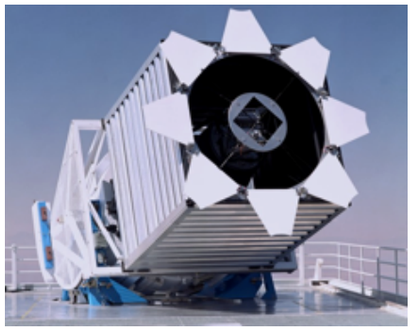
Data Management and Processing for SDSS-III and SDSS-IV

Adam S. Bolton
University of Utah

ING -- 2013-11-05

SDSS-IV: Massive Astronomical Survey Spectroscopy 2014-2020

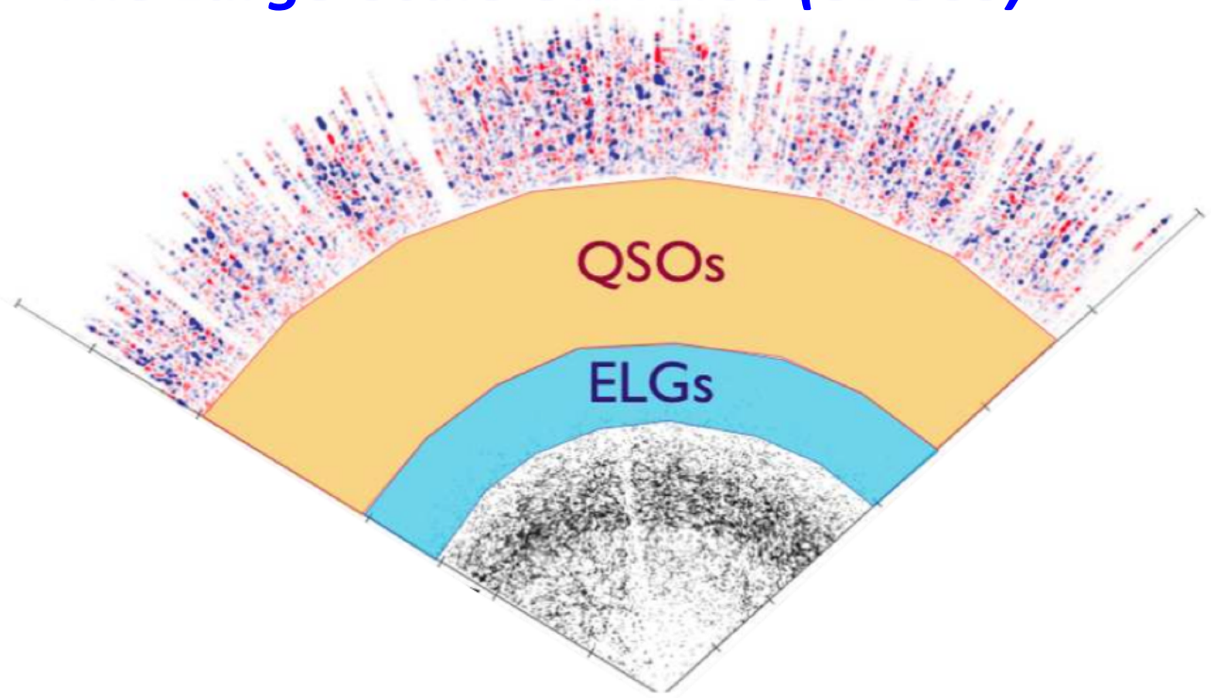
Collaboration forming now!



2.5-meter Sloan Digital Sky Survey Telescope (Apache Point, NM)

2.5-meter duPont Telescope (Las Campanas, Chile)

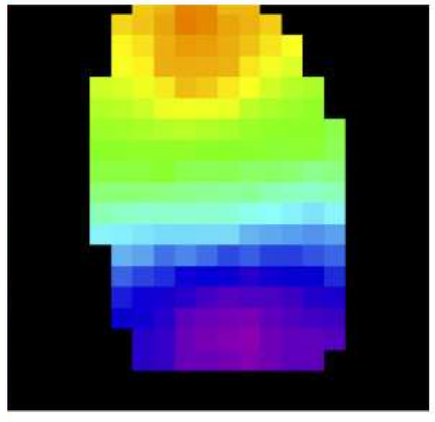
The Large-Scale Universe (eBOSS)

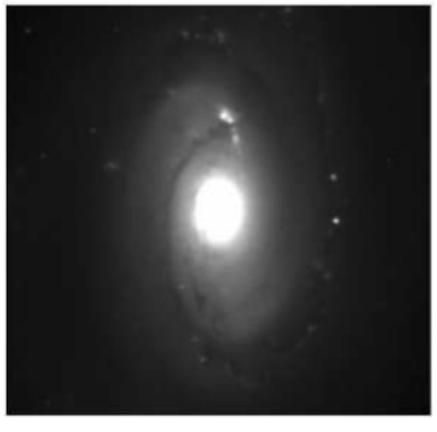


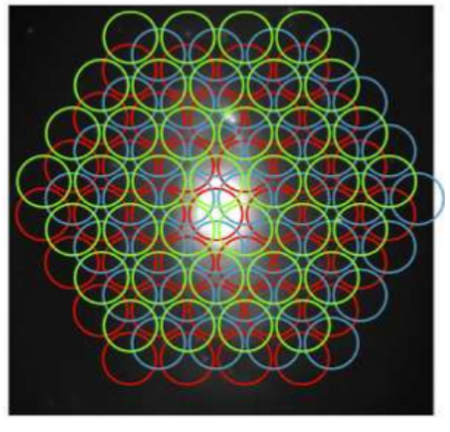
Mapping galaxies & quasars and measuring dark energy

The Local Universe (MaNGA)

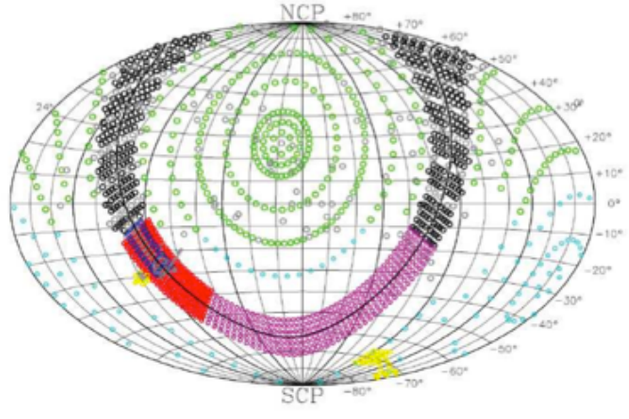
Discovering the architecture of nearby galaxies

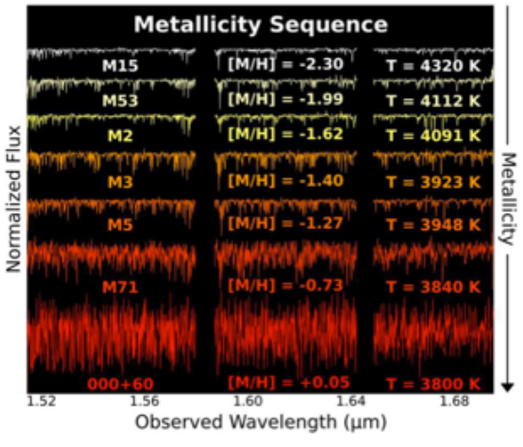






Our Galaxy (APOGEE)





Dynamical structure and chemical archaeology of the Milky Way

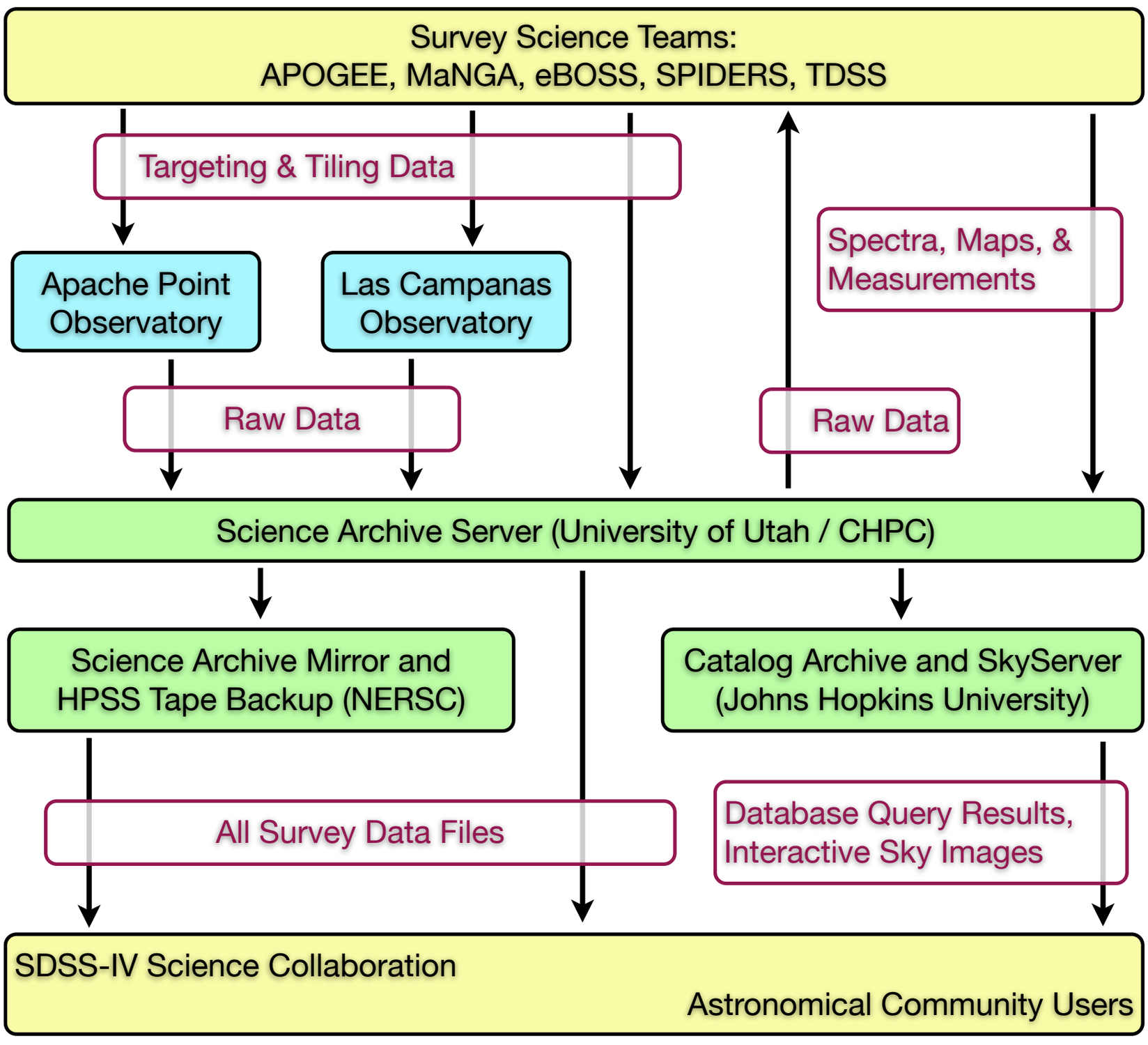
SDSS-IV Data Management Mission

- Ensure correct distribution, documentation, and archiving of all raw and reduced survey data products
- Support development, operations, & version control of software for core survey data reduction & analysis
- Determine and coordinate major project IT needs
- Facilitate electronic collaboration and communication within the project
- Integrate data flows with core survey operations
- Reconcile data-related needs of multiple project stakeholders
- Create & execute plans for public data release and long-term archiving

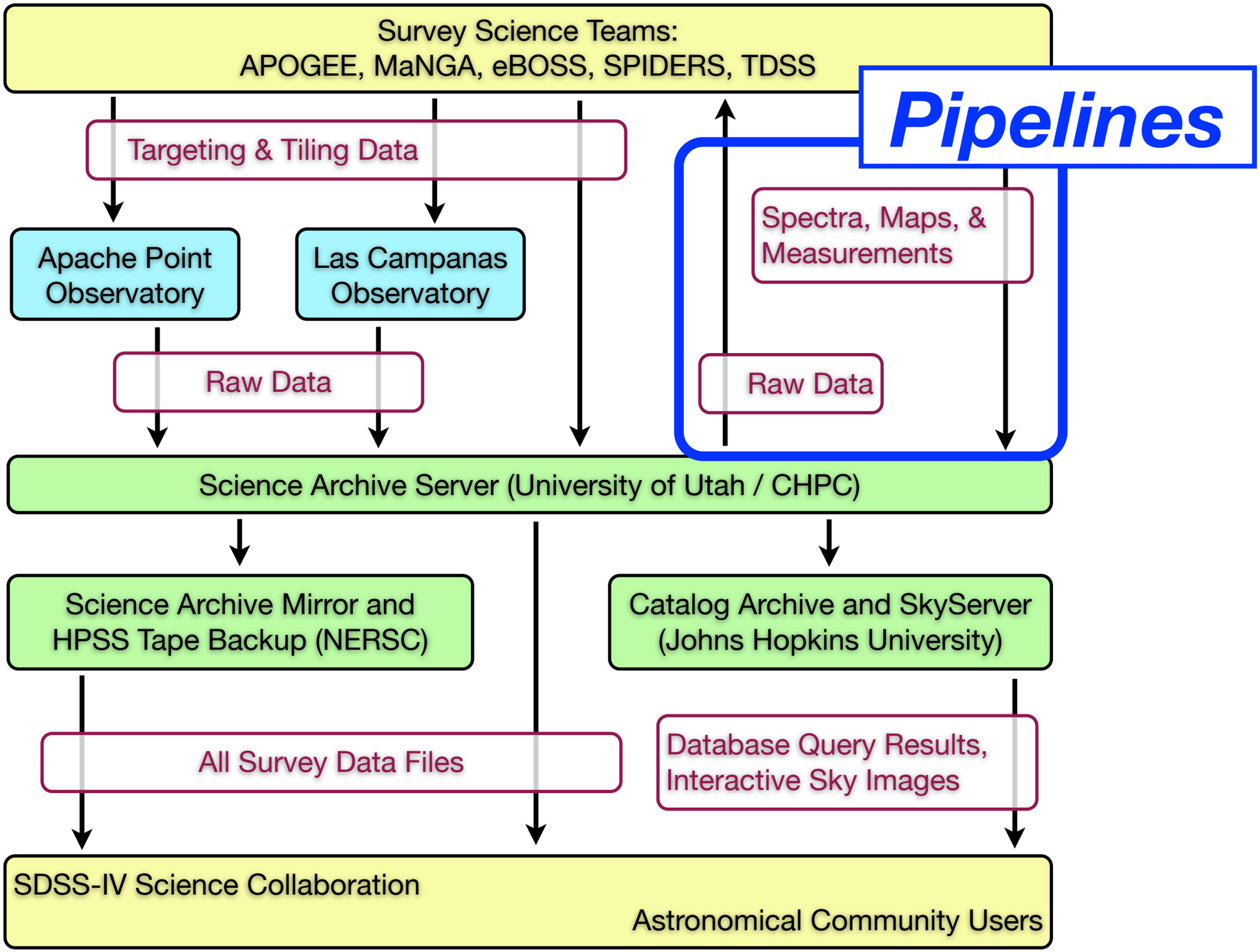
Core SDSS-III data management systems

- Science Archive Server (SAS)
  - Central online repository of raw & reduced SDSS data
  - Support for data transfers by http, rsync, bbcp, Globus
- Science Archive Mirror (SAM)
  - Online mirror of all SAS data and metadata content
- High-Performance Storage System (HPSS) tape backup
  - Off-line long-term backup & disaster recovery
- Documentation & Collaboration Server (DCS)
  - Wiki, software repository, code browser, ticketing system, mailing lists, publication lists, project announcement system
- Catalog Archive Server (CAS) & SkyServer
  - Public-facing DB server with powerful query functionality
  - Interactive browsable sky front-end to SDSS database

