STScI | SPACE TELESCOPE SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

The Evolution of Science Data Pipelines: HST, JWST, WFIRST

Megan Sosey

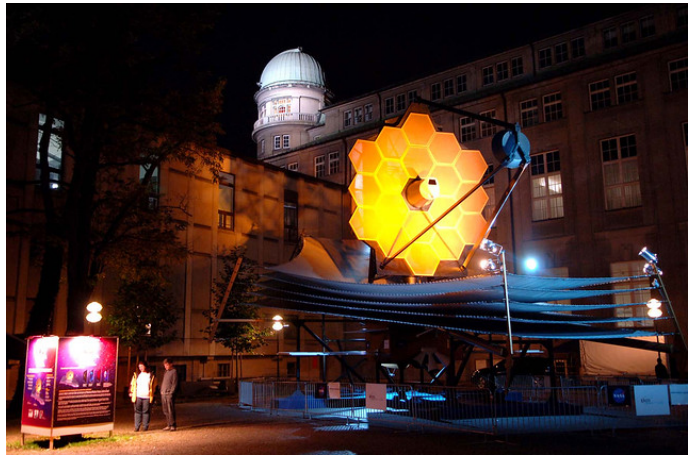
WFIRST Technical Lead, Data Management Division

ADASS 2019 - Session 3a – October 7, 2019

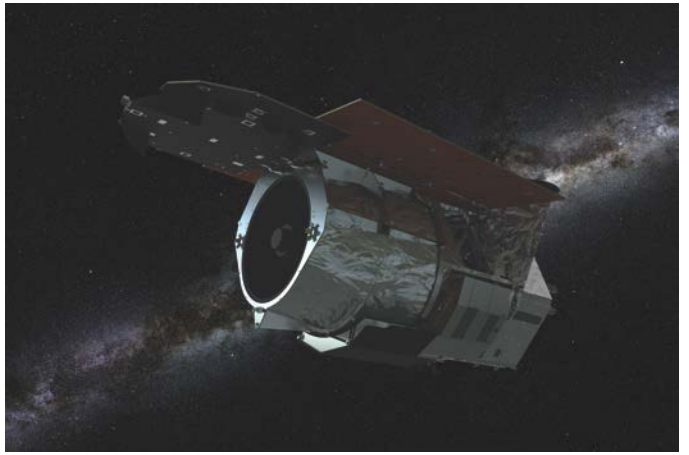
HST (1990), JWST (2021) , and WFIRST (2025)

Large community missions often require complex pipelines to support science calibration and analysis needs for the mission and the community. Their design and architecture evolve with instrumentation, mission requirements, and lifetime.

