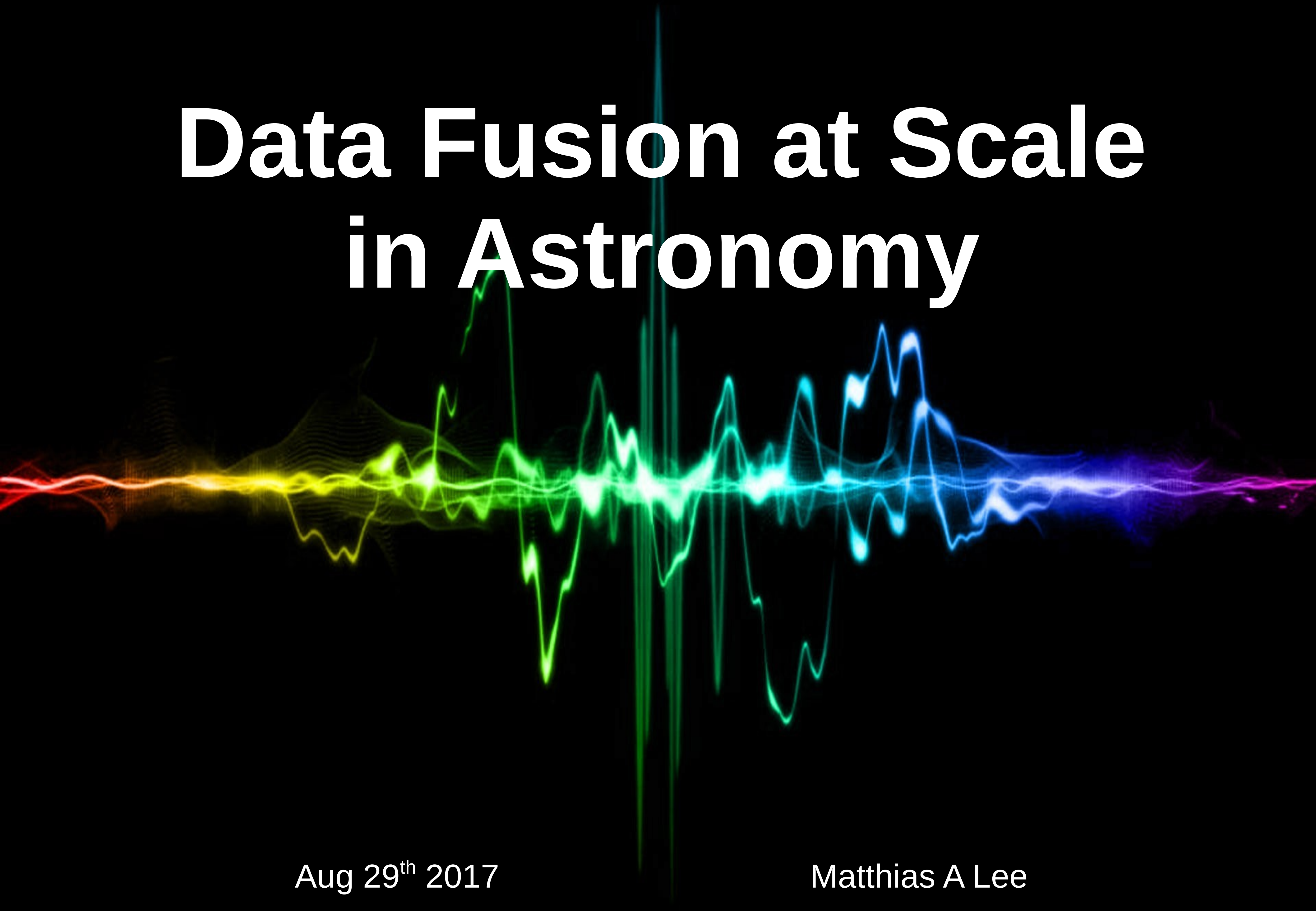
Data Fusion at Scale in Astronomy



Aug 29th 2017

Matthias A Lee

Lots of Data, Many Sources

- World is exploding with data
- Data collection grows with Moore's law
  - Over 100 PB of astronomy data in 10 years
    - Pioneer in big data
    - Sloan Digital Sky Survey in house 35 TB
    - Large Synoptic Survey Telescope expected 60 PB[1]
  - Big data everywhere
    - Sciences: genomics, particle physics, astronomy
    - Industry: finance, social media, machine learning

[1] LSST Data Management: https://www.lsst.org/about/dm

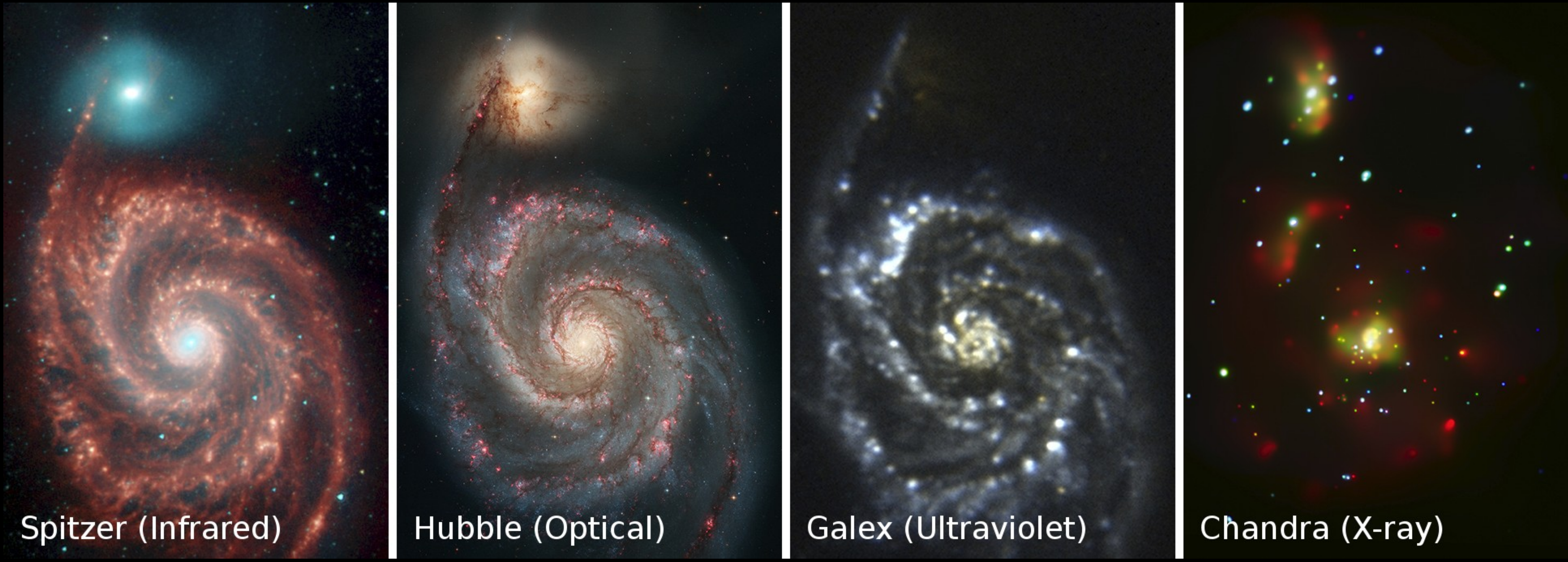
Making sense of it All

- Biggest challenges, extracting meaning
- Difficult to process, transfer and store
- Need new methods
  - Deal with lots of data
  - Scalable
  - Combine data, extract meaning.
- Parallel to Scale
  - GPUs and Clusters
  - Amdahl’s Law

Fusing Data at Scale

- Cross-Matching Astronomy Catalogs
- Optimal Image Coaddition
  - Sharpening Images, Doubling resolution
- Extracting Color from monochrome images

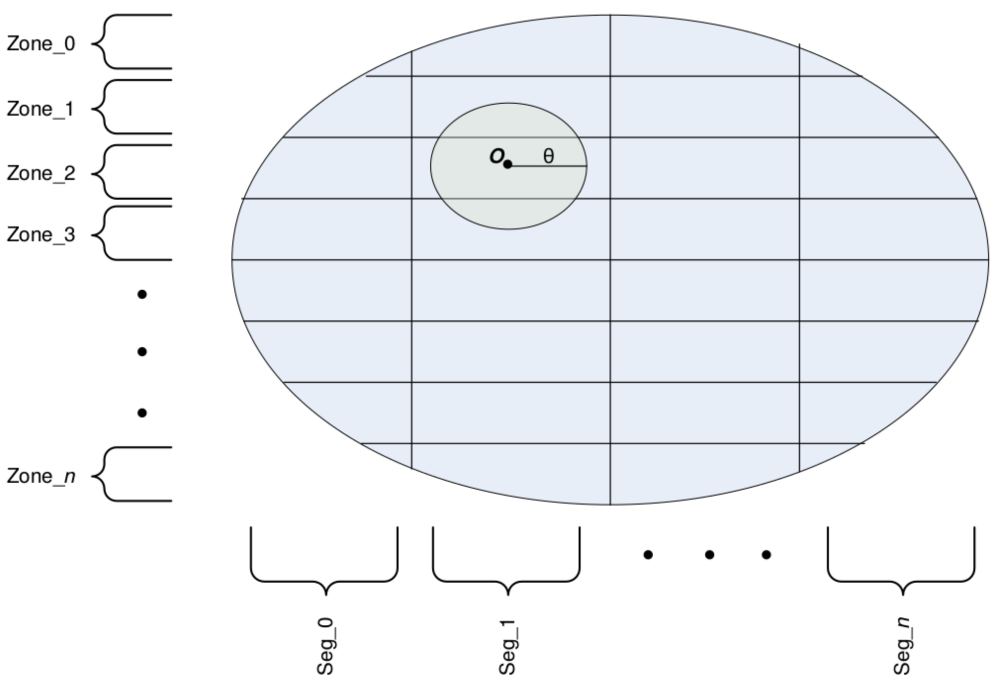
Fusing Astronomy Catalogs



Thesis Work

Divide, Conquer & Parallelize

- Pair of objects
  - Compute simple distance metric
- Challenge is in efficient Parallelism
  - Segments
  - Sorting/Thrust
  - Worker Jobs
  - Zones Algorithm[1]
  - Multi-GPU

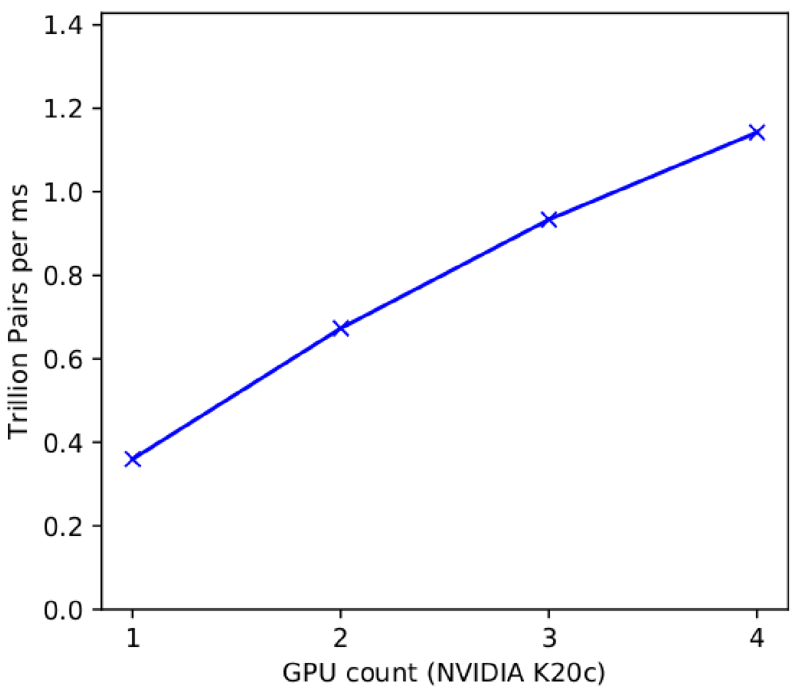


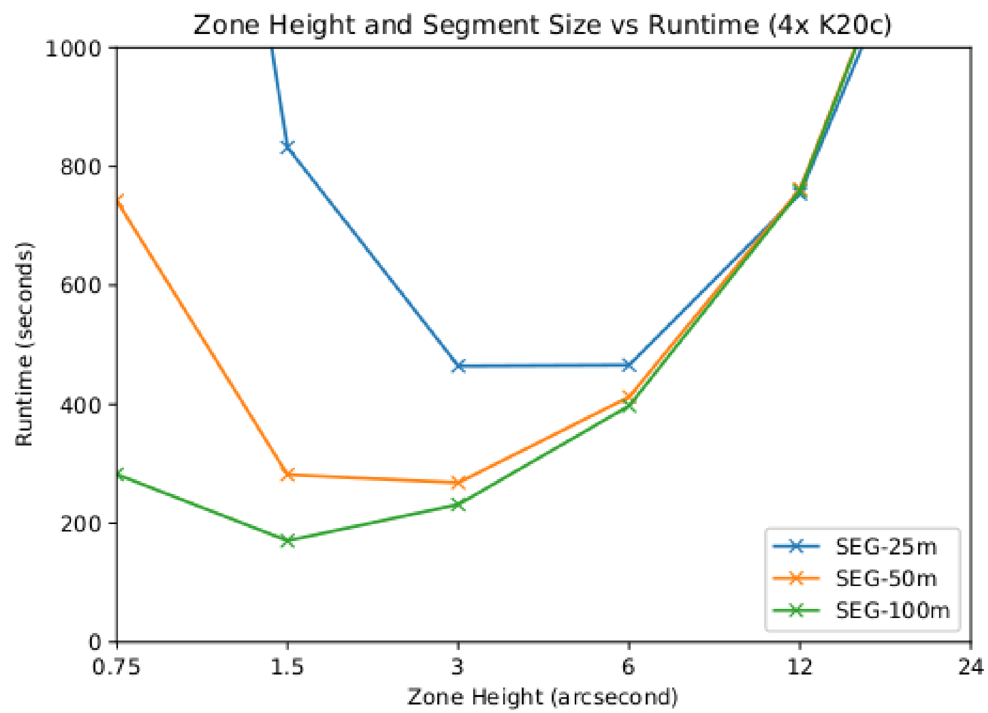
[1] "There Goes the Neighborhood: Relational Algebra for Spatial Data Search" Gray, Szalay, Fekete (2004)

Unprecedented Speed

Thesis Work

- Optimization yields high matching performance
  - 50M × 150M in 50 seconds
  - 450M × 450M in 3 minutes vs 45mins !





