STScI | SPACE TELESCOPE SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

Hit the Ground Running: Data Management for JWST

Mark Kyprianou

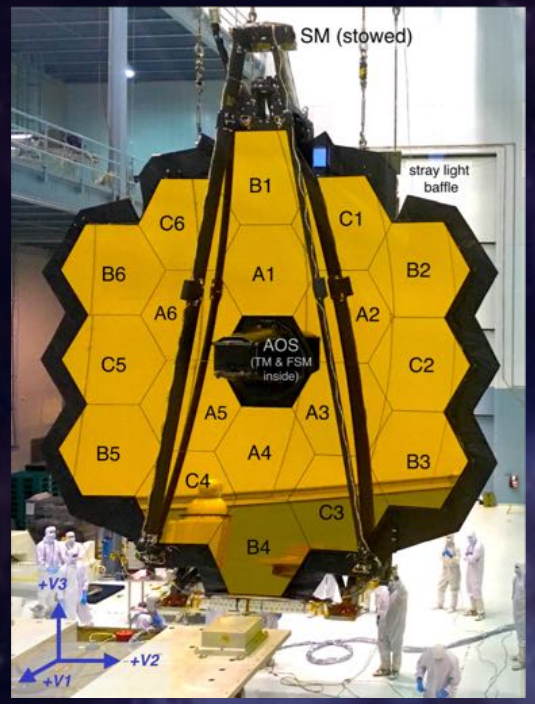
JWST DMS Leads:

Anastasia Alexov, Mark Kyprianou, Katie Kaleida

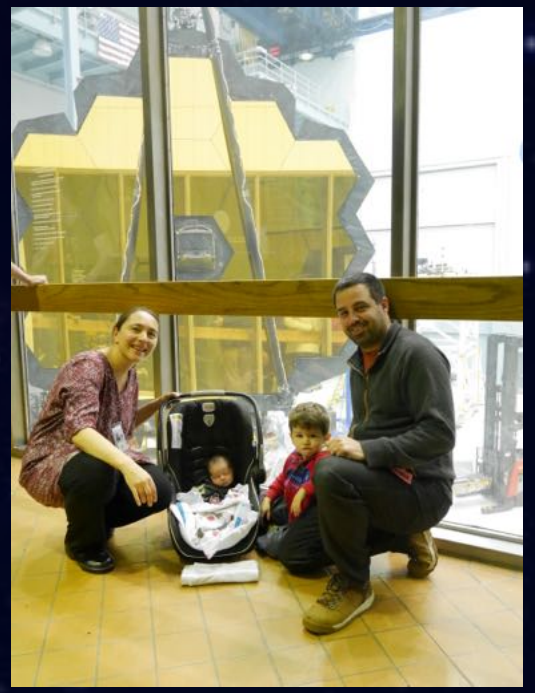
November 15, 2018

ADASS XXVIII

Katie Kaleida



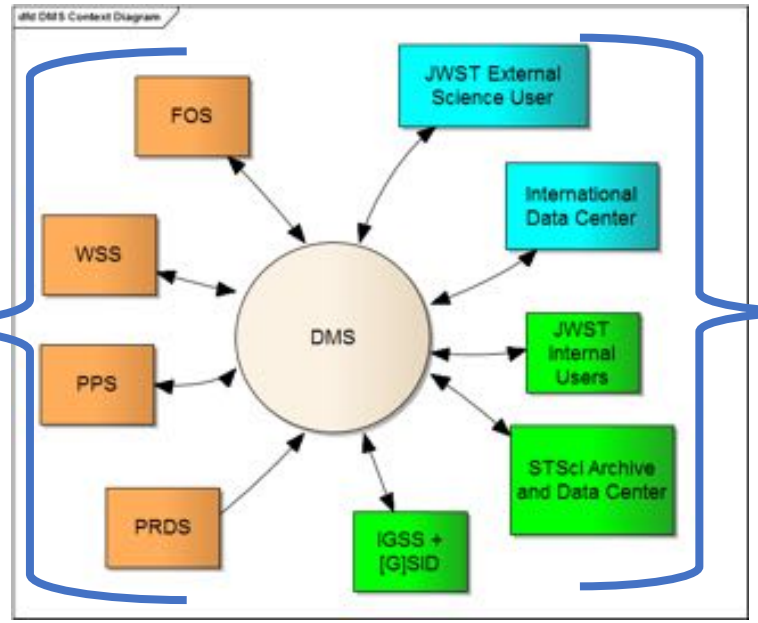
JWST DMS Introduction



Quick-n-Dirty Acronyms

- S&OC – Science and Operations Center (pronounced as “sock”)
  - FOS – Flight Operations Subsystem
  - PRDS – Project Reference Database Subsystem
  - WSS – Wave Front Sensing and Control (WFS&C) Software Subsystem
  - PPS – Proposal Planning Subsystem
  - OSS – Operations Script Subsystem
  - DMS – Data Management Subsystem

Internal S&OC interfaces



External & user interfaces

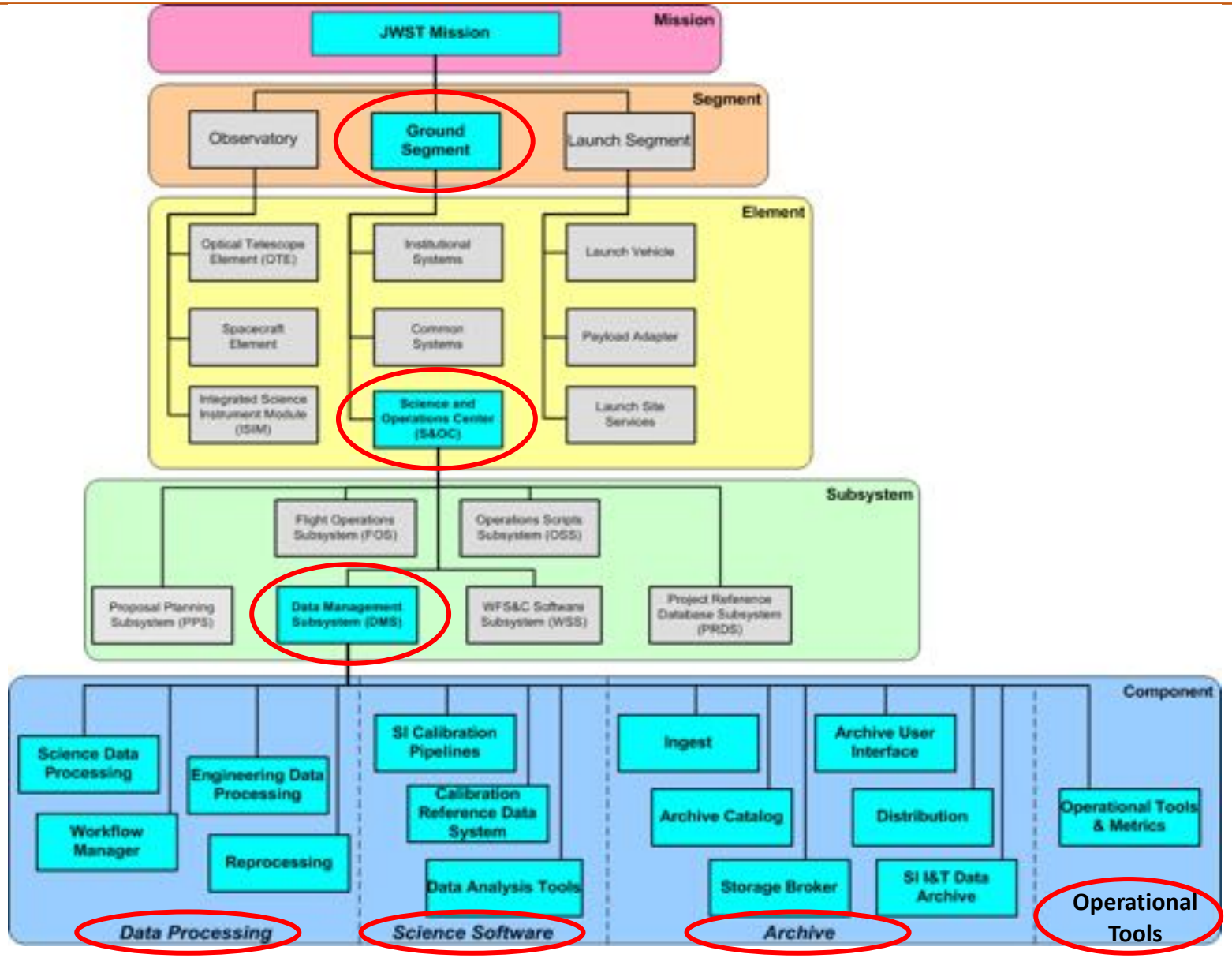
JWST DMS - Who are we?