Initial release of HST



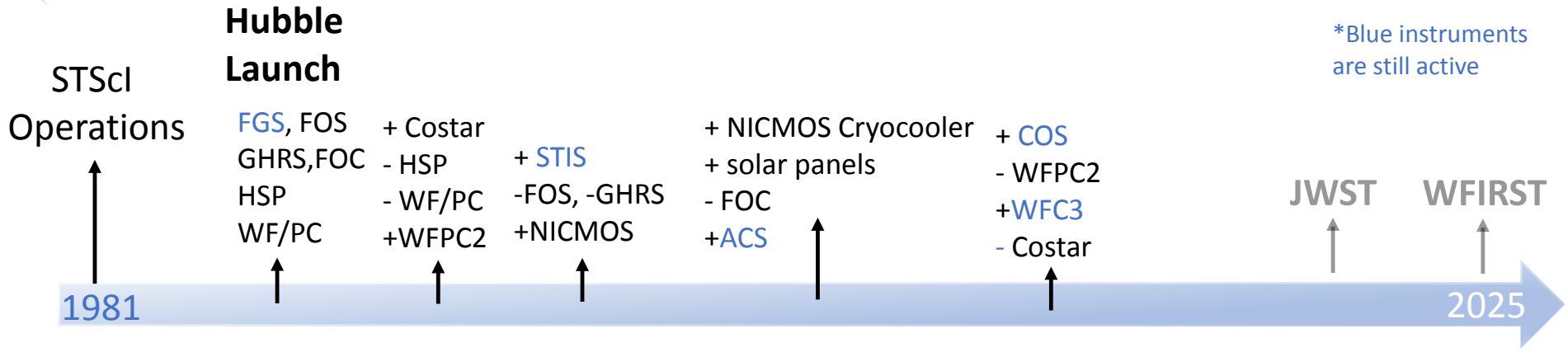
Full-size JWST in Germany



Artist rendition of WFIRST

Requirements that drive operational design, software development, and collaboration, are evolving to enable complex, maintainable, extensible, and yet still automated systems.

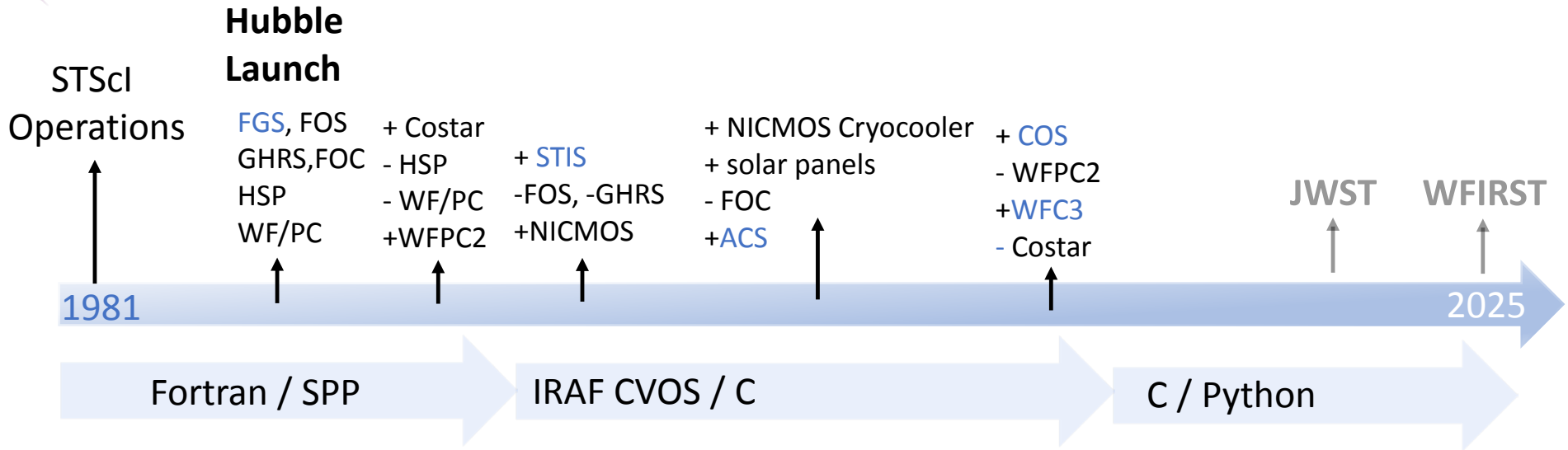
The Hubble Space Telescope – Launched April 24, 1990



High-level workflow manages *monolithic* science calibration pipelines

- Sci-Cal developed on a per-instrument basis, each evolves uniquely with time
  - However, automated pipelines were a new and novel concept at the beginning
- A range of detector types and wavelength regimes required different reduction methods
- Continued development of new instruments, software, and retirement of old systems
- General algorithm complexity and stringent calibration needs increasing with time
- Science teams develop algorithms, STScI software teams implement and maintain them
- Extreme changes in technology over the course of the mission
- Varied software/hardware environment, delivery, and testing over time

