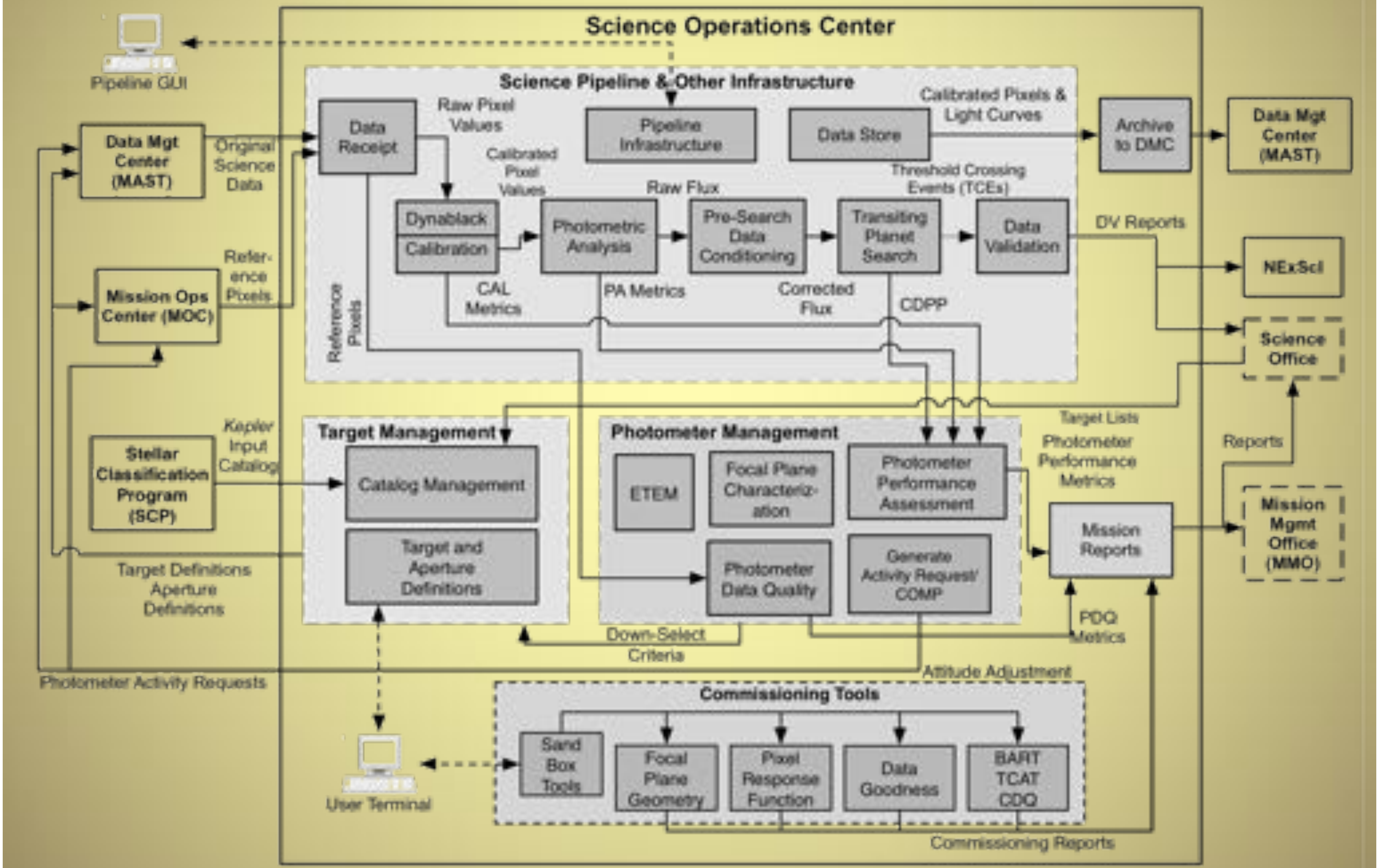
The Science Operations Center: What did it take?