SDSS-IV high-level data flow schematic



SDSS-IV high-level data flow schematic



Current SDSS-III Pipeline Model

APO

APOGEE pipeline team

APOGEE compute nodes

BOSS pipeline team

BOSS compute nodes

MARVELS pipeline team

MARVELS compute nodes

SEGUE-II pipeline team

SEGUE-II compute nodes

All Raw SDSS-III Data

Raw APOGEE data

Reduced APOGEE data

Raw BOSS data

Reduced BOSS data

Raw MARVELS data

Reduced MARVELS data

Raw SEGUE-II data

Reduced SEGUE-II data

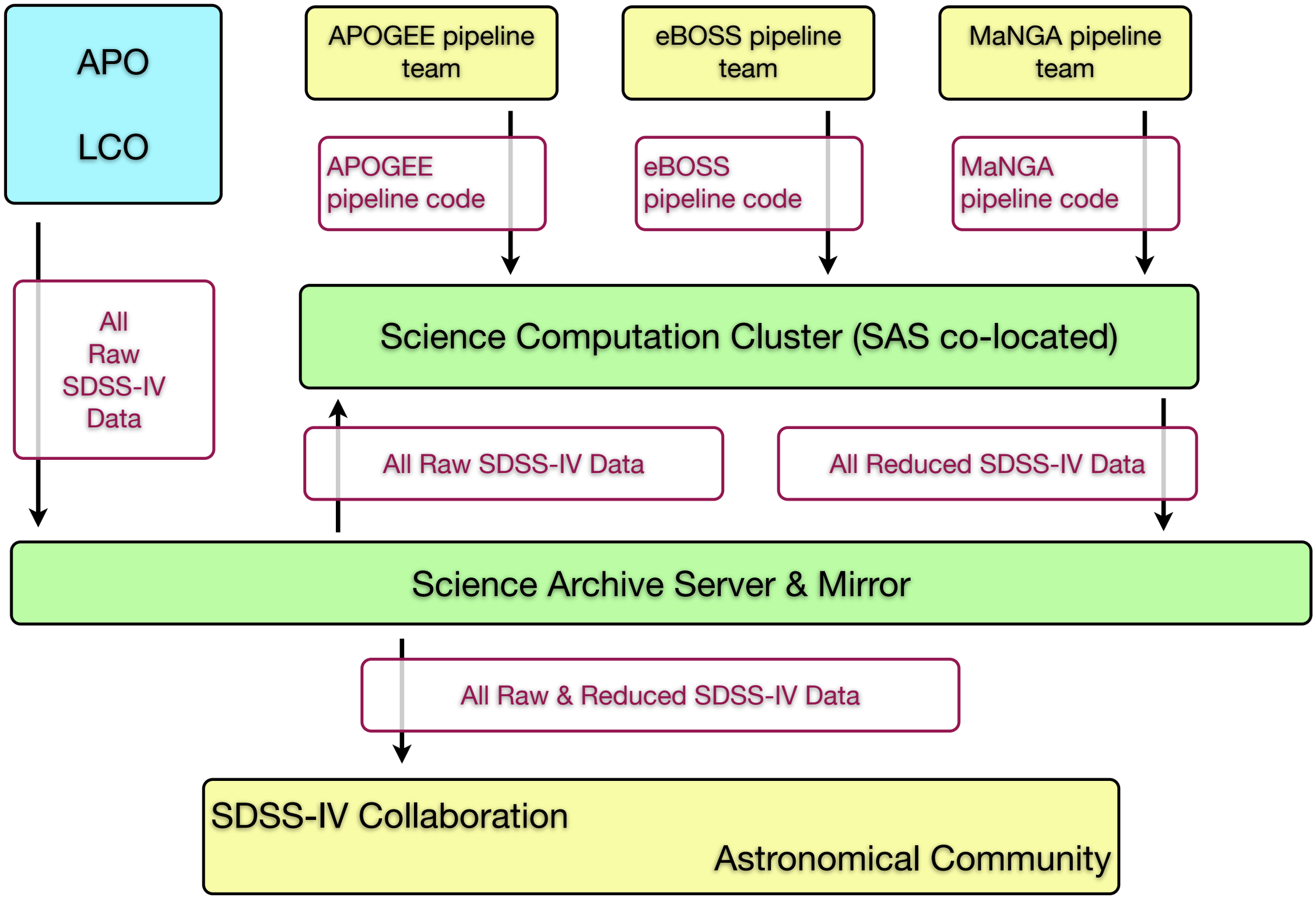
Science Archive Server & Mirror

All Raw & Reduced SDSS-III Data

SDSS-III Collaboration

Astronomical Community

Future SDSS-IV Pipeline Model

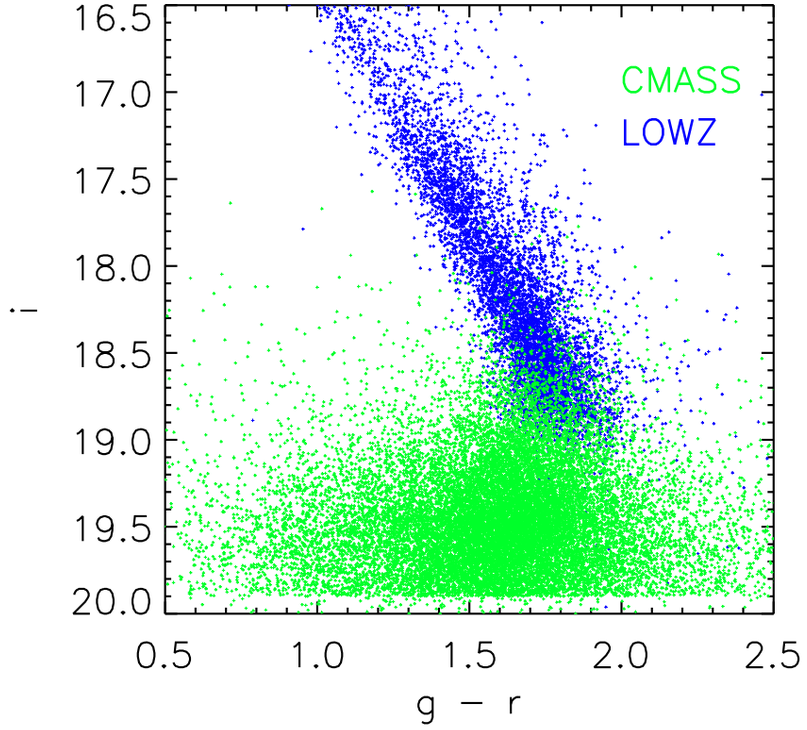


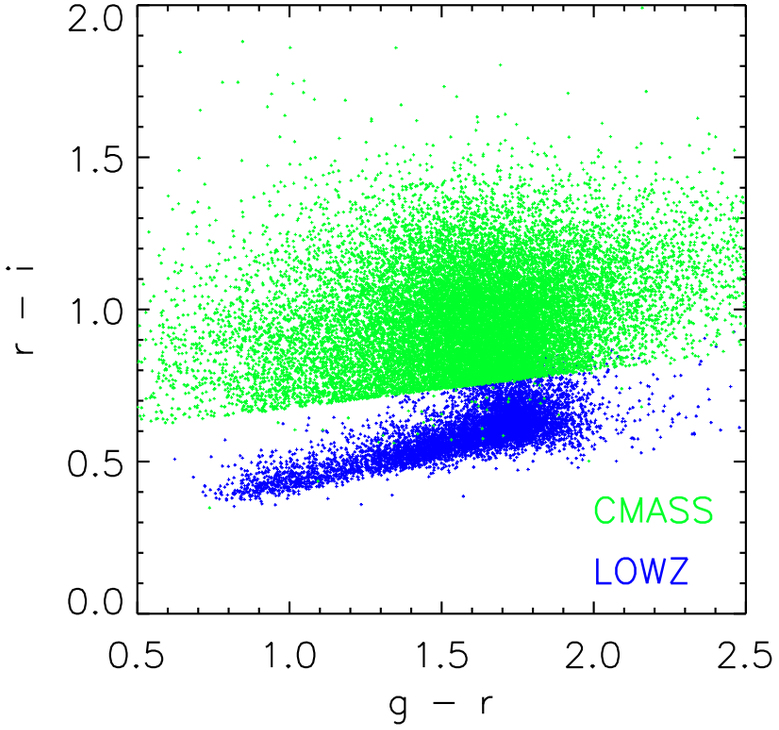
BOSS redshift survey in a nutshell

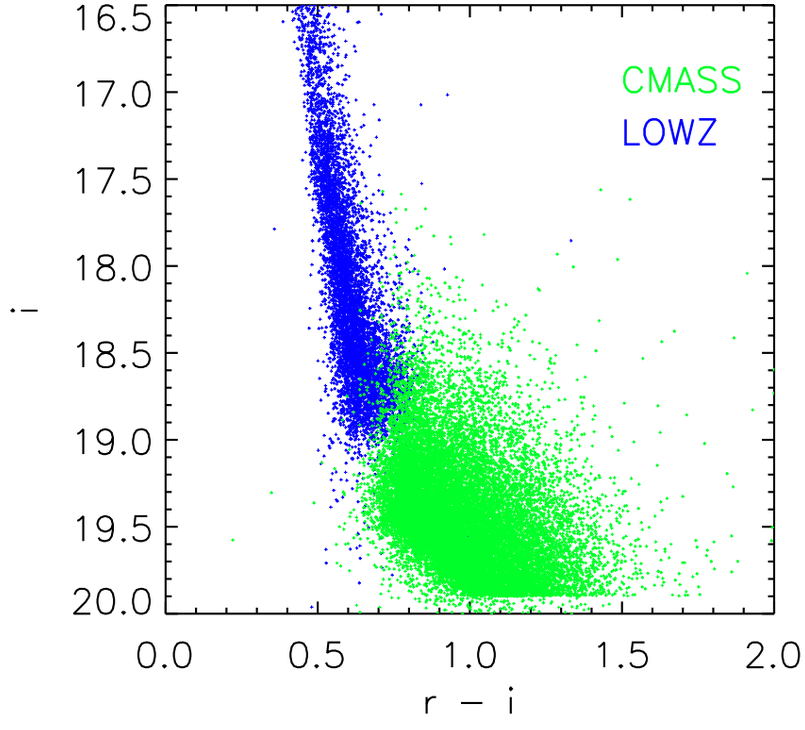
Select targets photometrically

CMASS = BOSS higher-redshift galaxy targets

LOWZ = BOSS lower-redshift galaxy targets



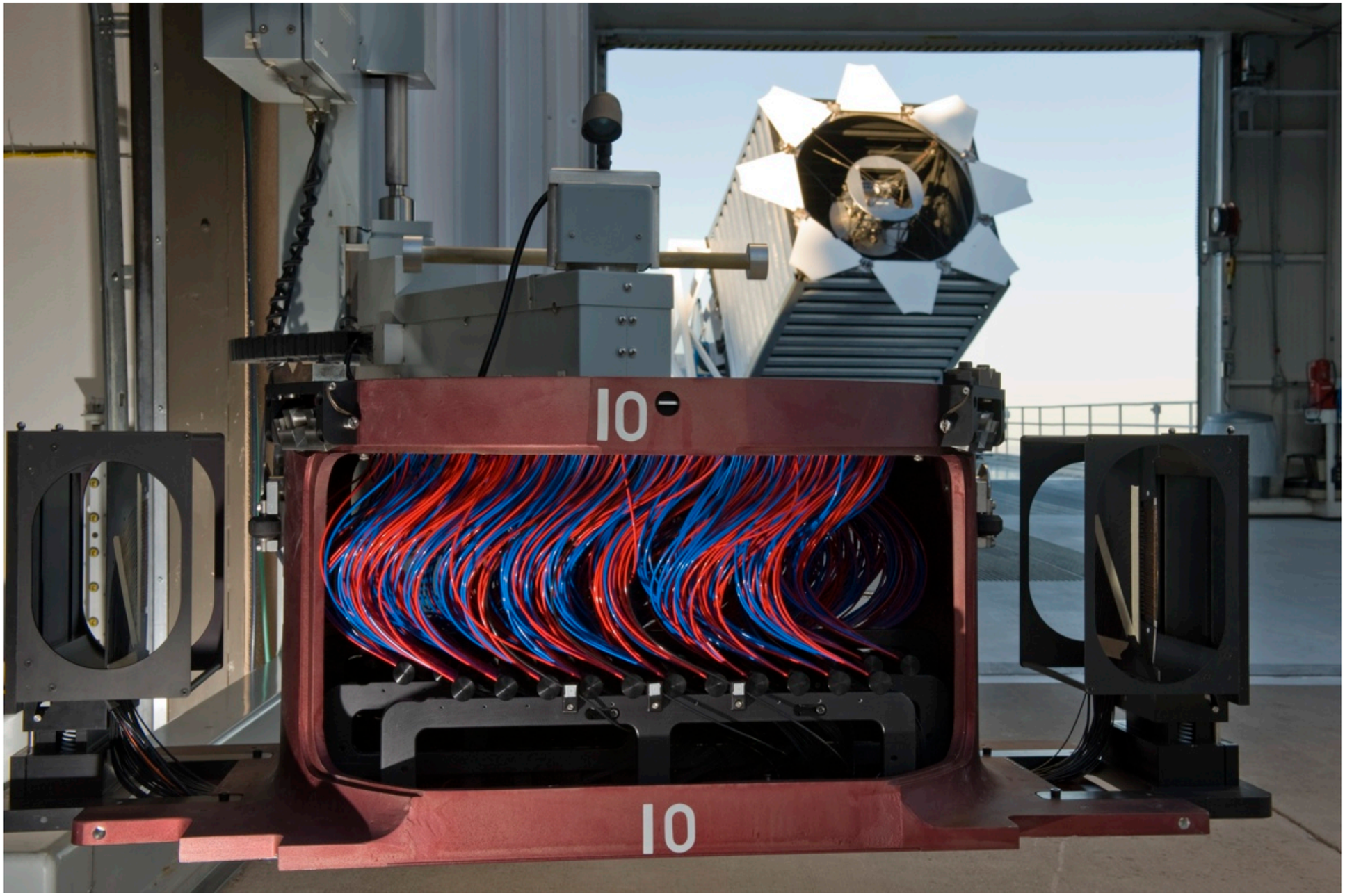




(N.B.: don't forget quasars!)