The Hubble Space Telescope – Launched April 1990



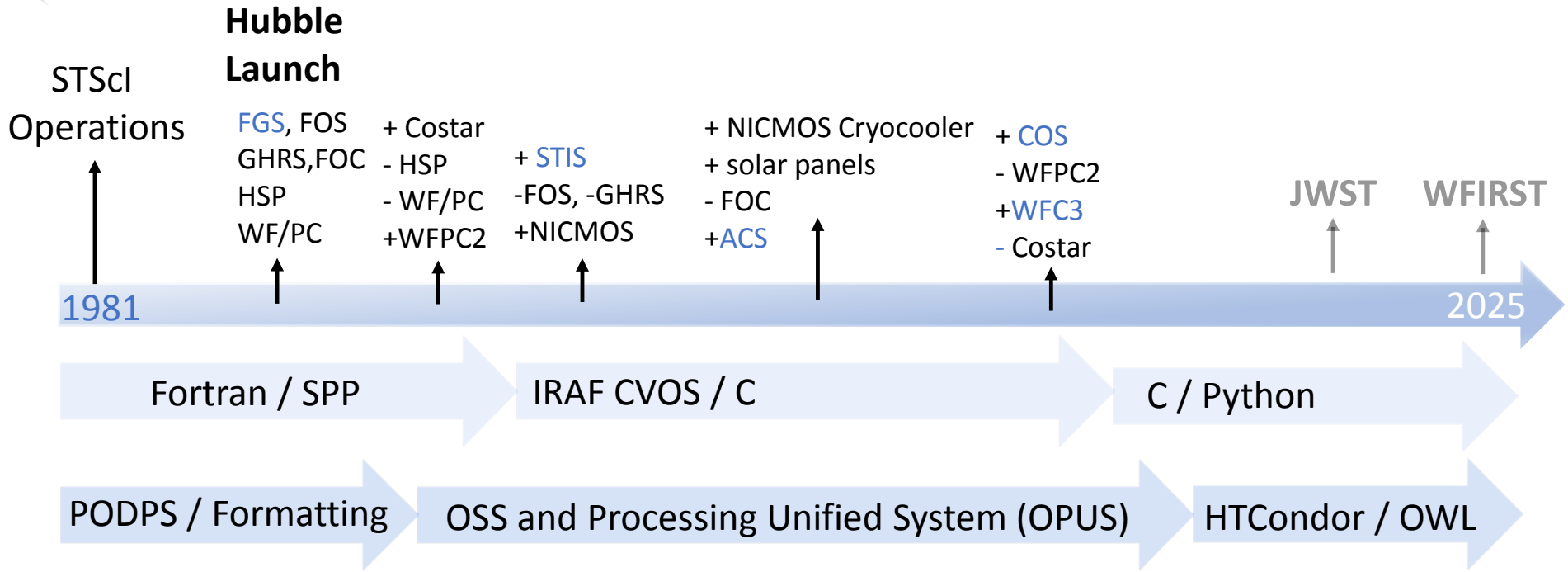
Platforms that the software is running on have changed too:

VAX VMS → Alpha VMS → Alpha Tru64 → Redhat Linux

And now we support science users running on select OSX and Linux platforms
vms → solaris → linux + osx (and to a lesser extent windows)