A Search for Earth-size Planets

- 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

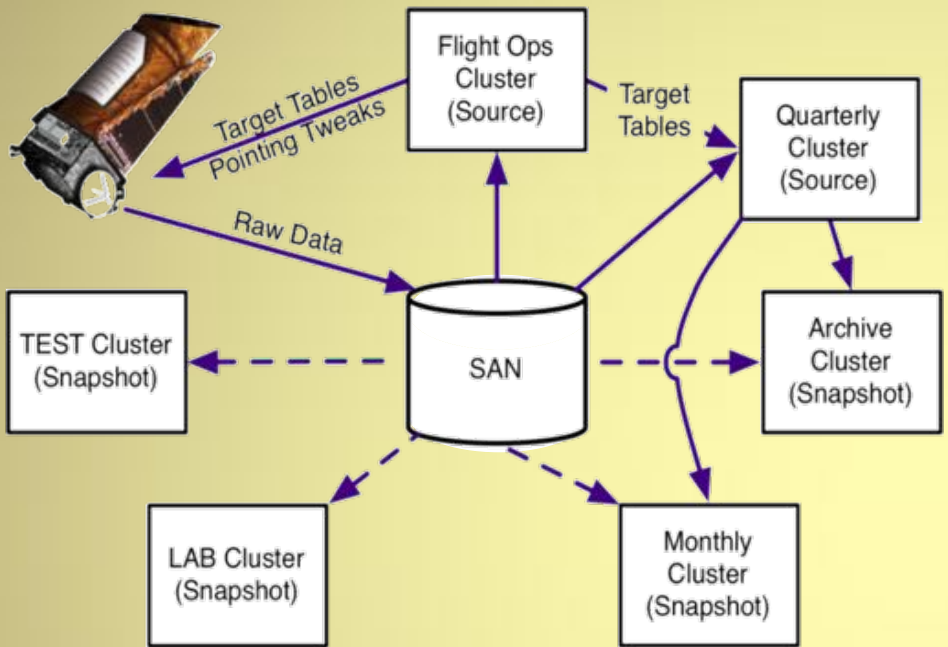
Science Operations Center Architecture

A Search for Earth-size Planets

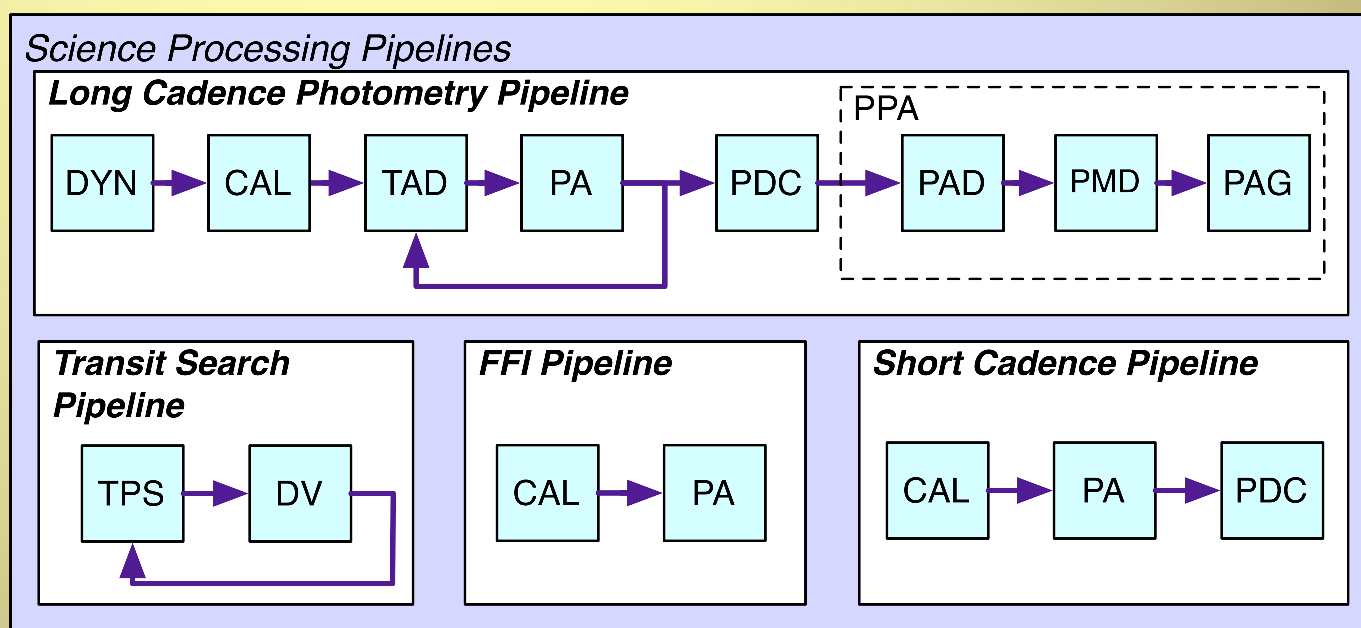


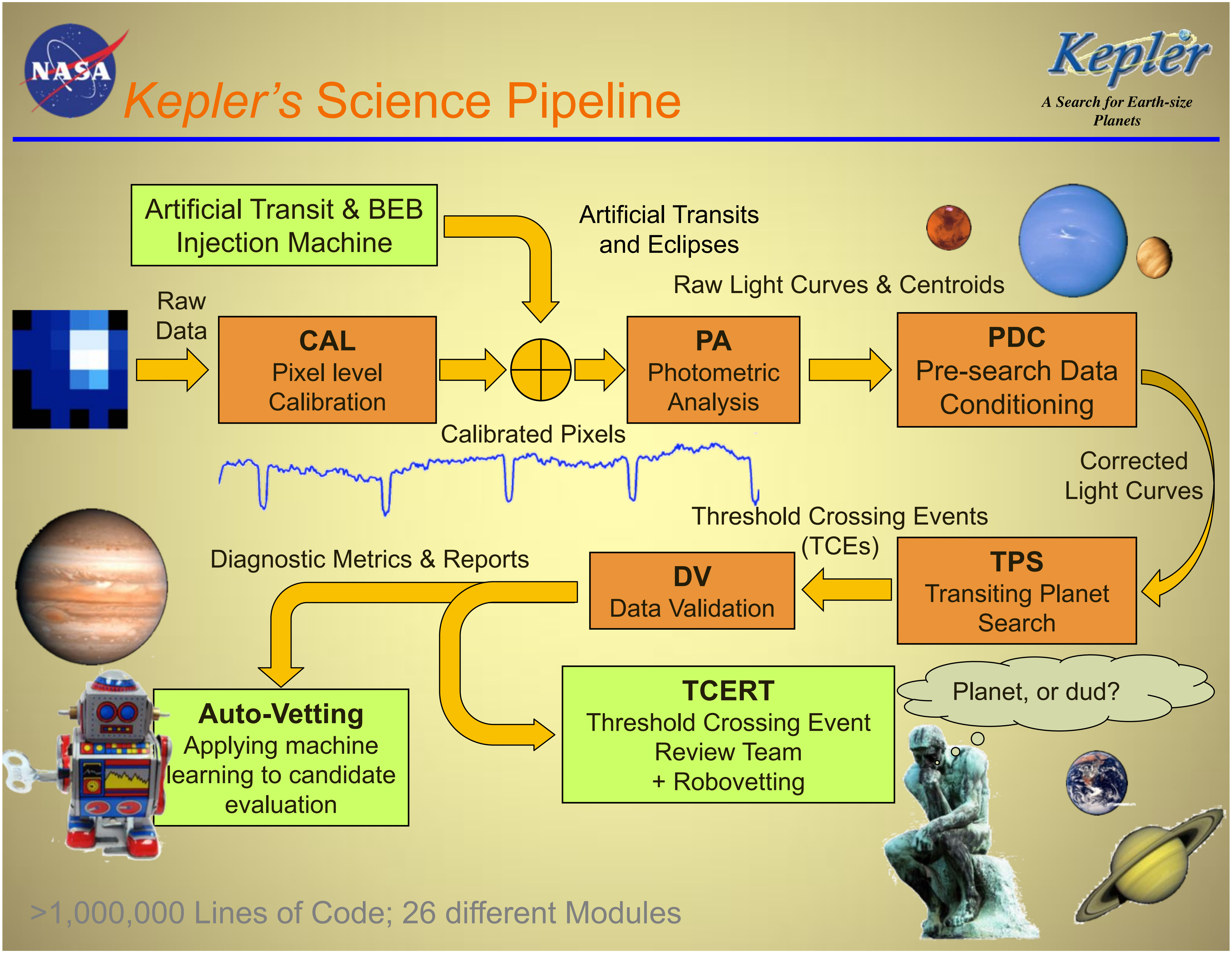
SOC Cluster Architecture

Kepler
A Search for Earth-size Planets



6 Clusters:
4 Operations Clusters:
Flight Ops, Quarterly, Monthly & Archive)
2 Test Clusters:
LAB & TEST




NASA
Kepler's Science Pipeline
Kepler
A Search for Earth-size Planets
Artificial Transit & BEB Injection Machine
Artificial Transits and Eclipses
Raw Data
CAL
Pixel level Calibration
Calibrated Pixels
PA
Photometric Analysis
Raw Light Curves & Centroids
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