"Faster catalog matching on Graphics Processing Units" M. Lee, T. Budavári (2017)

Catalog Crossmatch

- Scientifically Important Problem
- Scalable Acceleration
  - Parallelized across multi-GPU
- 15x faster vs State of the Art
  - 3 minutes vs 45 minutes

Observing the Sky

SDSS telescope at night
by Patrick Galume

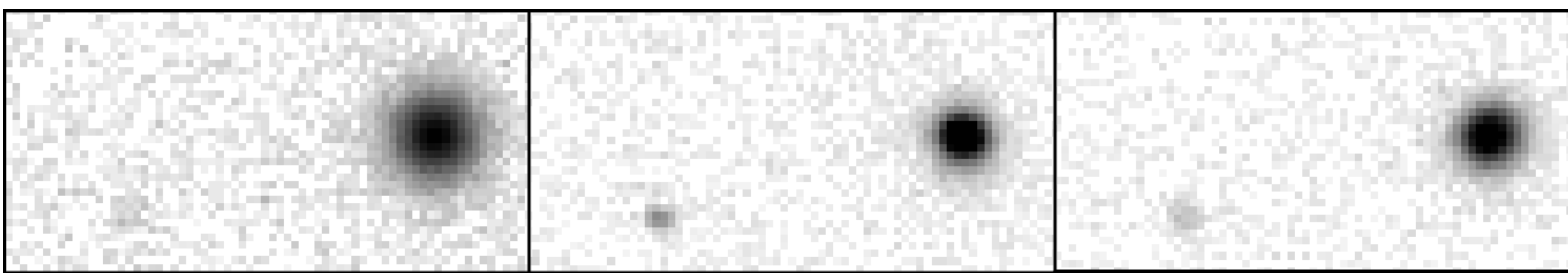
Multiple Exposures

- Sloan Digital Sky Survey (SDSS)
  - Stripe 82 has 70x coverage
- Large Synoptic Survey Telescope (LSST)
  - Will have 200x coverage

SDSS FRAMES

Traditional Solutions

- Lucky Imaging
  - Keep only the best/sharpest 1% of the images
- Coadding
  - Higher Signal-to-Noise Ratio
  - Worst acceptable blur (PSF)

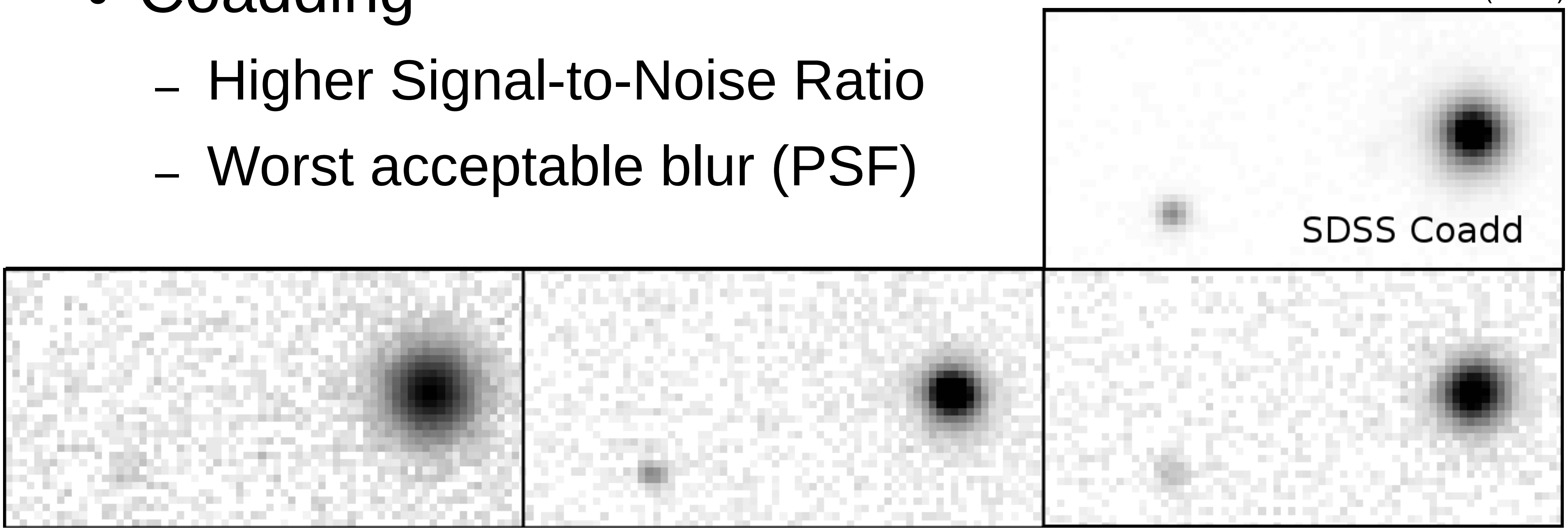


SDSS FRAMES

Traditional Solutions

- Lucky Imaging
  - Keep only the best/sharpest 1% of the images
- Coadding
  - Higher Signal-to-Noise Ratio
  - Worst acceptable blur (PSF)

Annis et al. (2011)



Next-Generation Processing

- Traditional methods are sub-optimal
  - Naive assumptions yield wrong results
- Computational Optics
  - Best possible signal-to-noise ratio
  - Sharper & deeper images
  - Higher resolution

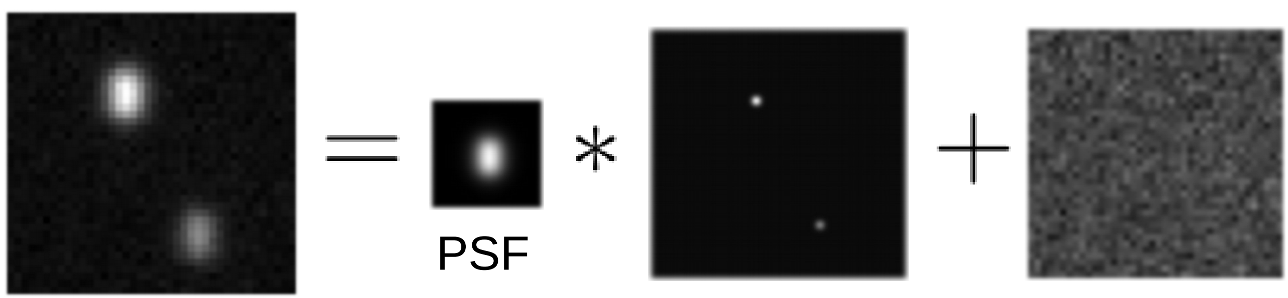
Computational Optics

- Single frame solutions before
  - Correcting Hubble optics
  - Classic Richardson-Lucy deconvolution
    - White (1994), Starck+(1994), Fish+ (1995)
  - Limited by information in single frame
- Multiple frames provide new opportunities
  - Breaks degeneracy of PSF and the “true” image

Linear Model for Exposures

- “true” image convolved with unknown PSF
- Plus some noise

$$y_t = f_t * x + \epsilon_t$$



- Solve for *x*

  … and f

Streaming Deconvolution

- We solve for the underlying “true” image
- Gaussian likelihood function yields quadratic minimization

$$|y_t - Fx_t|^2$$

- Multiplicative updates
  - cf. Richardson-Lucy

$$x_{t+1} = x_t \odot \frac{F^T y_t}{F^T F x_t}$$

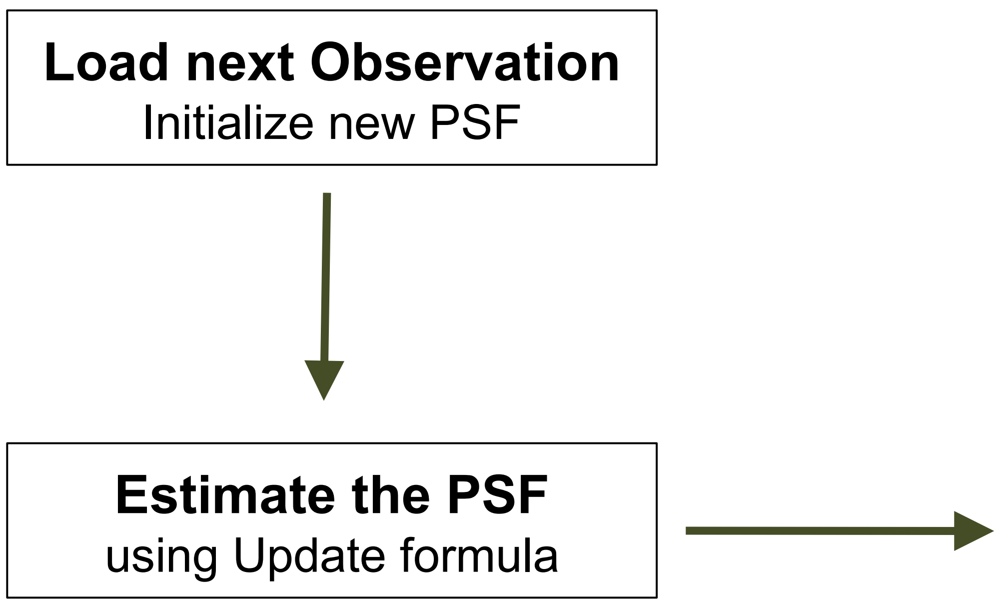
Multi-Frame Blind Deconvolution

- Iterative approach:

Load next Observation
Initialize new PSF

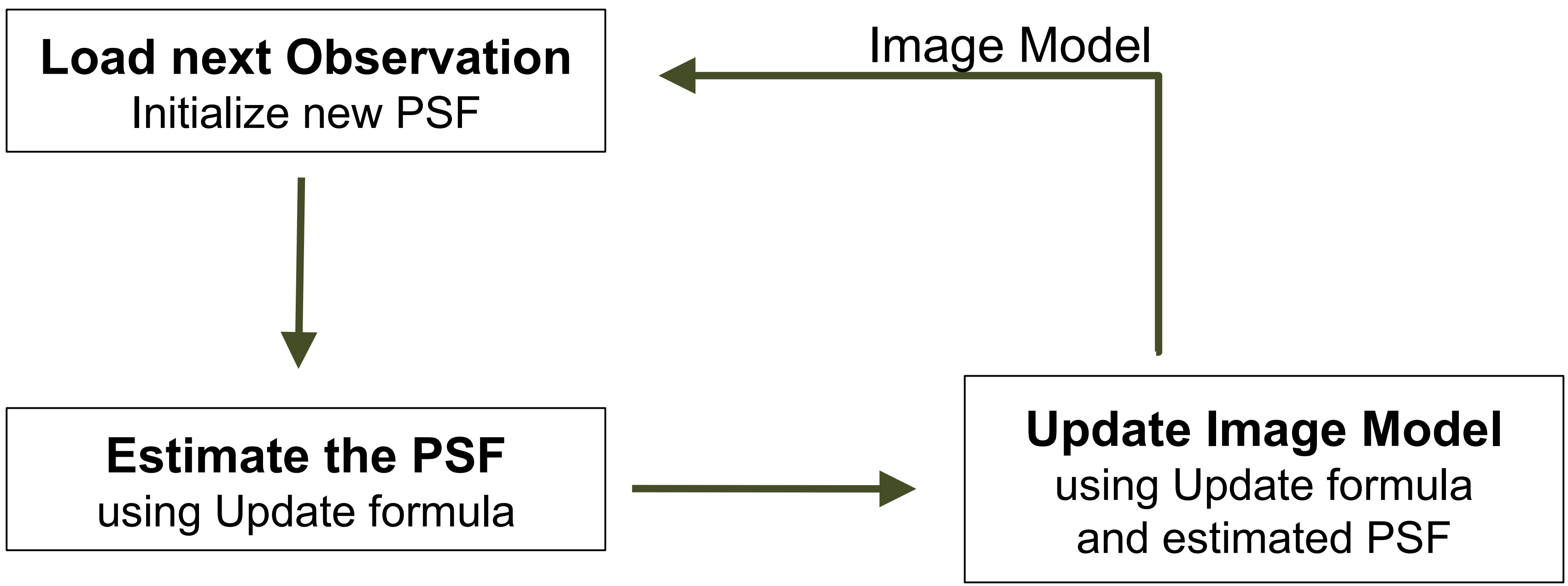
Multi-Frame Blind Deconvolution

- Iterative approach:



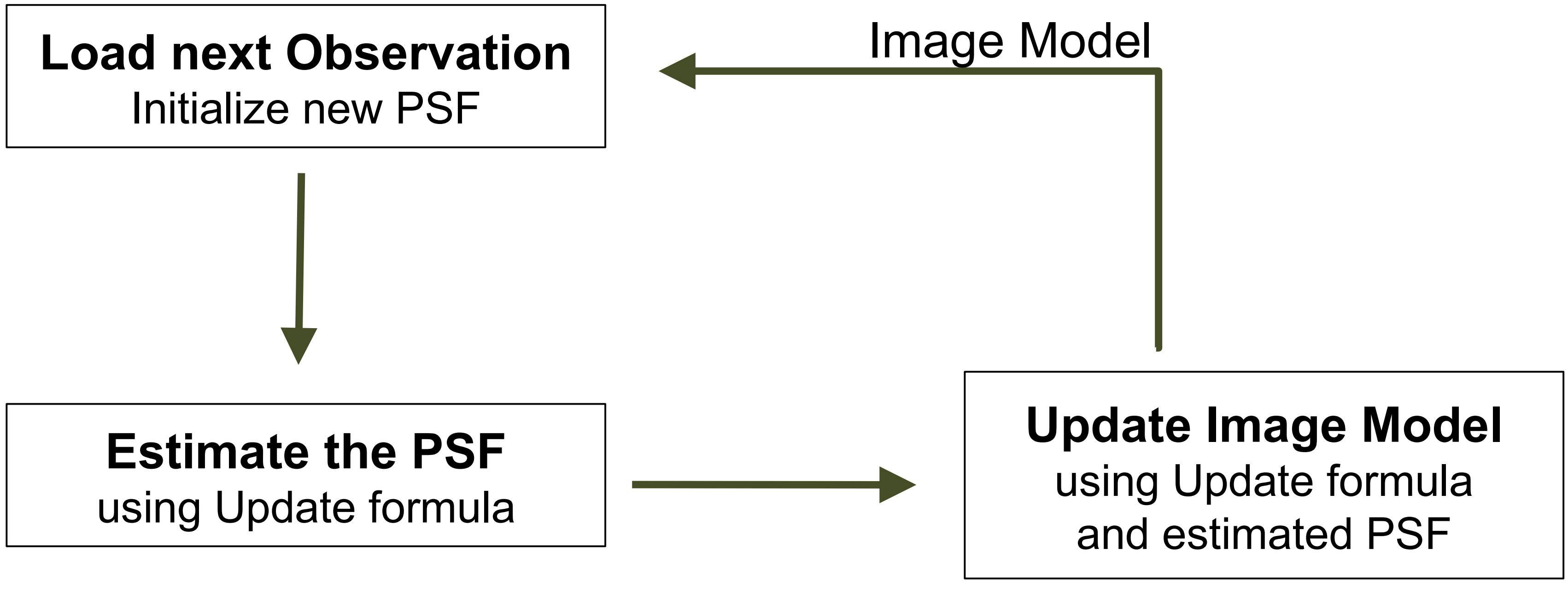
Multi-Frame Blind Deconvolution

- Iterative approach:



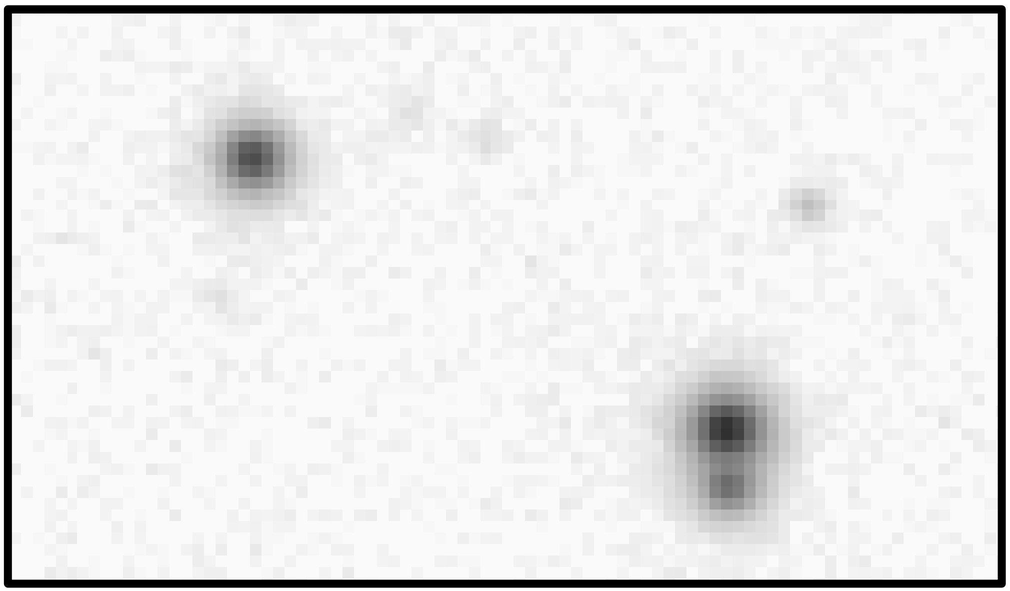
Multi-Frame Blind Deconvolution

- Iterative approach:

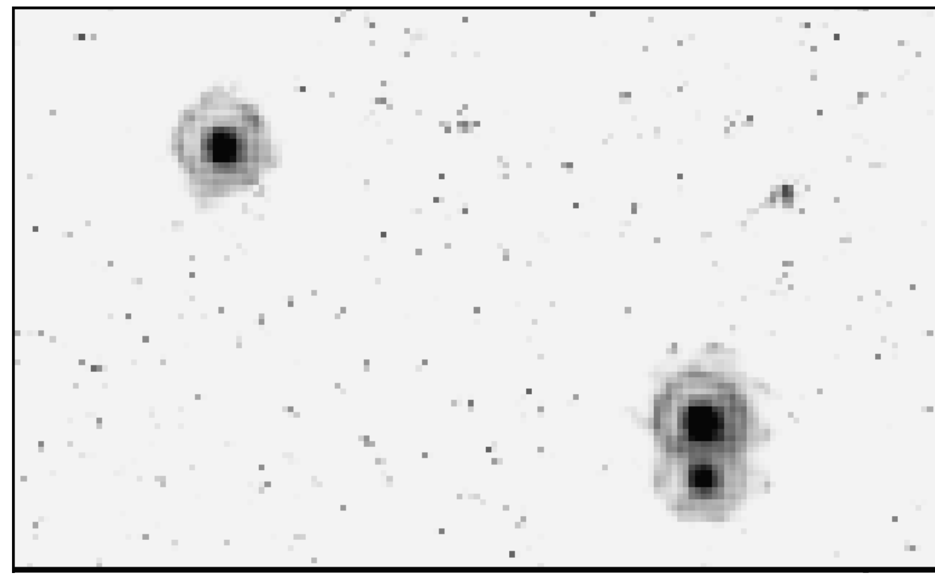


The devil is in the details!

Textbook Deconvolution

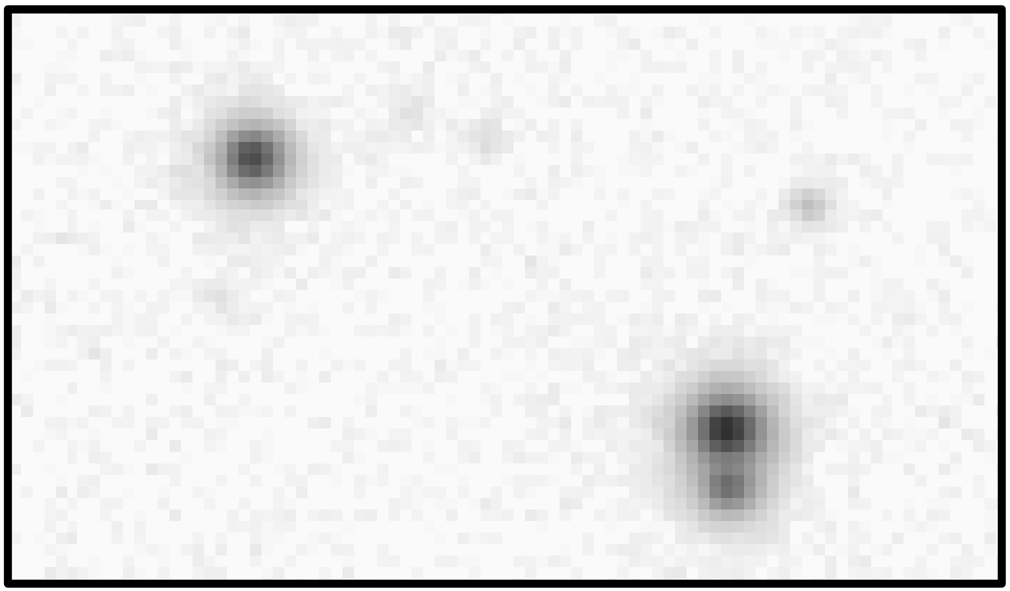


Annis Coadd (2011)

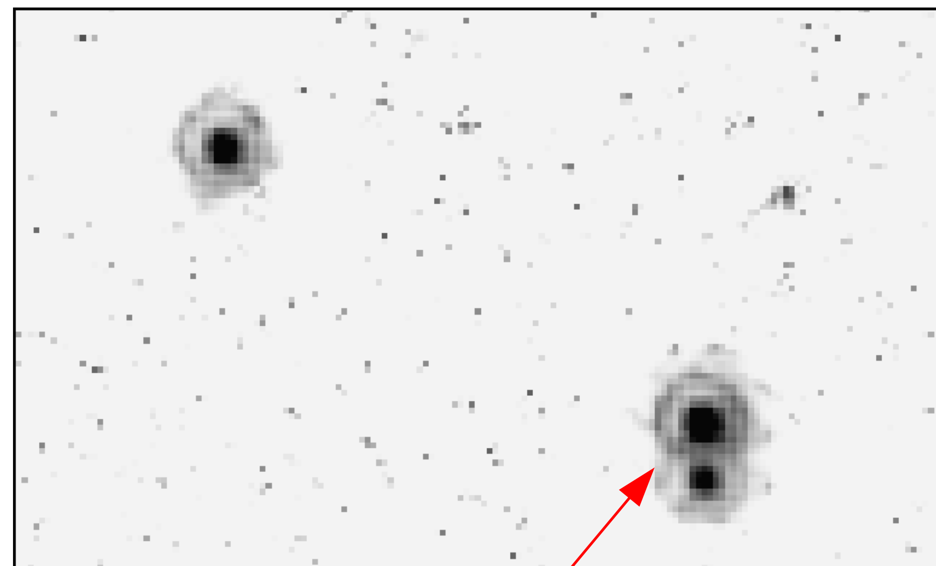


unmodified deconv

Textbook Deconvolution



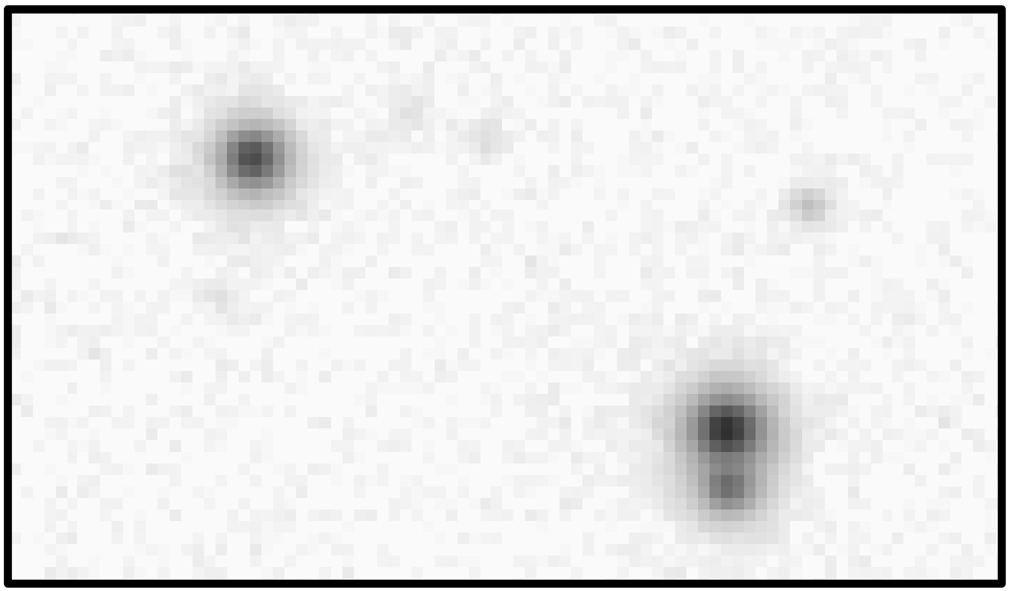
Annis Coadd (2011)



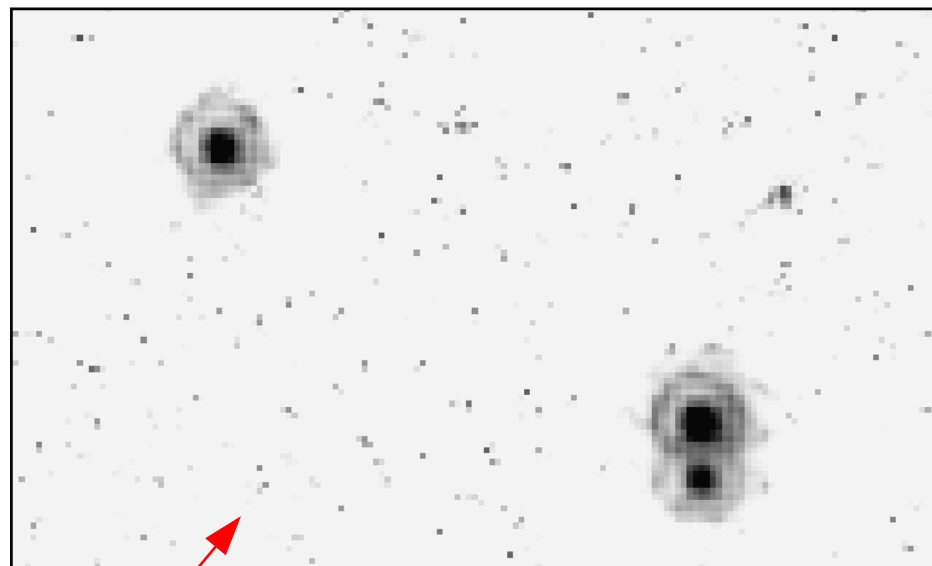
unmodified deconv

- Ringing artifacts

Textbook Deconvolution



Annis Coadd (2011)