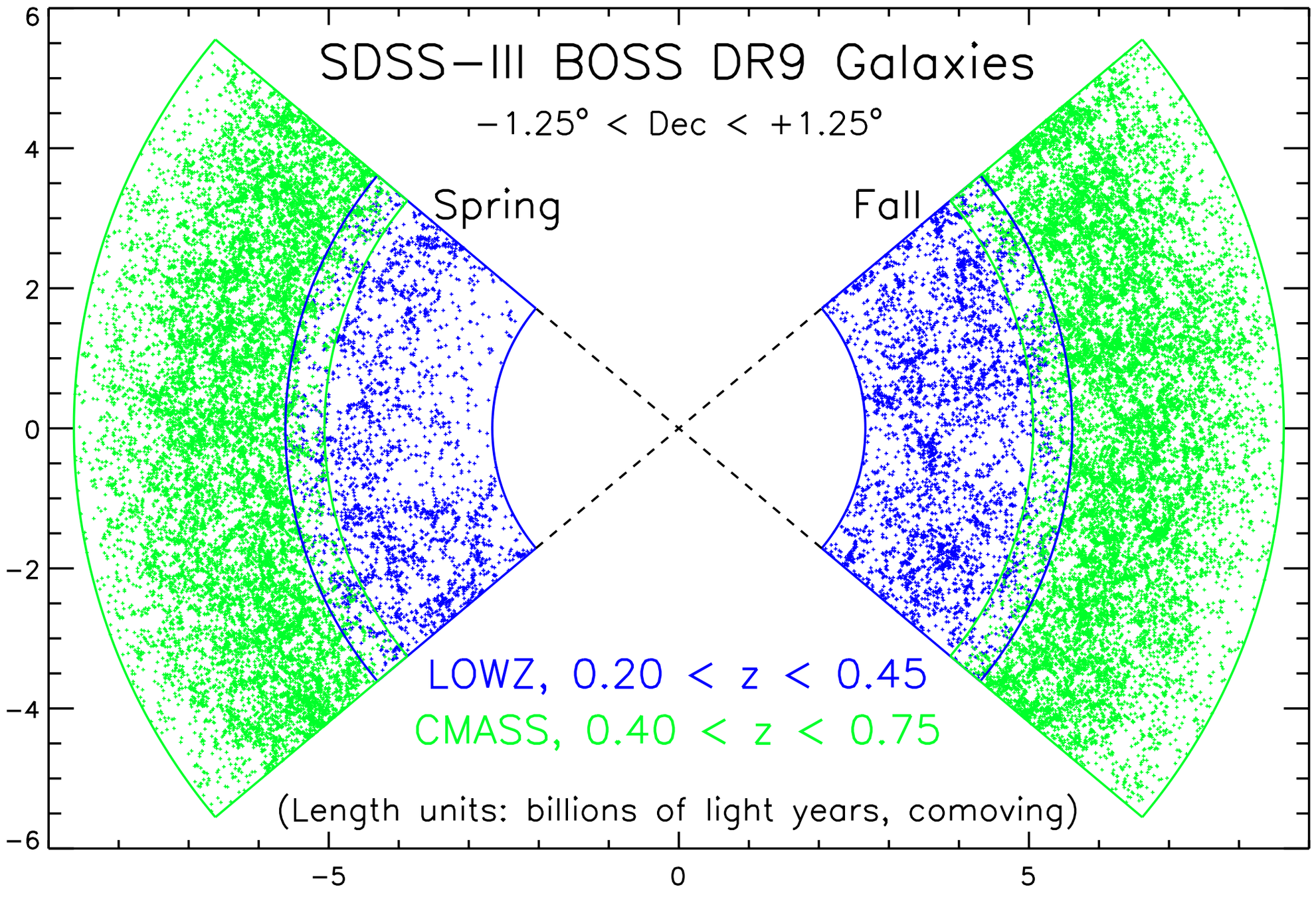
BOSS redshift survey in a nutshell

Observe them spectroscopically



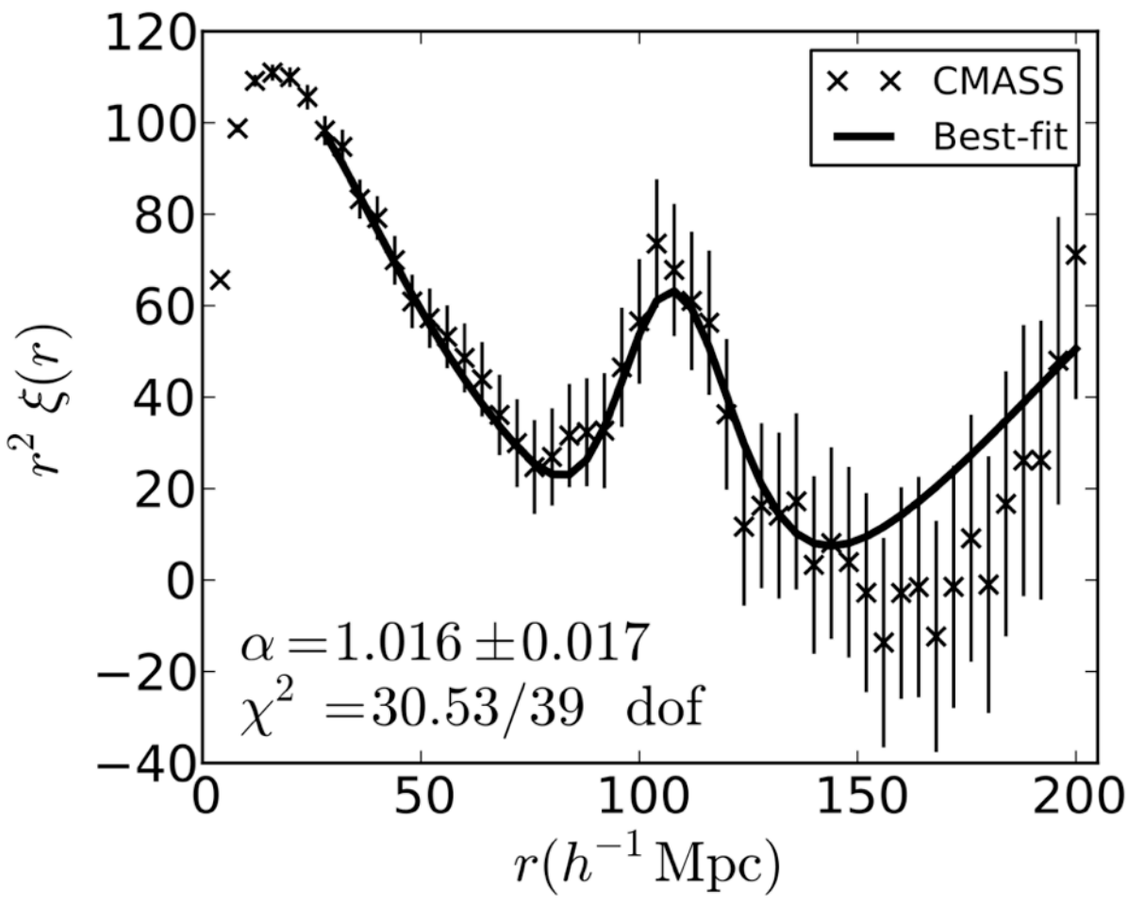
BOSS redshift survey in a nutshell

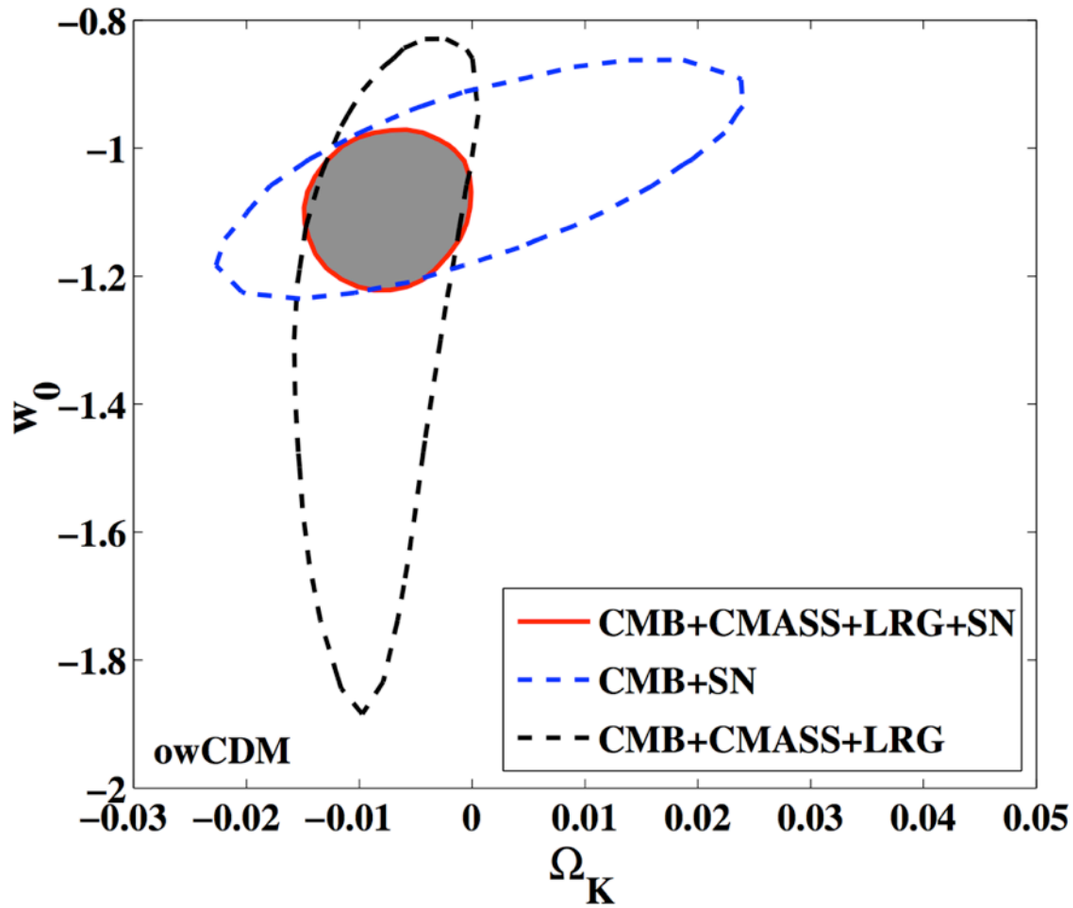
Measure their redshifts to make a 3D map



BOSS redshift survey in a nutshell

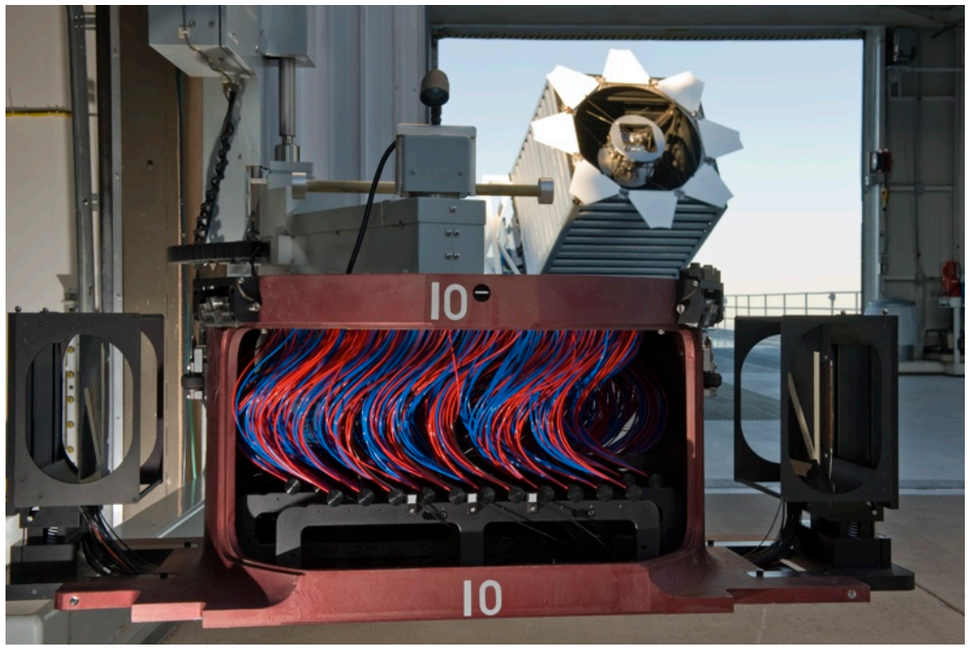
Measure 2-point correlation function and constrain dark energy parameters

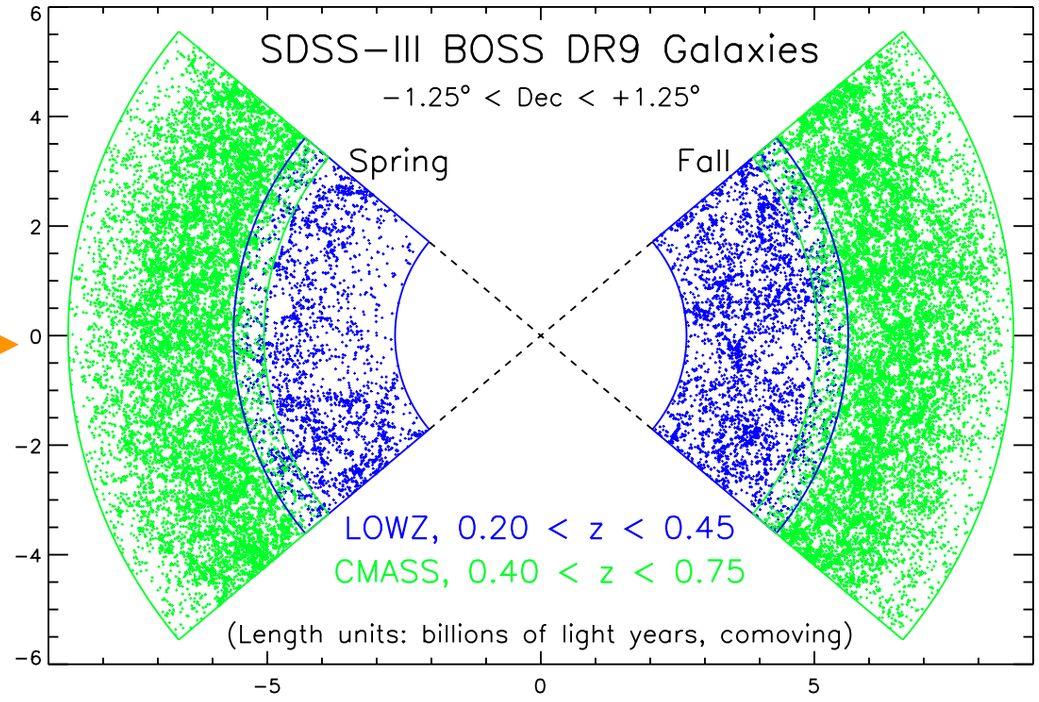




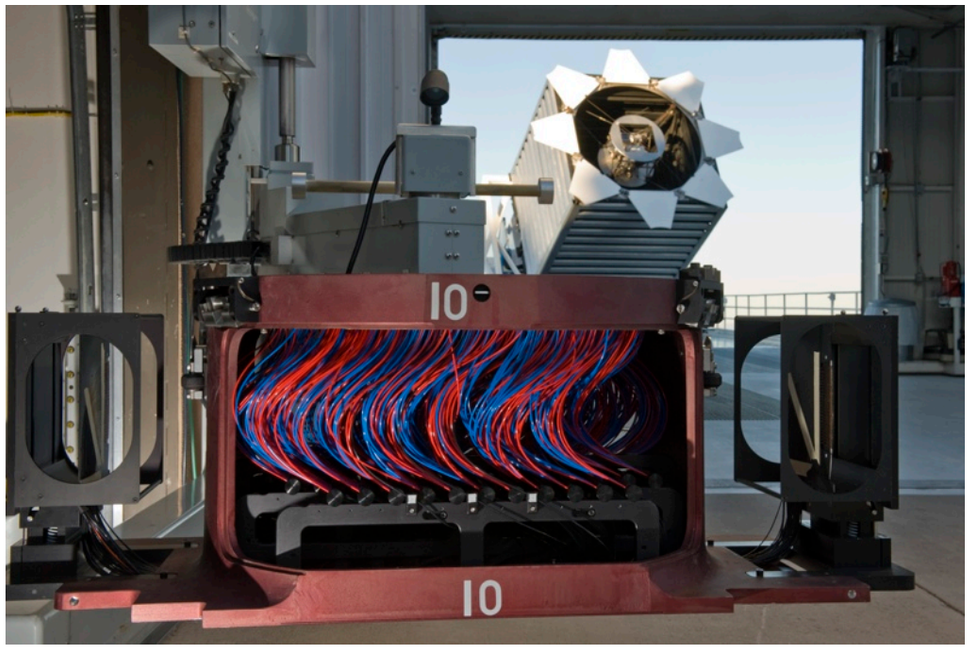
Anderson et al. 2012

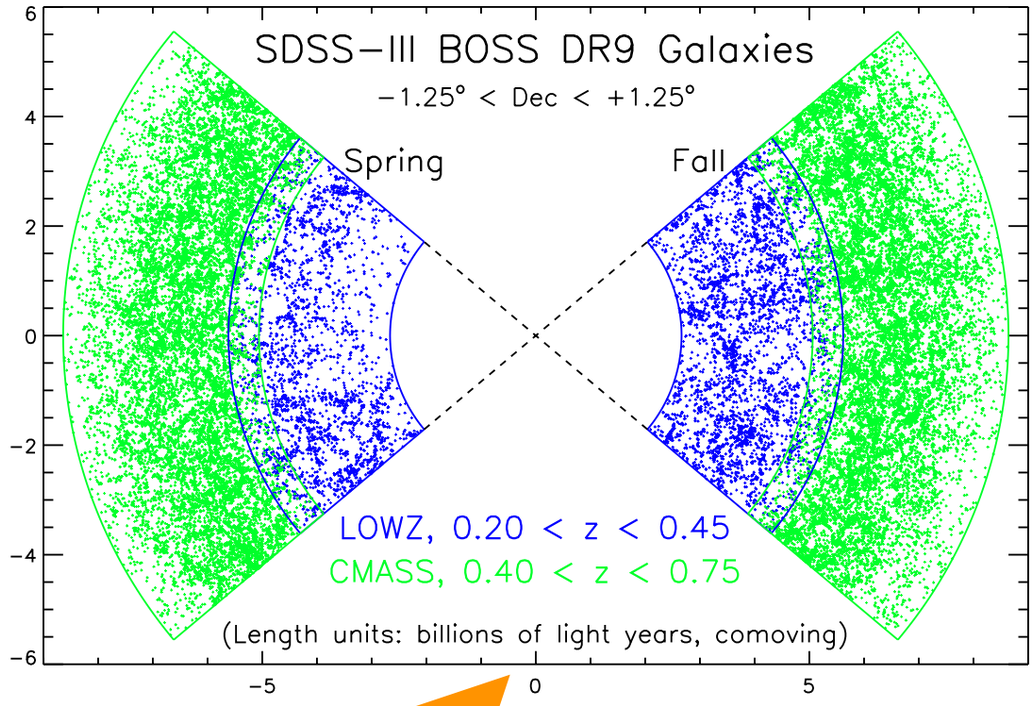
What really goes on in this step?