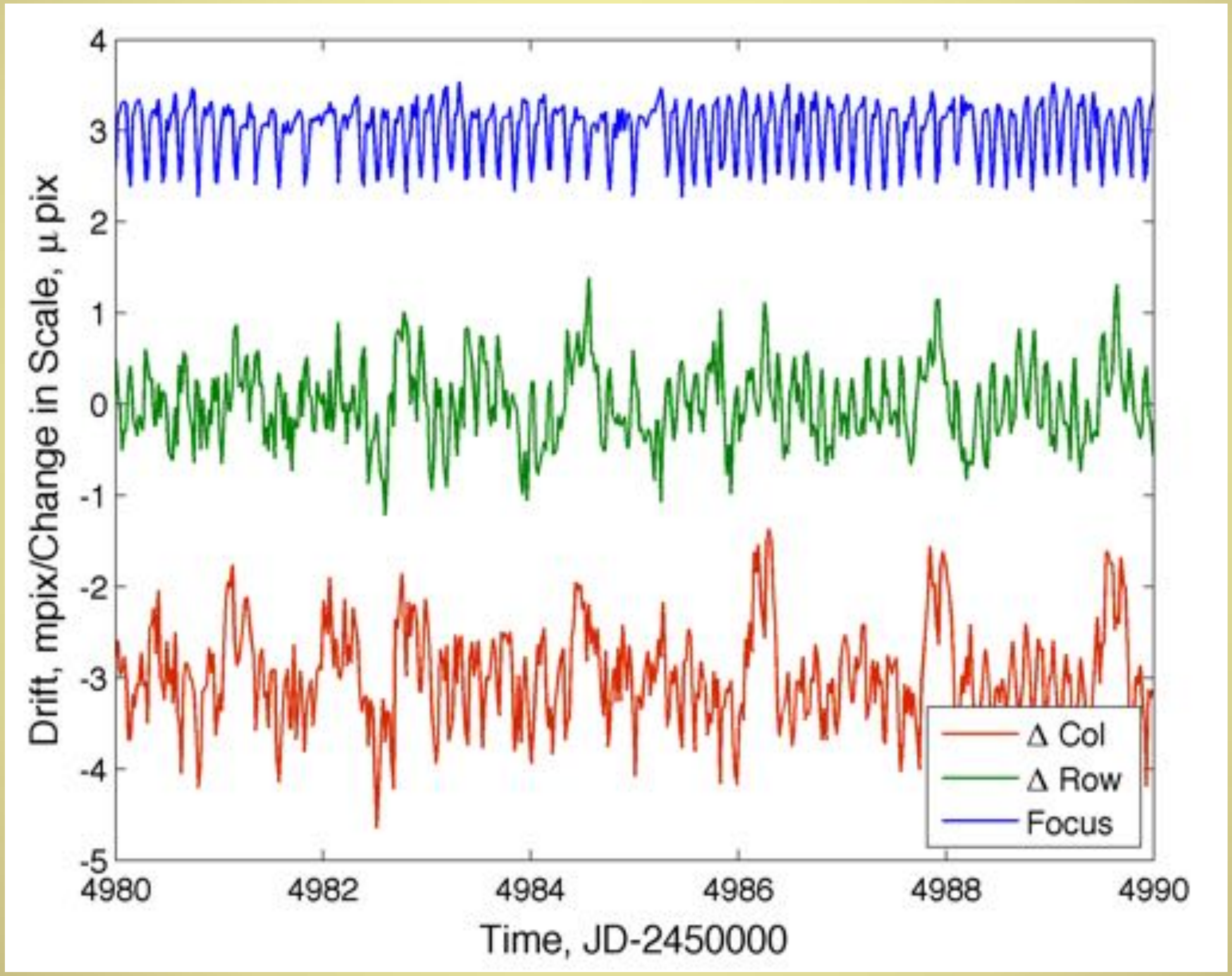
Data Validation

- Difference image analysis
- Ghost Diagnostic + other metrics

~80% of the science code was re-written

Short Timescale Instrumental Errors

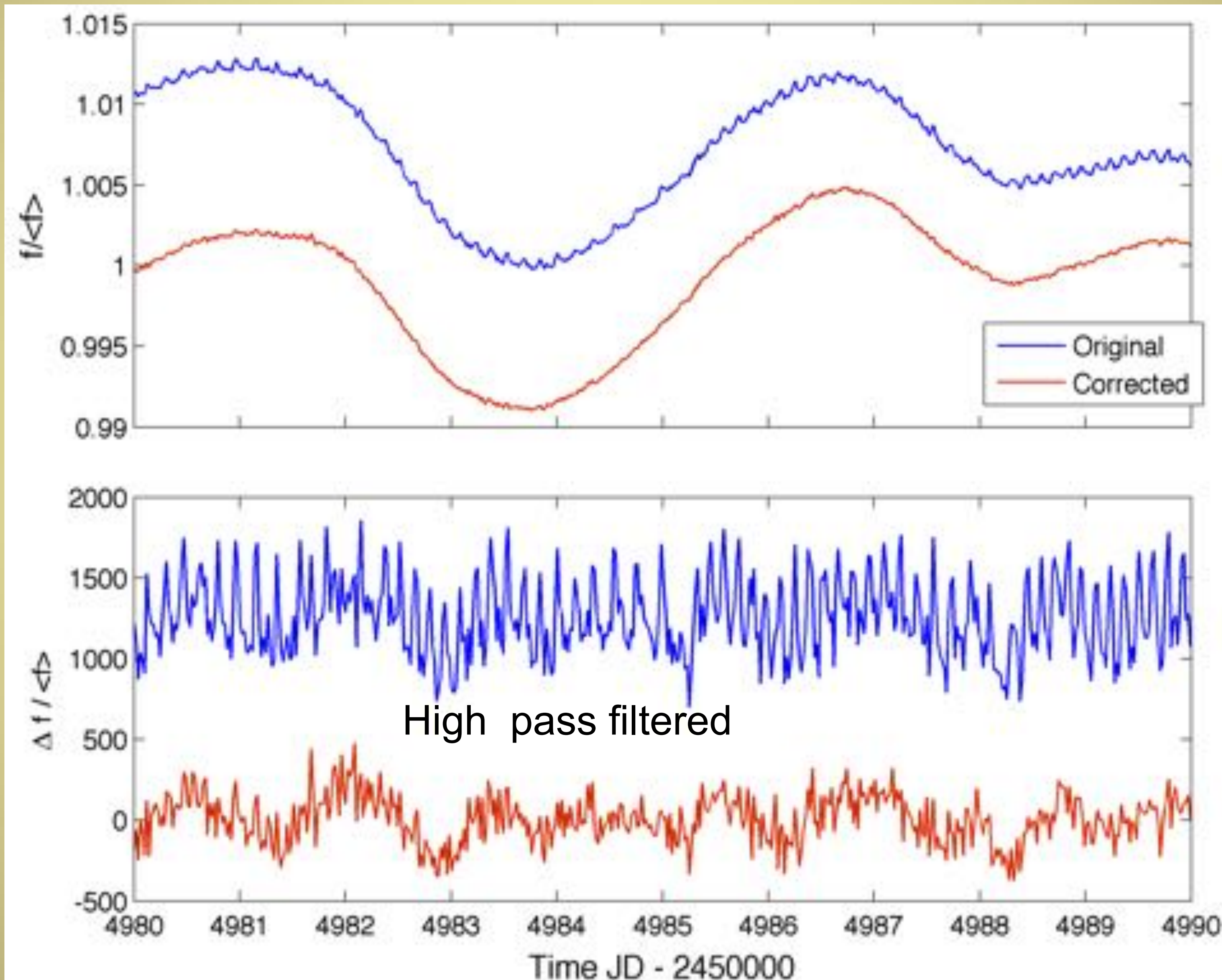
Signature of a heater cycling on the reaction wheels 3/4