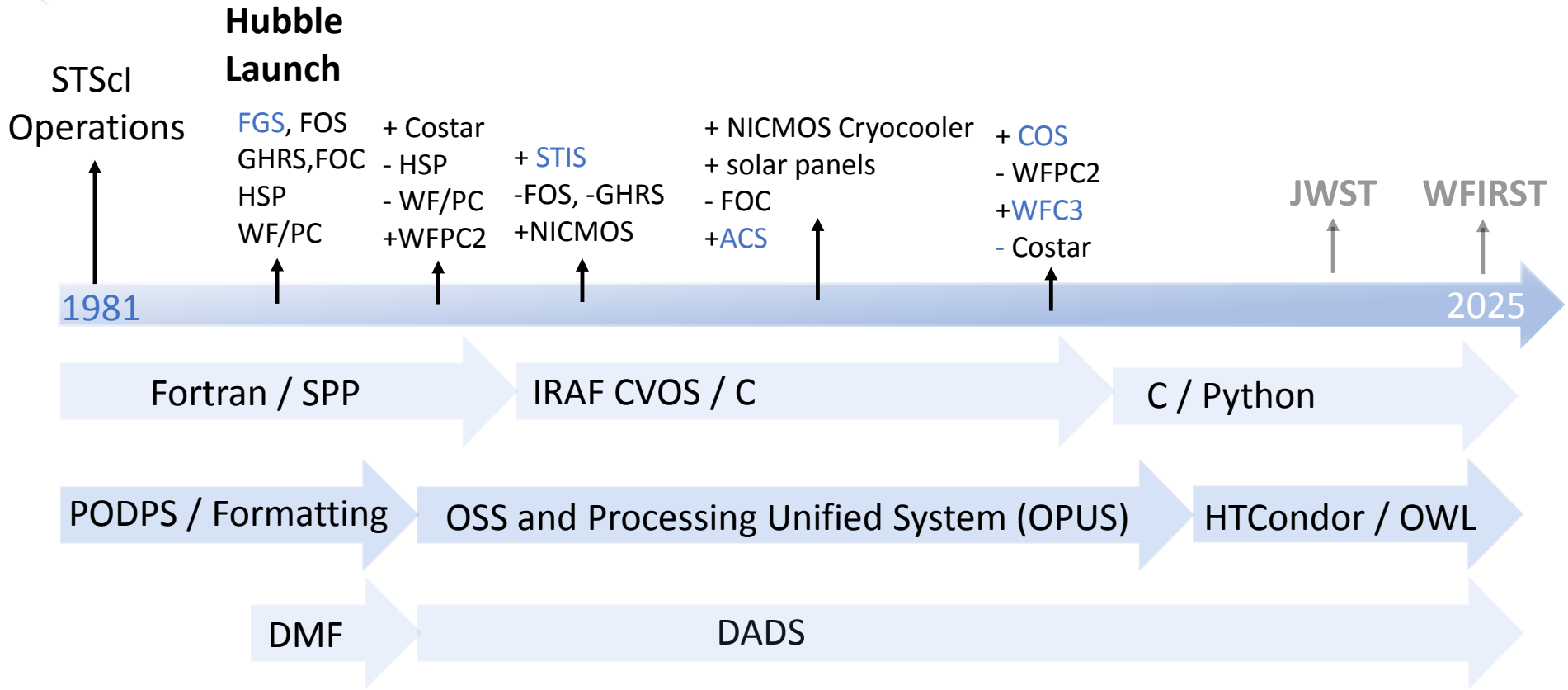
Software Language Evolution

The Hubble Space Telescope – Launched April 1990



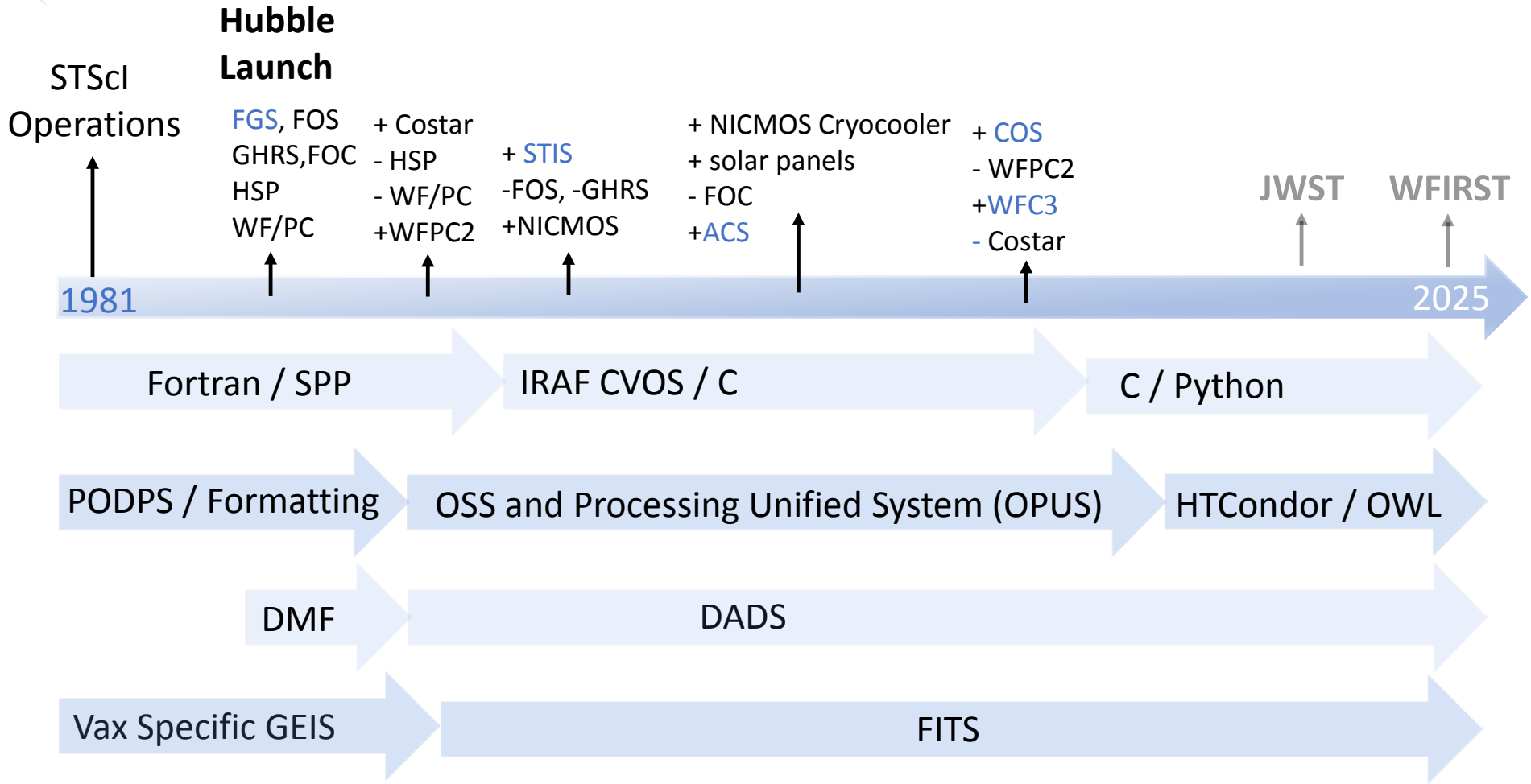