JWST Data Management Subsystem (~39 FTEs)

- Focused on development work on JWST end-to-end data processing, calibration, archiving, data access services and data analysis tools
- Work is spread over 5 Engineering Branches with over 100 staff
- All roads from the S&OC subsystems meet at DMS, requiring massive coordination
  - There are 2.5 FTEs for Technical and Project Management
  - Interface with >120 staff across other JWST Subsystems and STScI Divisions & Branches
  - Spend our time in 20+ meetings per week
  - Coordinate a lot of people, topics, cross-component & cross-subsystem interfaces
- Moving from Waterfall to Agile methodology using 2-week Sprint cycles
- Moving towards Continuous Integration (CI)

Science and Operations Center (S&OC) Overview

JWST Mission Hierarchy for DMS



14 DMS Groups / Components

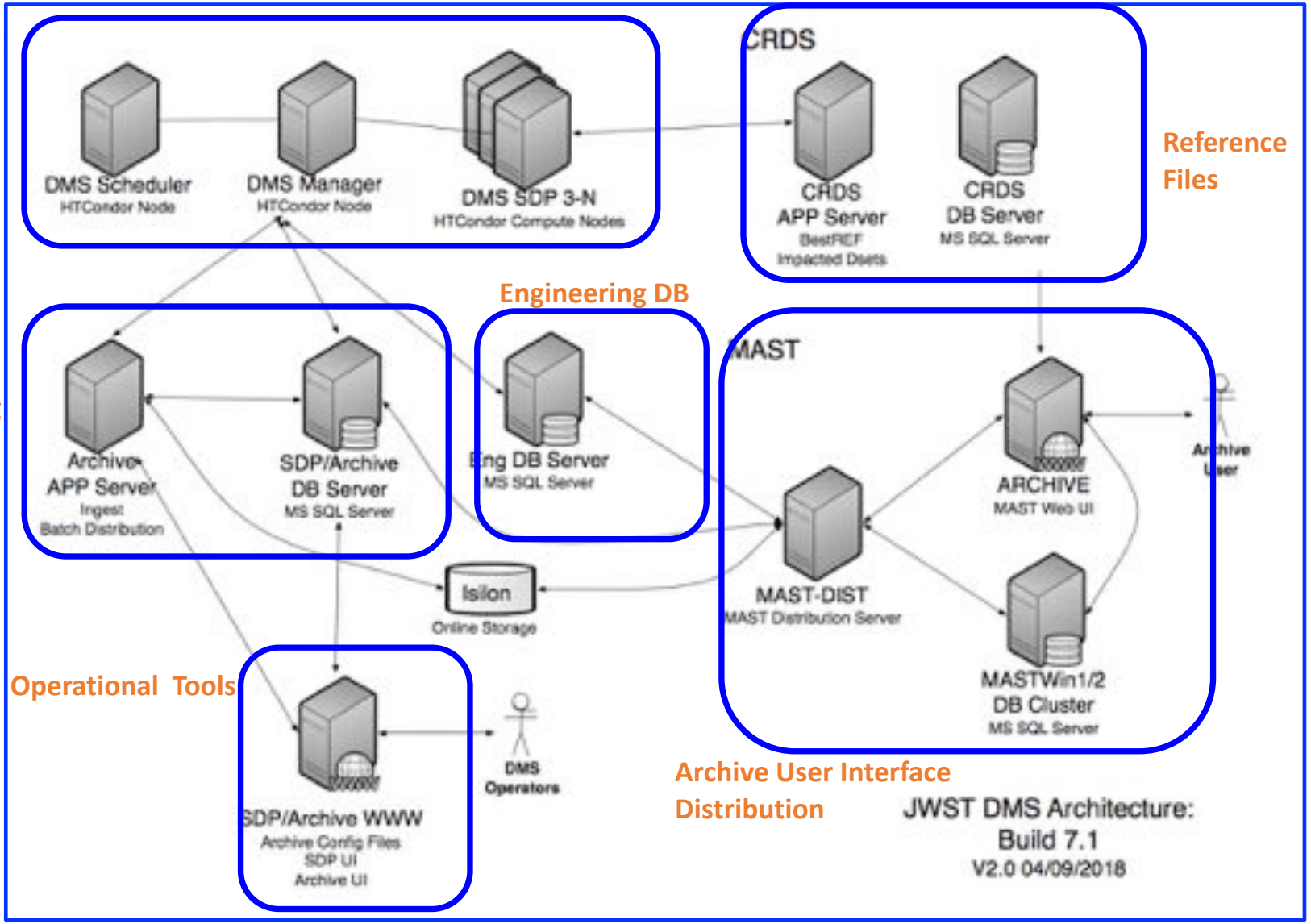
DMS Requirements & Software Build Deliveries

- DMS has 700 requirements (97% complete!)
  - DMS is using cryovacuum test data, but these data do not cover “typical” science use cases
  - DMS needs test data from the end-to-end simulator to complete the remaining requirements to the best of our ability prior to launch
- Requirements are delivered as software functionality in large “Builds”
  - There are 14 software components, or “groups”, with specific functionality
  - All software groups combine into a large integrated software build
  - Builds are installed ~2 x year (6 years)
  - Each DMS build comes on a String/Environment containing nearly 20 servers
    - Maintain 4 sets of Strings

DMS Server Diagram = "String" of 20 VMs

Science Data Processing
Work Flow Manager
Reprocessing
Calibration Pipelines

Ingest
Archive Catalog
Storage Broker
Distribution



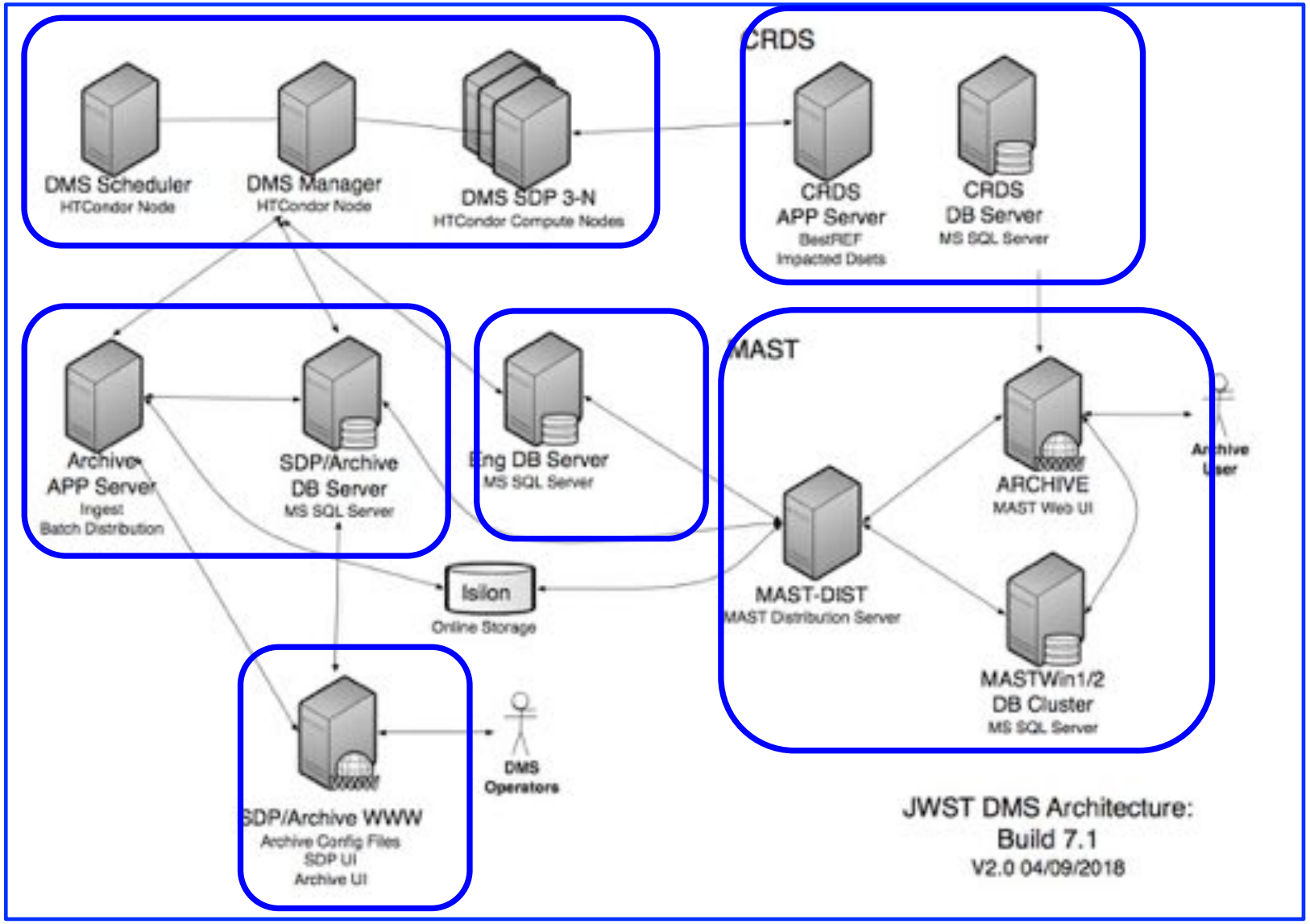
DMS Components

Software/Tech:

HTCondor
Python
Django, Apache
Astropy, numpy
(Astro)Conda
MySQL, SQLite
MS SQL Server
C#, Javascript
Hybernate
Java, Scala
Confluence