Kepler is sensitive to its thermal environment

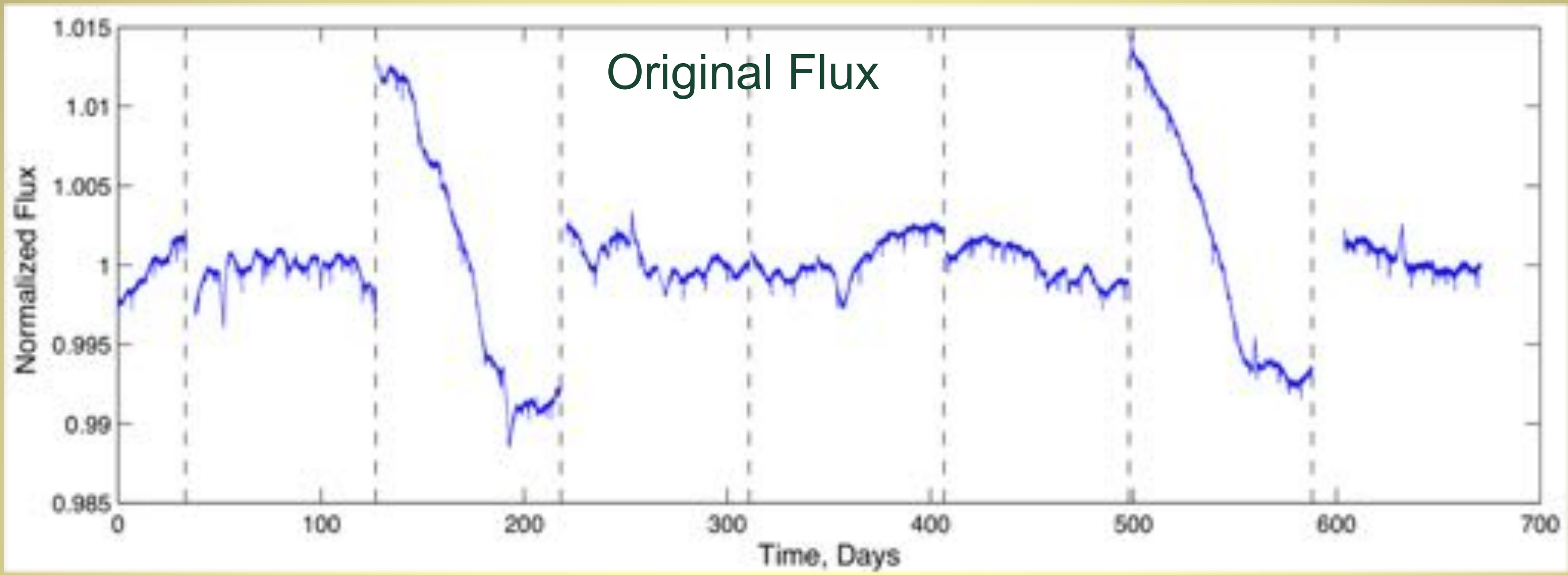
Instrumental Effects in Photometry

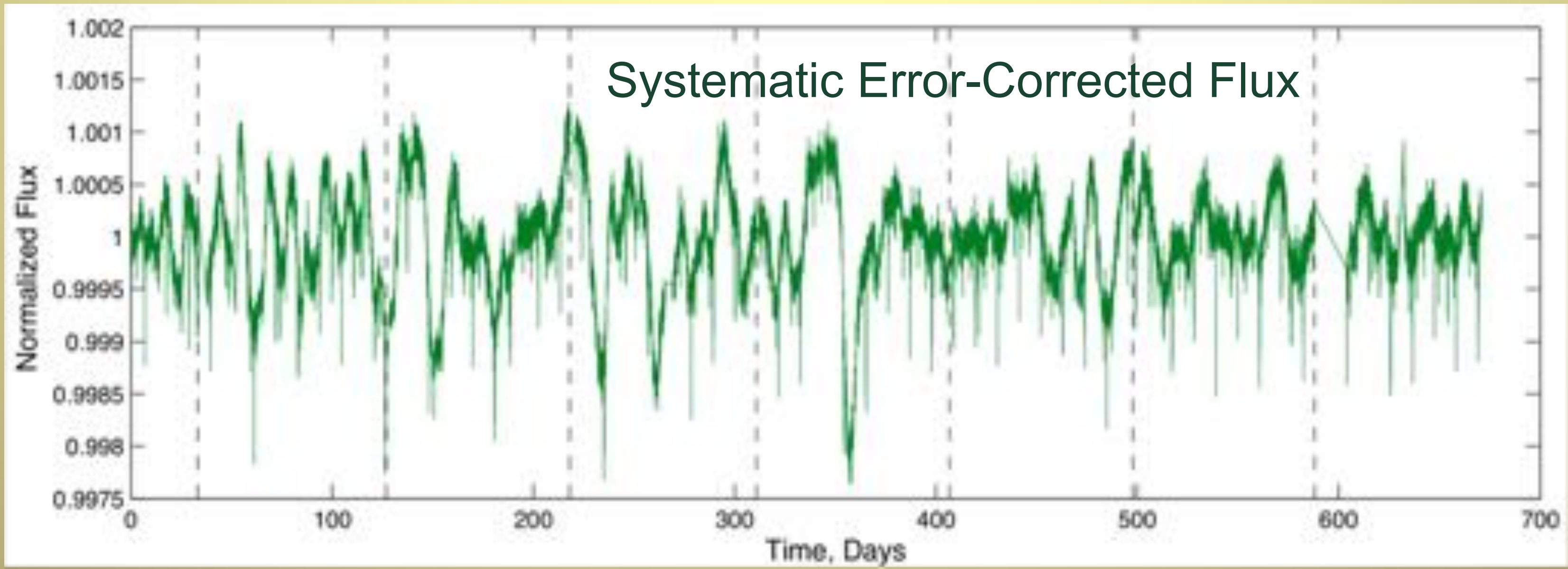
Kepler
A Search for Earth-size Planets



Correcting Systematic Errors

Kepler
A Search for Earth-size Planets



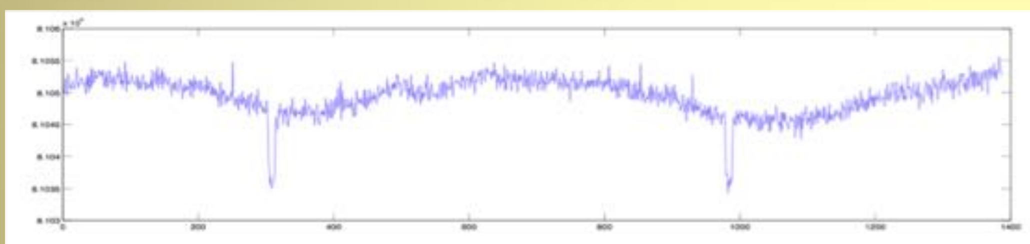


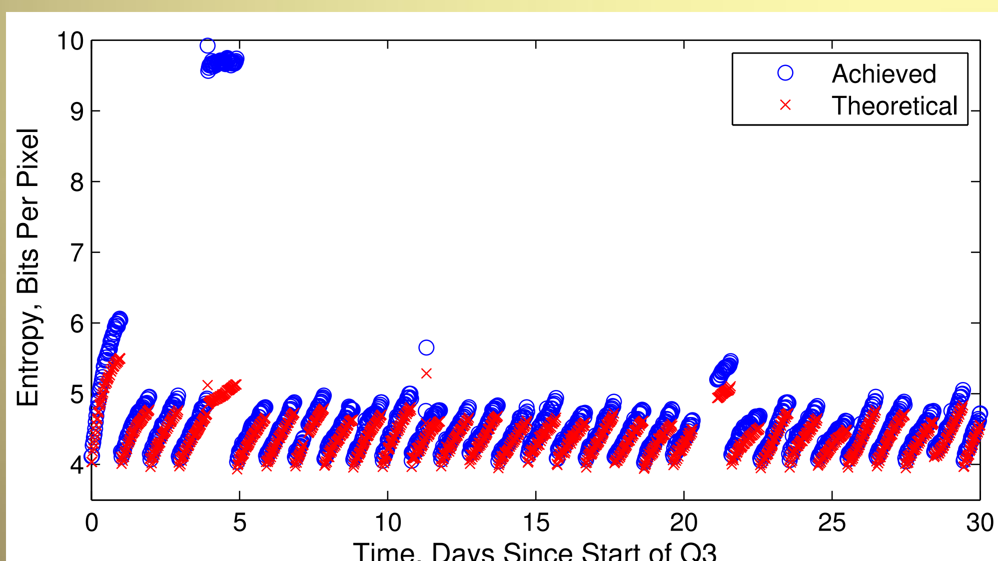
We apply a Maximum A Posteriori approach as per Stumpe et al. 2014

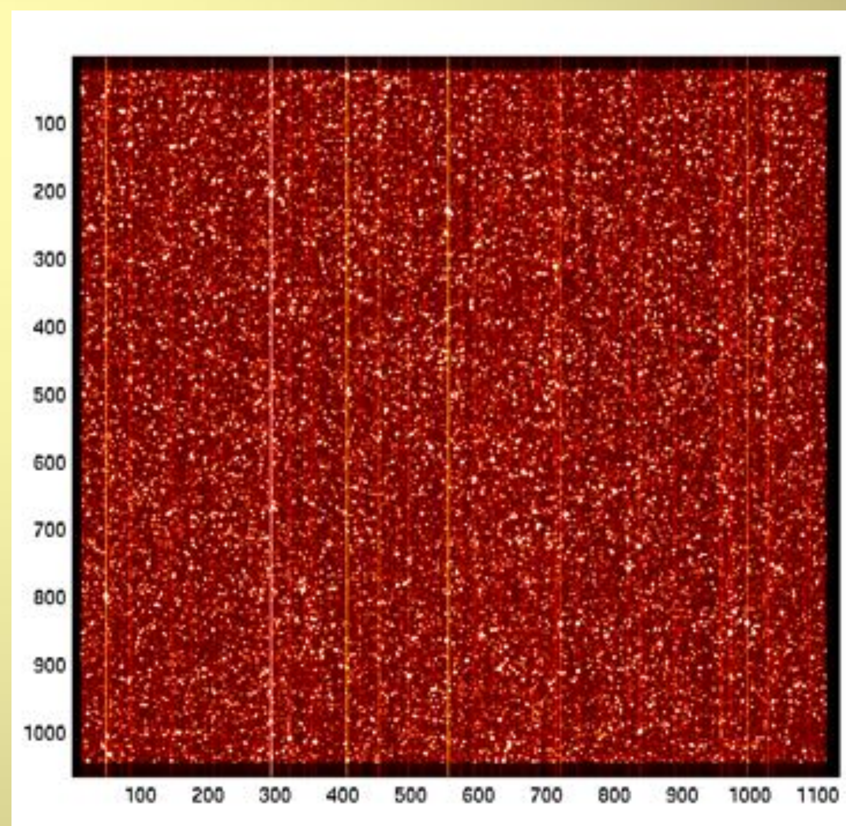
High Fidelity Simulations are Indispensable