What really goes on in this step?





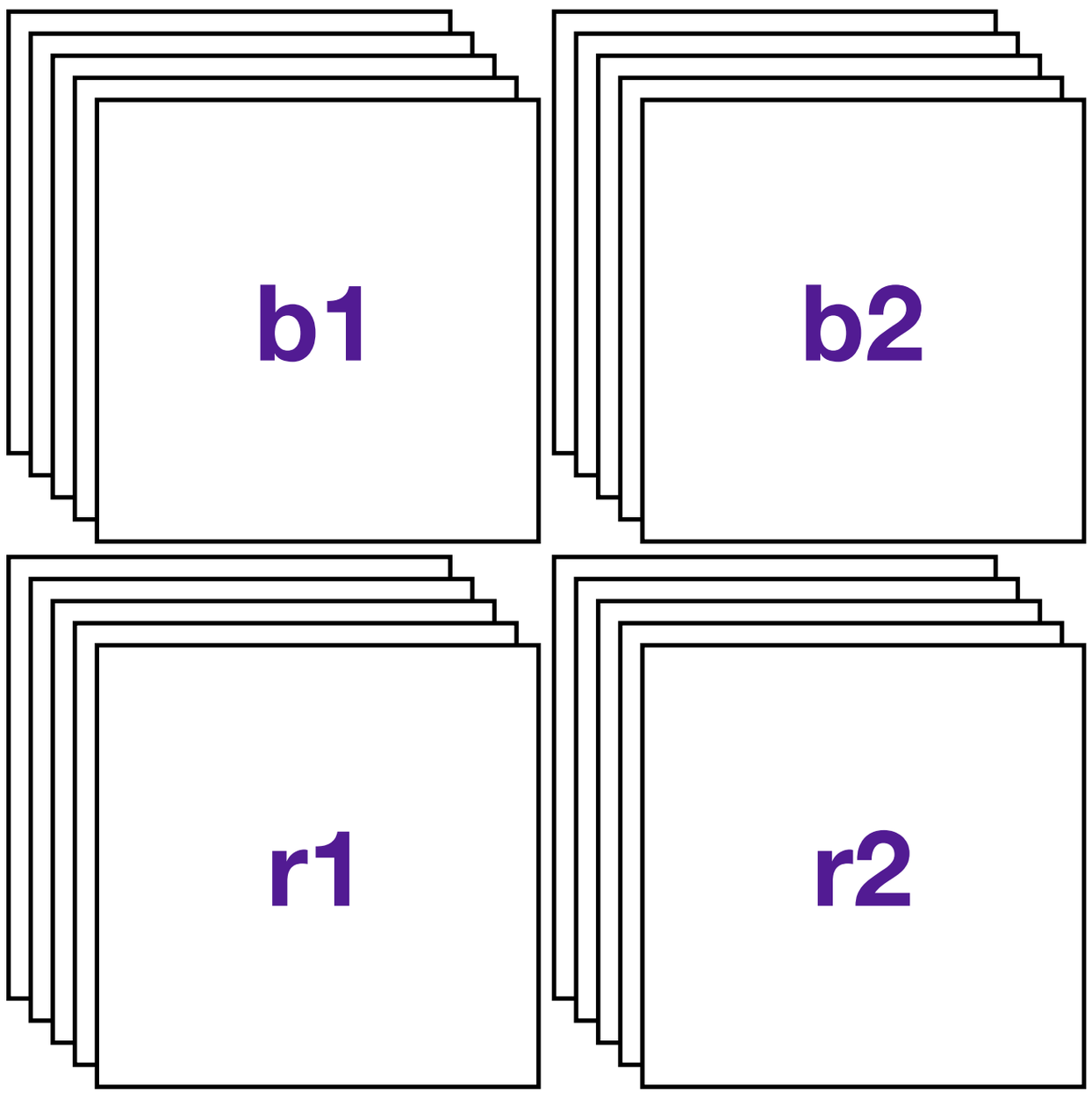
BOSS SPECTROSCOPIC PIPELINE

BOSS Spectroscopic Pipeline Overview

- Fully automated IDL software pipeline
- Inherited & modified from SDSS-I/II
- Cluster-parallelized by plates of 1000 fibers
- Extensive execution logging and QA plot generation
- Raw pixel errors estimated from first principles
- Errors propagated to redshifts and classifications
- Exceeds BOSS redshift science requirements
- Schlegel et al. in prep, Bolton et al. 2012, AJ, 144, 144

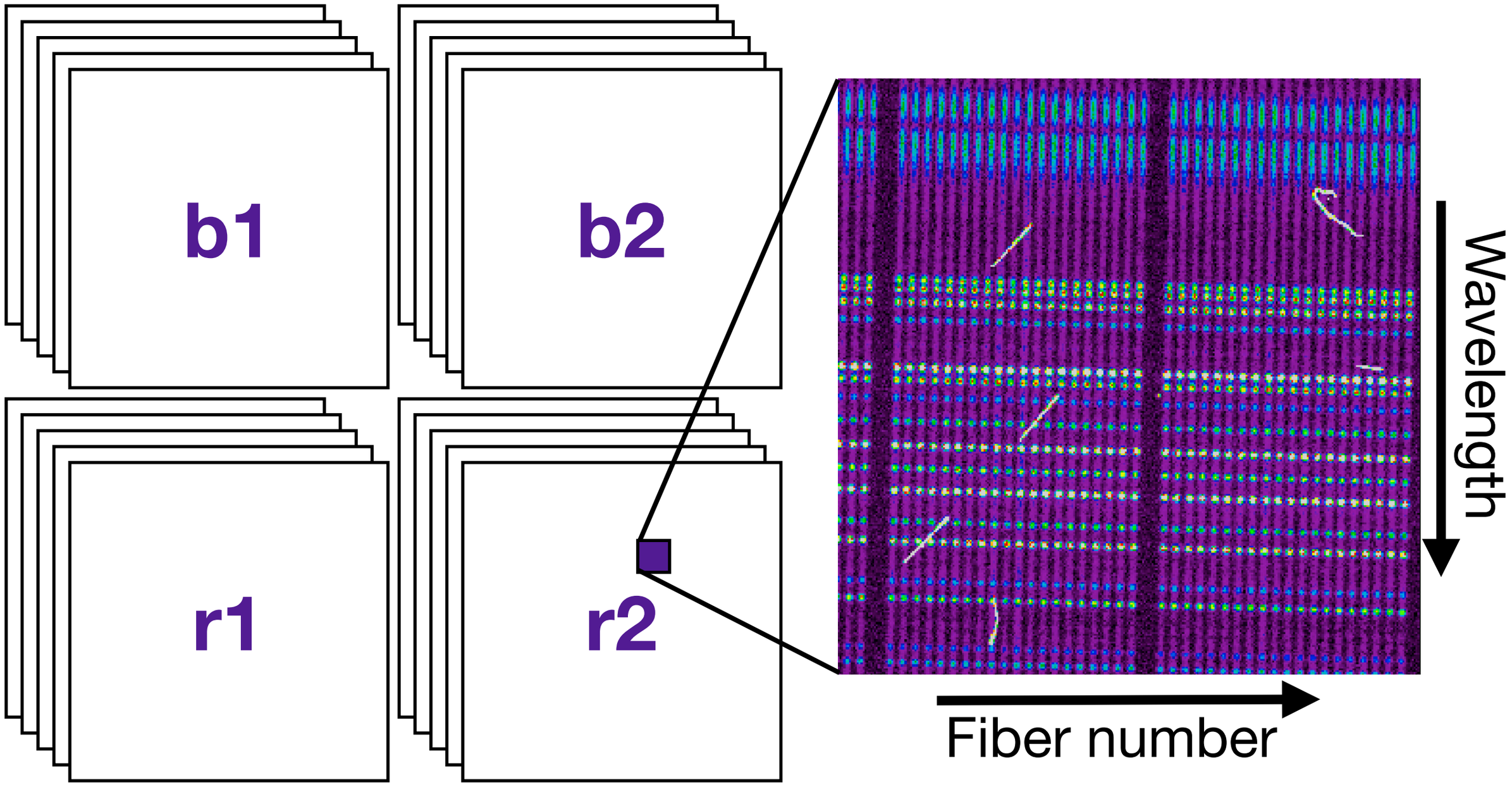
Science Frames

~5 exposures X 4 cameras X ~4,000x4,000 pixels



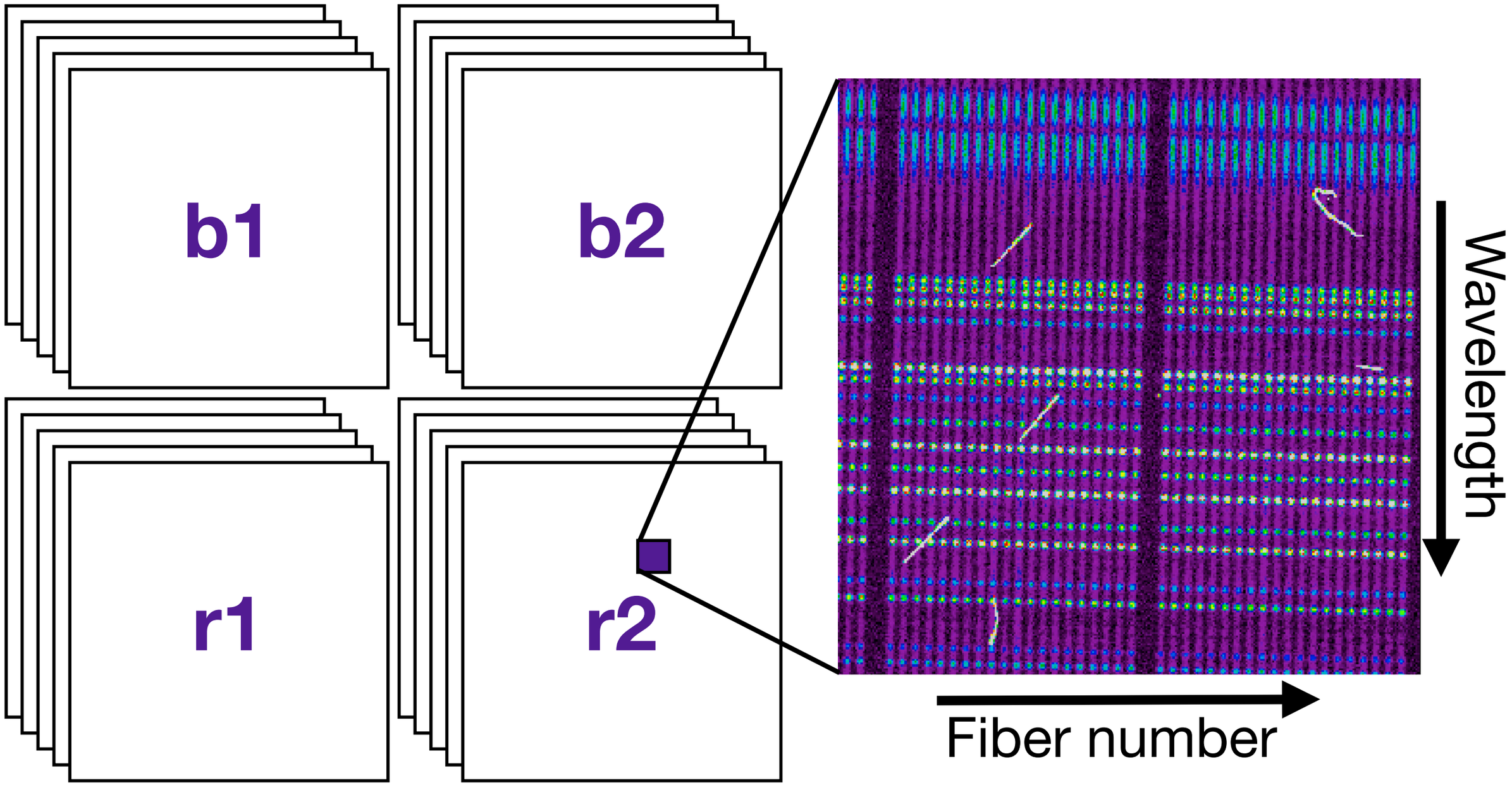
Science Frames

~5 exposures X 4 cameras X ~4,000x4,000 pixels



Science Frames

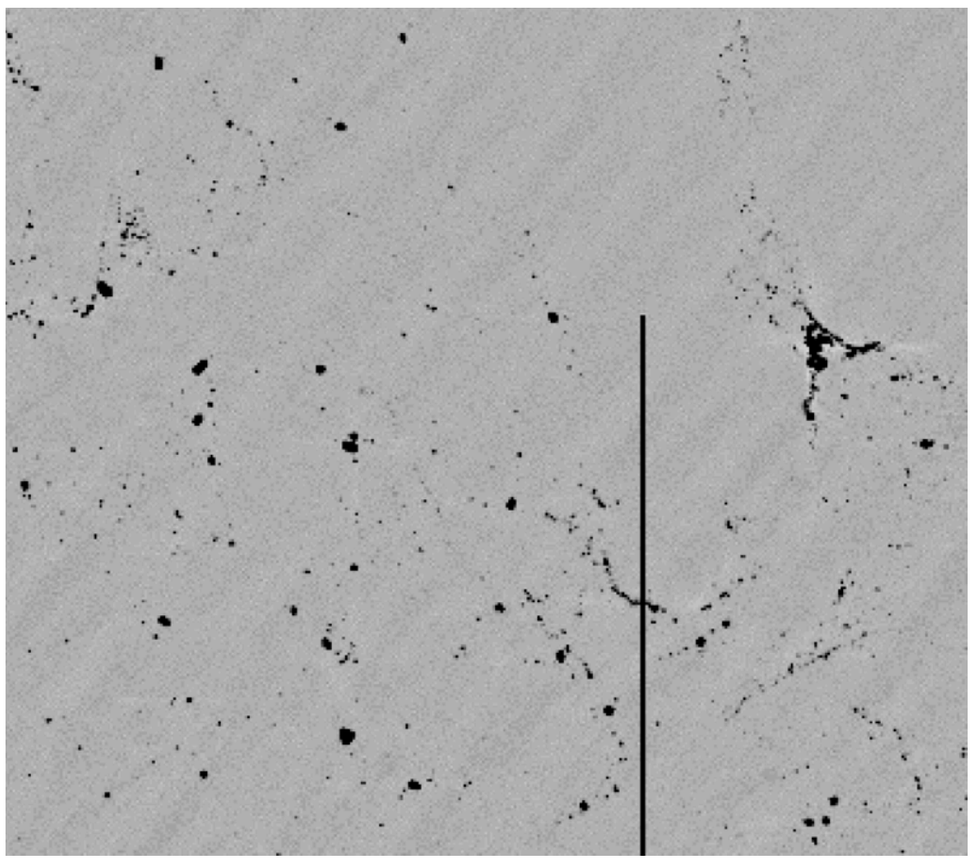
~5 exposures X 4 cameras X ~4,000x4,000 pixels

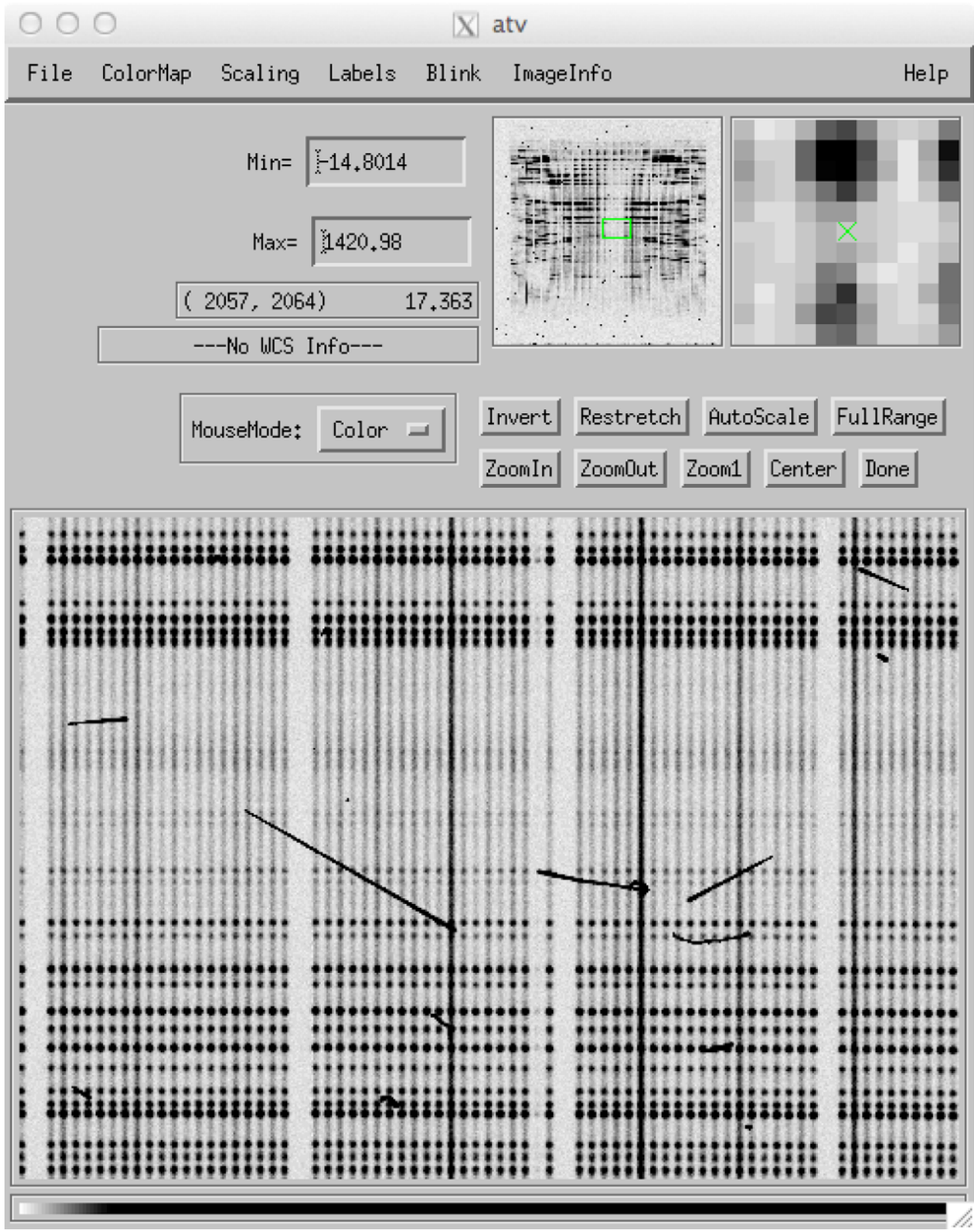


Most counts are night-sky photons!