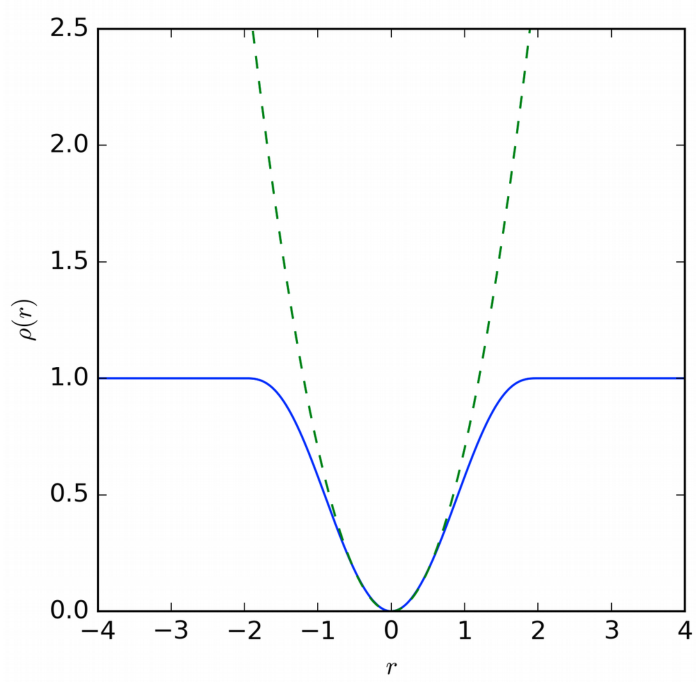


unmodified deconv

- Ringing artifacts
- Speckled background

Robust Statistics

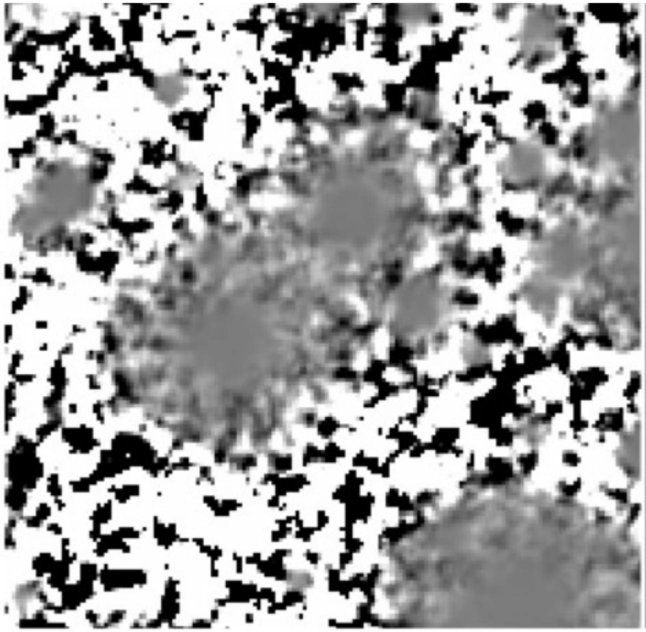
- Quadratic cost function dominated by bad pixels
  - Poor convergence across image
- Apply Robust $\rho(r)$
  - Quadratic for small residuals
  - Down-weights large
- Iterative re-weighting
  - Integrate with streaming



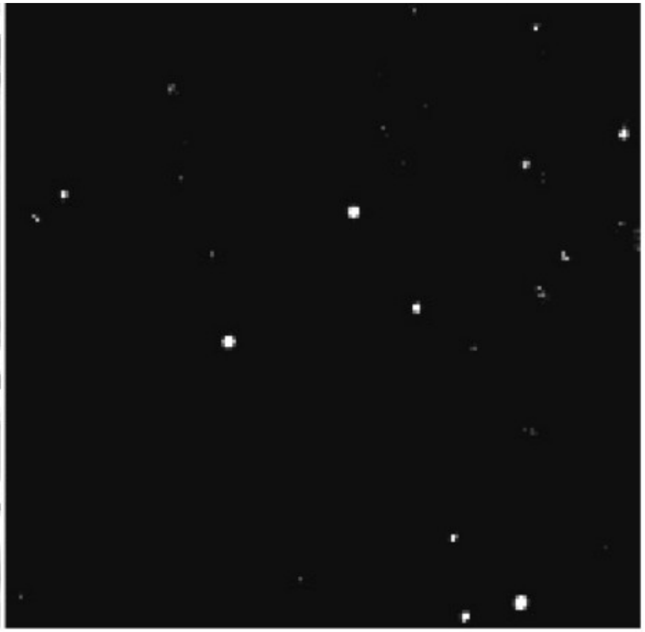
Thesis Work

Careful Updates

- Artifacts from nowhere
  - Large updates of tiny values
- Limit the influence of updates
  - E.g., no more than 2x



Update Image

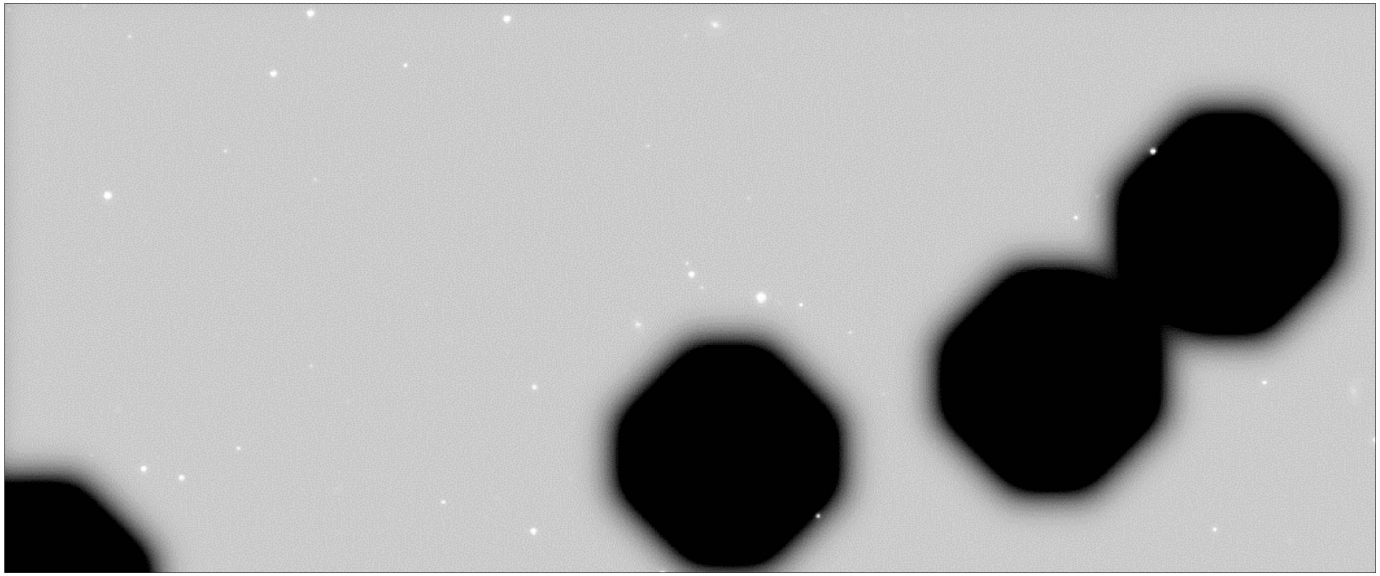


Model Image

Thesis Work

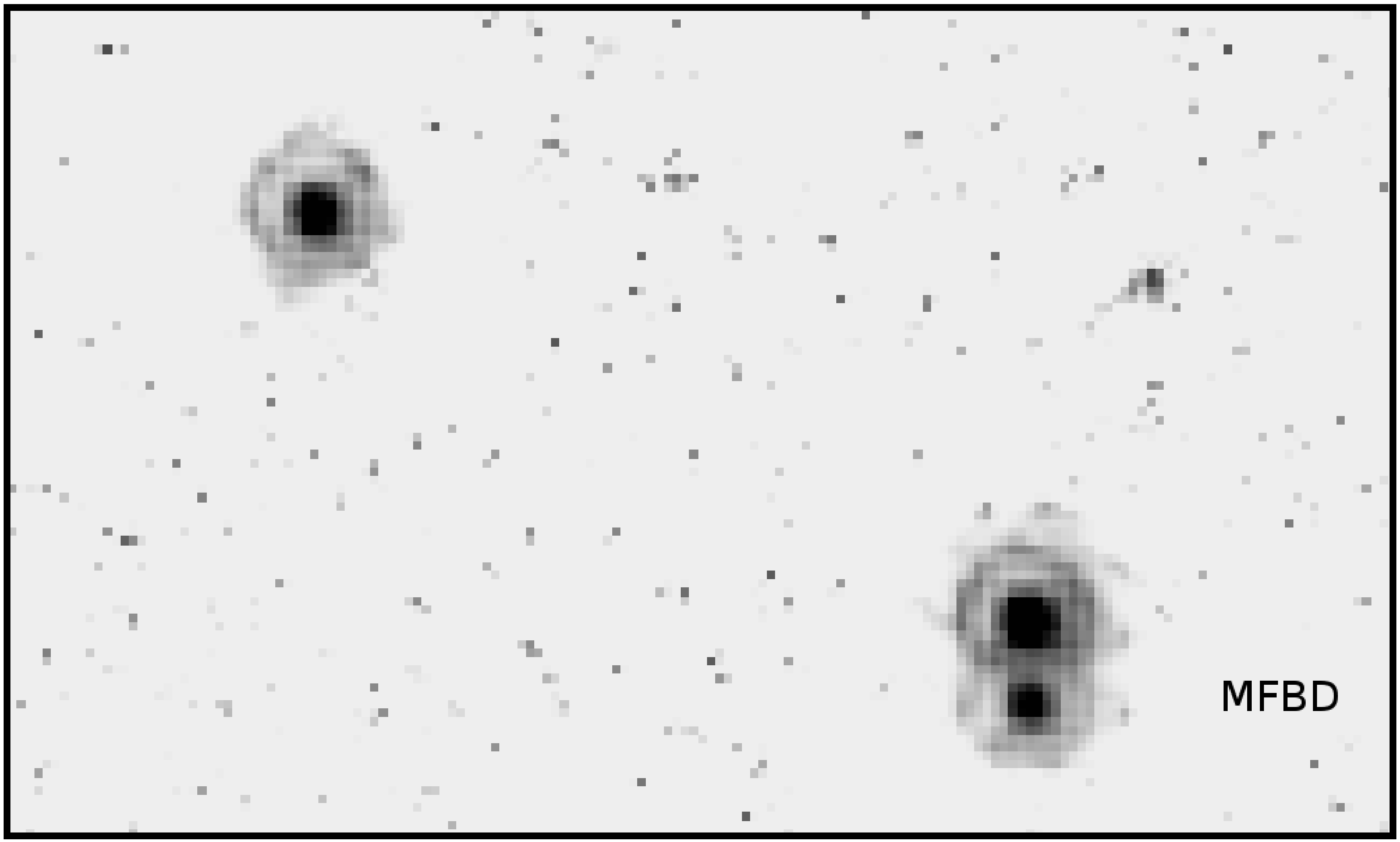
Masking Pixels

- Ignore gaps as well as bad or saturated areas

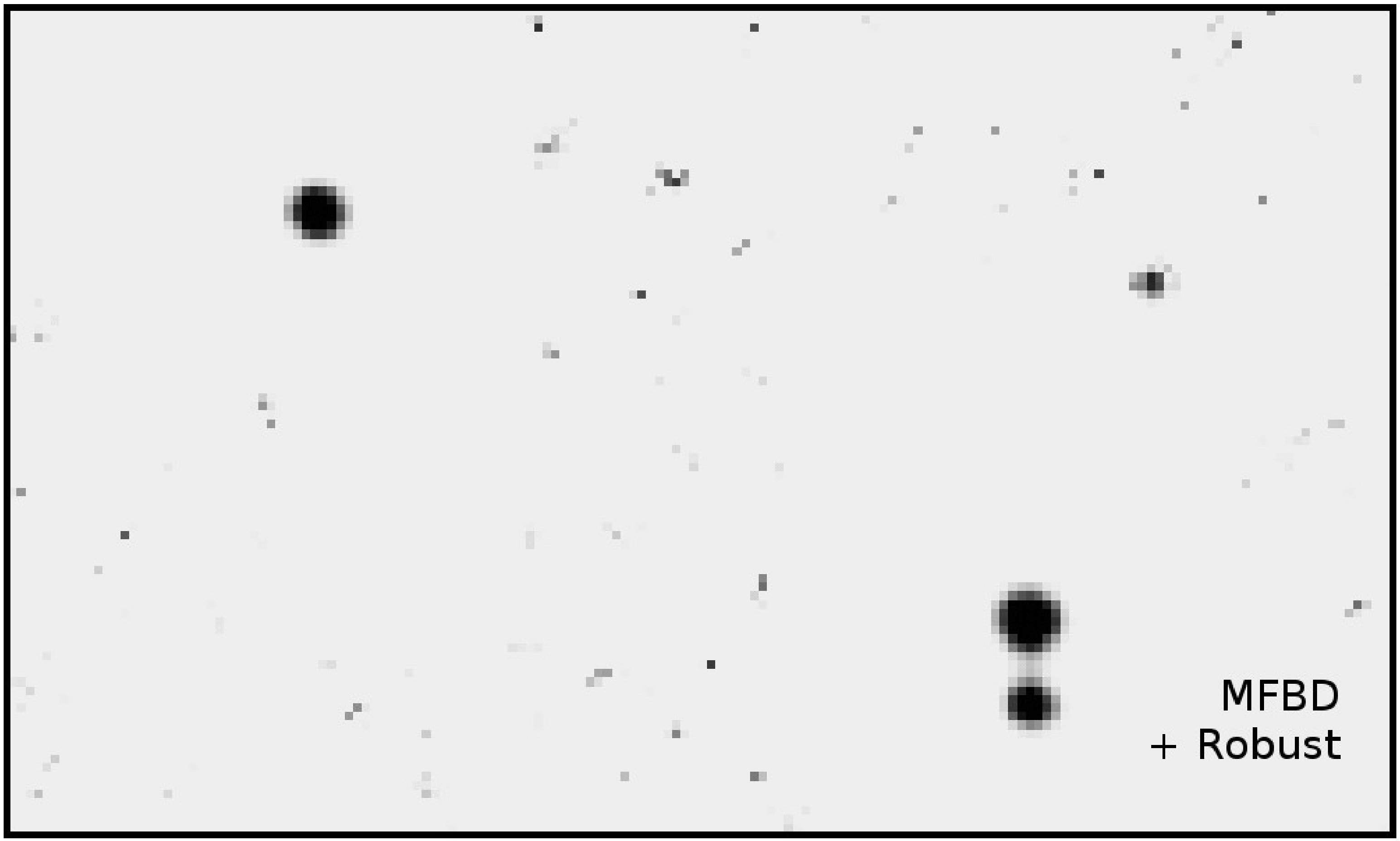


- But we also solve for the missing areas!