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







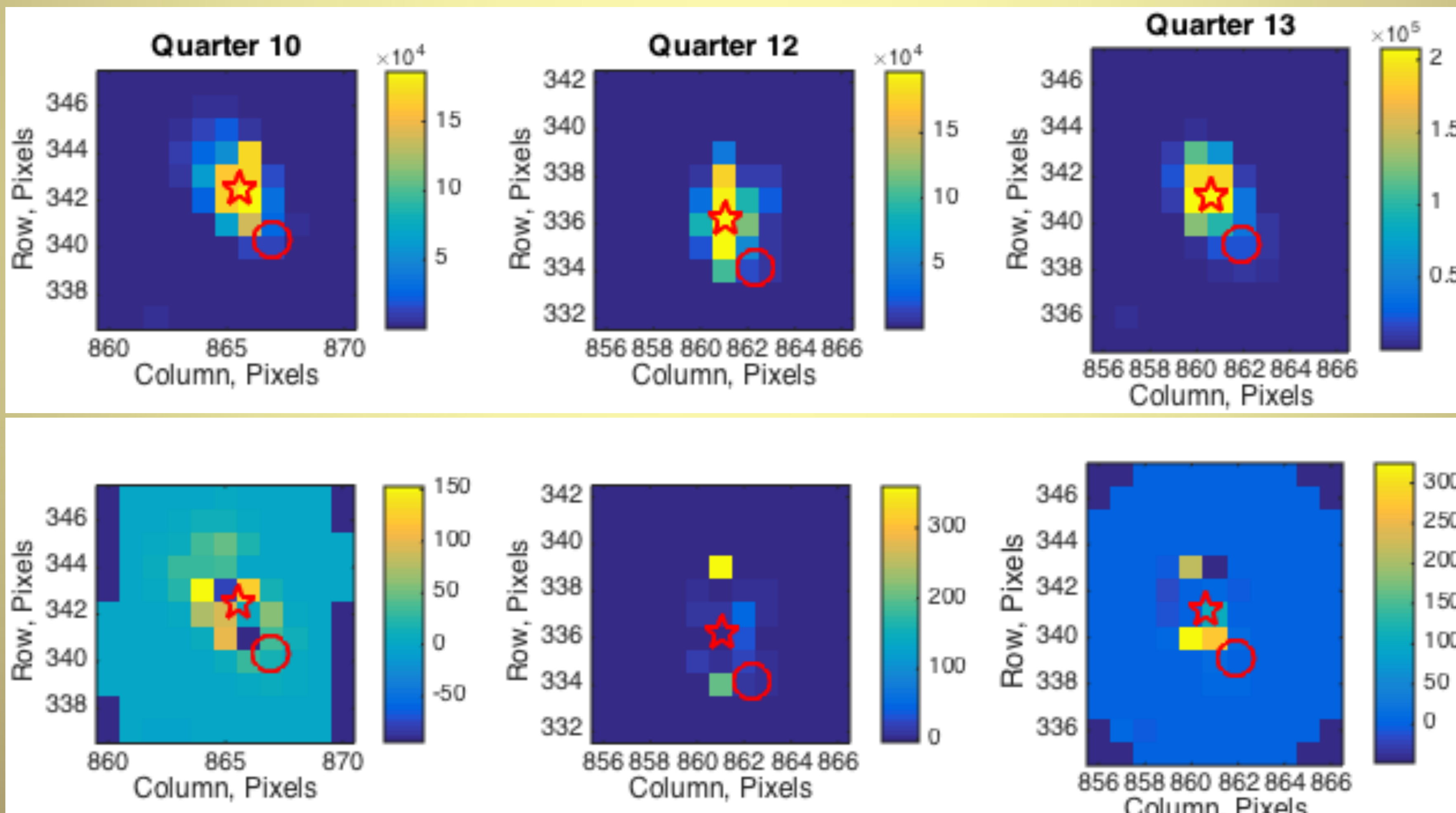
Difference Image Analysis

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

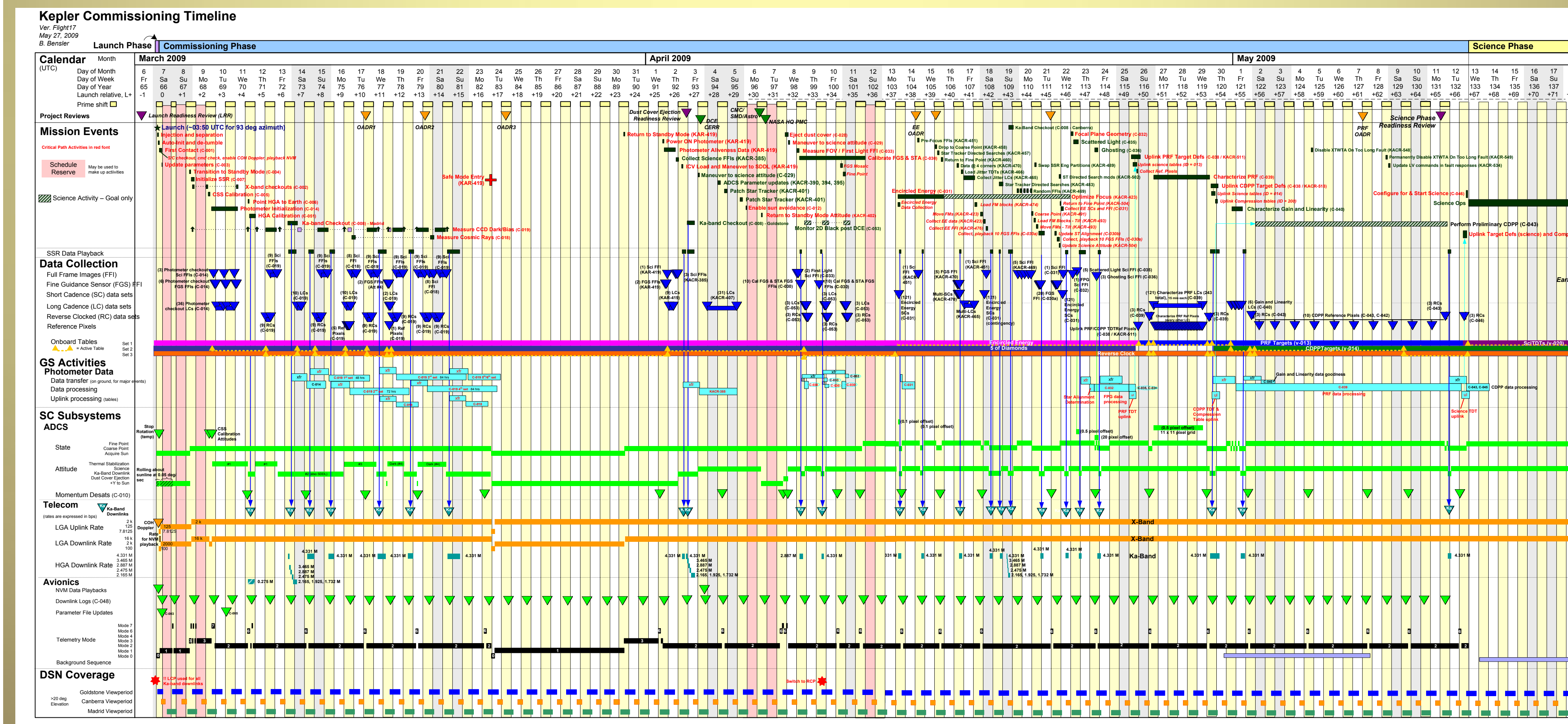


Rappaport et al.2017, arxiv1708.06069

Commissioning, Commissioning, Commissioning!

A Search for Earth-size Planets

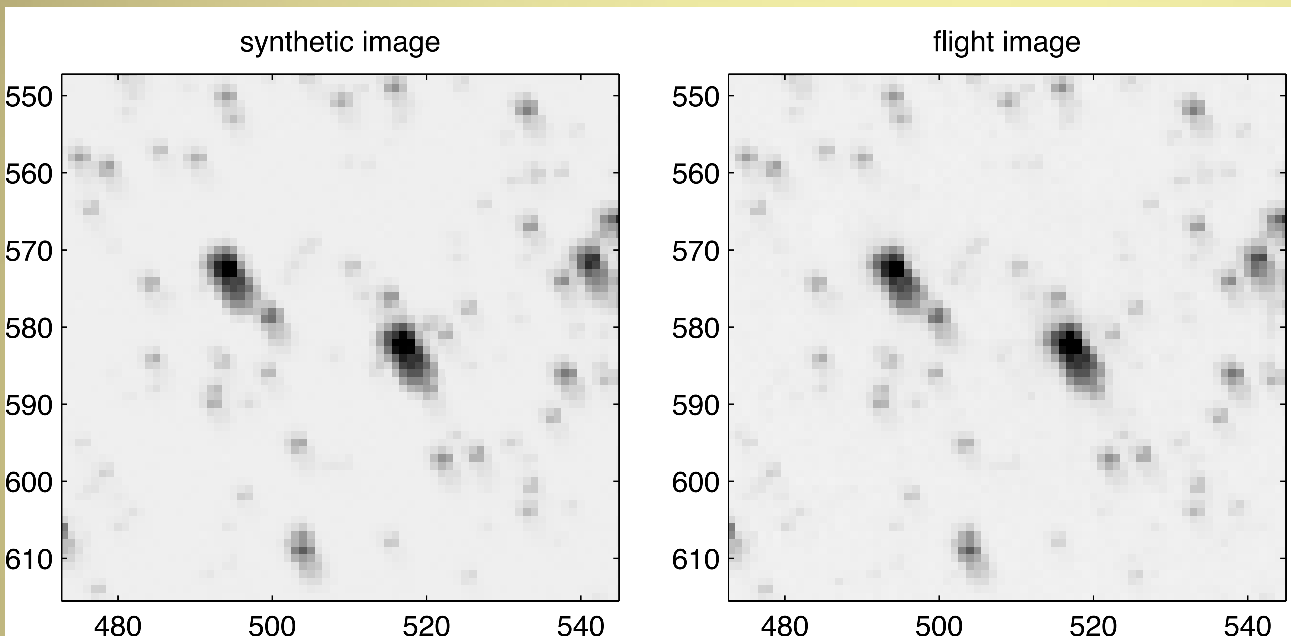
- Commissioning tools require special attention and data sets
- Effort for commissioning tools may be as great as that for major science pipeline modules
- Don't leave commissioning tool development to the last



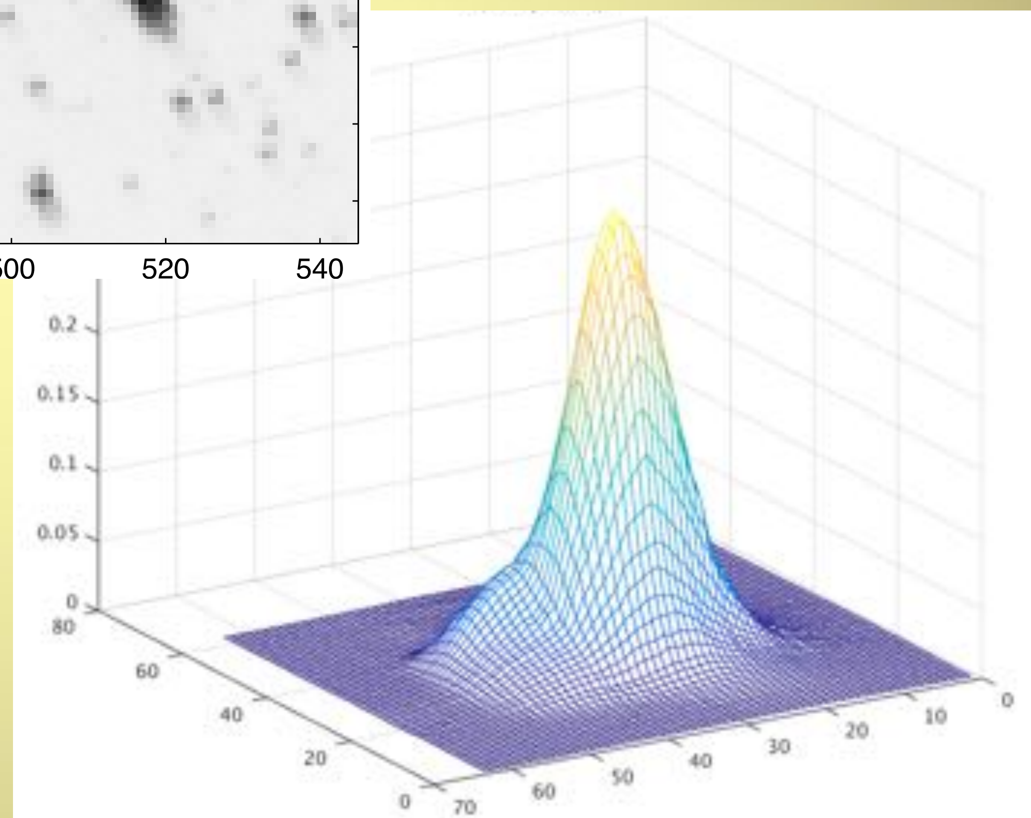
A Search for Earth-size Planets

Pixel Response Function Characterization

Kepler PRF



TESS PRF

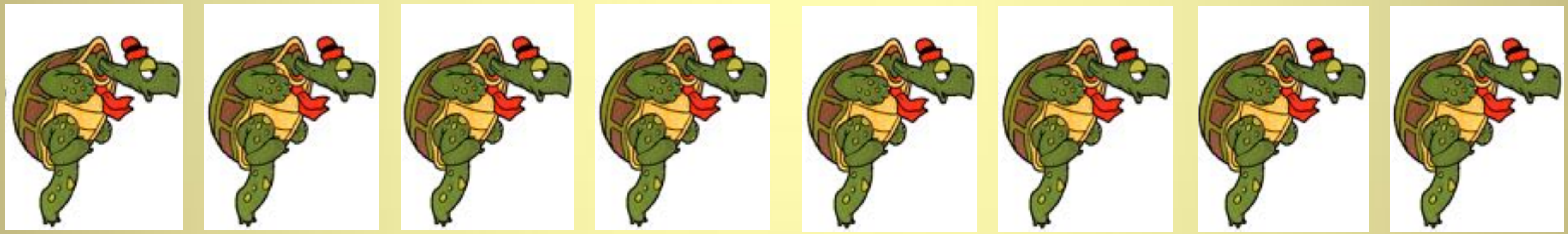


Keeping Up with the Data

Improving the Throughput

Some fast code; Some slow code

Step 1: Parallelize all code



Step 2: Make slow code fast(er)