High-Level Workflow Management Evolution

The Hubble Space Telescope – Launched April 1990



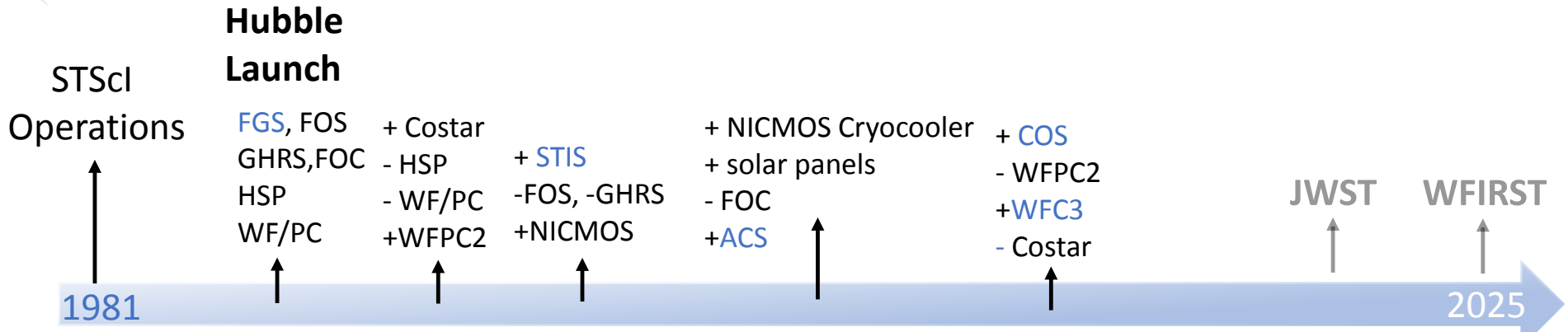
Data Archiving and Distribution Evolution