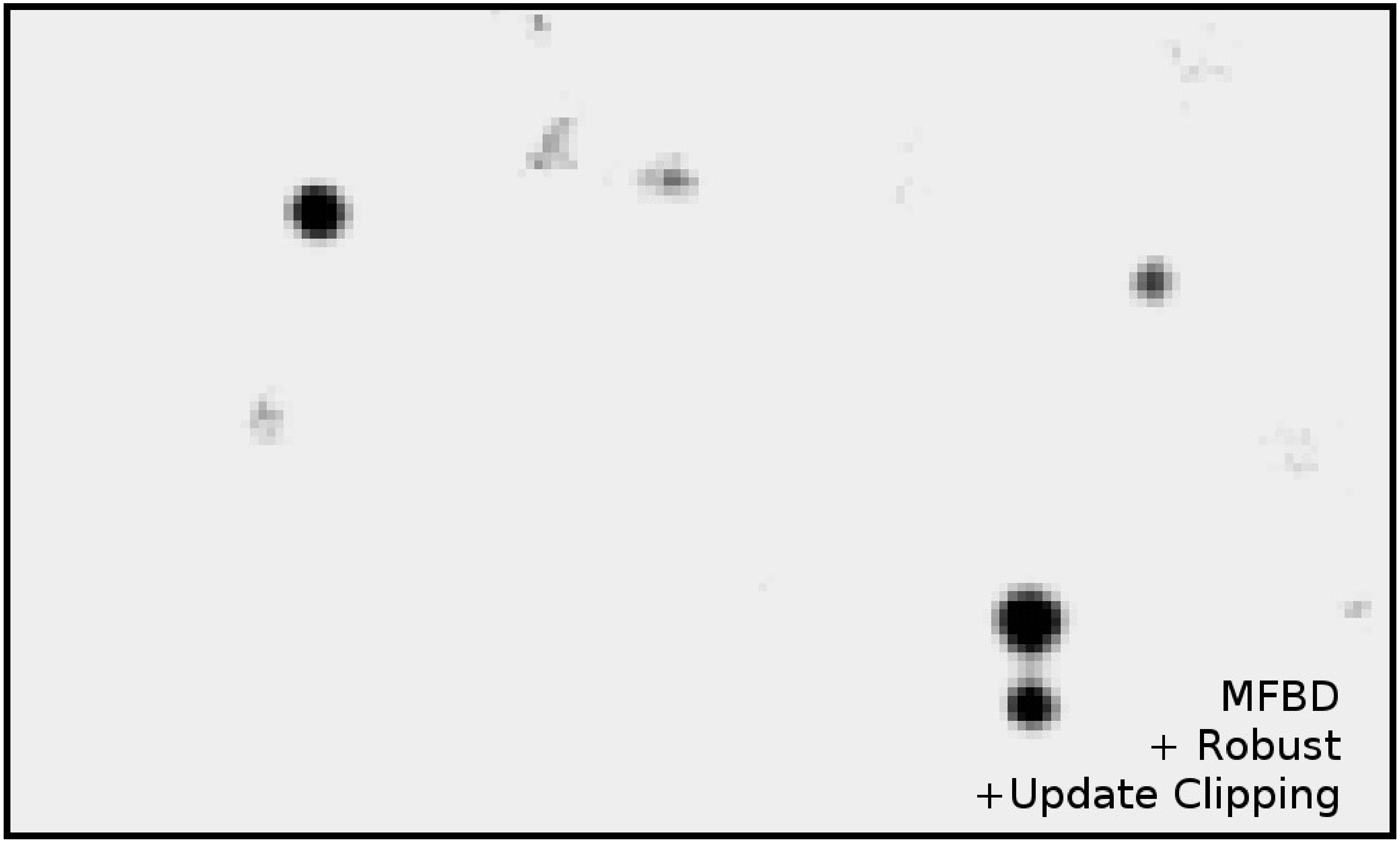
Enhance!



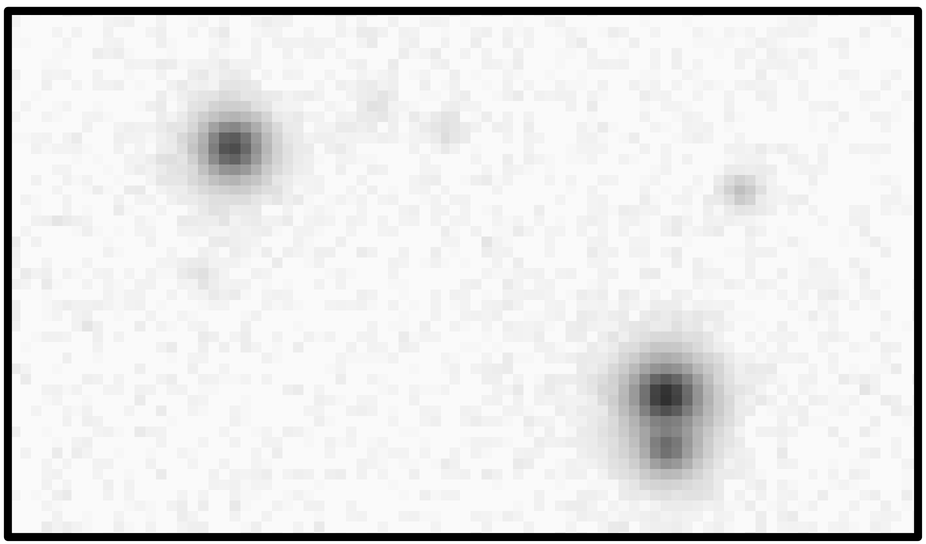
Enhance!



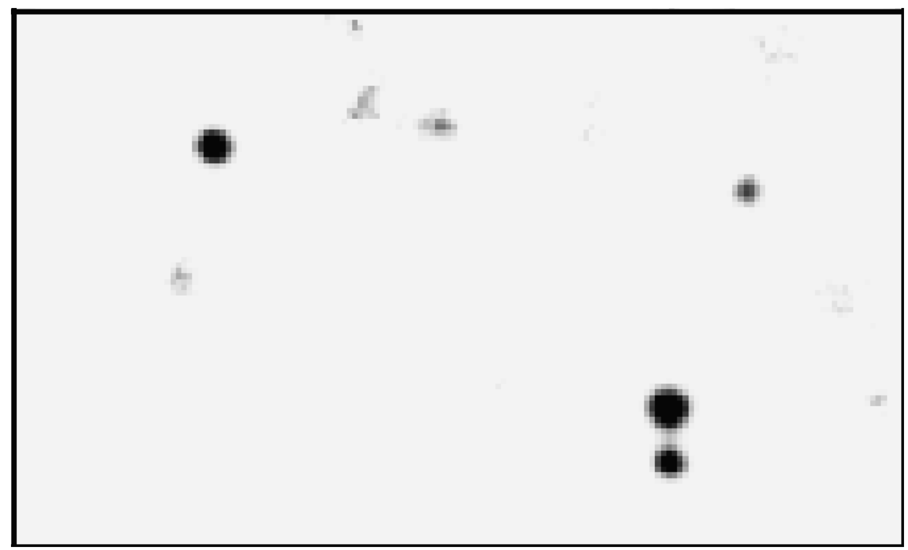
Enhance!



Enhance!



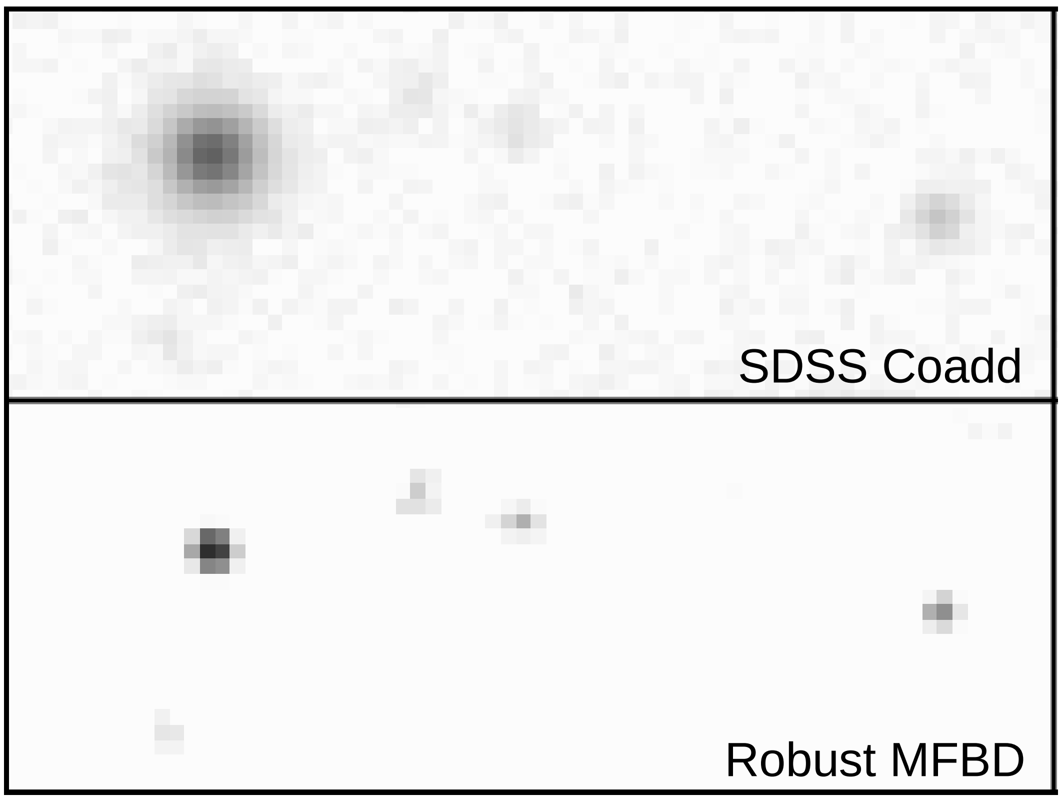
Annis Coadd (2011)



MFBD
+ Robust
+ Update Clipping

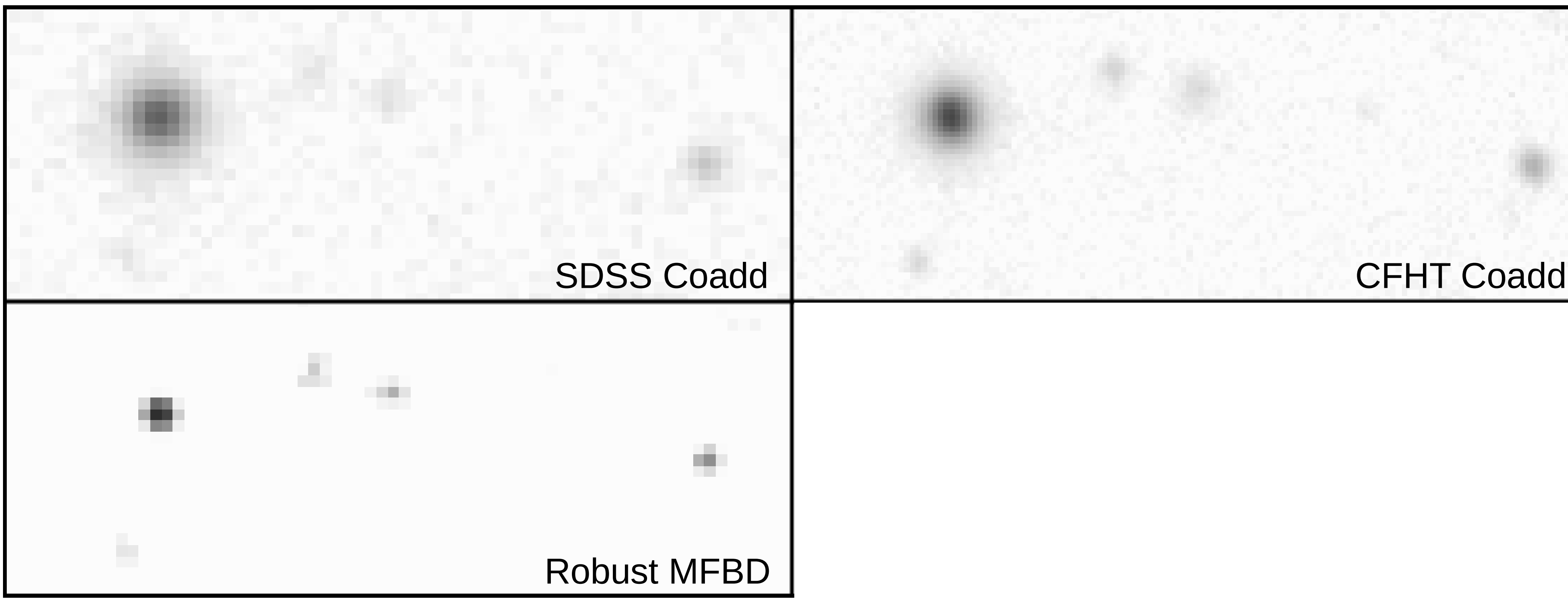
Super Resolution

- Upscale and Downscale Operator



Super Resolution

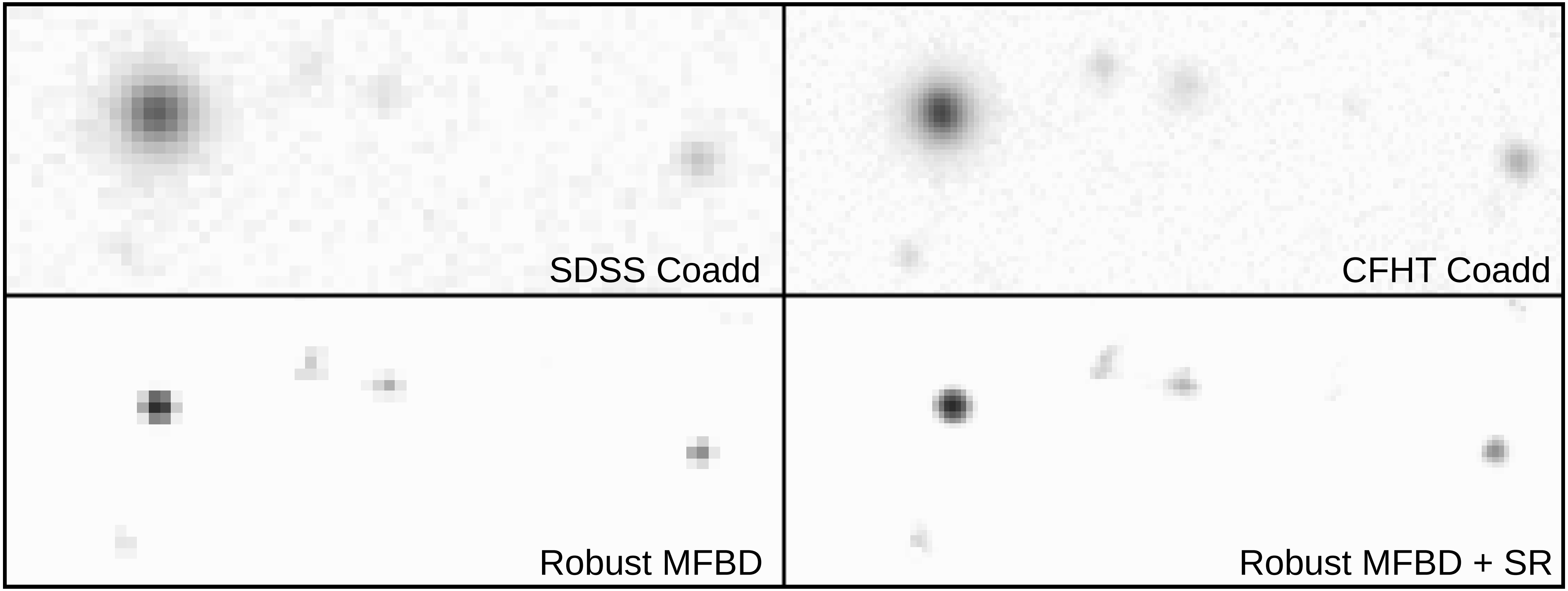
- Upscale and Downscale Operator
- CFHTLS ~ 2x resolution of SDSS