Hardware Architecture: *Kepler* Science Operations Center

A Search for Earth-size Planets

64 hosts, 712 CPUs,

3.7 TB of RAM,

148 TB of raw disk storage

Hardware Architecture: NAS Pleiades Supercomputer

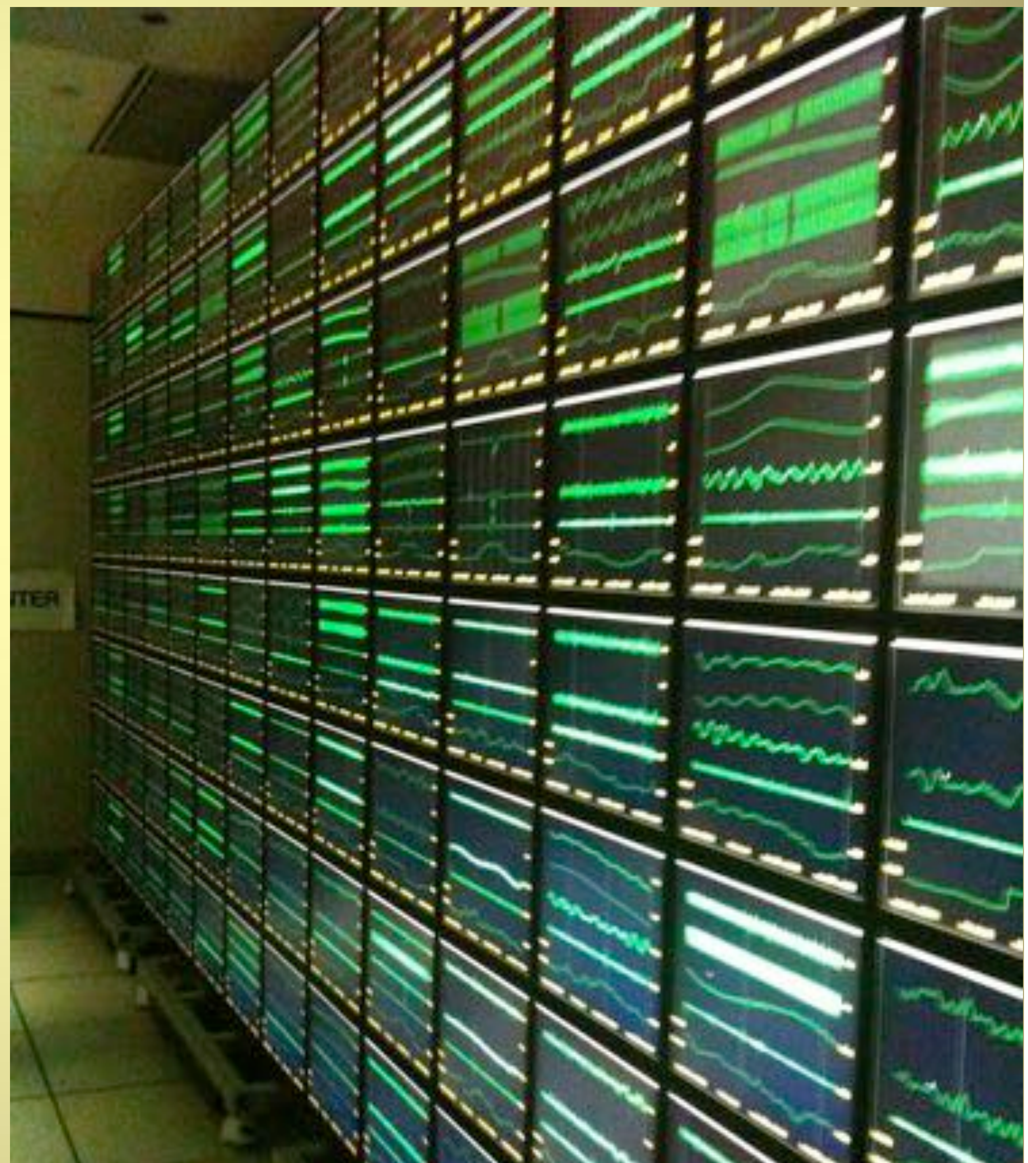
Kepler
A Search for Earth-size Planets

5.34 Pflop/s peak cluster

211,872 cores

724 TB of memory

15 PB of storage



η_{earth}: Mapping Completeness and Reliability

Kepler
A Search for Earth-size Planets

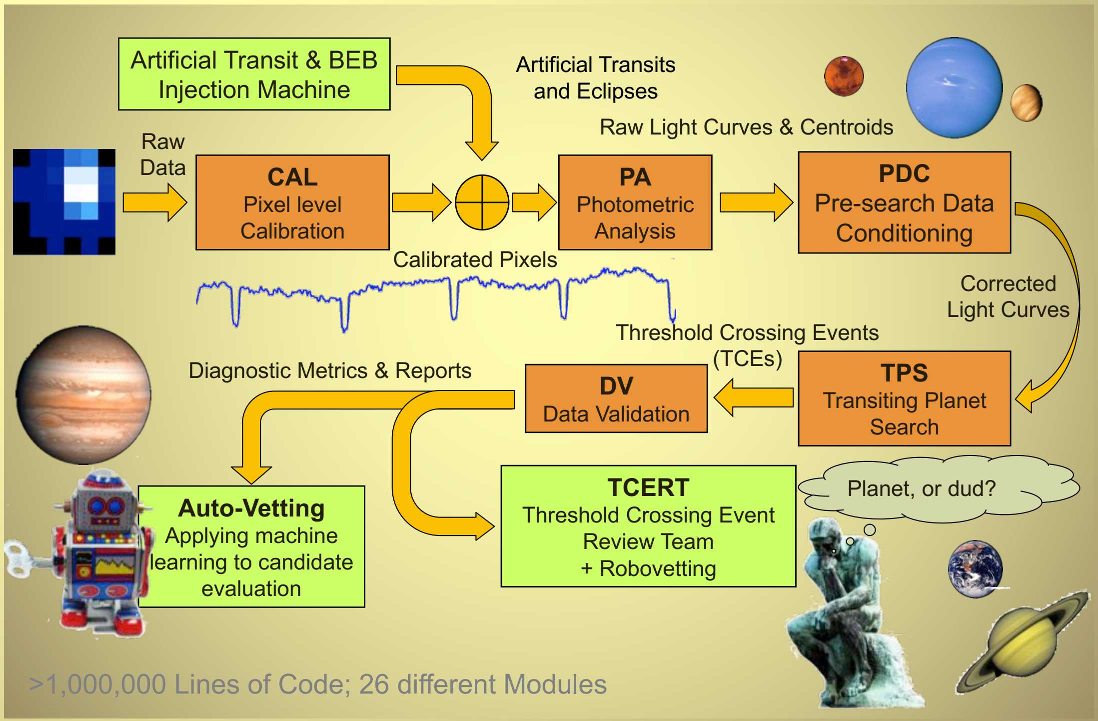
Characterizing completeness and reliability of software/people pipelines is extremely resource intensive

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

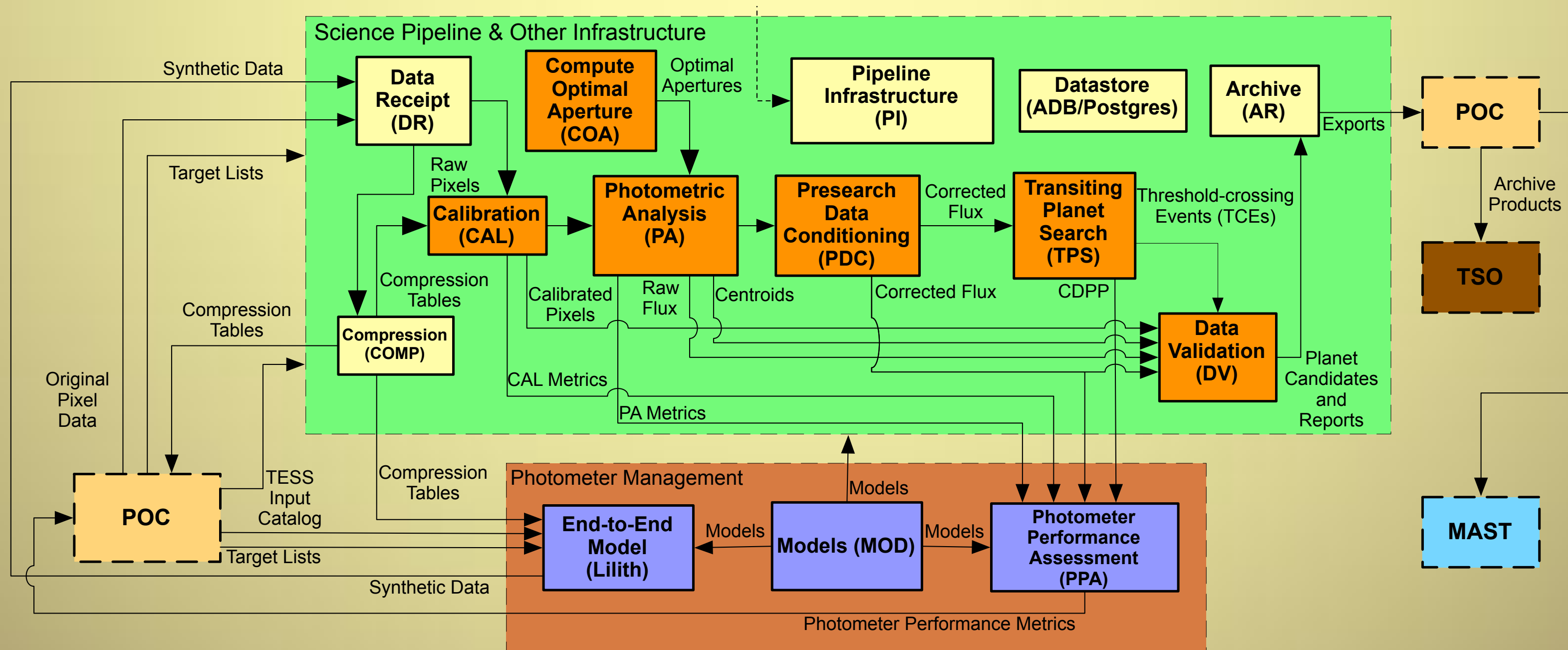
Recommendation: Pursue machine learning for conducting or modeling the candidate vetting process