Calibration, Part 1

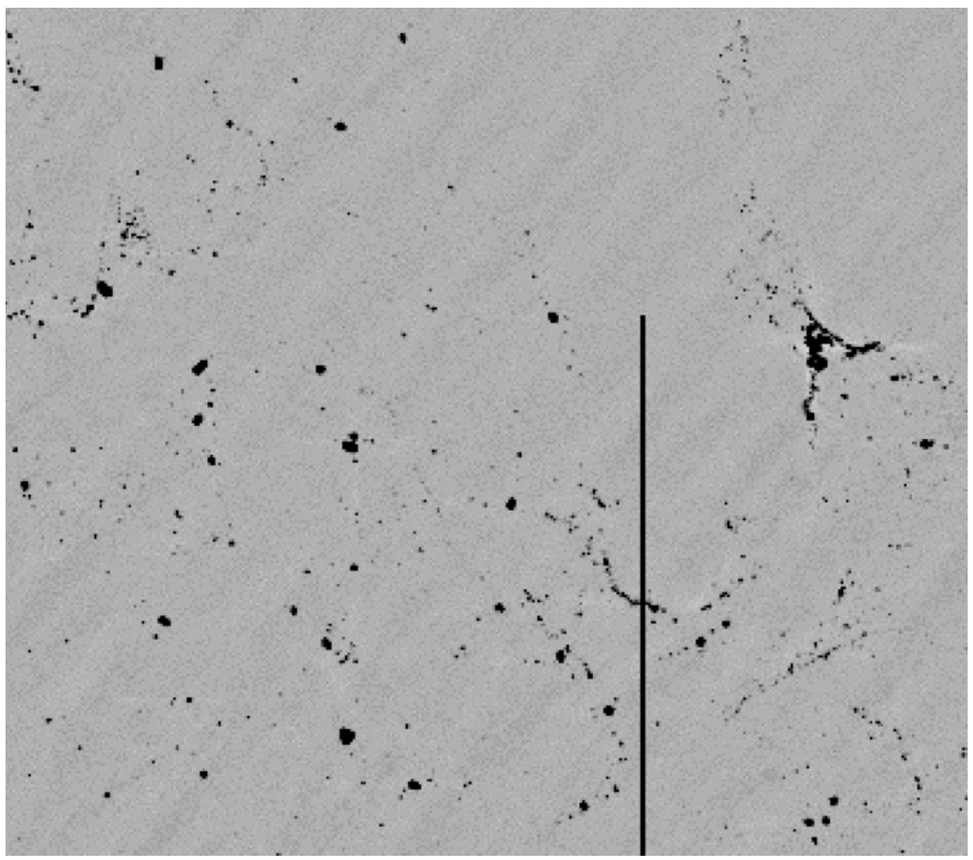
bias subtraction,
gain correction,
pixel-response correction,
bad-column masking,
cosmic-ray rejection,
pixel error estimation

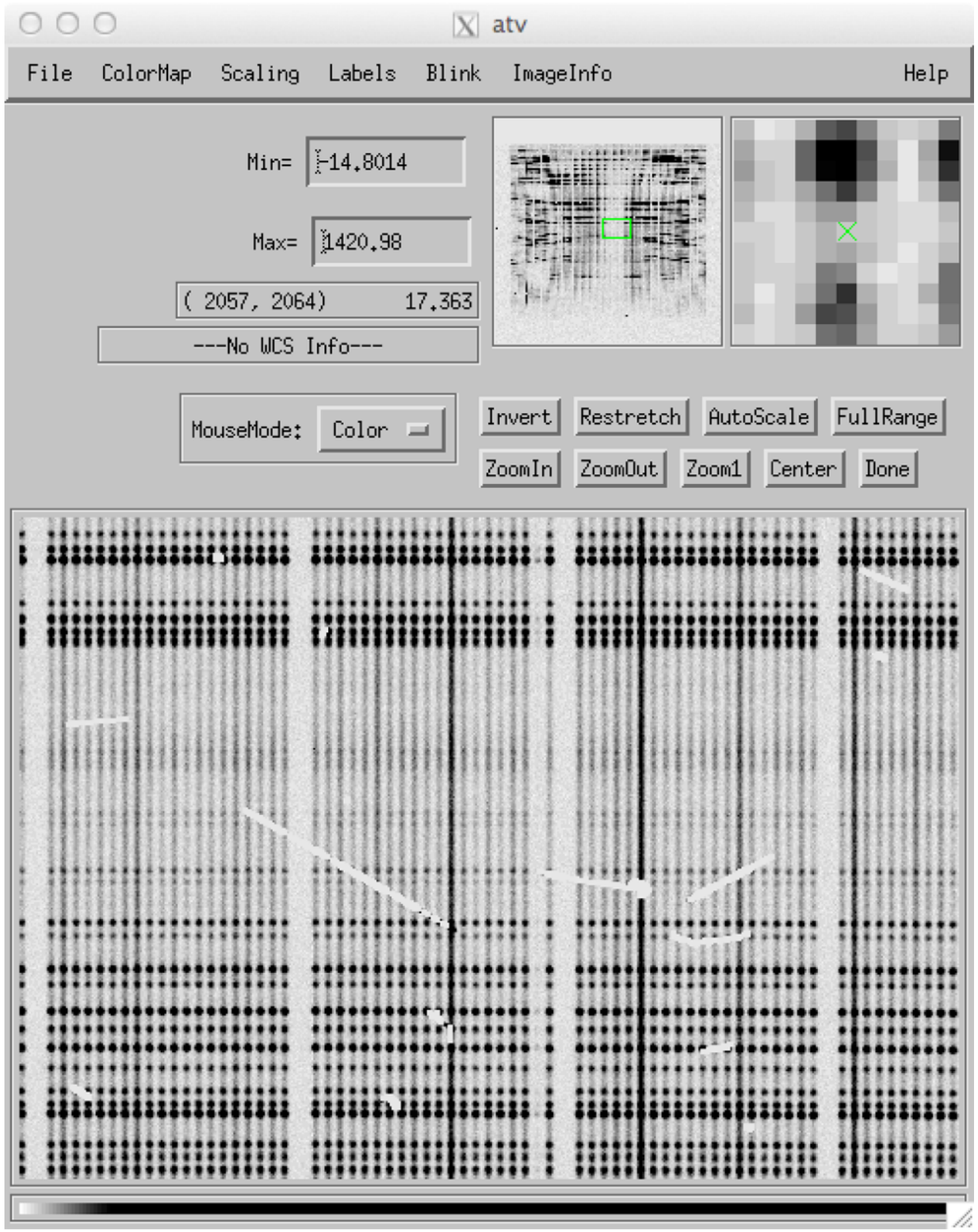




Calibration, Part 1

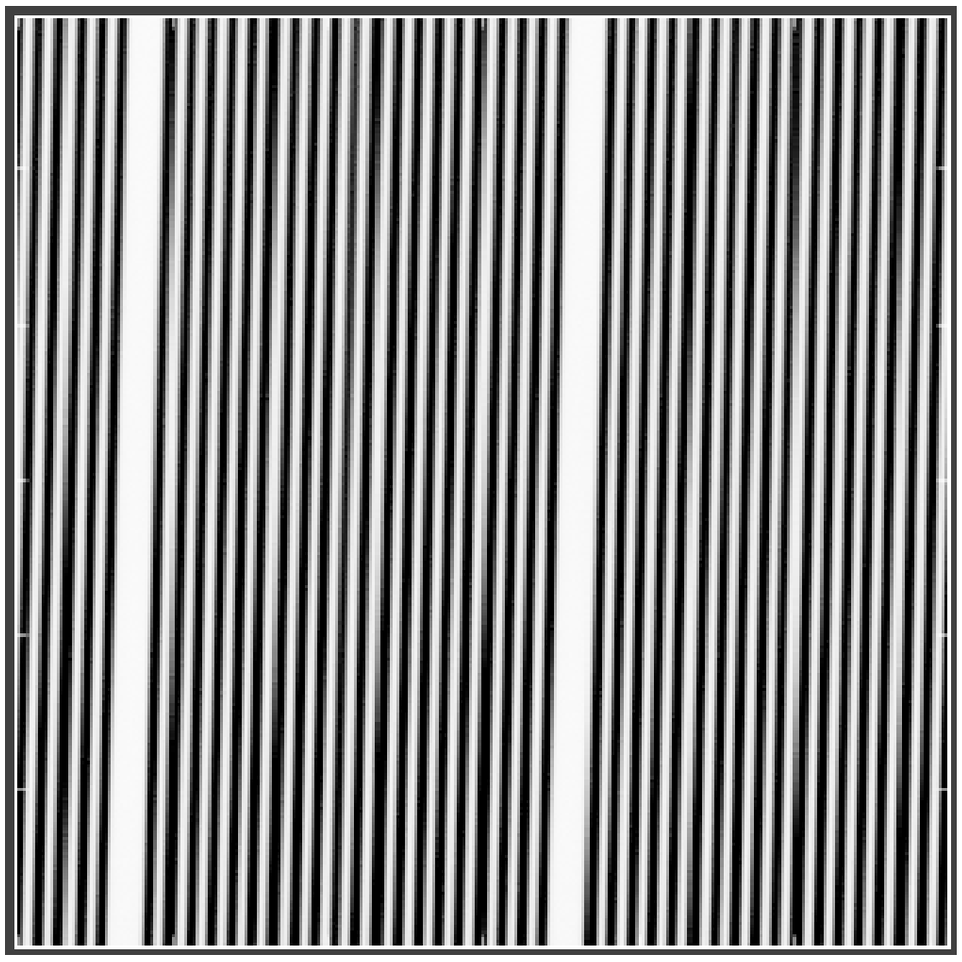
bias subtraction,
gain correction,
pixel-response correction,
bad-column masking,
cosmic-ray rejection,
pixel error estimation