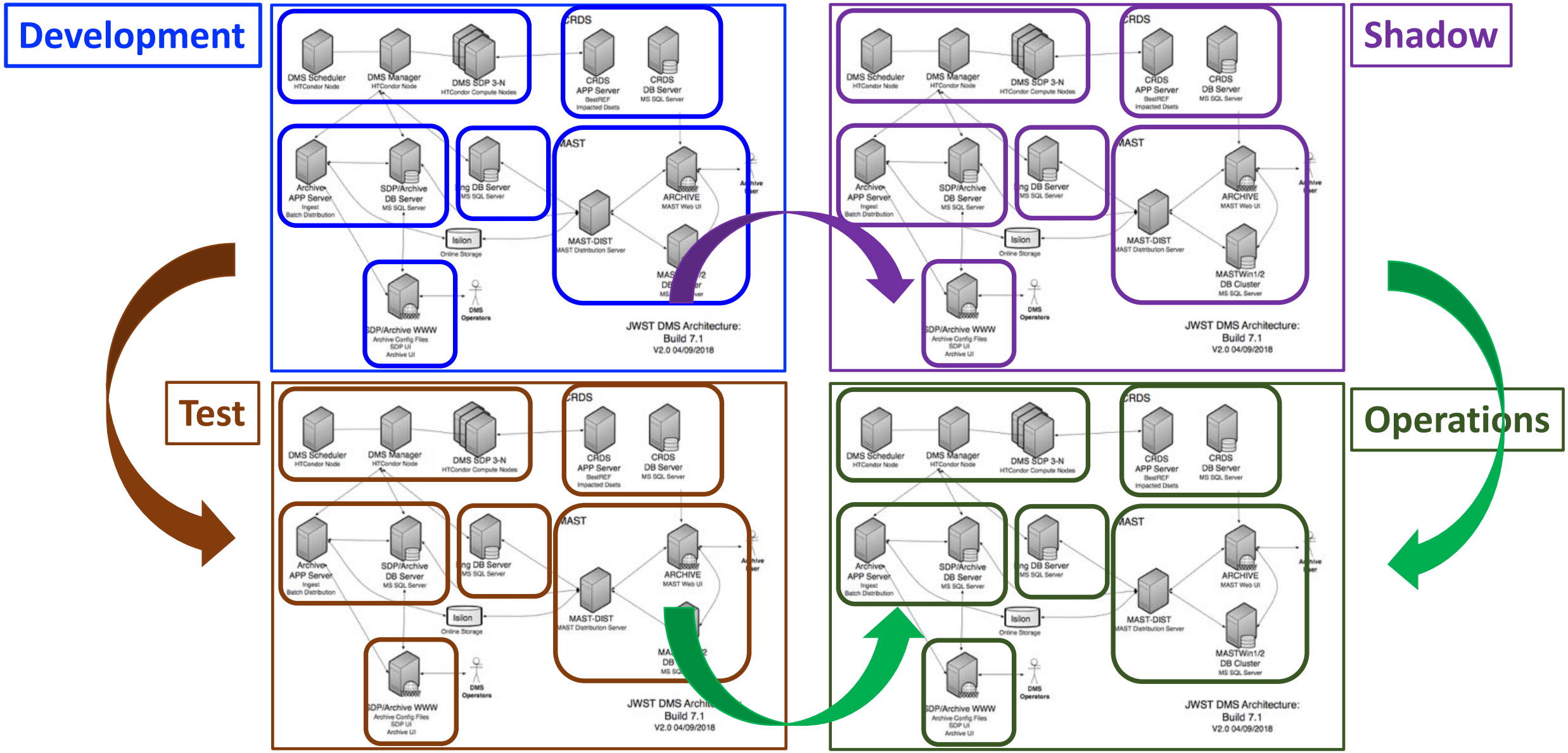
Isilon/SSDs
SSO/Shibboleth
IIS/Linux
Service-Service Arch
MAST API
VO-Compliant
Github
Jenkins, Artifactory
Selenium

DMS has 4 “Strings” to Configure & Manage



DMS has 4 “Strings” to Configure & Manage

Total Servers: 30

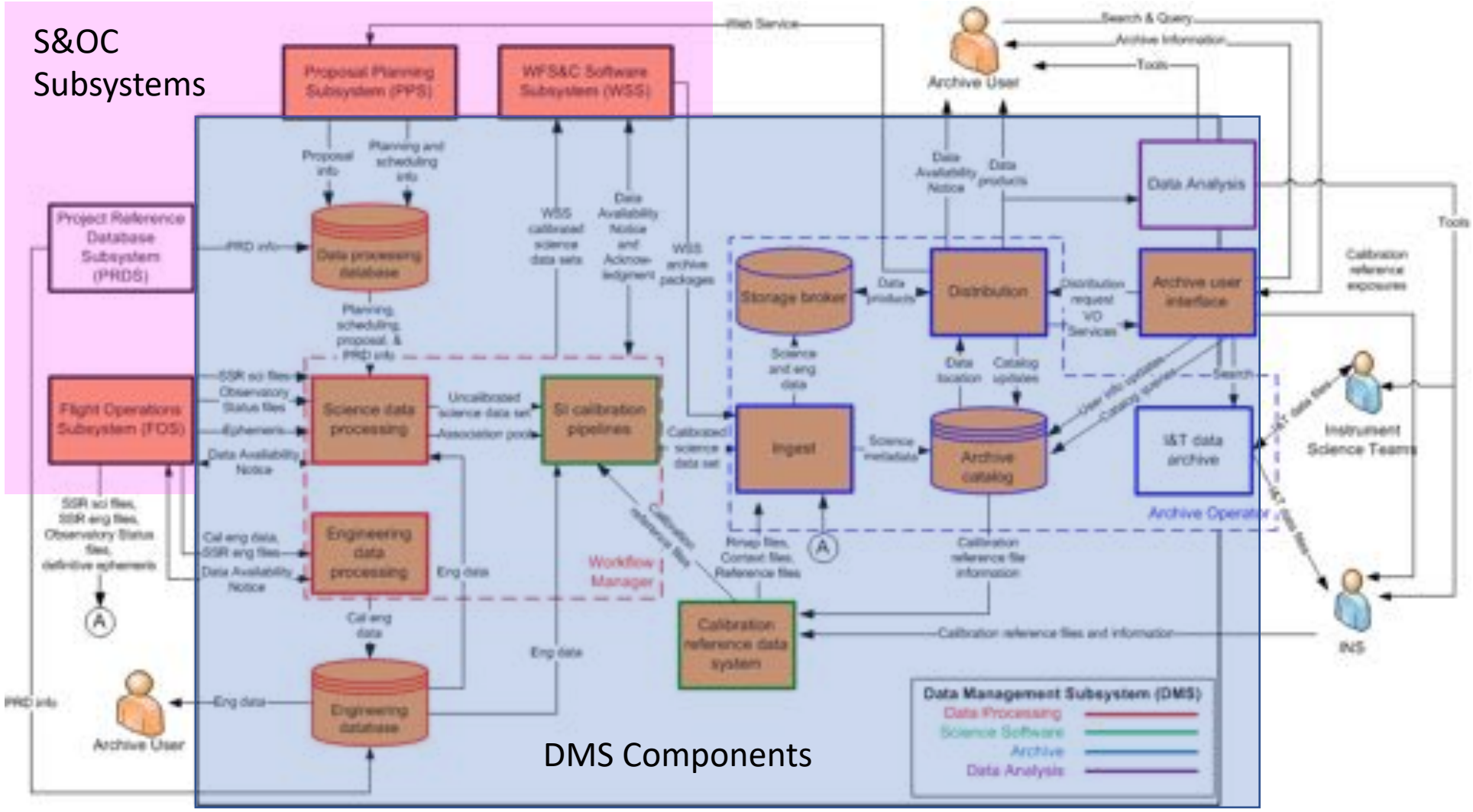


DMS Current Functionality

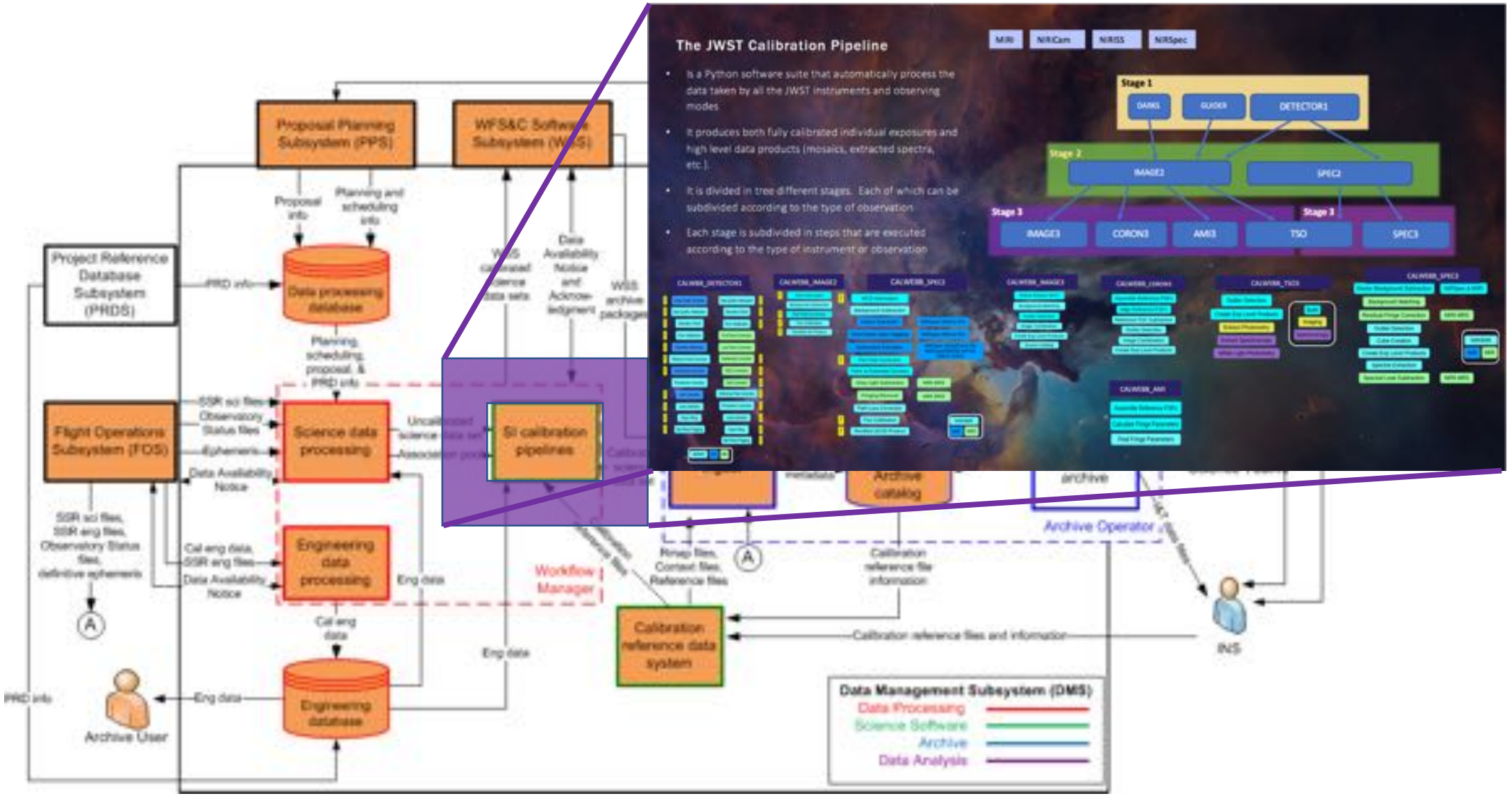
- What does the system do now?
  - Receives and archives Science and Engineering data
  - Processes Science data through calibration and in some cases combined products
  - Provides search and distribution to end users
  - Provides monitoring tools for operations
  - Provides tools for data re-calibration (if needed) and post-processing
  - Provides Data Analysis tools