The Hubble Space Telescope – Launched April 1990



Data Format Evolution

The Hubble Space Telescope – Launched April 1990



Fortran / SPP → IRAF CVOS / C → C / Python

PODPS / Formatting → OSS and Processing Unified System (OPUS) → HTCondor / OWL

DMF → DADS

Vax Specific GEIS → FITS

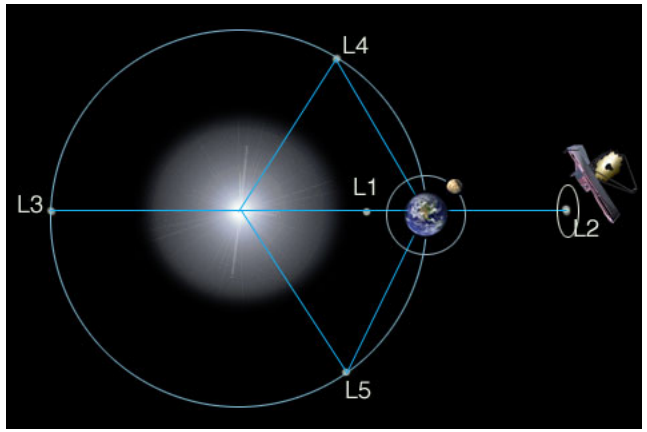
No version control system → RCS -> CVS -> SVN → Git / GitHub

Lessons Learned from HST

- FITS is inadequate
  - Unsigned 16-bit data has been used in FITS but its implementation is a kludge
  - FITS has no clear interpretation of how to deal with shared metadata
  - FITS headers often end up as one giant namespace with cryptic keyword names (due to the 8 character limitation)
  - Explicit grouping of data involves keywords, documentation, and specialized software
- Workflows should be parameter driven, not metadata driven with switches in the data itself
- Dealing with the WCS for un-resampled data is a major problem (accuracy requirements are very stringent for ACS and WFC3, and issues with FITS makes this extremely hard to manage and maintain)
- Monolithic and evolving pipelines made it hard to share common code as well as prevent inconsistent behavior between pipelines.
  - This also made customization difficult
  - Underlying algorithmic code becomes inaccessible to astronomers
  - The software could be difficult to run remotely at observer home sites
- The system to determine the right reference files, update them for the data, and to obtain them locally, was very awkward and needed to be improved

JWST Mission Summary

- International Collaboration
  - NASA
  - European Space Agency (ESA)
  - Canadian Space Agency (CSA)
- Launching in 2021 on an Ariane 5 rocket from Kourou, French Guiana
- 5 year primary mission with a possible 5 year extension
- Primary mirror 6.5 meters (21 ft 4 in) with 18 segments
- Orbit at L2, not expected to be serviced
- Complement of Near-Infrared and Mid-Infrared Instruments
  - **Near-Infrared Camera,** coronagraph, grisms, imaging
  - **Near-Infrared Spectrograph**, w IFU, MSA, fixed slits
  - **Mid-Infrared Instrument**, w 4 IFUs, coronography, imaging, prism
  - **Fine Guidance Sensor/ Near InfraRed Imager and Slitless Spectrograph,** grism, imaging
- Science: First-light & Reionization, early galaxies, planetary systems