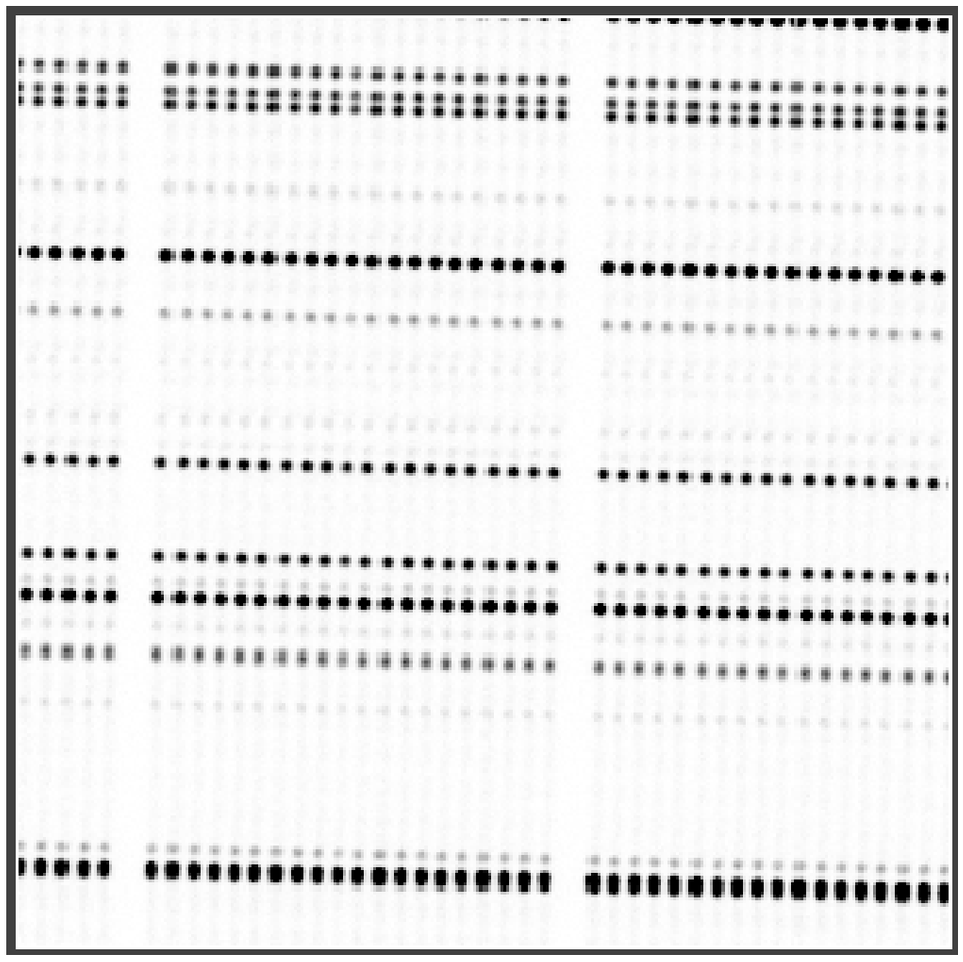
Calibration, Part 2

Flat Frames



trace positions
fiber-to-fiber throughput
cross-sectional profile

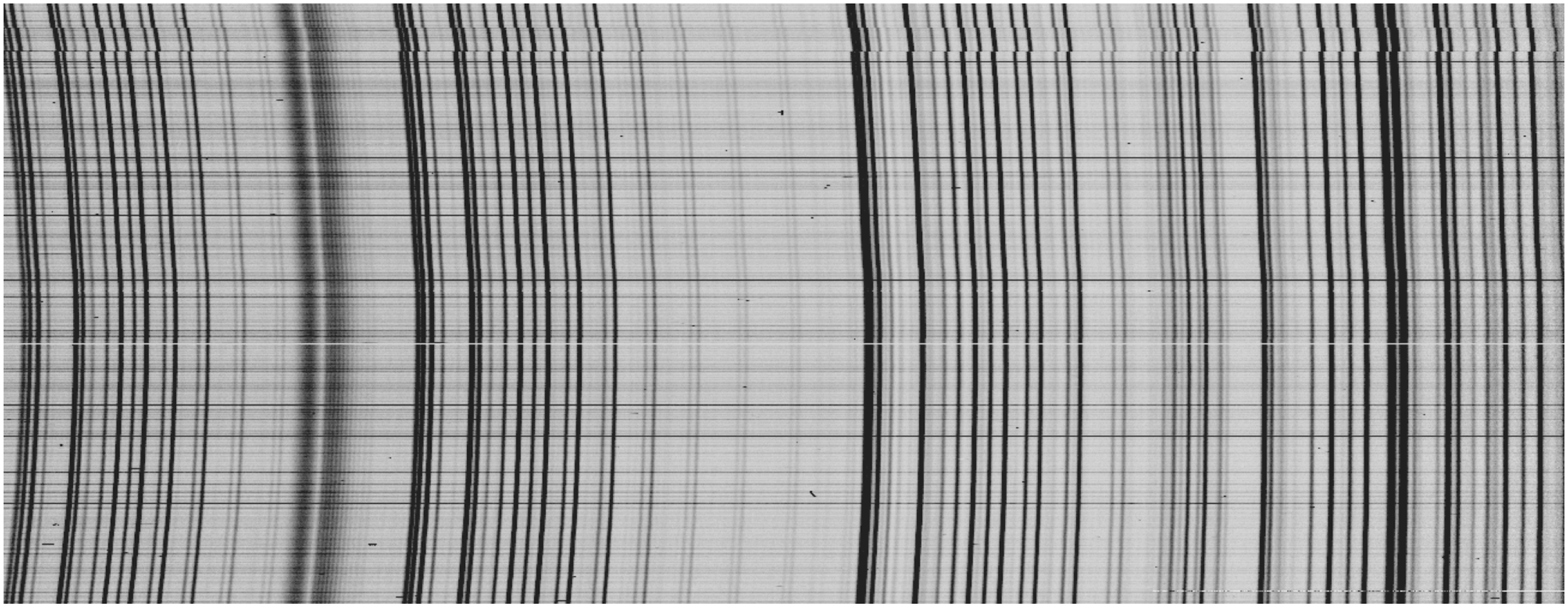
Arc Frames



wavelength solution
spectral resolution

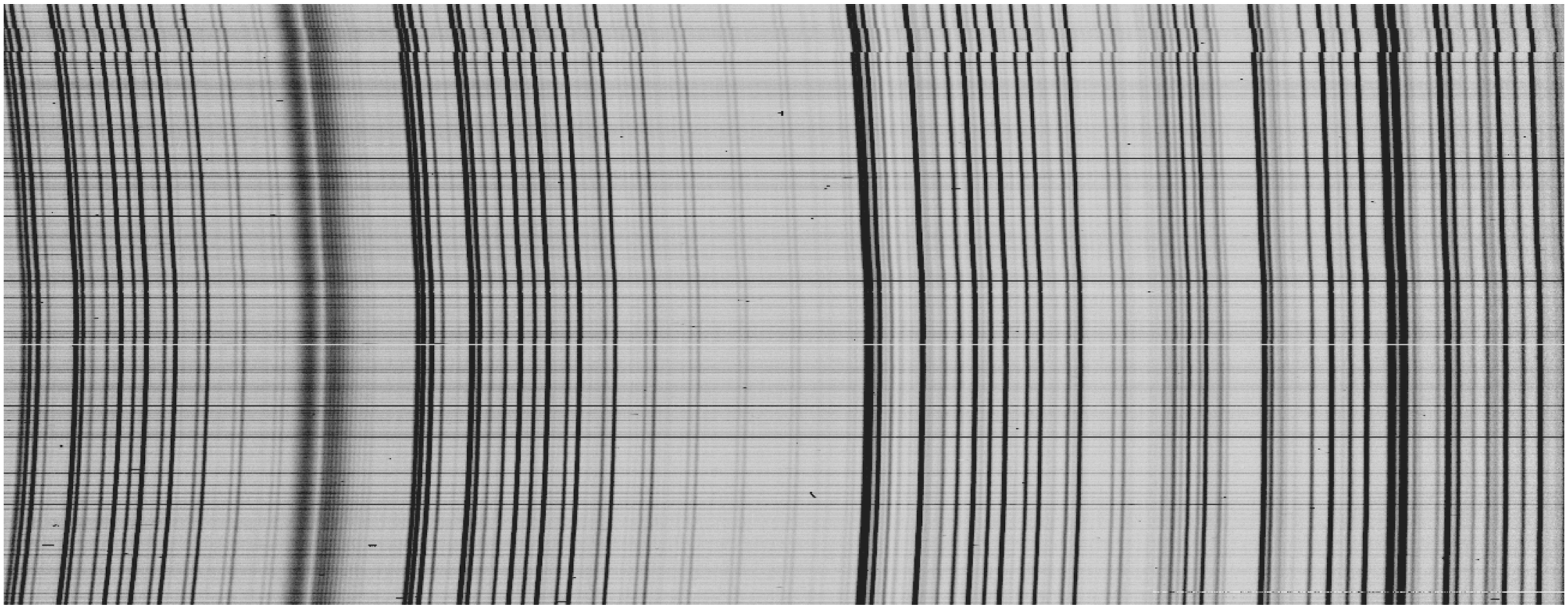
Extraction

Profile fitting “optimal extraction”, a la Horne 1986



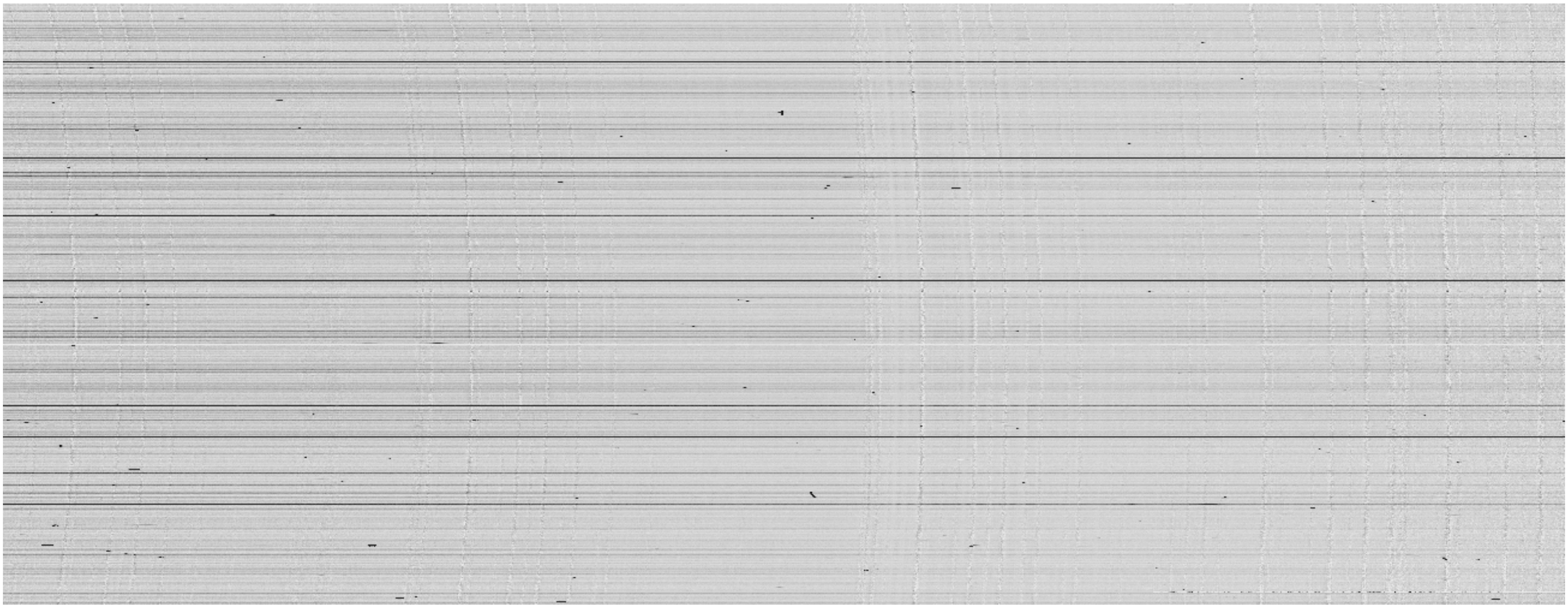
Sky Subtraction

B-spline sky model in un-rebinned pixel space



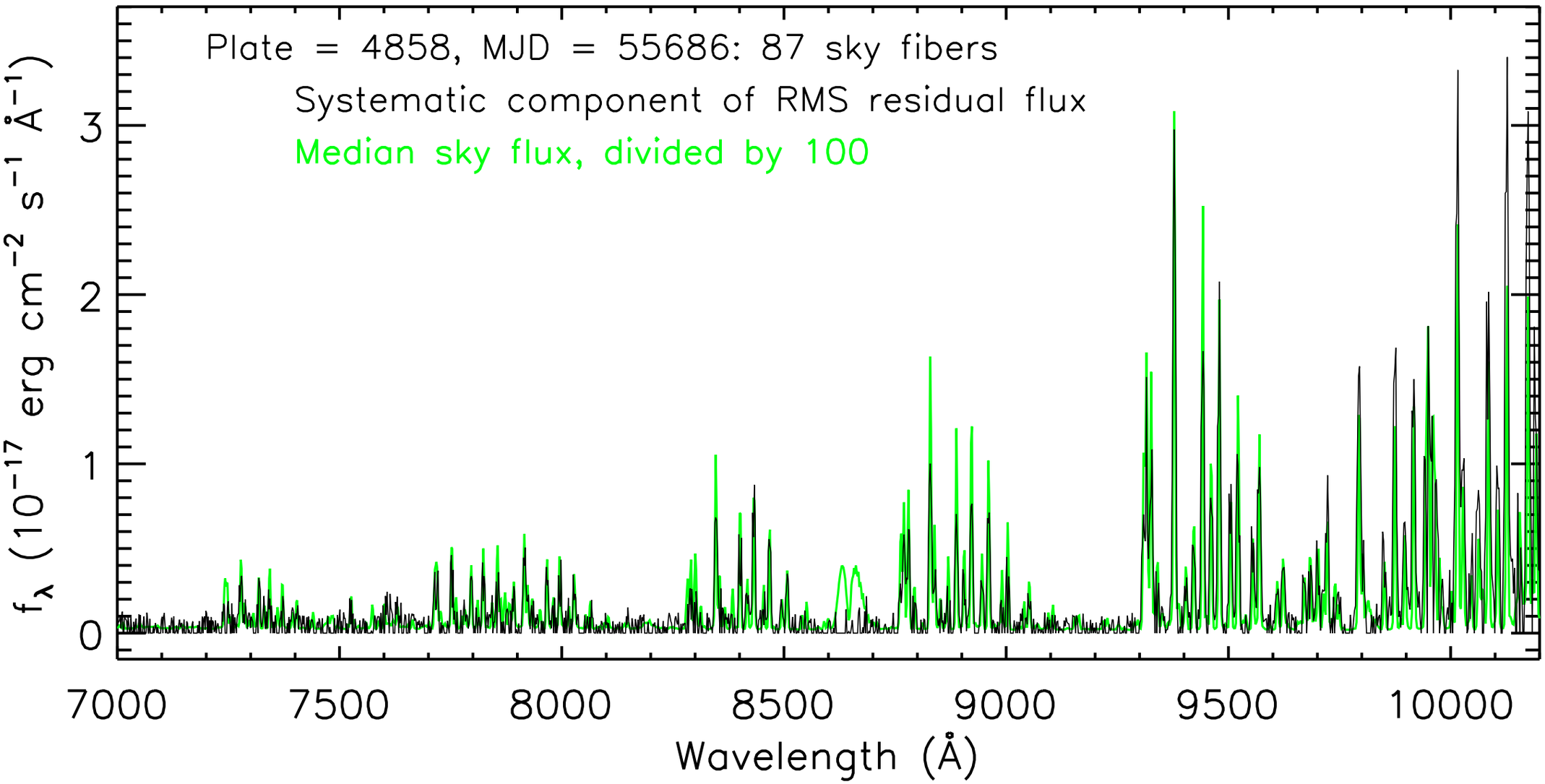
Sky Subtraction

B-spline sky model in un-rebinned pixel space