DMS Data Flows

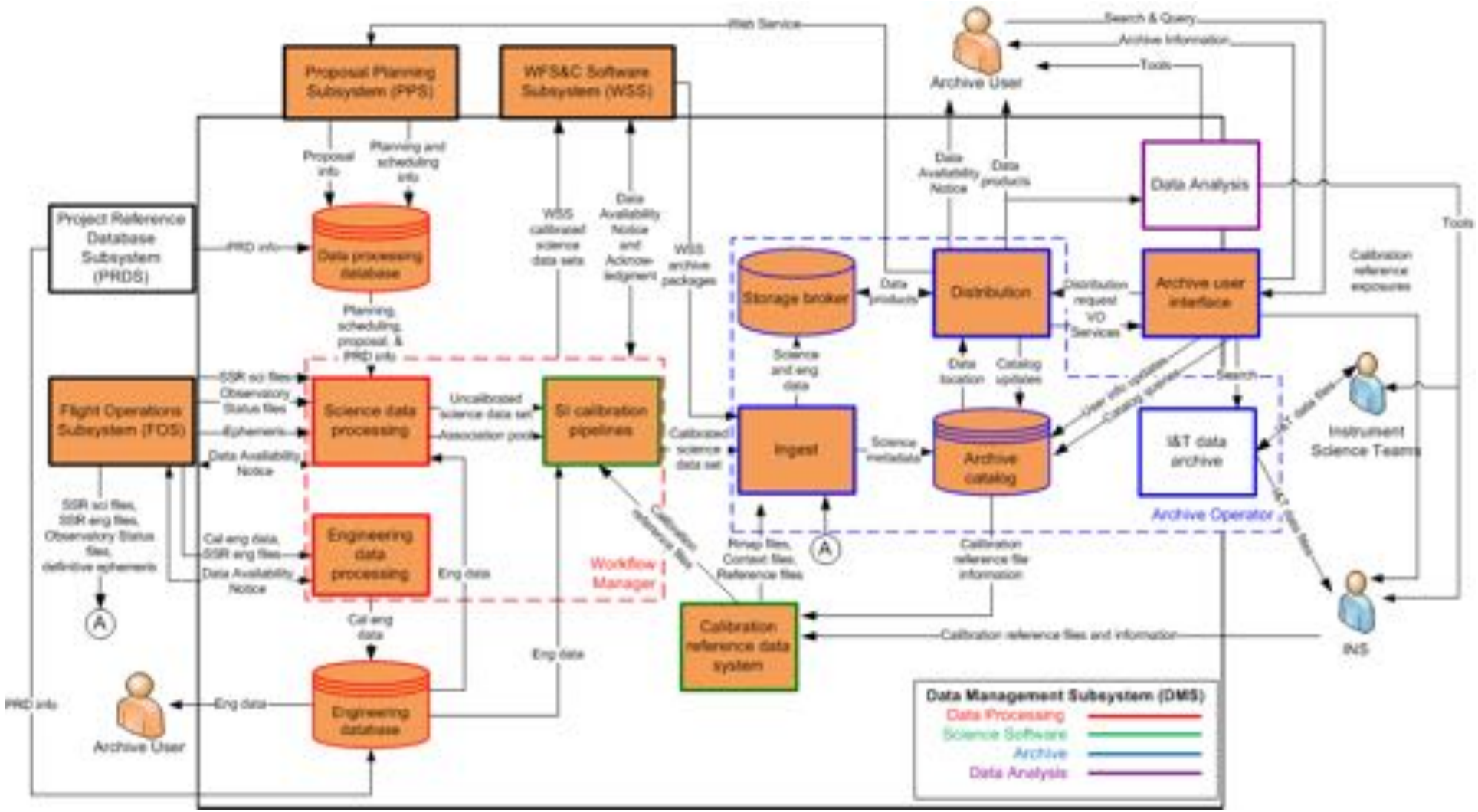
Wave Front Sensing and Control (WFS&C) Processing Flow #1: Exposure Processing



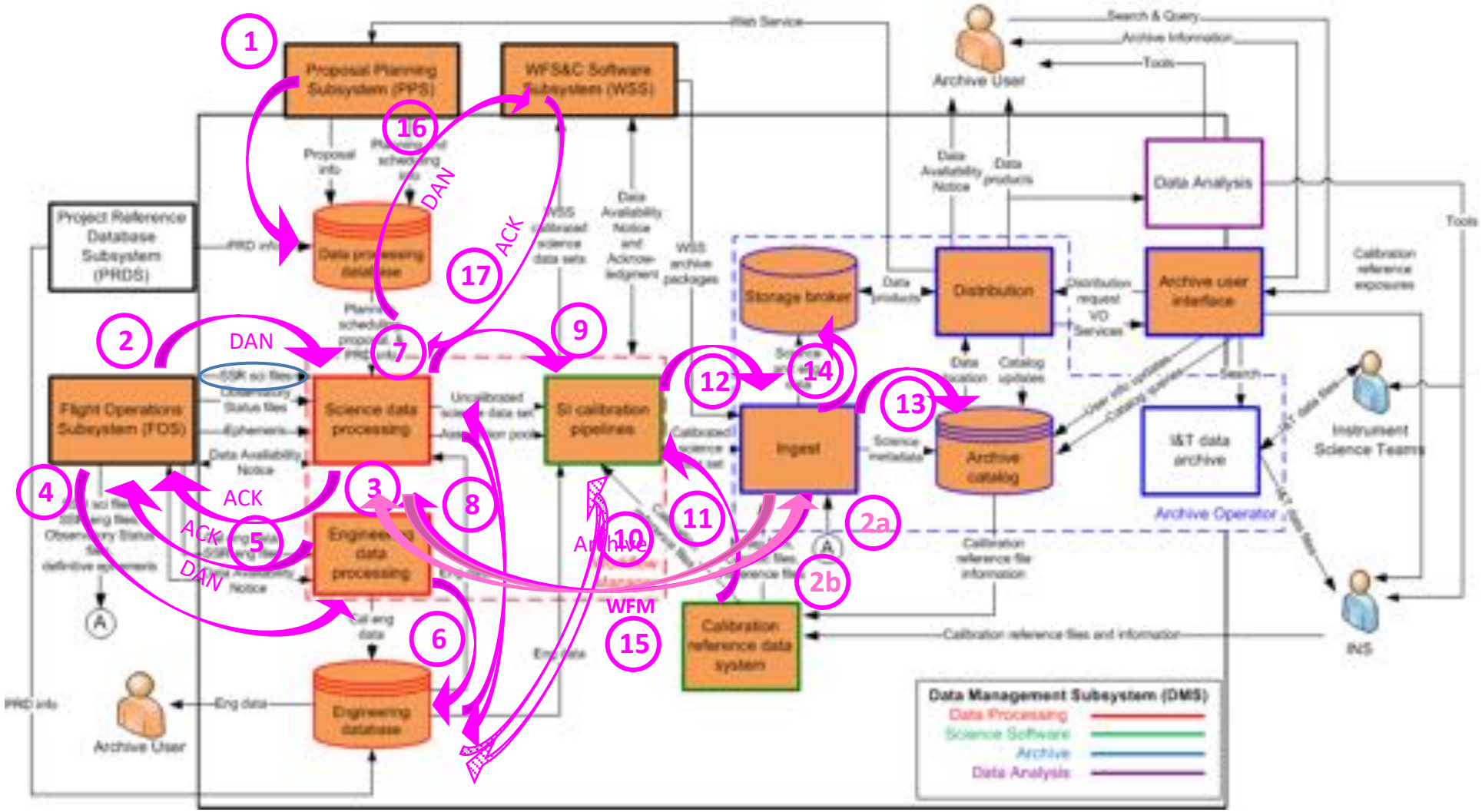
Wave Front Sensing and Control (WFS&C) Processing Flow #1: Exposure Processing



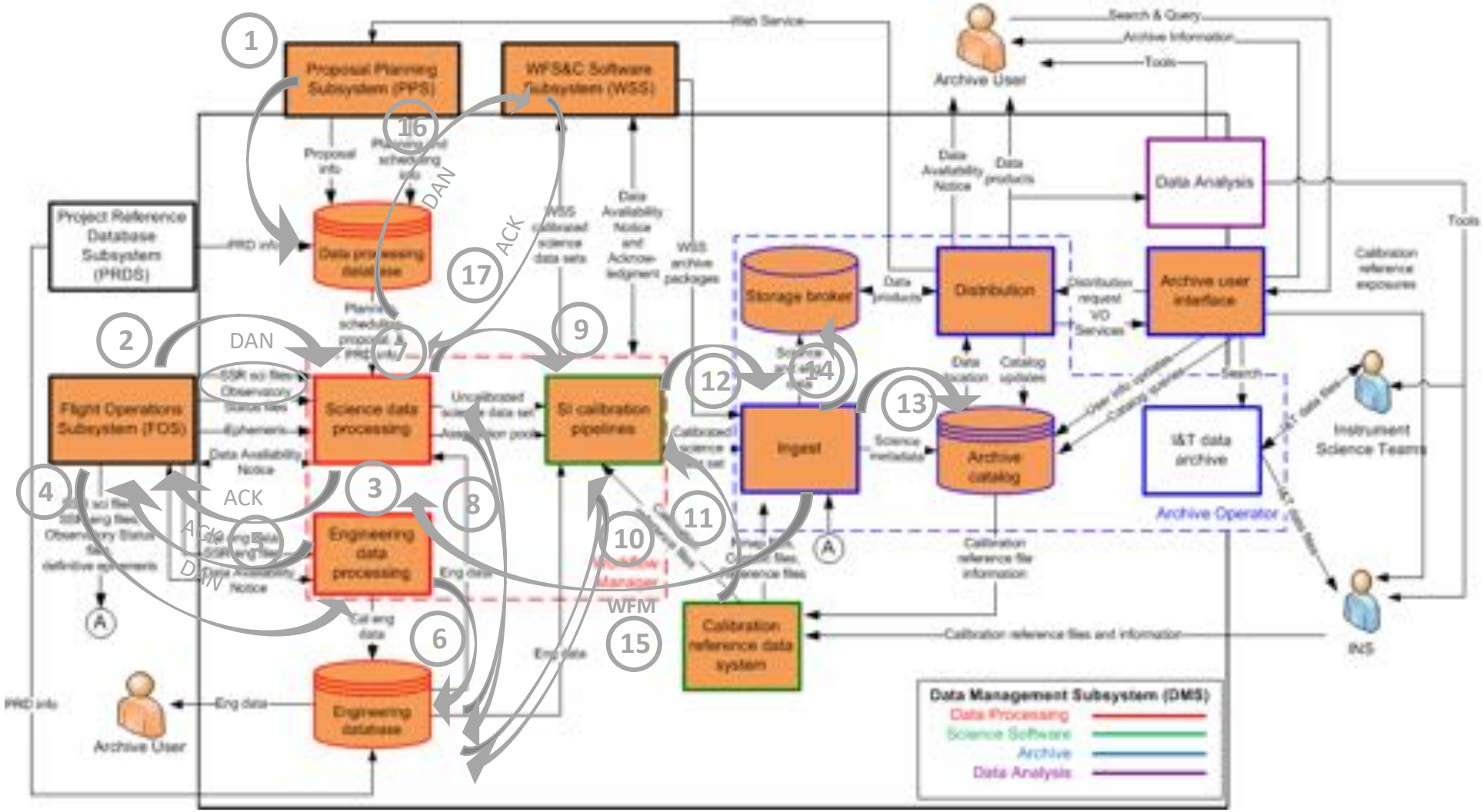
Wave Front Sensing and Control (WFS&C) Processing Flow #1: Exposure Processing