James Webb Space Telescope (JWST) – Scheduled Launch 2021

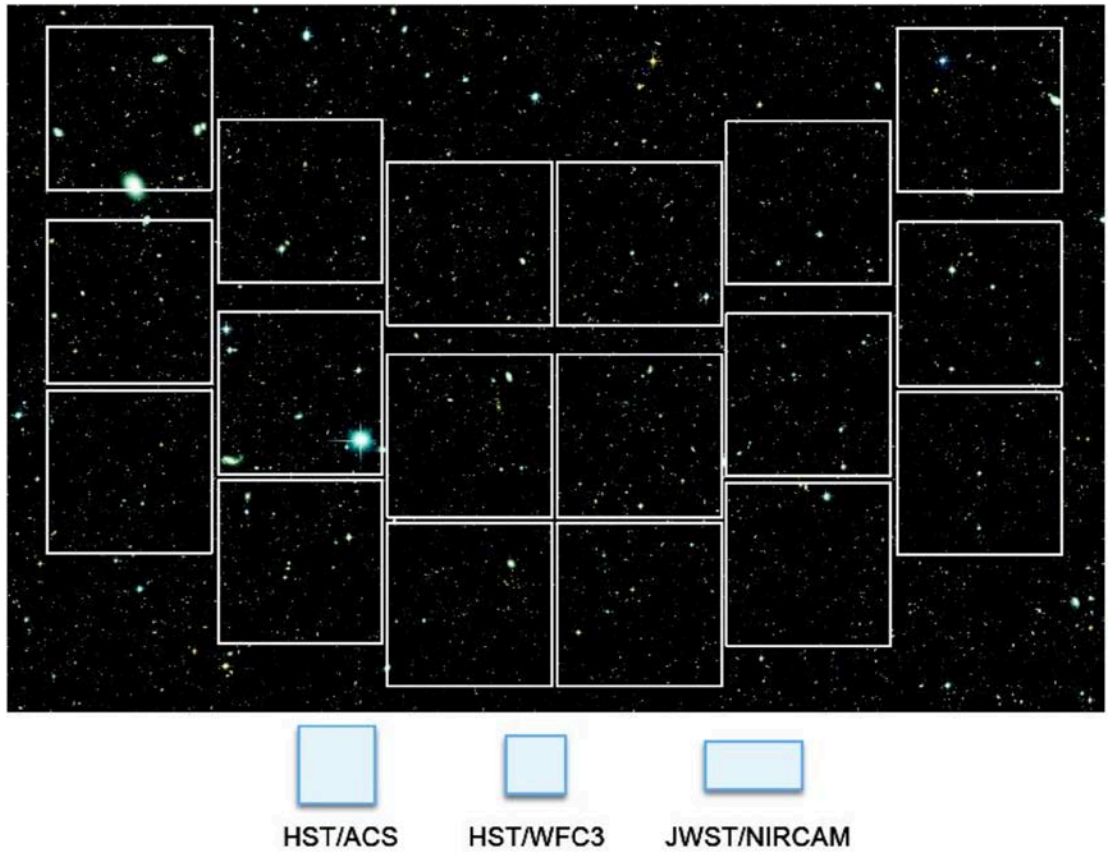
High-level workflow manages a *unified and modular* science calibration pipeline

- Sci-Cal developed as a library that is shared among all instruments, written in Python and C
  - Allows end-user interaction with the data as it proceeds through processing
  - Gives end-users the flexibility to reconfigure a pipeline to fit their needs
  - Shared development of common core routines and utilities, enables easy reuse of software
  - Open development and community software support, including Astropy core development
  - Calibration Reference Data System (CRDS) allows assigning best reference files at processing time
  - STPIPE controls data flow, I/O, and allows for data models to abstract away from the storage format
  - Complexity of the instruments and algorithms imposes new requirements on the data formats in use
    - Science calibration and CRDS support both FITS and ASDF data formats
    - Generalized WCS used for all raw (un-resampled) imaging products and spectroscopy modes

- Automated Testing and Continuous Integration
  - Unit and regression testing of individual software modules
  - Testing of the end-to-end data management system with ground-test and simulated datasets
  - Enabled using tools such as Jenkins, Artifactory, Selenium, Travis-CI, Github