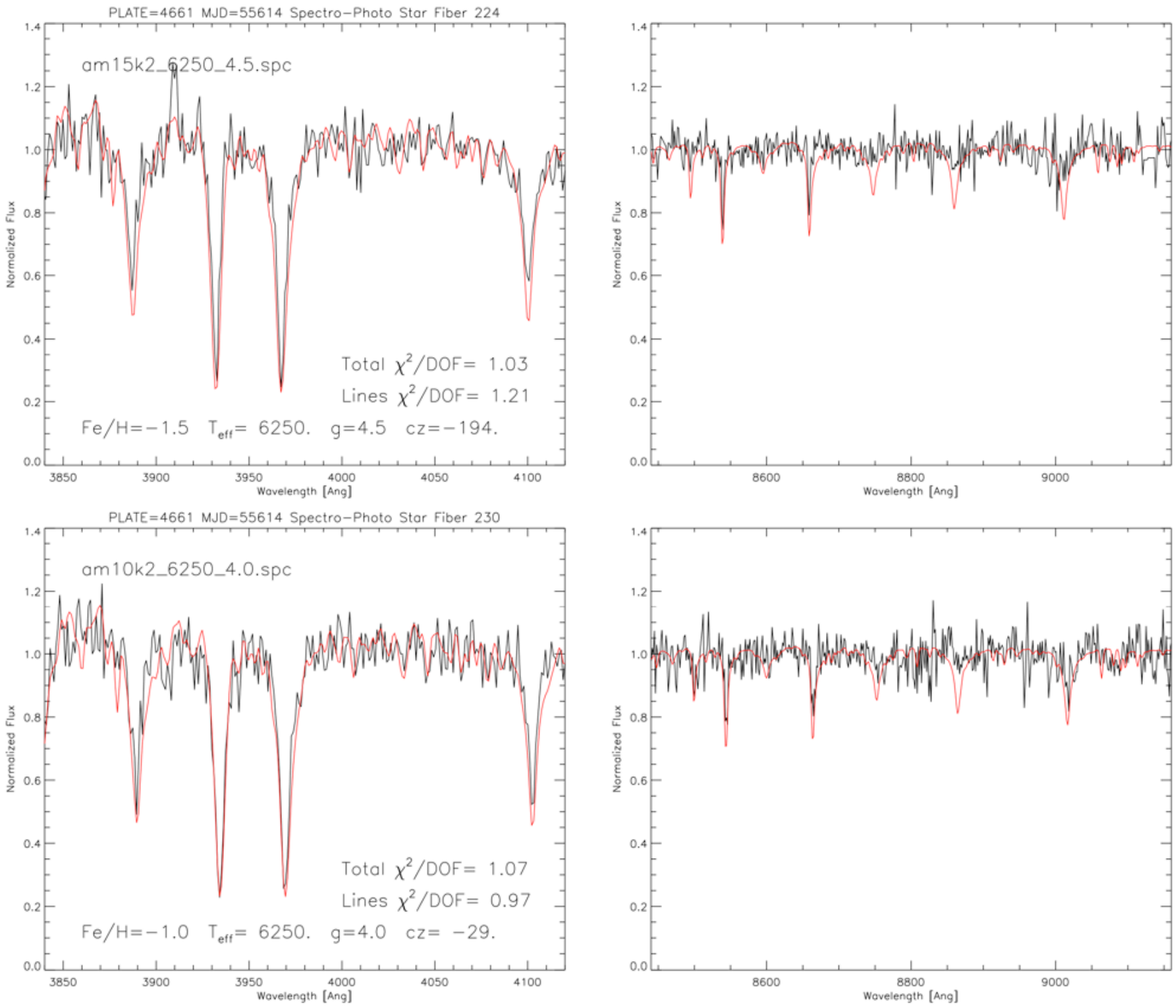
Sky Subtraction Quality

Systematic residuals < ~ 1% even on sky lines



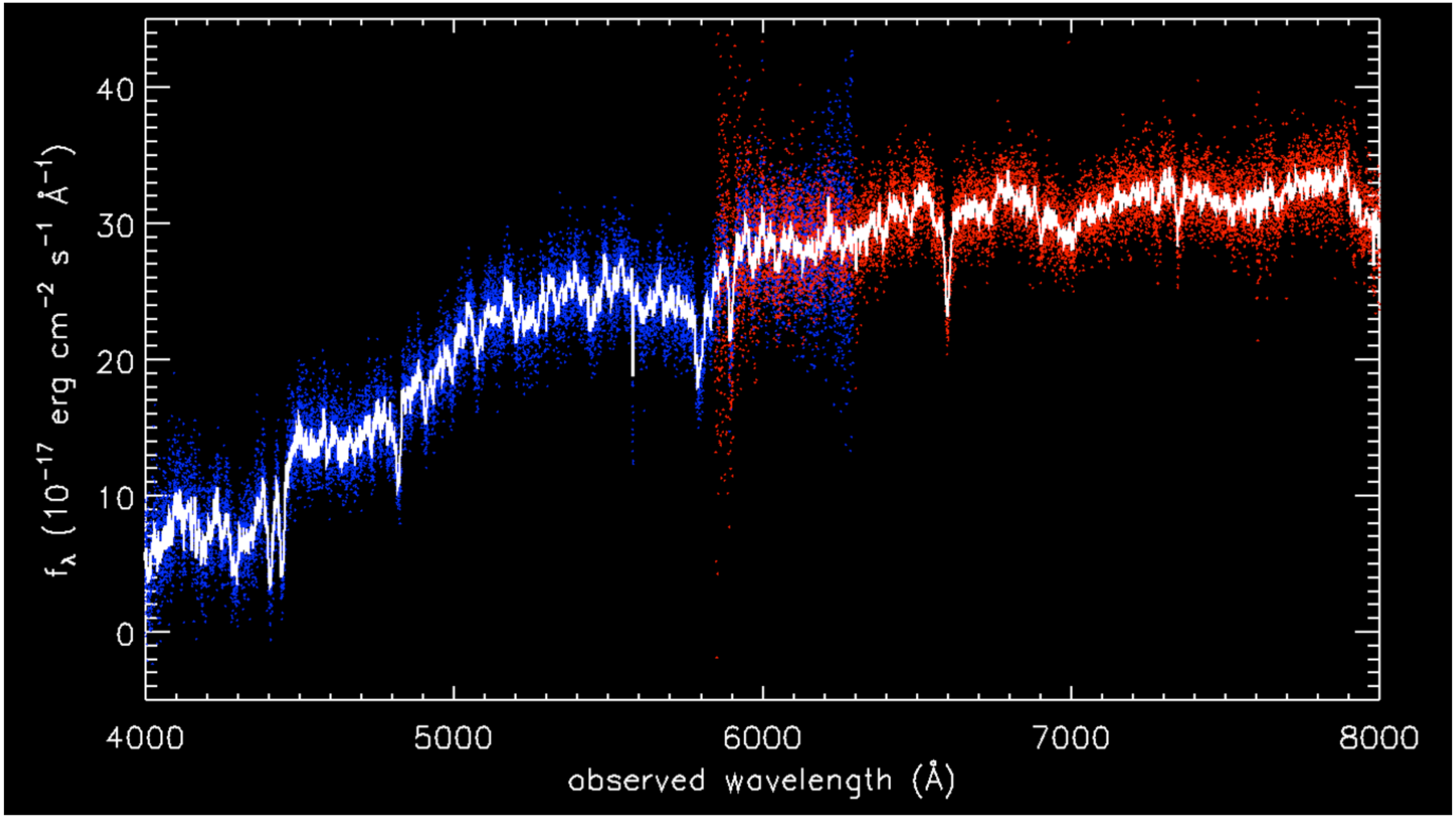
Calibration Part 3: Spectrophotometry

F-star model atmospheres selected by line fitting



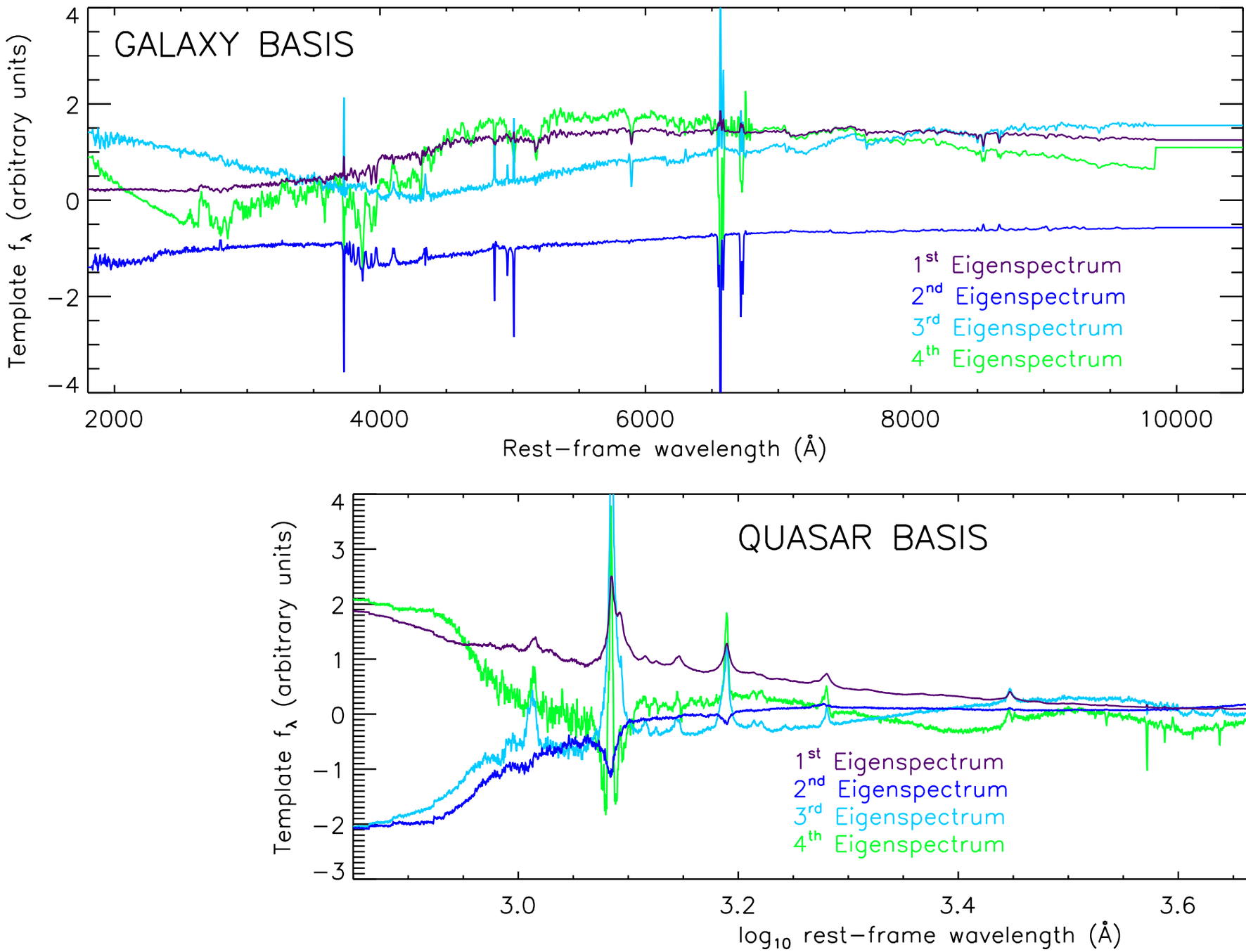
Coaddition

B-spline model for output coadd:
fitted to all fluxed frames,
evaluated at dlogλ = 10^{-4}



Redshift Measurement and Classification

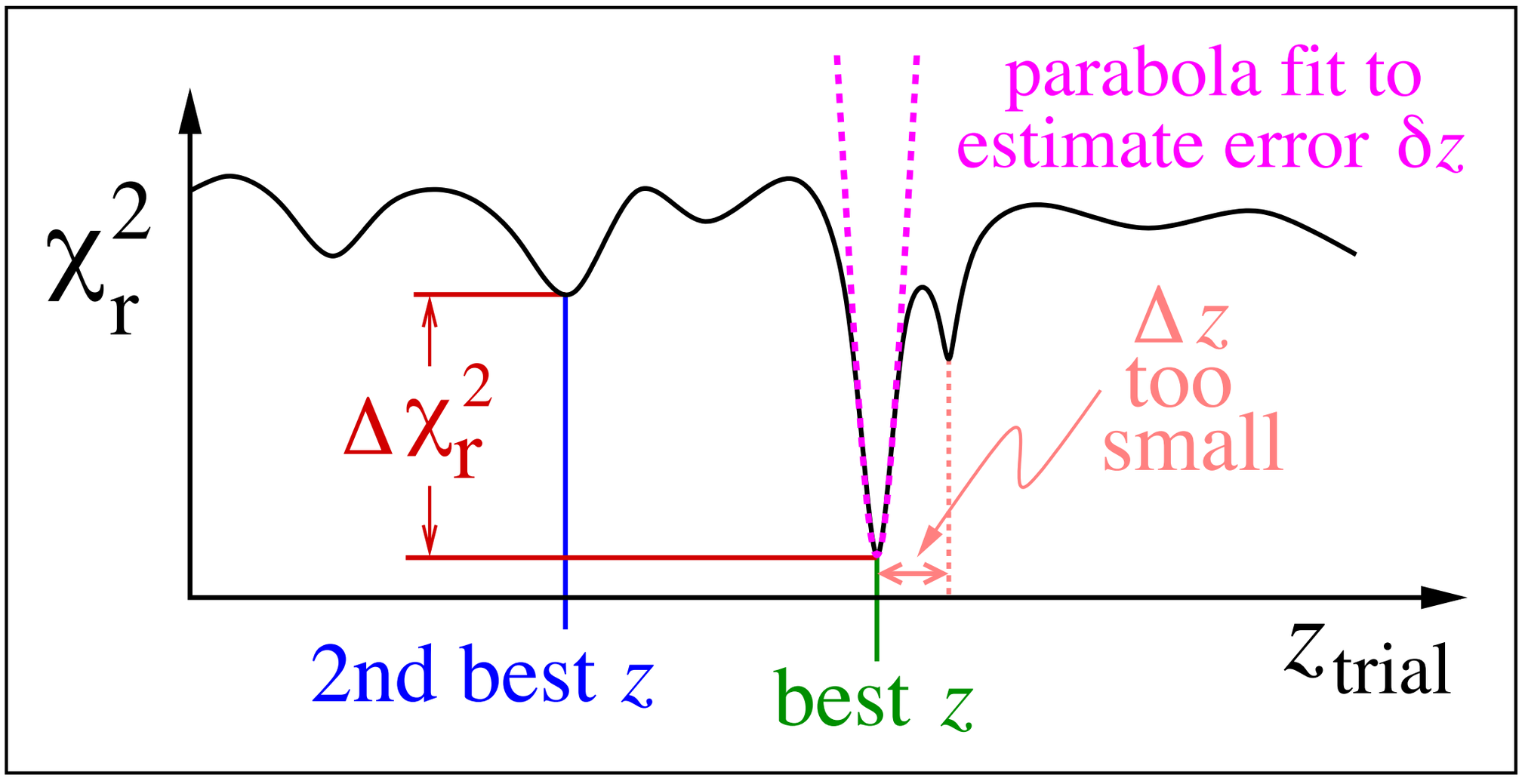
Fitting done using bases of PCA Eigenspectra