Thesis Work

Super Resolution

- Upscale and Downscale Operator
- CFHTLS ~ 2x resolution of SDSS



Thesis Work

Performance

- Performance is important!
  - GPU-accelerated using pyCUDA
  - 140 images 2k by 2k: < 5 min
  - + Super Resolution, 4k by 4k: ~ 10 min

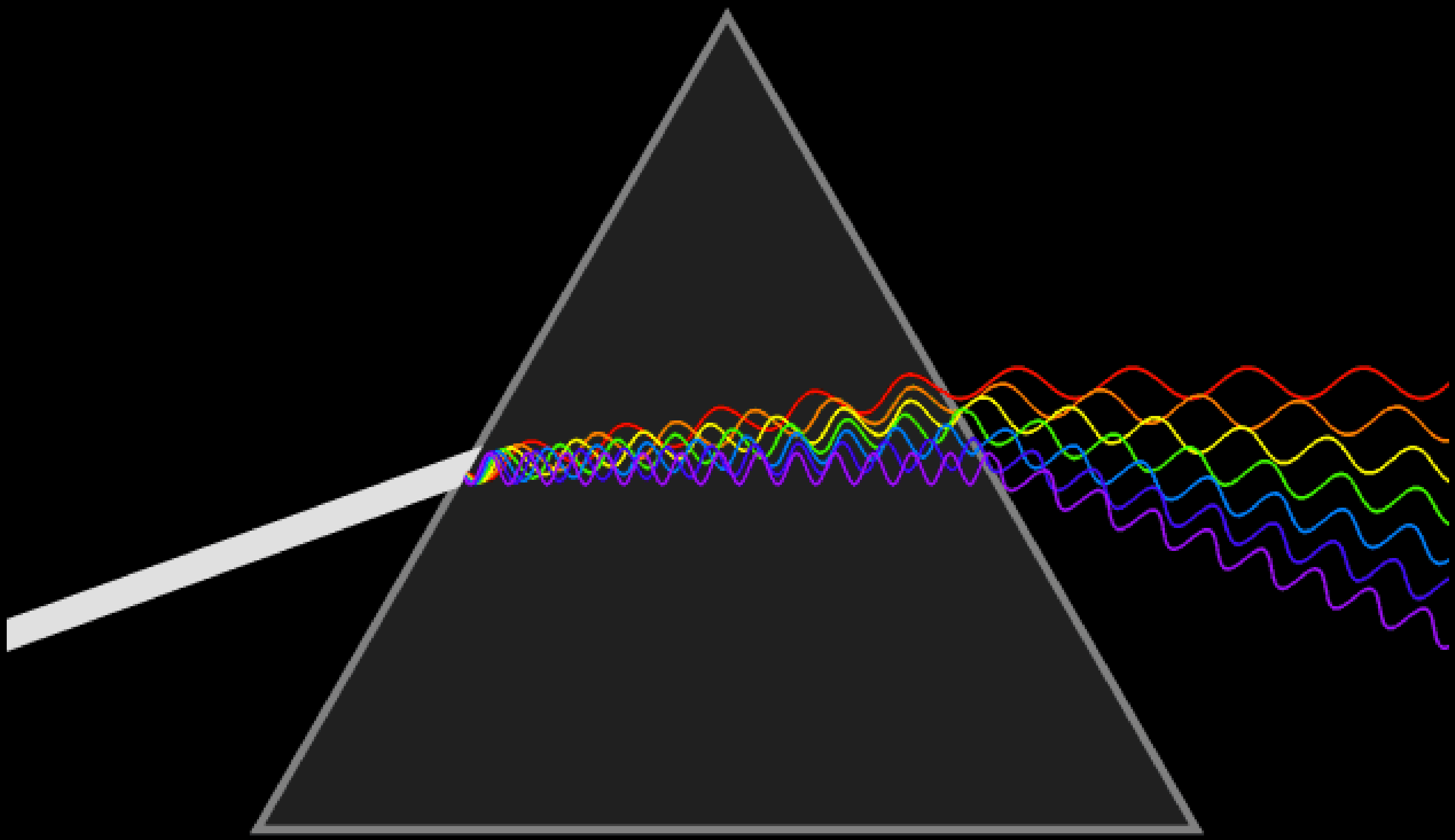
- Python, fast prototyping for experimentation
- Built Pipeline for processing Survey on MARCC

Beyond the Optimal Coadd

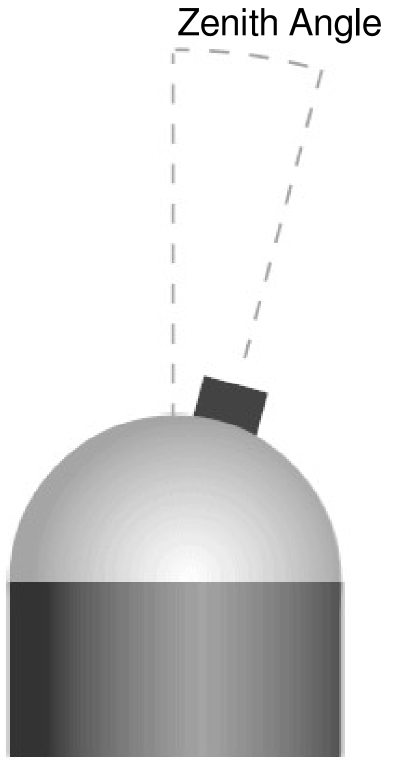
- Combine multiple images
  - Increase quality: SNR
  - Increase clarity: deblurring
  - Increase resolution: Super Resolution

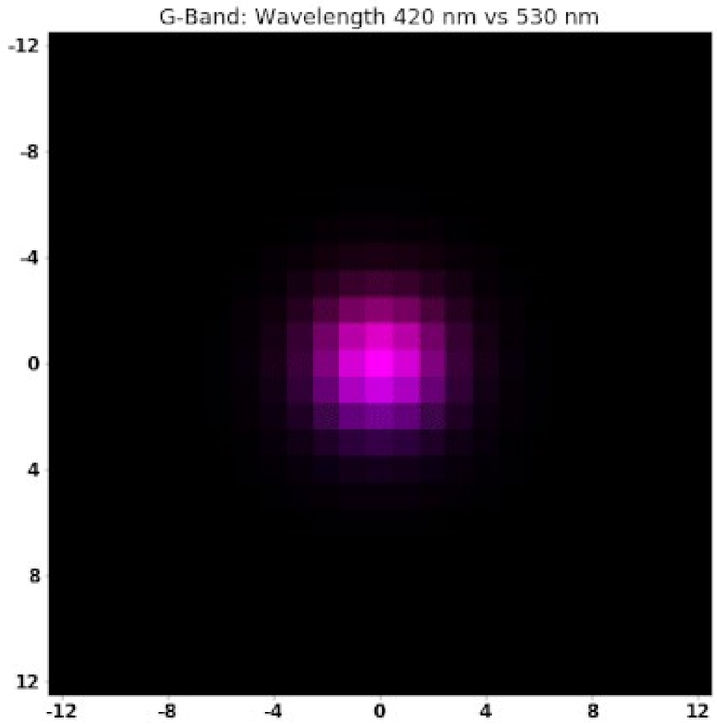
- Color estimation

Differential Chromatic Refraction

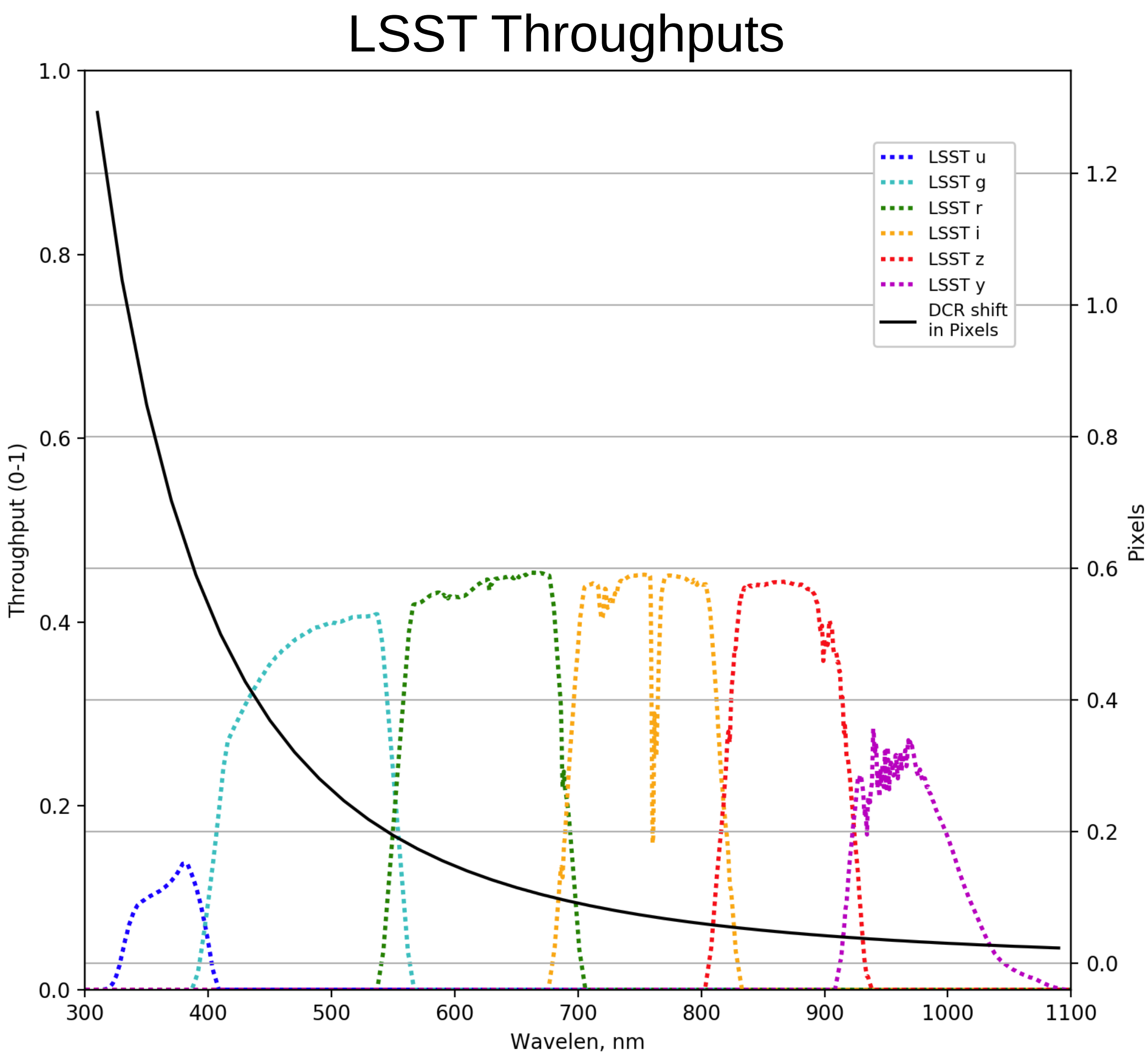


Differential Chromatic Refraction





Differential Chromatic Refraction



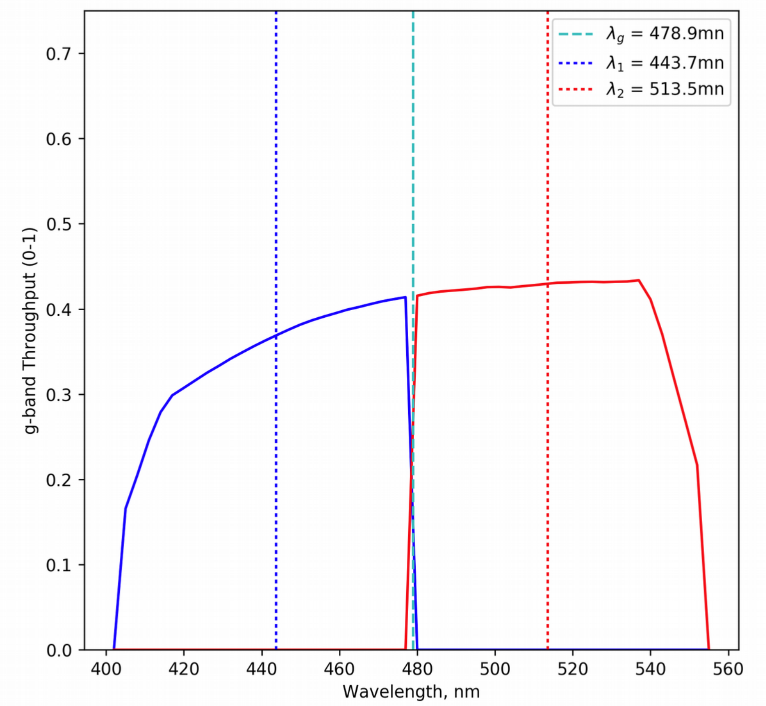
Thesis Work

DCR: Multiple PSFs

- Old model:

$$y_t = f_t * x + \epsilon_t$$

- Generalized model:

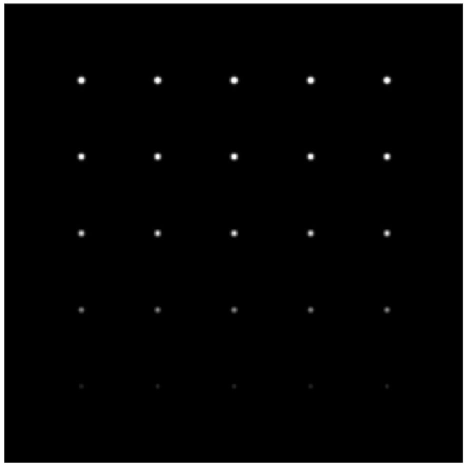


$$y_t = \underbrace{f'_t * x'}_{\textit{subband 1}} + \underbrace{f''_t * x''}_{\textit{subband 2}} + \epsilon_t$$

Thesis Work

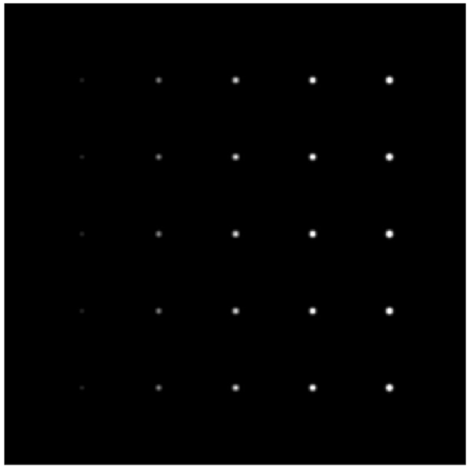
Simulate Observations

- LSST StarFast Simulator
- 2 Discrete wavelengths
- PSF of varying Zenith angle



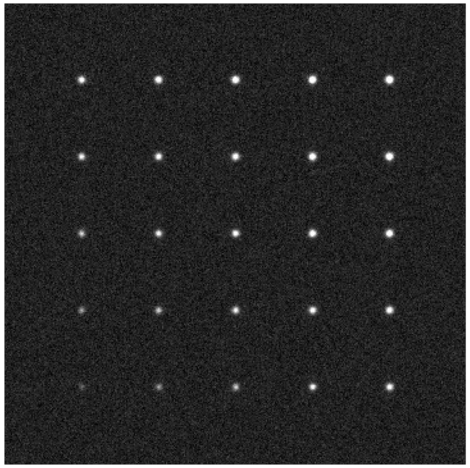
subband 1

$+$

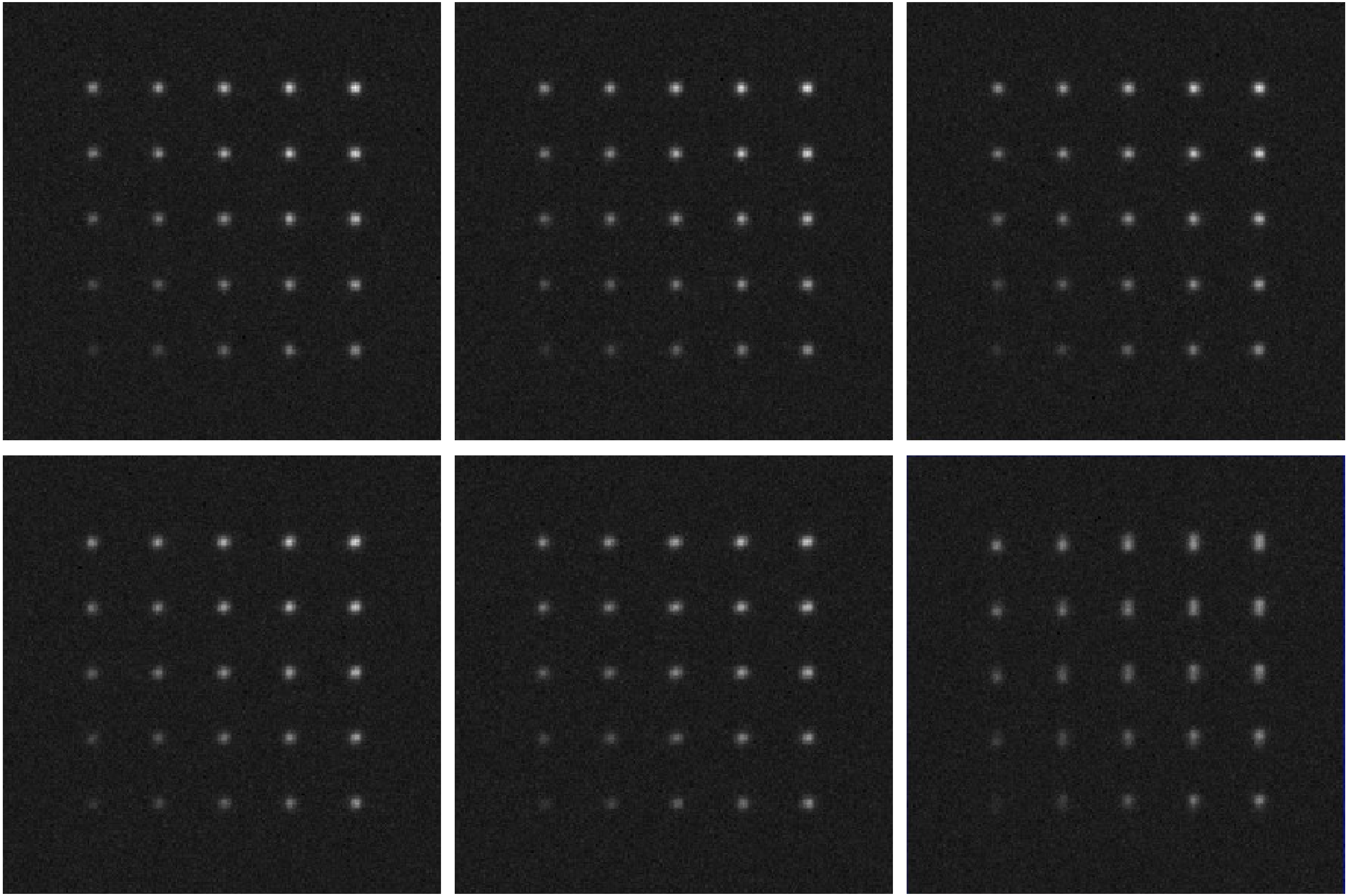


subband 2

$+ \quad \epsilon_t \quad =$



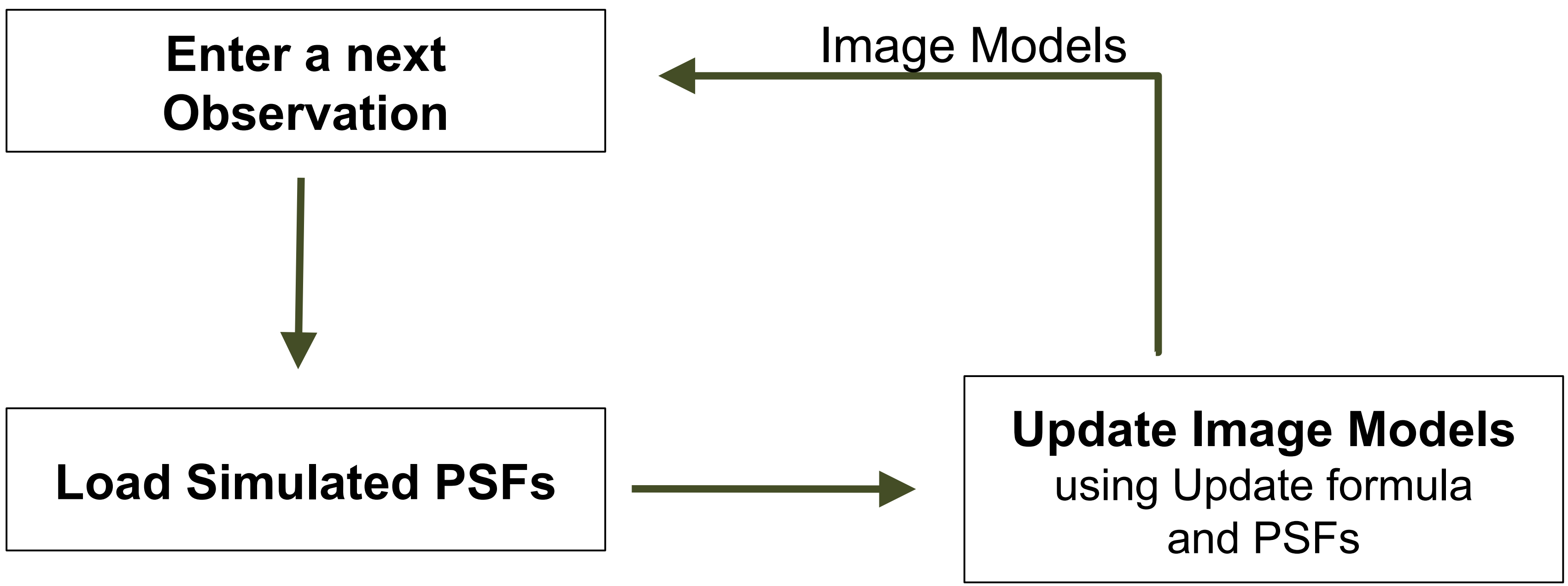