Wide Field Infrared Survey Telescope (WFIRST) - Mission Summary

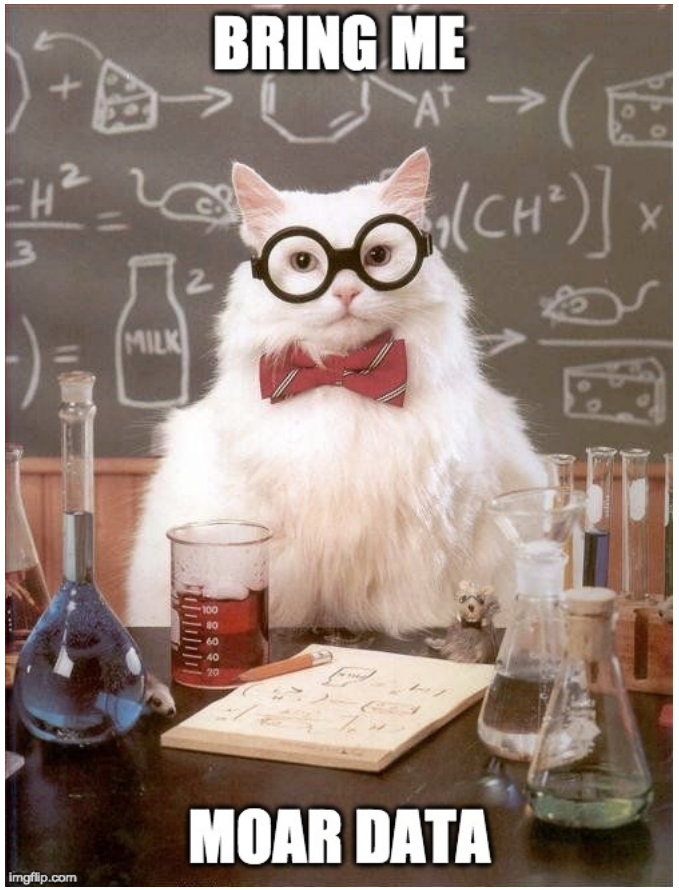
- Launching in 2025
- 5 year primary mission, with possible 5 year extension
- Primary mirror 2.4 meters
- 0.281 deg^2 field of view (100 times larger than HST)
- Orbit at Sun-Earth L2, same as JWST
- Science
  - Dark Energy
  - Exoplanets
  - infrared astrophysics
- Pre-selected mission survey observation program
- GO/GI competed science program
- Complement of Instruments
  - Wide Field Imager (WFI) – 18 4k x 4k H4RG NIR detectors, imaging and spectroscopy
  - Coronographic Imager (CGI), technology demo
- Evolved Expendable Launch Vehicle (EELV) out of Cape Canaveral.



Archives

HST

- Generates 30 GB per day; 11 TB per year
- 166 Terabytes of HST Science Data exists
- Processed data are also synchronized at ESAC and CADC



JWST

- STScI provides operations and archive services
- Expected to generate 114 TB of science data per year
- 650 TB during the primary mission

WFIRST

- STScI provides science operations and archive services
- Generates approximately 9 TB / day; Possible 24 Petabyte total archive holdings after 5 years
- Bring the astronomers to the data and provide a data-local analysis environment

WFIRST – Scheduled Launch 2025

Shared mission responsibilities and processing **between STScI, IPAC, GSFC, and Science Teams**

- Complex ground system
  - Mission Operations located at Goddard Space Flight Center (GSFC)
  - Science Operations Center and Mission Archive located at STScI (SOC)
  - Science Support Center located at IPAC (SSC)
  - Science Teams collaborating on science requirements and software creation for data processing (ST)
- Science Data Calibration
  - Pixel level WFI processing is similar to JWST pipeline and infrastructure, performed by STScI
  - WFI Combination and catalog processing shared between STScI and IPAC (also does CGI)
  - Publicly available source code and automation
- Shared testing and continuous integration infrastructure between mission partners
- Some portion of the data processing and data storage will utilize a cloud environment
- To support the large data volume, and to enable shared software and resources, a data analysis environment will be provided for all investigators that lives next to the data

Thankyou!

Be sure to check out the following posters!

P9.16 “JWST Science Data End-to-End Validation Framework”

P10.44 “CALACS: The Seventeen Year Evolution of a Space Telescope Data Pipeline”

P10.48 “HSTCosmicrays: A python package for analyzing cosmic rays in HST calibration data”

P10.5 "The James Webb Space Telescope Quicklook Application (JWQL)”

P11.8 “JWST NIRISS Data Simulations and How to Use Them”

P11.17 “COSMOdern: An HST COS Monitoring System for the Contemporary”