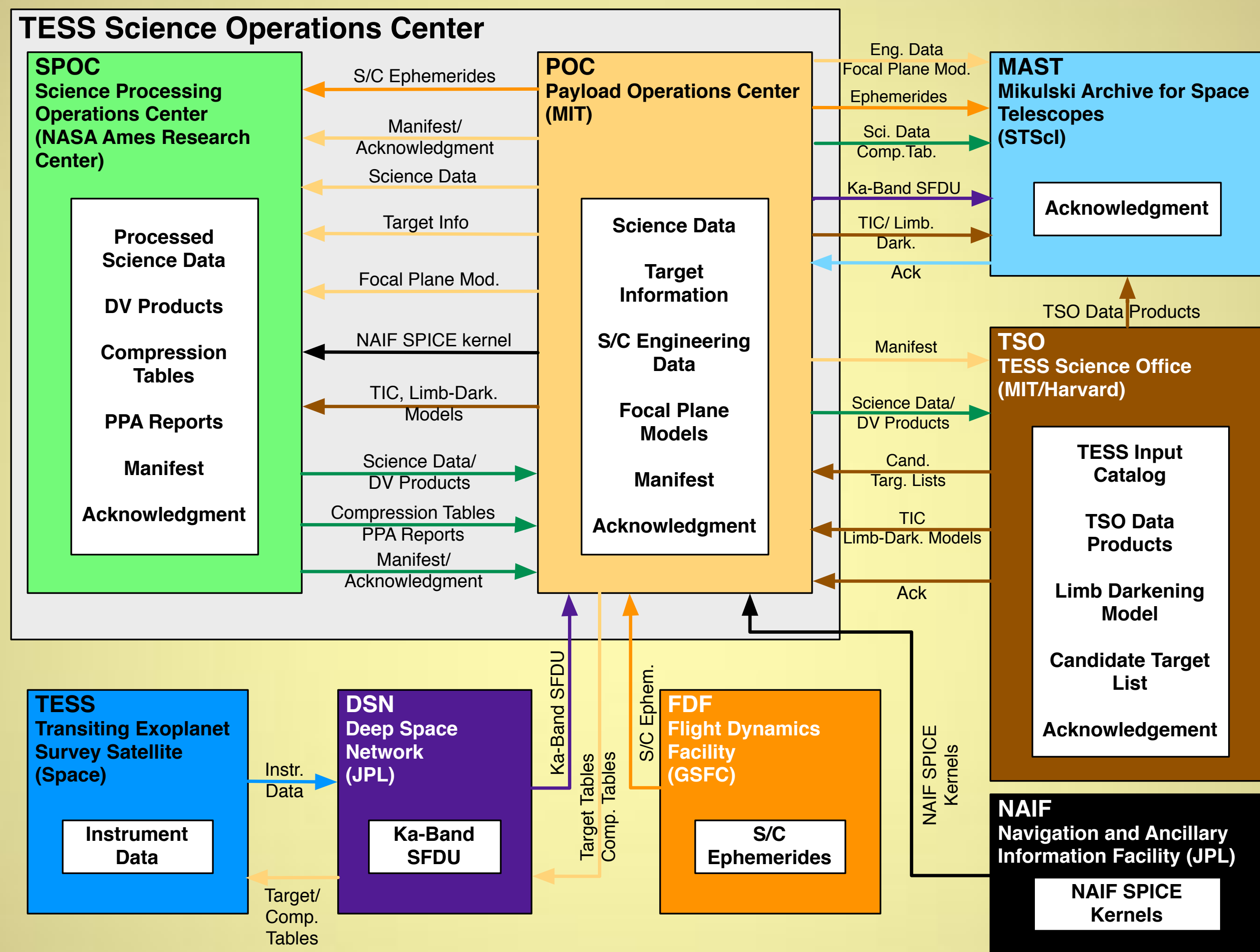
Developing the TESS Pipeline

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



Communication is Key



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

Managing Kepler's Data Volume

A Search for Earth-size Planets

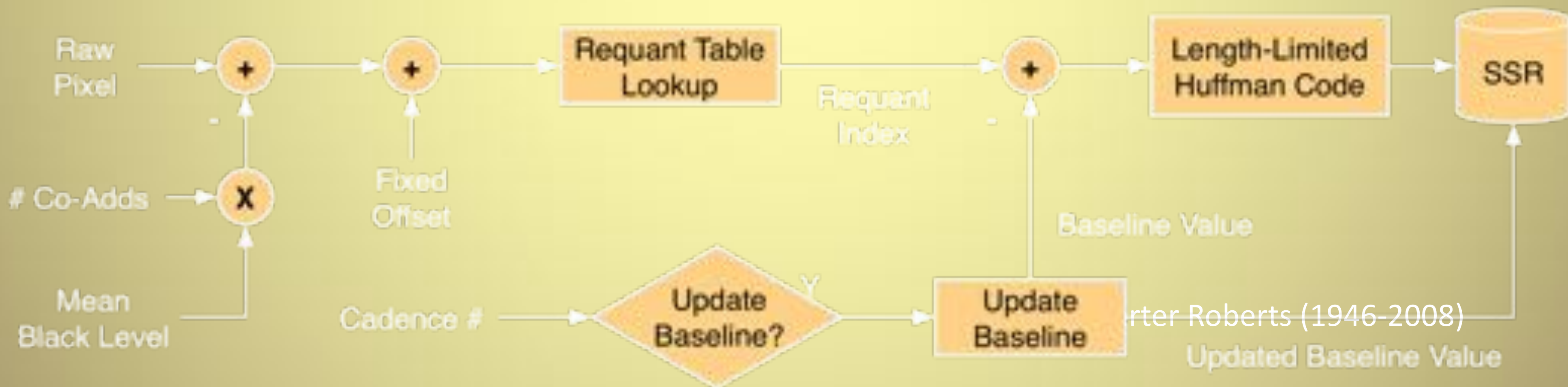
Key Requirements:

Collect and store image data for up to 170,000 stars for up to 66+ days

Downlink 31 days of data in less than 24 hours

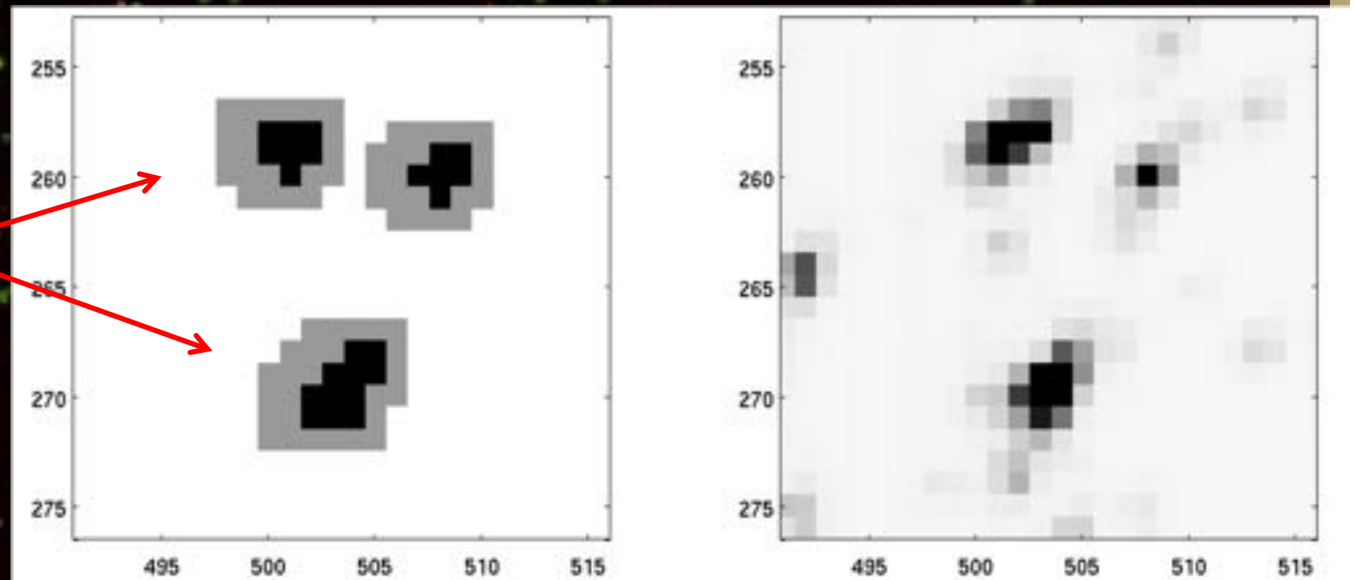
Solution:

- Collect ***only*** pixels of interest
- Compand data to manage quantization noise as a term in noise budget
- Detrend the data to centralize the measurements to be transmitted
- Entropically encode the data before CCSDS packetization (must be robust)