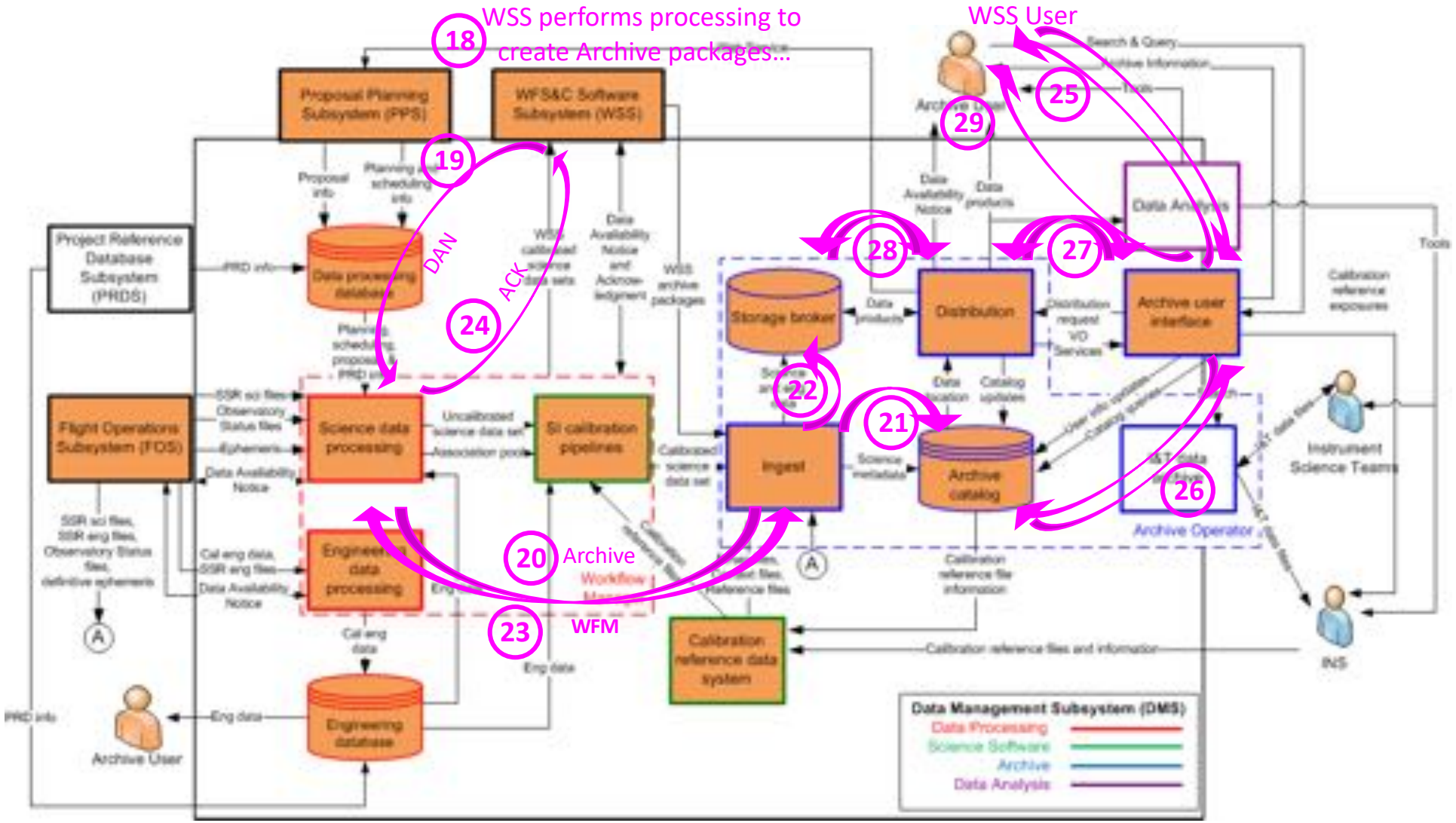
Wave Front Sensing and Control (WFS&C) Processing Flow #1: Exposure Processing



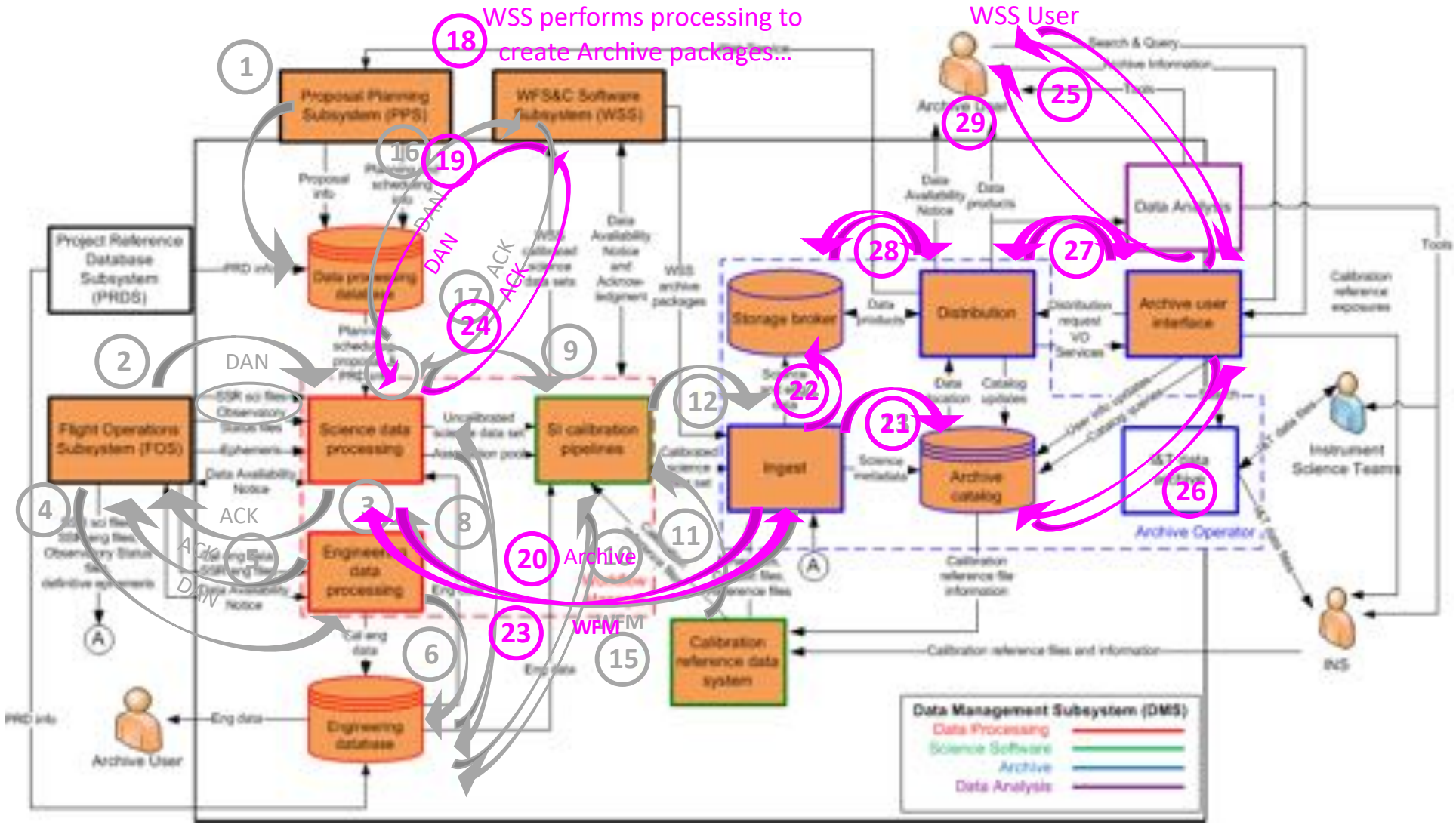
Wave Front Sensing and Control (WFS&C) Processing Flow #1: Exposure Processing