Plus: stellar archetype spectra

Redshift Measurement and Classification

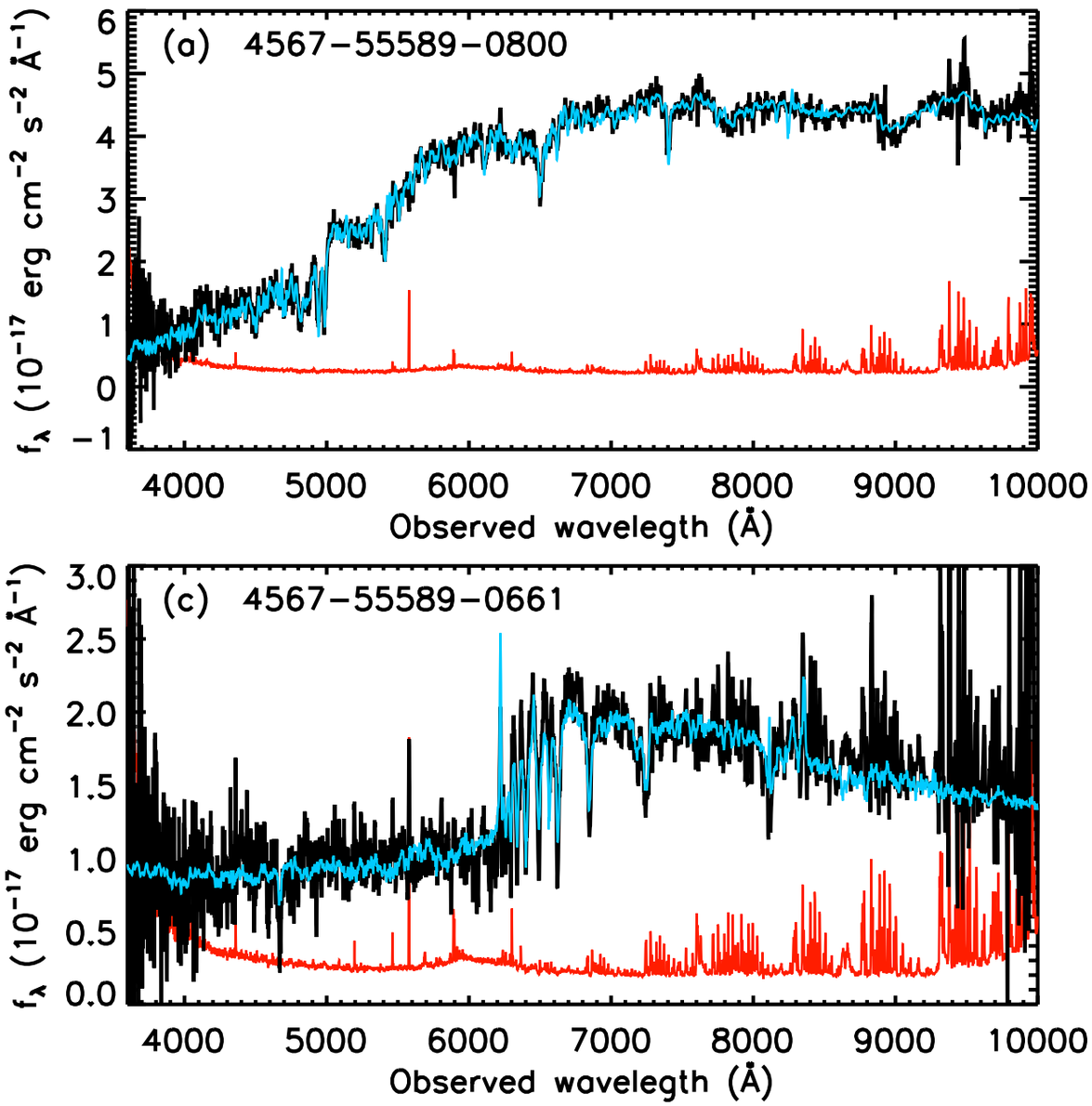
Schematic illustration of the algorithm



Forward modeling allows objective model comparison and error estimation

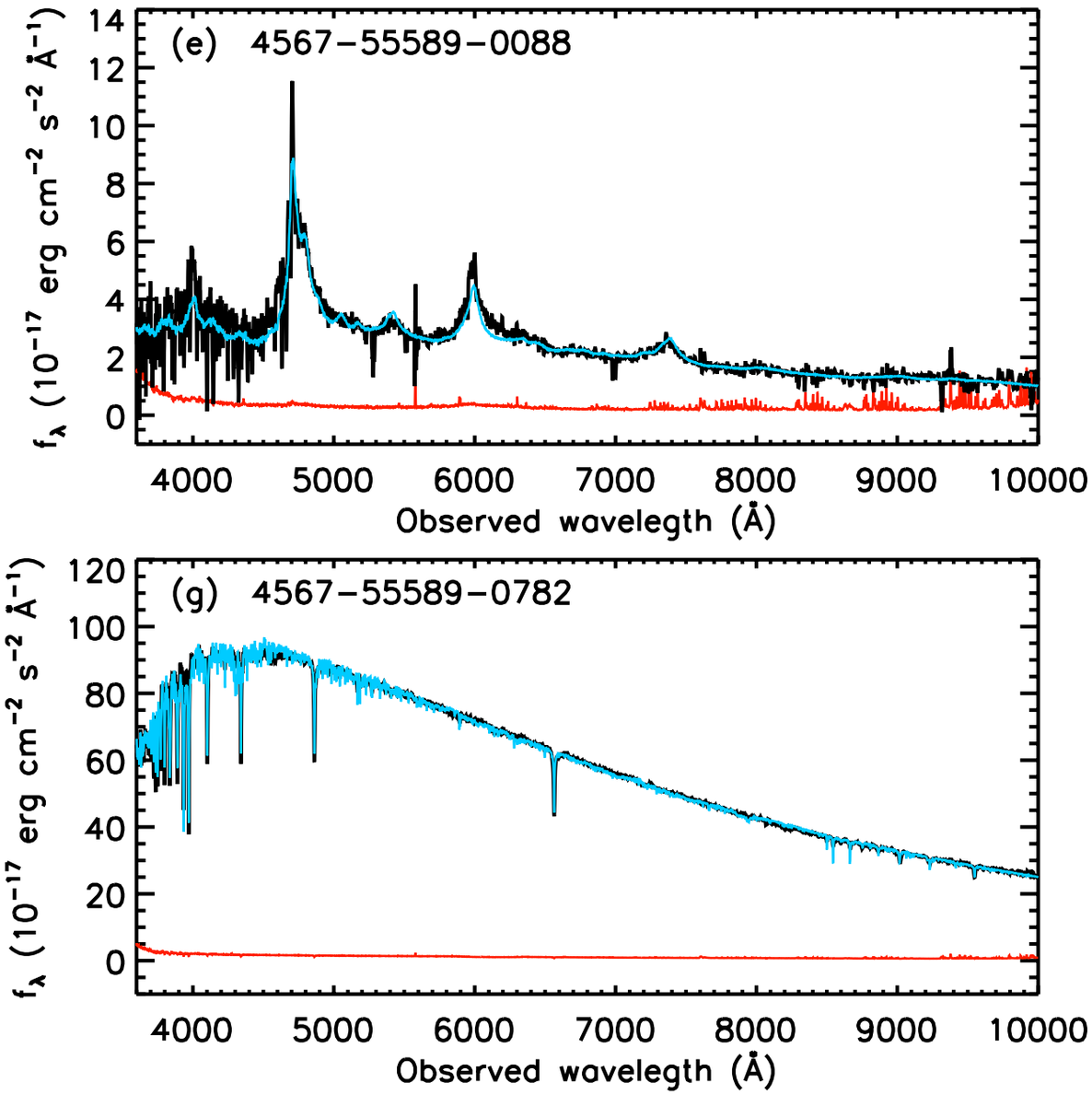
Redshift Measurement and Classification

Example redshift model fits: galaxies



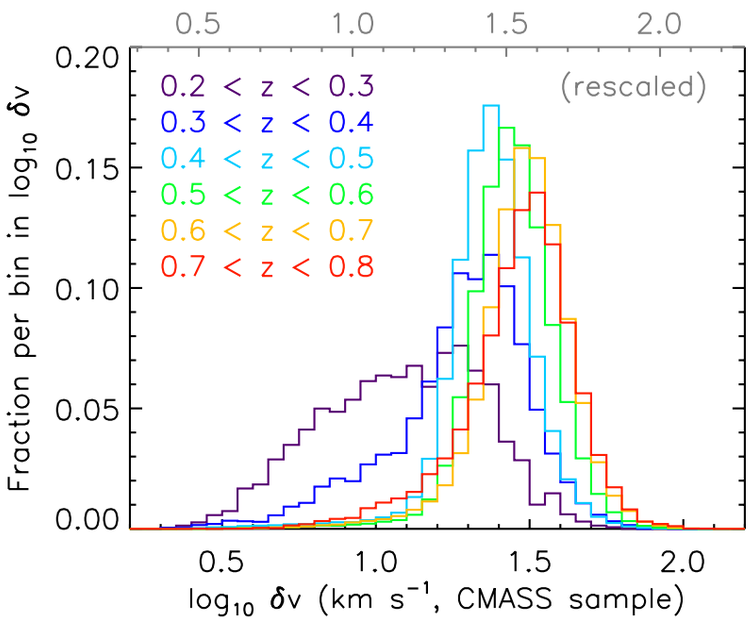
Redshift Measurement and Classification

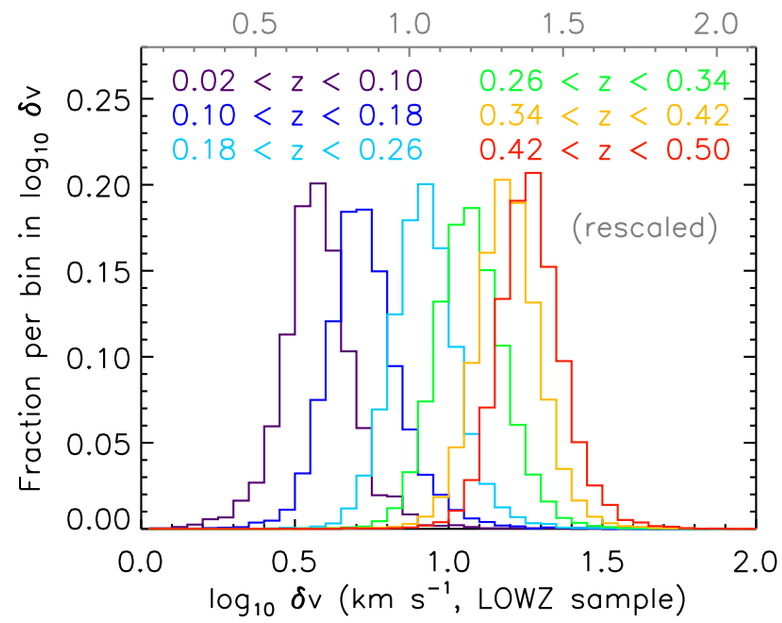
Example redshift model fits: quasars and stars

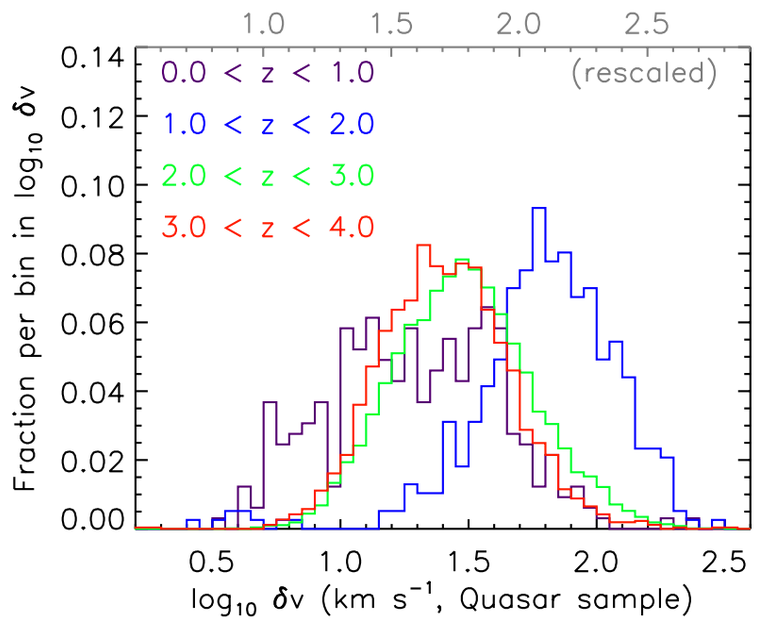


Redshift Measurement Precision

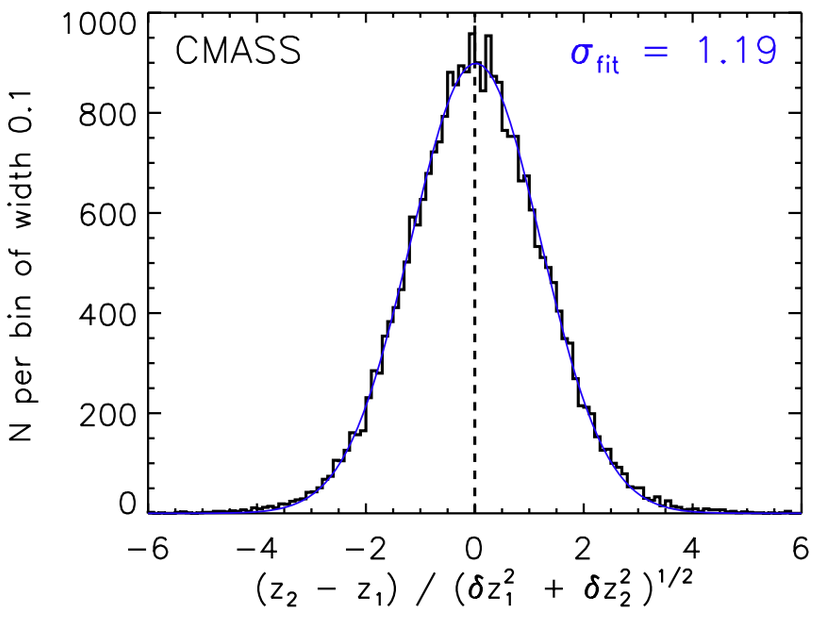
Typical redshift errors: few tens of km/s (statistical):

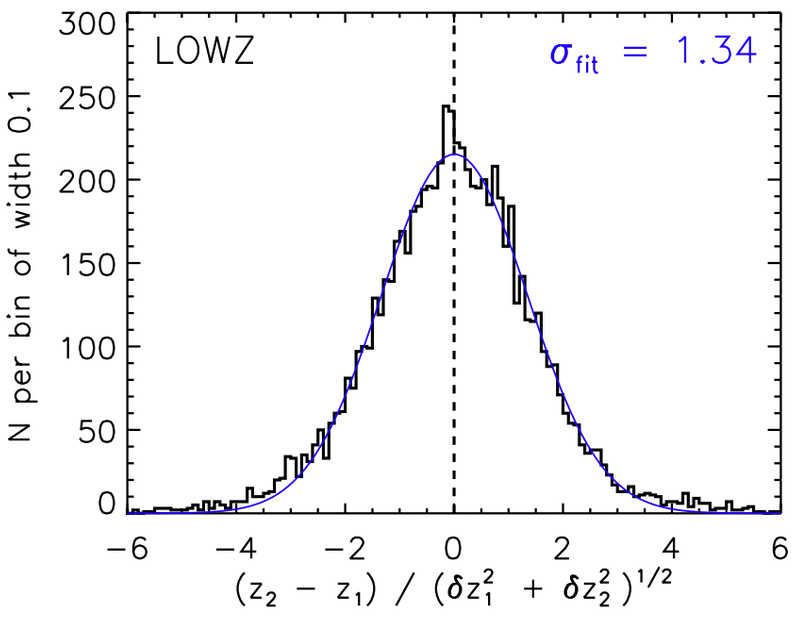


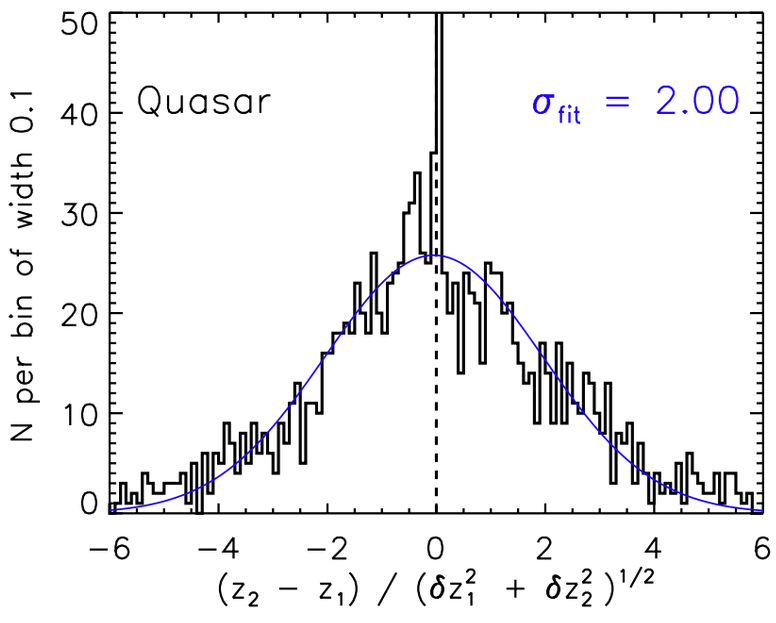




Statistical error estimates substantially accurate:



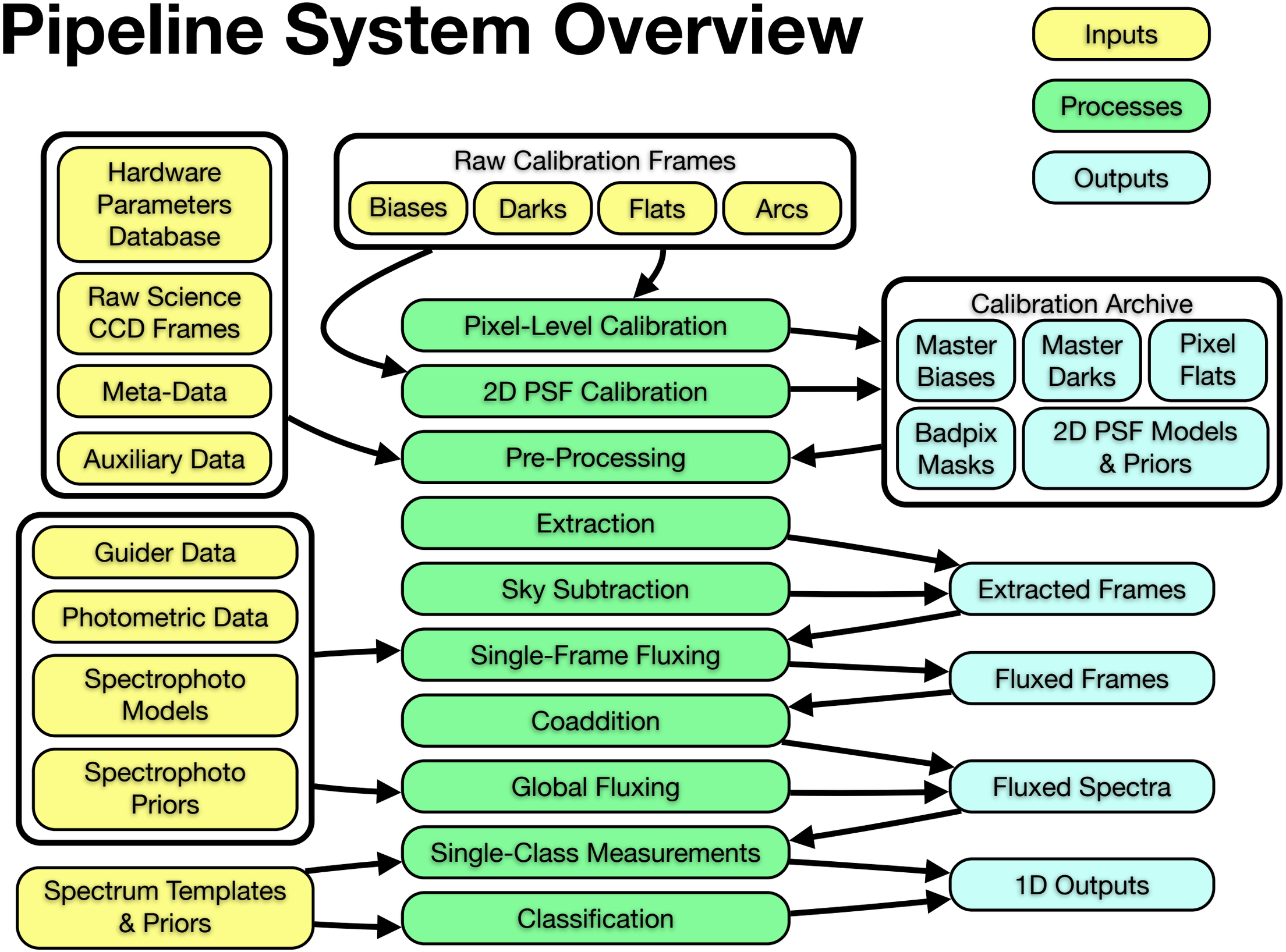




BOSS Pipeline Performance Statistics

Target Sample	CMASS	LOWZ	QSO (2.2 < z < 3.5)
Sample classification success rate	98.7%	99.9%	79%
Target object confirmation rate	95.4%	99.2%	33.6%
Completeness vs. visual inspections	99.6%	100%	98.3%
Purity vs. visual inspections	99.8%	100%	99.4%

Pipeline System Overview



Current and Future Development

- Archetype-based redshift measurement
- More flexible template-class management
- Joint fit to spectroscopic and photometric data
- Unifying redshift and parameter measurement
- Hierarchical Bayesian priors for redshift distribution and physical parameter distributions
- Greater quasar redshift accuracy control
- Performance analysis and optimization for emission-line galaxy (ELG) target class
- Next-generation extraction algorithms