Step 1: Selecting Pixels of Interest

Pixels of interest specified for downlink



Generate synthetic images of each star and its background scene

Perform SNR calculation to identify pixels of interest for photometry

Add a 1-pixel buffer ring

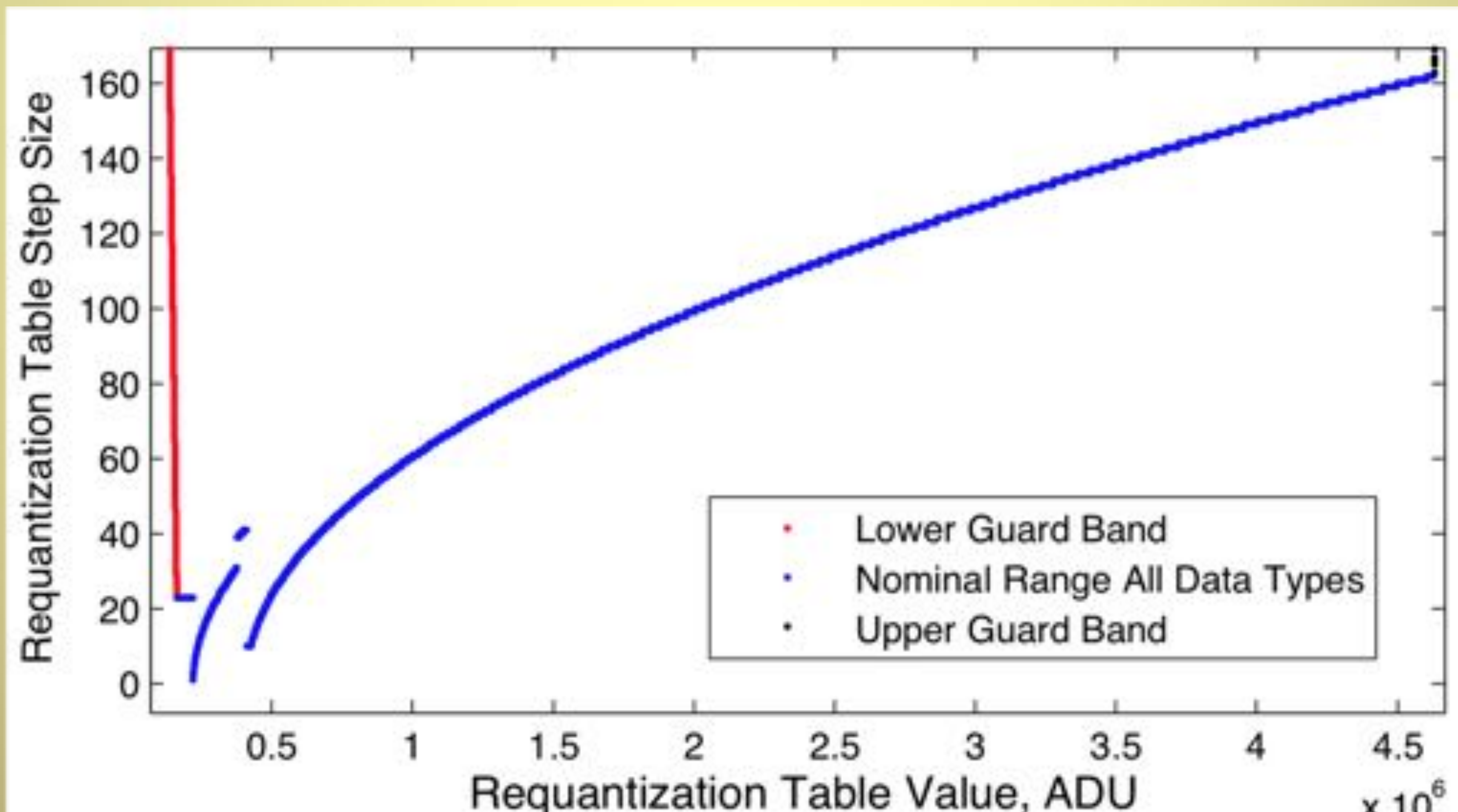
Choose smallest of 1024 masks that will cover POI

Only 4% excess pixels collected

Step 2: Requantization

A Search for Earth-size Planets

- Manage quantization noise in noise budget – ¼ of intrinsic measurement noise
- Must account for dispersion of read noise, gain, and bias across 84 channels!
- Raw pixel measurements are represented with 23 bits
- Requantization compresses the data to 16 bits



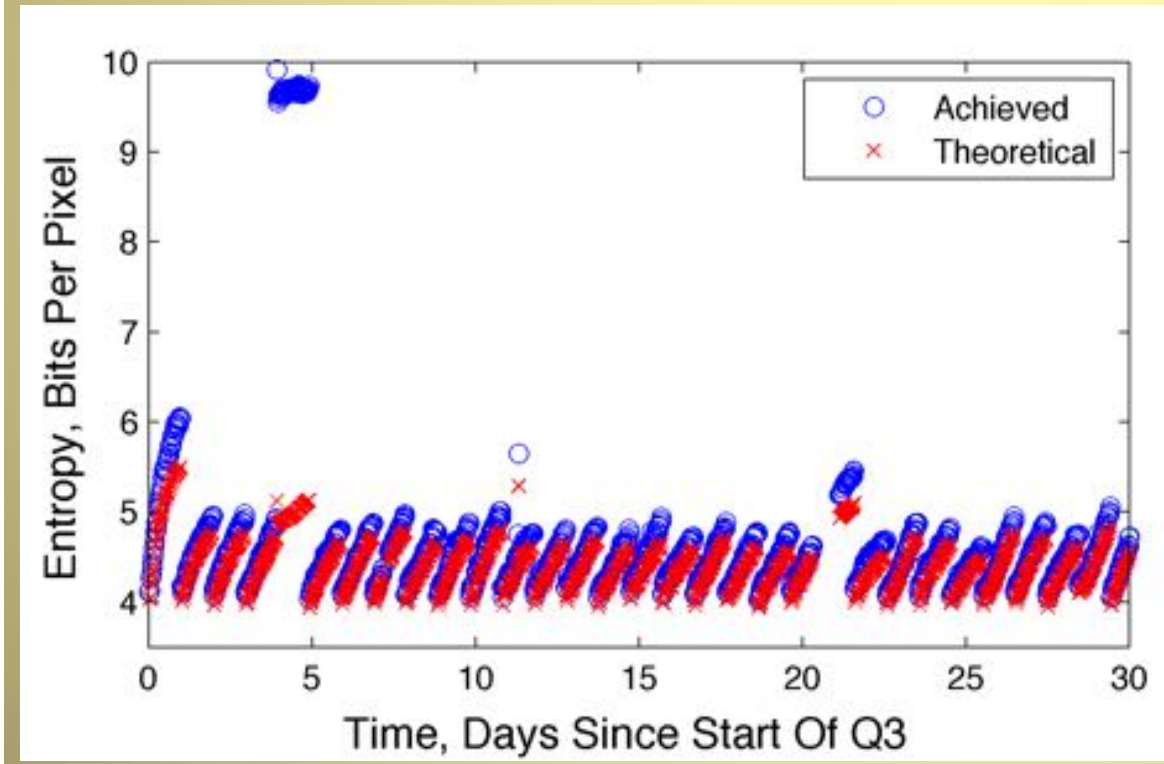
2008)

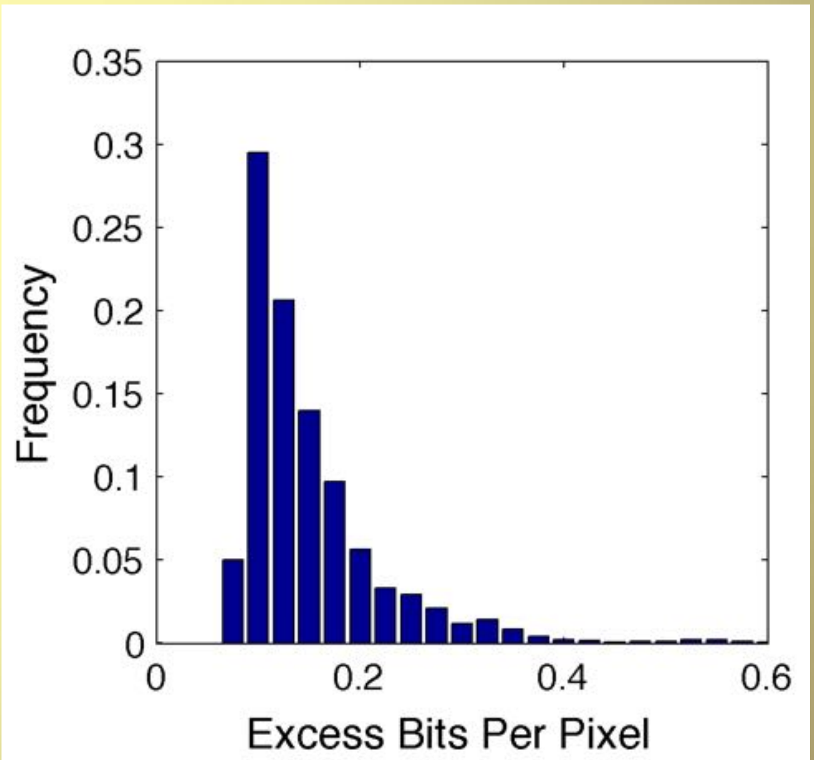
Step 3: Huffman Coding

Subtract a baseline measurement (changes on a daily basis)

Entropically encode residuals with a Length-Limited Huffman Code

Compresses the data to 4.4 bits/symbol





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