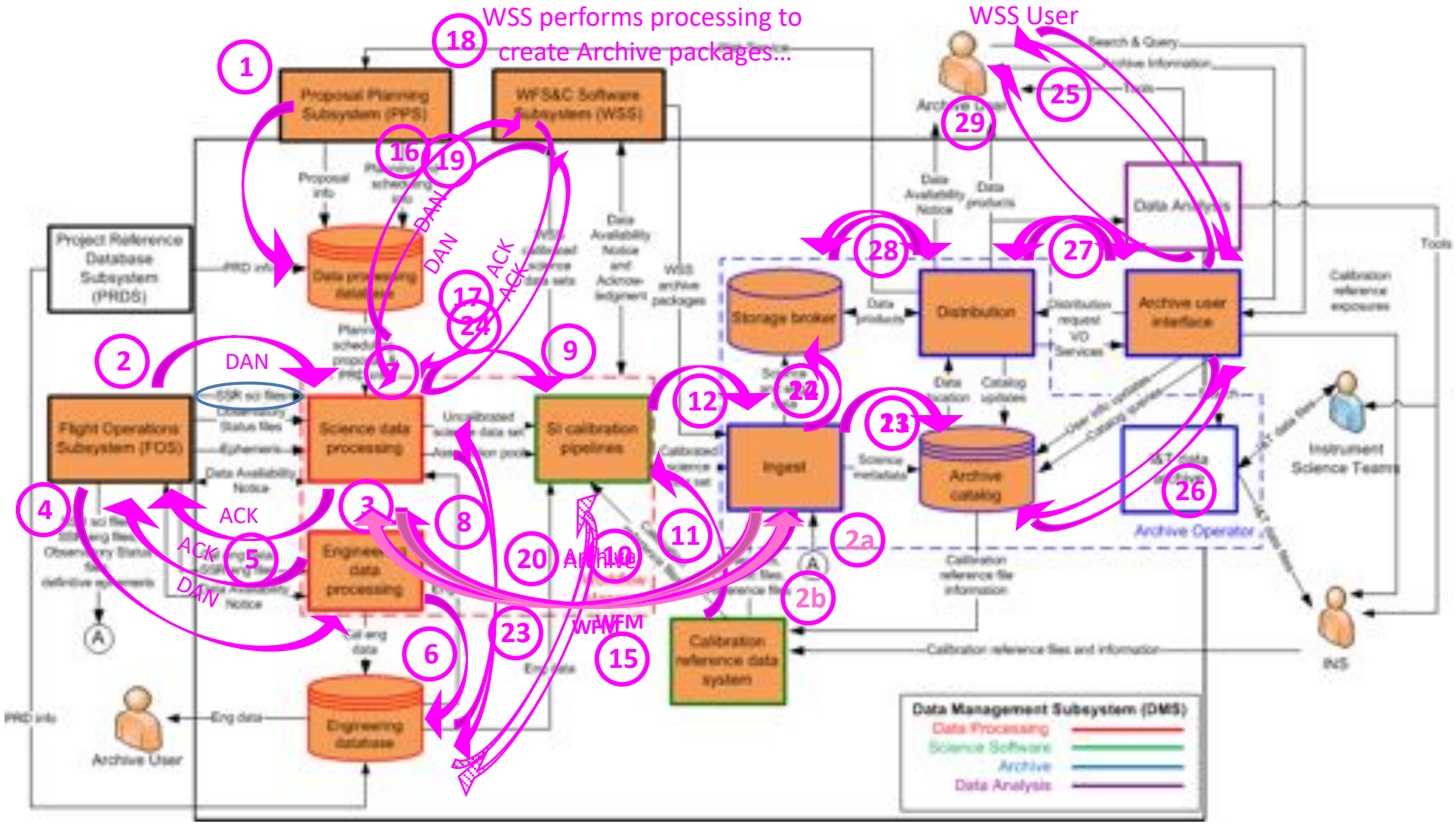
Wave Front Sensing and Control (WFS&C) Processing Flow #2: Archive Package Processing



Wave Front Sensing and Control (WFS&C) Processing End to End



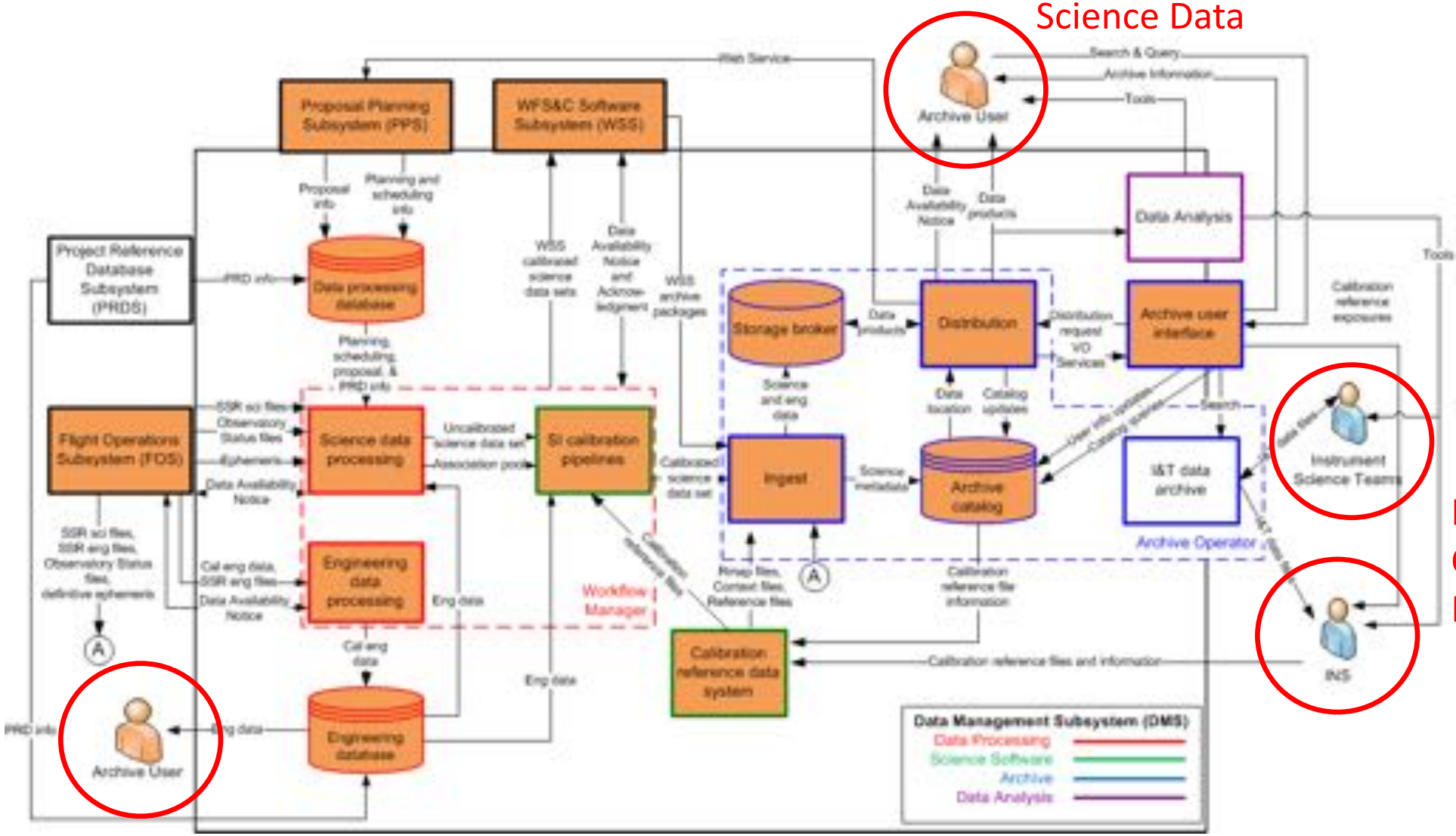
Wave Front Sensing and Control (WFS&C) Processing End to End



Archive User Interface(s)

JWST Data Access Points

Science Data



Internal Data, Calibration Reference Files

Engineering Data

Archive User Interface (AUI), aka "The MAST Portal"

- The MAST Portal (https://mast.stsci.edu) will be used to search and retrieve JWST data
  - Science Data
    - Search by program, position, or use advanced filtering to hone in on the dataset(s)
    - Several download options are available: ZIP, cURL, Batch, etc.
  - Science Instrument Tables
    - Search by filtering on instrument metadata parameter(s)
  - Engineering Data
    - Engineering data will be made public
    - Search for mnemonic and date range
- Astroquery: MAST API (https://astroquery.readthedocs.io)
- MAST is VO-compliant
- MAST demos on https://www.youtube.com/user/STScIMAST

MAST: Barbara A. Mikulski Archive for Space Telescopes

The MAST Portal lets you search multiple collections of astronomical datasets all in one place. Use this tool to find astronomical data, publications, and images.

What's New

Recent Highlights Include:

- **Access to Calibration Observations**: Calibration data is now available in the Portal. Use the "Observation Type" filter in Advanced Search to include the calibration observations in the search results.
- **HST Subscriptions and Notifications**: Logged-in users may now subscribe to HST observations to be notified of status changes, such as when an observation is reprocessed or is made public.
- **Download Basket History**: The Download Basket now keeps track of download history so data can be re-downloaded without having to search again.
- **Disk Detective HLSP Beta**: The Disk Detective HLSP catalog is available as a public beta release. Select "MAST Catalogs" from the collection dropdown and choose "Disk Detective (Beta)" from the secondary dropdown to preview the catalog.
- **Programmatic Access to the MAST Portal**: A MAST service is now available in Astroquery. General MAST API information is here.
- **Release Notes**: A history of software release notes may be found here.

Quick Start:

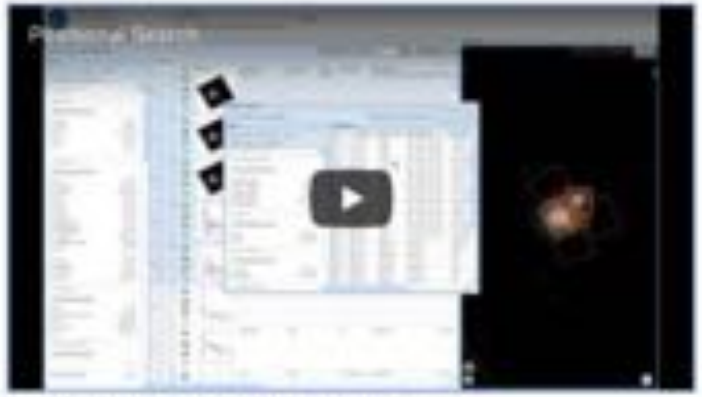
1. Select a collection and enter a new search target OR upload an existing list of targets.
2. Use the filters and analysis tools to find the exact data for which you're looking.
3. Add files to the download basket to control your download options.

See the User's Guide for more detailed documentation and video tutorials.

Currently available data collections:

- MAST Observations: Millions of observations from Hubble, Kepler, GALEX, IUE, FUSE, and more.
- Virtual Observatory: Search thousands of astronomical data archives from around the world for images, spectra, and catalogs.
- Hubble Source Catalog: A master catalog with a hundred million measurements of objects in Hubble images.
- MAST Catalogs: Access to catalog data such as Gaia and TESS Input Catalog, with more coming soon.

Featured tutorial: Conducting a positional search



See all video tutorials on the MAST YouTube channel.

MAST is managed by Space Telescope Science Institute.
For more collections, visit archive.stsci.edu.
Information about acknowledging the use of this resource may be found here.

Data Analysis Tools & Visualization

JWST Data Analysis Tools

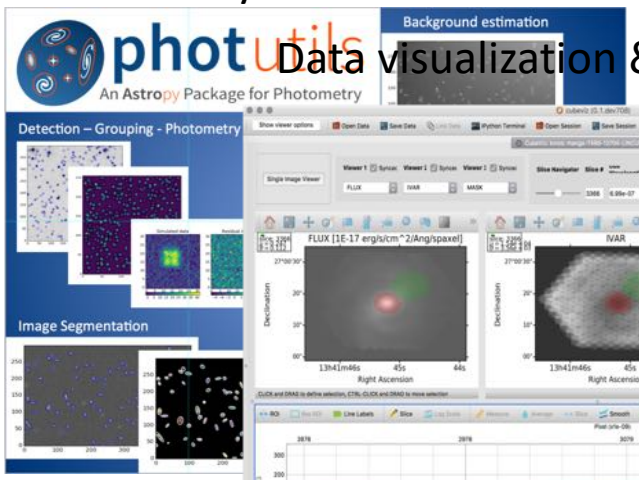
Current Situation:

- There are no flight JWST datasets
- Community needs to ramp up on Python
- Users have data from other observatories and would like to use the new tools

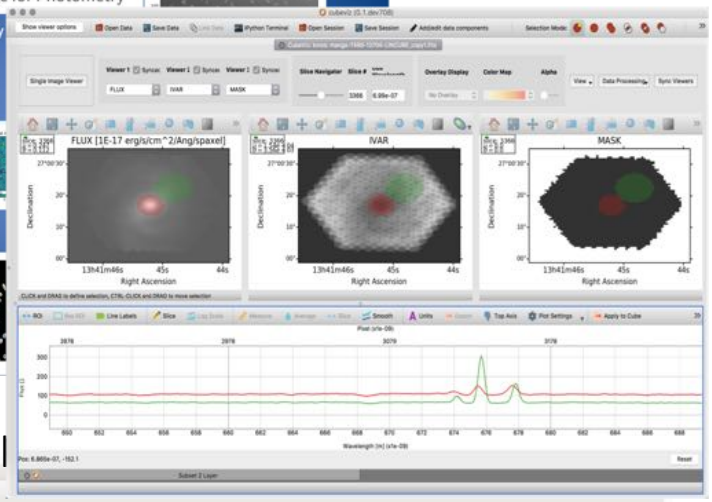
STScI DMS Data Analysis Tools:

- Library infrastructure (i.e. Astropy)
- Spectral analysis (MOSViz, SpecViz)
- Photometry (photutils)
- Visualization (Glue, stginga, CubeViz)
- Training & Documentation

Photometry tools

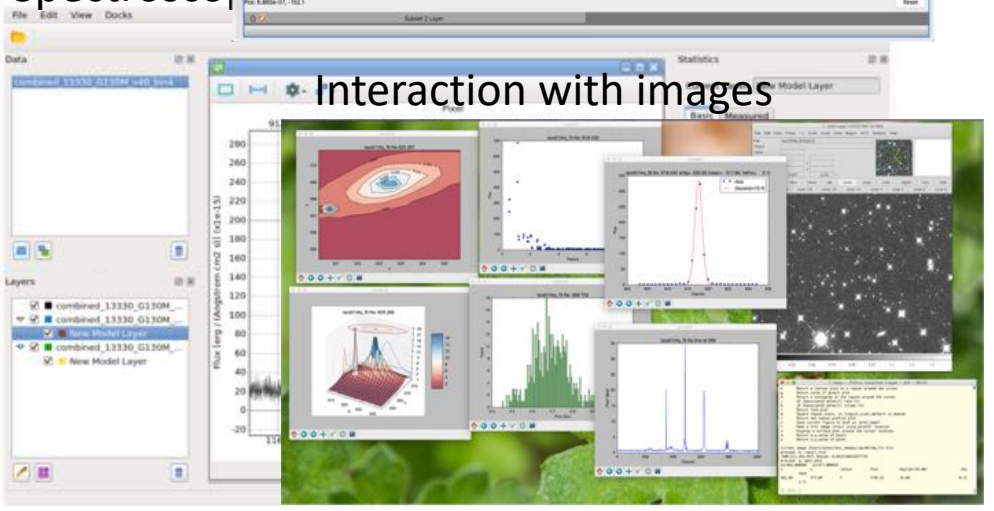


Data visualization & exploration



Spectrosco

Interaction with images



Lessons Learned

Credit: Pinterest

JWST DMS: Lessons Learned (I)

- High fidelity, end-to-end simulators to create data are vital for development and to exercise the S&OC
- Automated testing is essential to catch problems:
  - In the code (Python)
  - Using tools such as Jenkins, Artifactory, Selenium
  - Using end-to-end regression testing with simulated data
- Effort spent up front in Interface Control Docs (ICDs) is well worth it
- Integrate your components/systems as early as possible and often!
- Align your release process across components/systems so that they can be tested together and in-step

JWST DMS: Lessons Learned (II)

- Have a well-defined installation and patch process with designated sign-off
  - Detailed plan with work assigned for installations/patches
  - Have a plan to revert in case of emergency
  - Use Test or Shadow to pre-test a release or flip between Shadow and OPS for operations
- Get the baseline ("vanilla") accomplished first! Fend off scope creep - calibration is never-ending and will be improved and tweaked endlessly to get "chocolate".
- Combined higher-level products (e.g. mosaics, dithers, background subtraction) are really hard! Simply your design when possible.
- If using Agile/Scrum for large systems, establish a Scrum of Scrums or similar for interdependencies
- Large systems are hard to manage; put in management structure, checkpoints, coordination and communication
- More interfaces/subsystems multiplies the dependencies and complexities; requires more resources for testing and coordination

Thank you to Mark Kyprianou and Katie Kaleida for being my partners in crime, and a huge **Thank You** to the cast of hundreds!

Archive Sciences Branch:

Clara Brasseur

Tom Donaldson

Theresa Dower

Michael Dulude

Chinwe Edeani

Scott Fleming

Peter Forhsay

Jonathan Hargis

Tim Kimball

Karen Levay

Steve Lubow

Sunita Malla
Tony Marston

Jacob Matuskey

Brian McLean

Christian Mesh

Susan Mullally

Oliver Oberdorf

Data Analysis Tools Branch:

Larry Bradley
Daniel D'Avella
Harry Ferguson

Perry Greenfield

Joe Hunkeler
Craig Jones
Pey Lian Lim

Christopher Moriarty

James Noss

Sara Ogaz
Matt Rendina

Tom Robitaille
Megan Sosey
Erik Tollerud

Information Technology Services Division:

Marty Boegner

Mark Calvin

Dante Frias

Data Processing & Archive Services Branch:

Faith Abney

Matthew Burger
Thomas Bair
Pat Brown

Hannah Drew-Moyer

Tracy Ellis
Forrest Hamilton

Brian Hayden

Brendan Hurst

Herb Kennedy

Debbie Kenny

Michael Ridgaway

John Scott
Hien Tran

Commanding Branch:

Elizabeth Barker

Kerry Clark

Ilana Dashevsky

Mike Uzzo

Data Systems Branch:

Hunter Brown
Andrew Cortese
David Davis

Jesse Doggett

Kim DuPrie

Lisa Gardner

Michael Gough

Chris Heller-Boyer

David Kauffman

Talya Kelley

Rich Kidwell

Melissa Major

Donald McLean

Alex Padgett

Chris Rahmani

Mary Romelfanger

John Schultz

Lisa Sherbert

Richard Spencer

Science Instrument Calibration Software Branch:

Steven Crawford

Nadia Dencheva

Howard Bushouse
James Davies
Michele De La Pena

Jonathan Eisenhamer

Dave Grumm

Warren Hack

Phil Hodge

Robert Jedrzejewski

Todd Miller

Jane Morrison

Bernie Simon

Project Reference Database Branch:

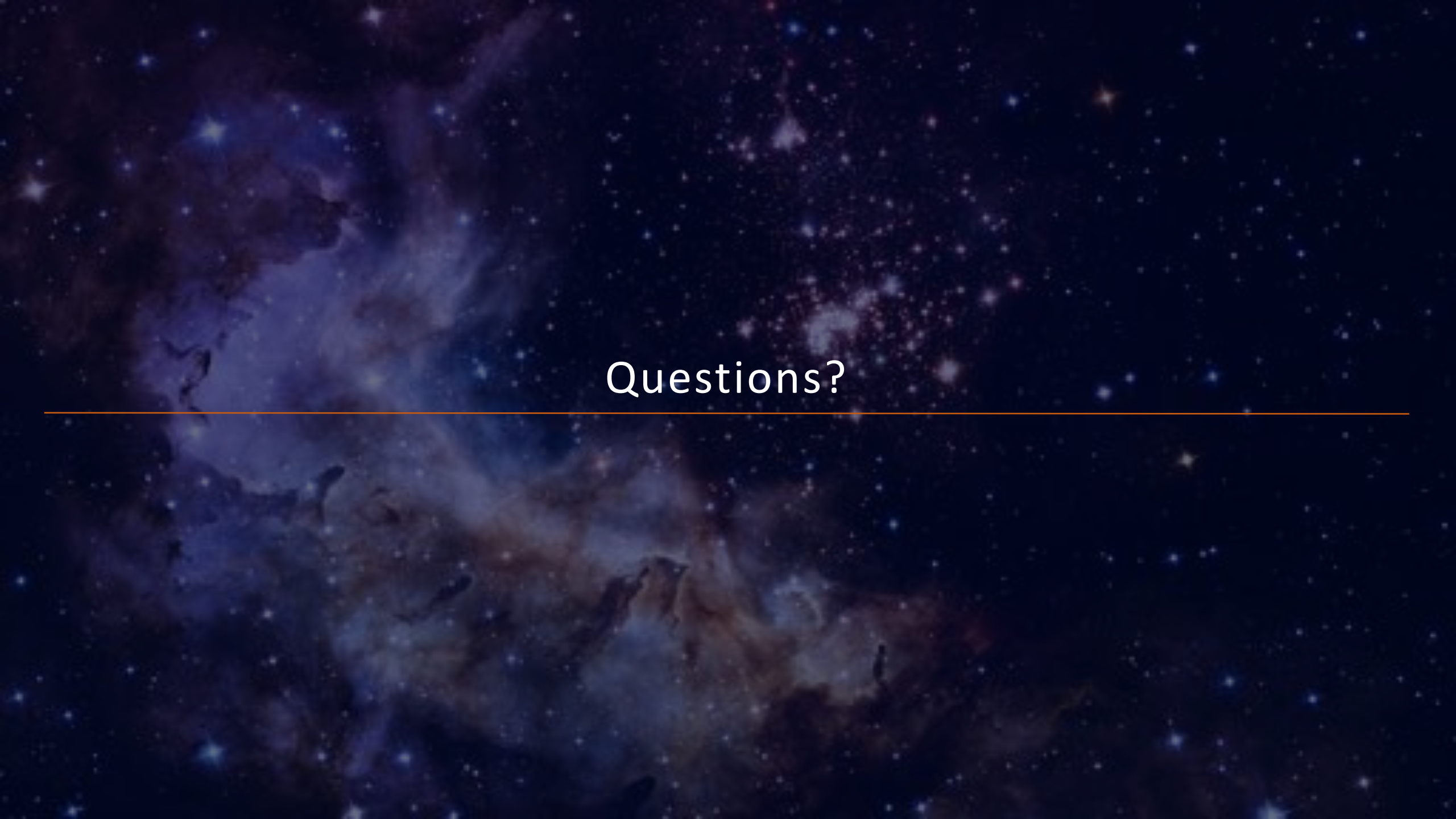
Andy Groebner

Fred Romelfanger

JWST Mission Office; Management:

Magan Beddard

STScI | SPACE TELESCOPE SCIENCE INSTITUTE


Questions?

Supporting/Backup Slides

14 DMS Component Definitions (14 'Groups')

Architectural Partition	DMS Component	Acronym
Data processing	Science data processing	**SDP**
	Engineering data processing	**EDP**
	Workflow Manager	**WFM**
	Reprocessing	**REPRO**
Science Software	Science instrument calibration pipelines	**CAL**
	Calibration Reference Data System	**CRDS**
	Data analysis tools	**DA**
Archive	Ingest	**ING**
	Archive Catalog	**AC**
	Storage Broker	**SB**
	Distribution	**DIST**
	Archive User Interface	**AUI**
	SI I&T Data Archive	**SID**
Operational Tools	Operational tools and metrics	**OPS**

- Receipt of science and engineering telemetry data
- Reformat, check quality, execute calibration pipelines, perform processing needed to prepare data for the archive
- Execute reprocessing

- Perform science calibration
- Ingest and serve reference files
- Provide user analysis/viz tools

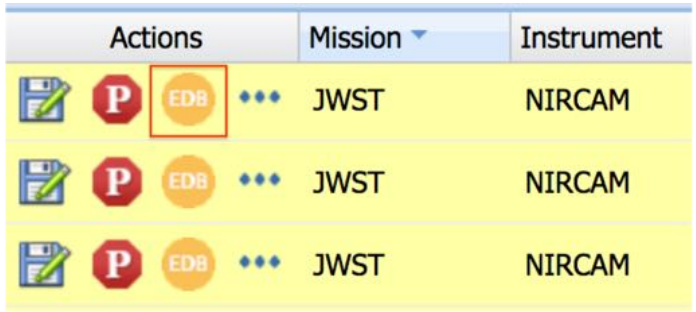
- Archive the data
- Provide user search & access tools
- Distribute data to the end-users
- Provide long term/offsite data storage
- Support contributed data (High Level Science Products)
- Support pre-flight test data

- Monitor/operate the system

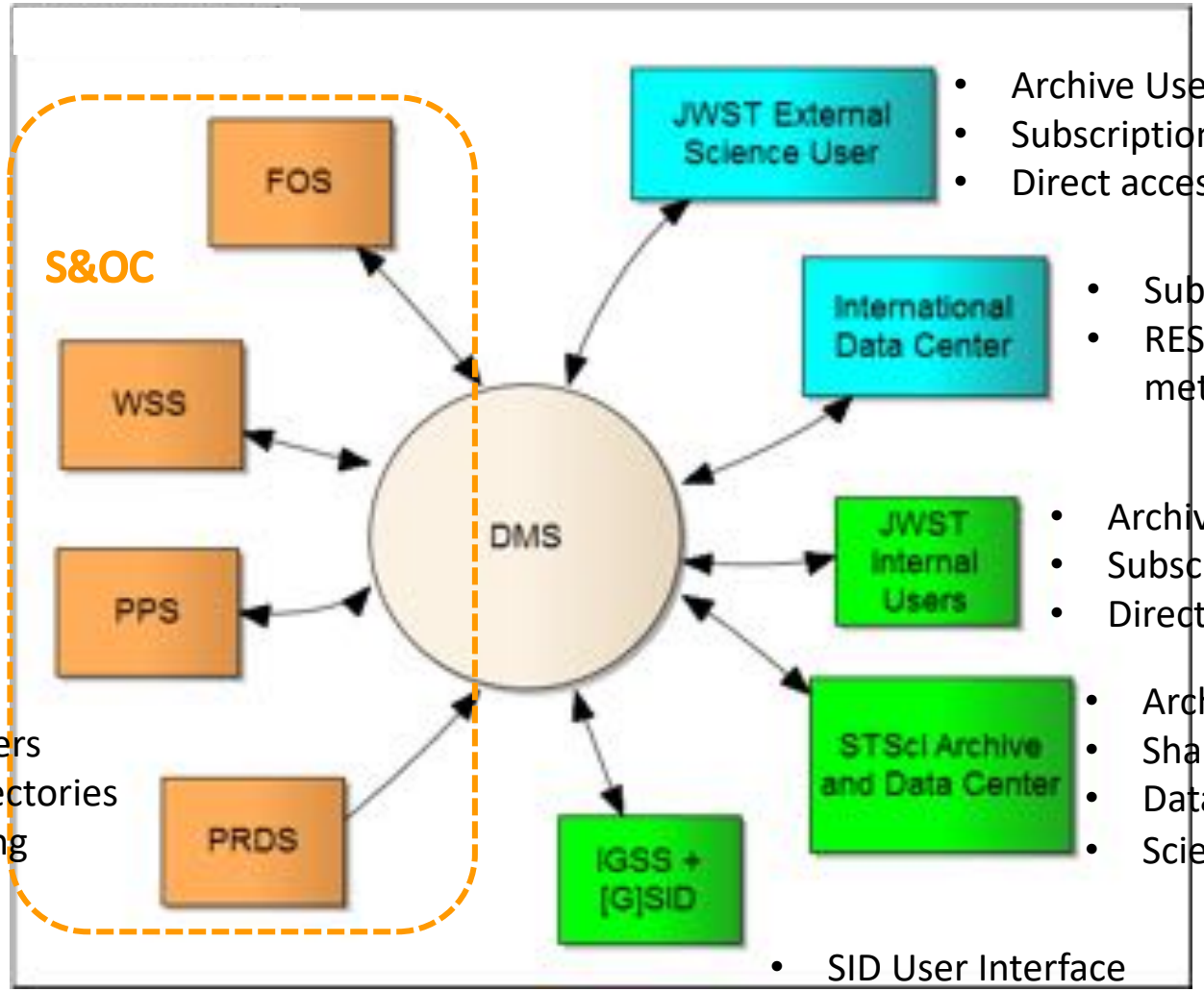
Archive User Interface (AUI), aka "The MAST Portal"

The MAST Portal will be used to search and retrieve JWST data

- Science Data
  - Search by program, position, or use advanced filtering to hone in on the dataset(s)
  - Multi-Mission searching available with JWST
  - Several download options are available: ZIP, cURL, Batch, etc.
  - Users must be authorized to download "exclusive access" data
- Engineering Data
  - Must log in during commissioning until embargo is lifted
  - Engineering data will be made public
  - Search for mnemonic and date range
  - Search using UI or via Portal EDB button
- Science Instrument Tables
  - Search by filtering on instrument metadata parameter(s)



DMS Primary Interface Context



- Archive User Interface
- Subscription Service
- Direct access scripts

- Subscription Service
- RESTFul interface to pull metadata & science data

- Archive User Interface
- Subscription Service
- Direct access scripts

- Archive User Interface
- Shared Disk
- Database
- Science Cluster

- DAN transfers
- Shared Directories
- Data Loading

- SID User Interface
- GSID retrieval thru OPS requests (ITAR)

Data Level Definition/Processing