Sample Observations

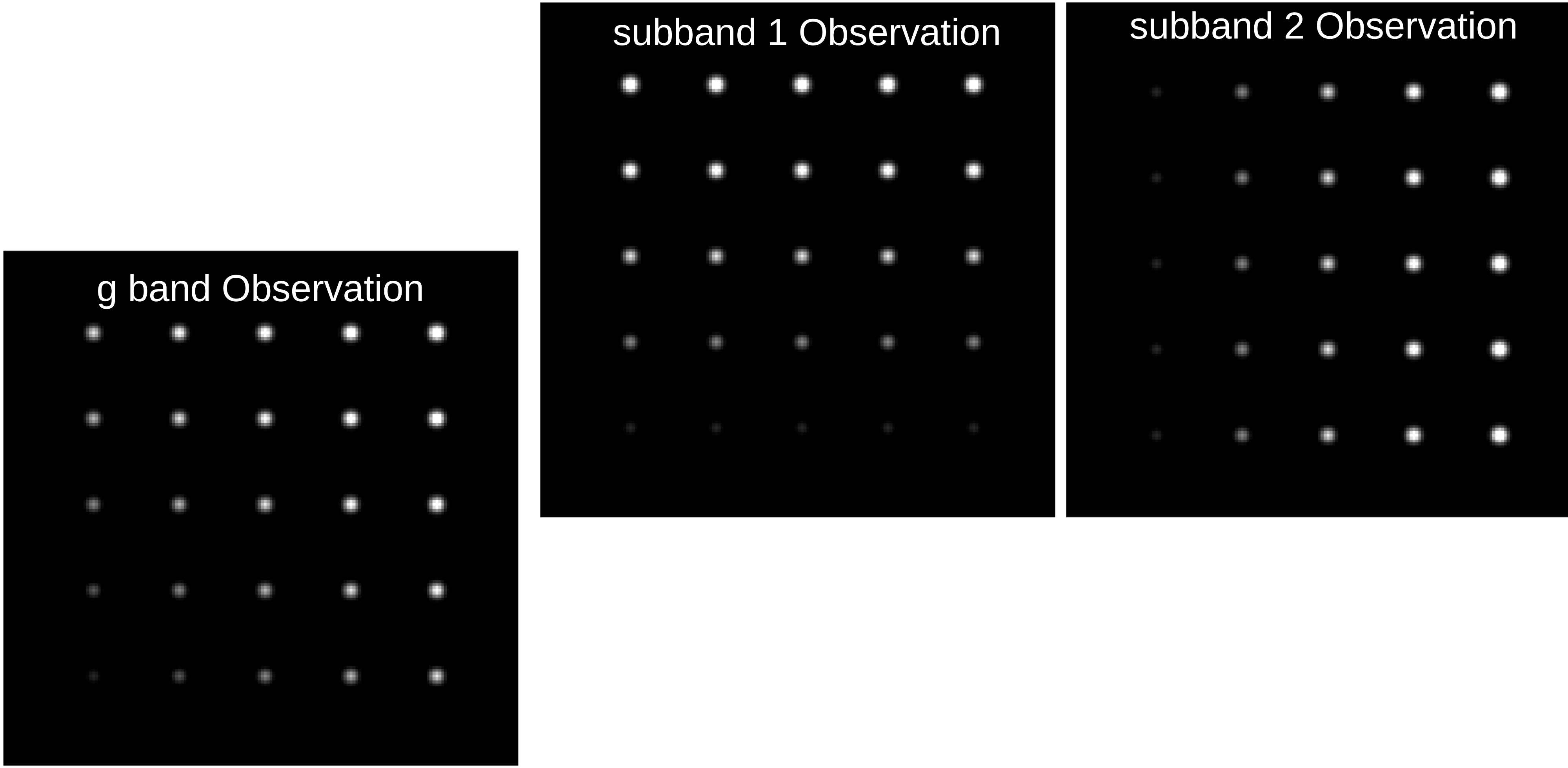


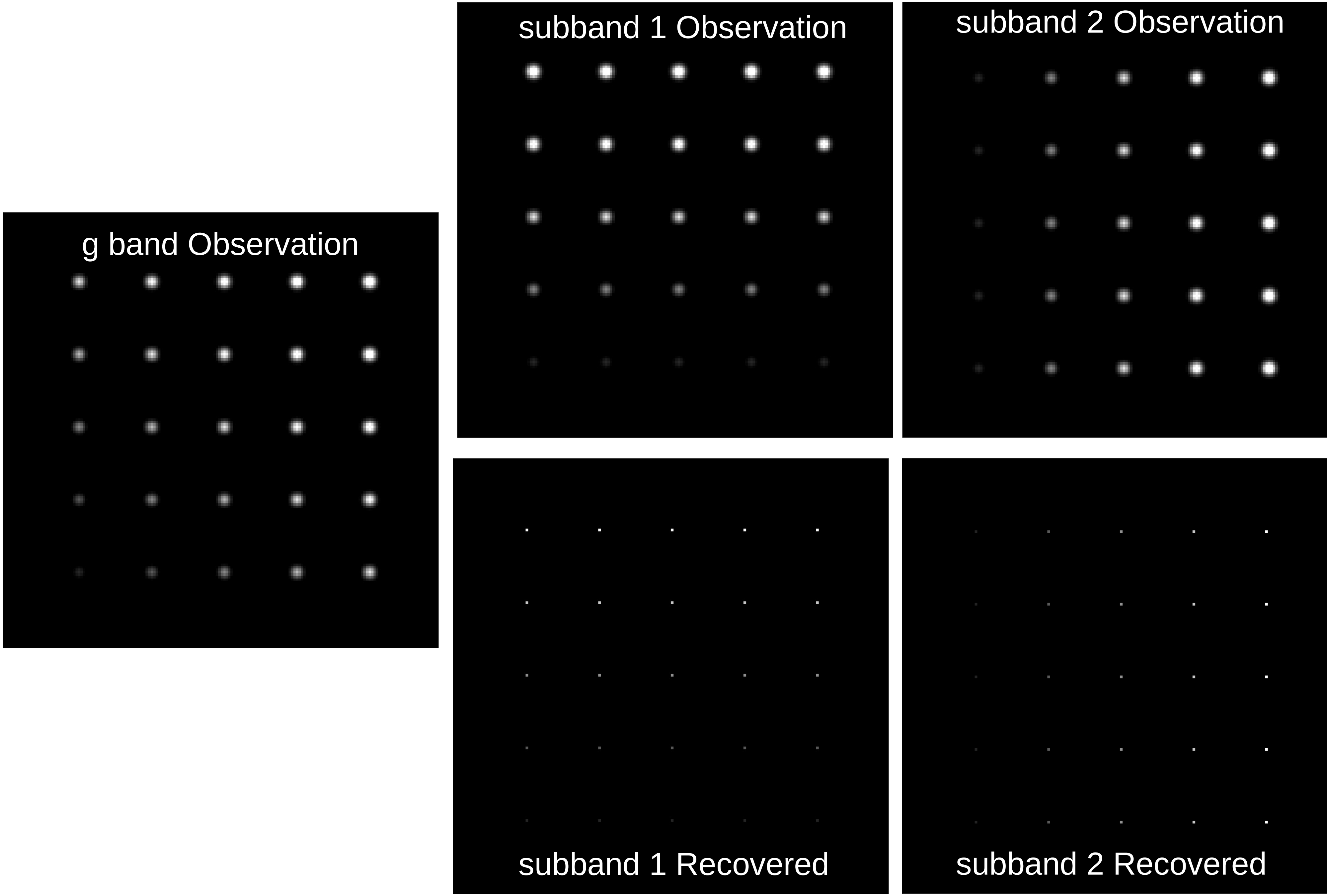
Thesis Work

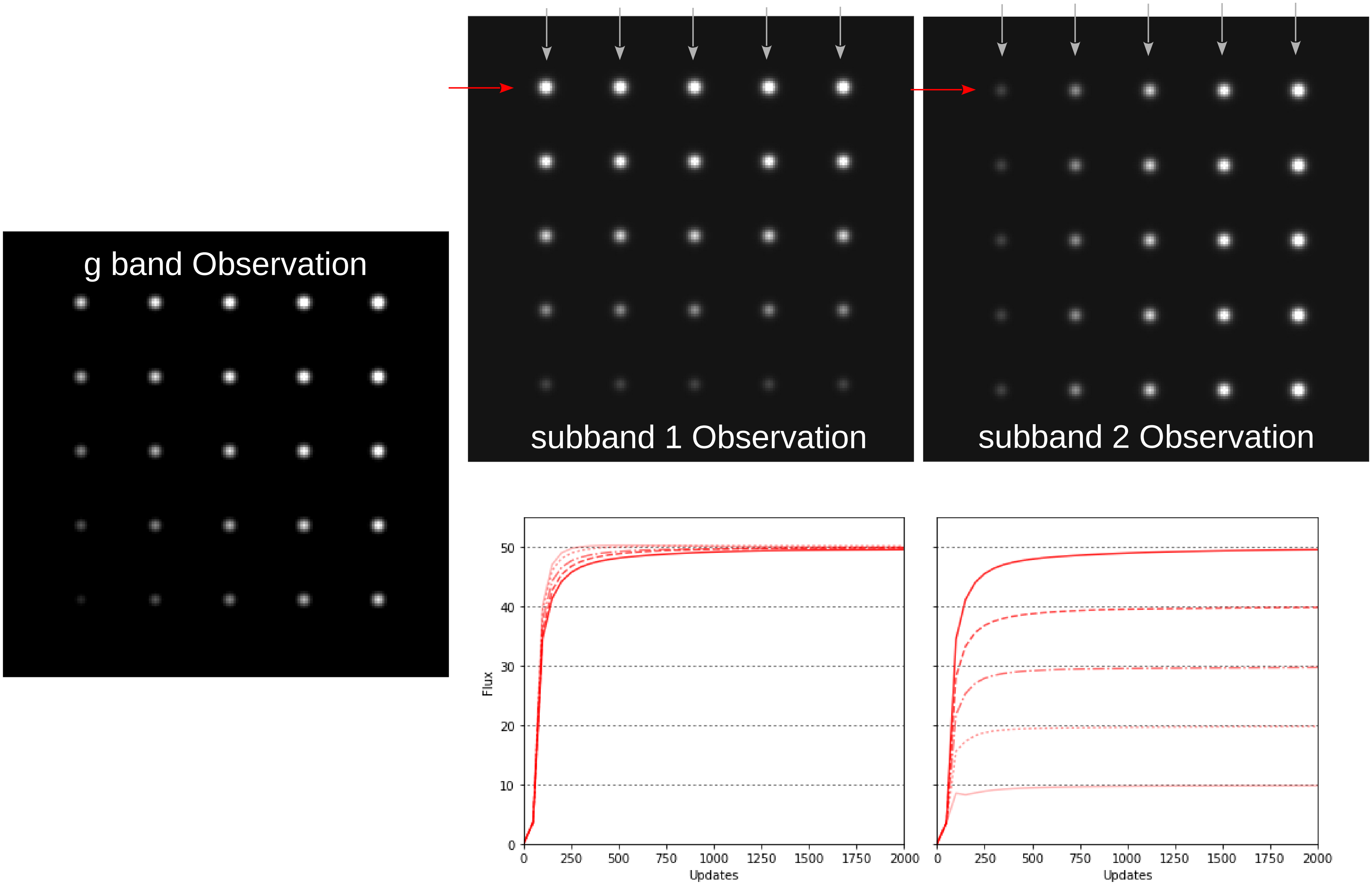
New Algorithm for Subbands

- Similar to MFBD
- Solving for multiple subband images

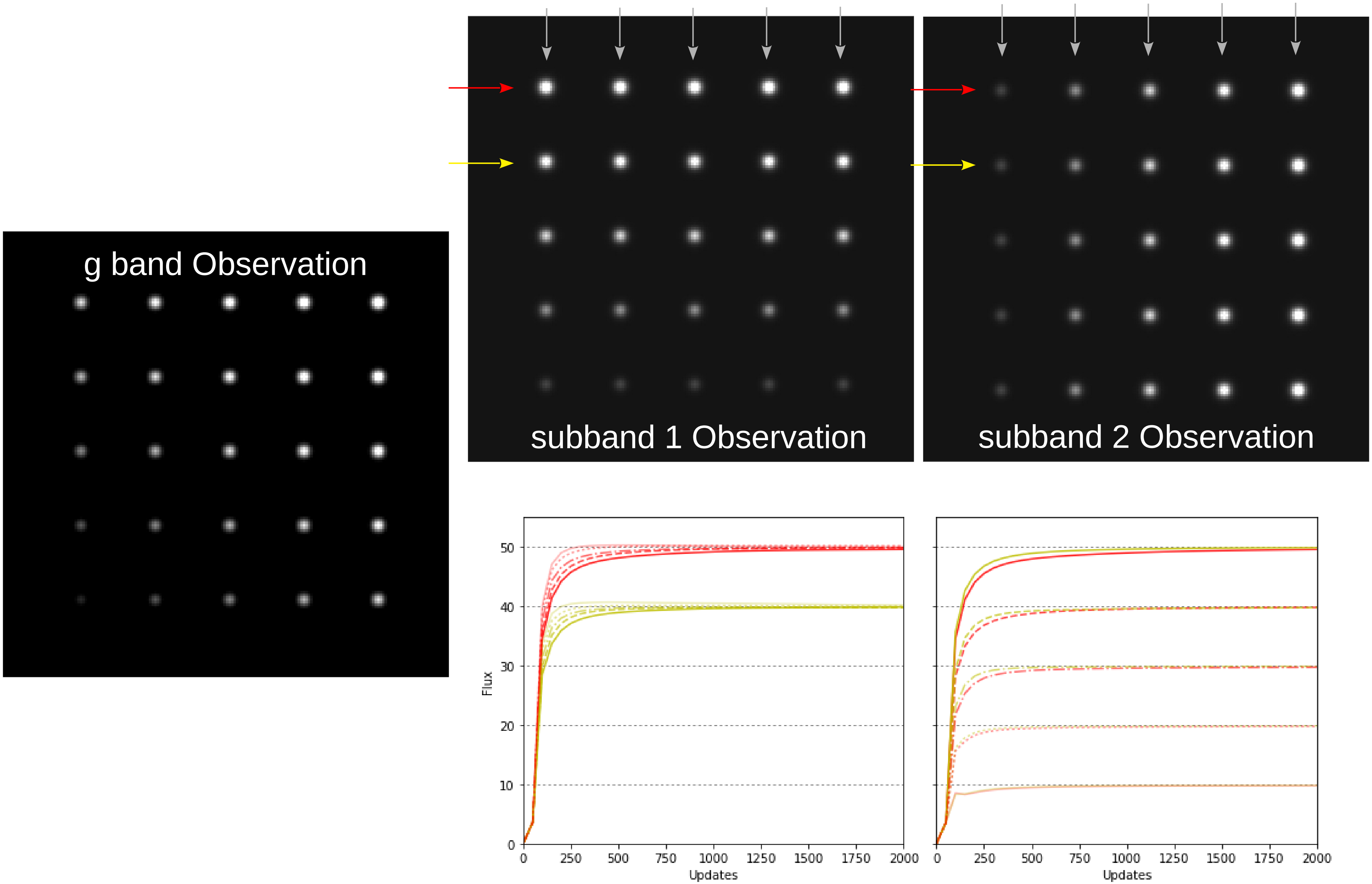



g band Observation
subband 1 Observation
subband 2 Observation


subband 1 Observation
subband 2 Observation
g band Observation
subband 1 Recovered
subband 2 Recovered

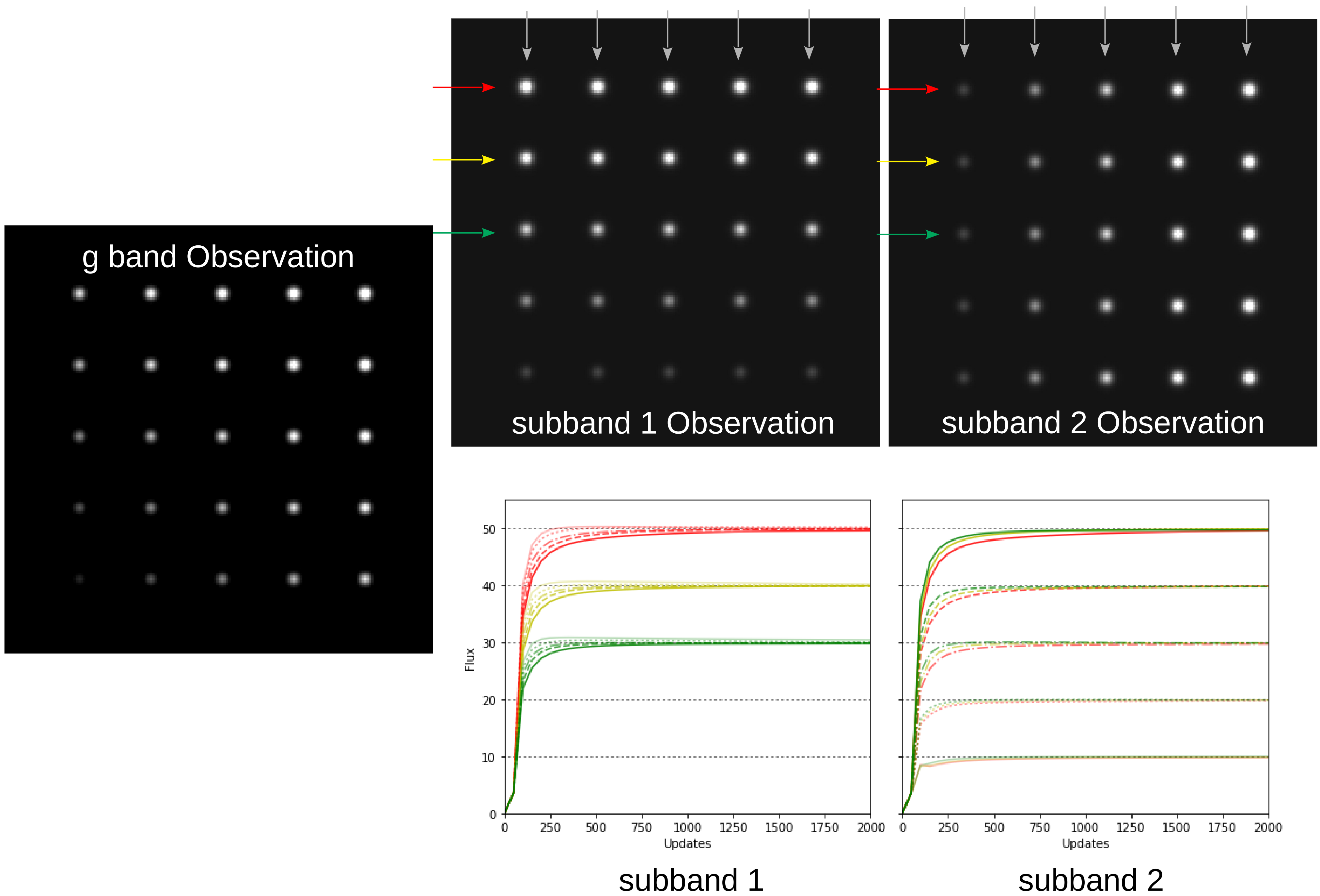


subband 1 subband 2

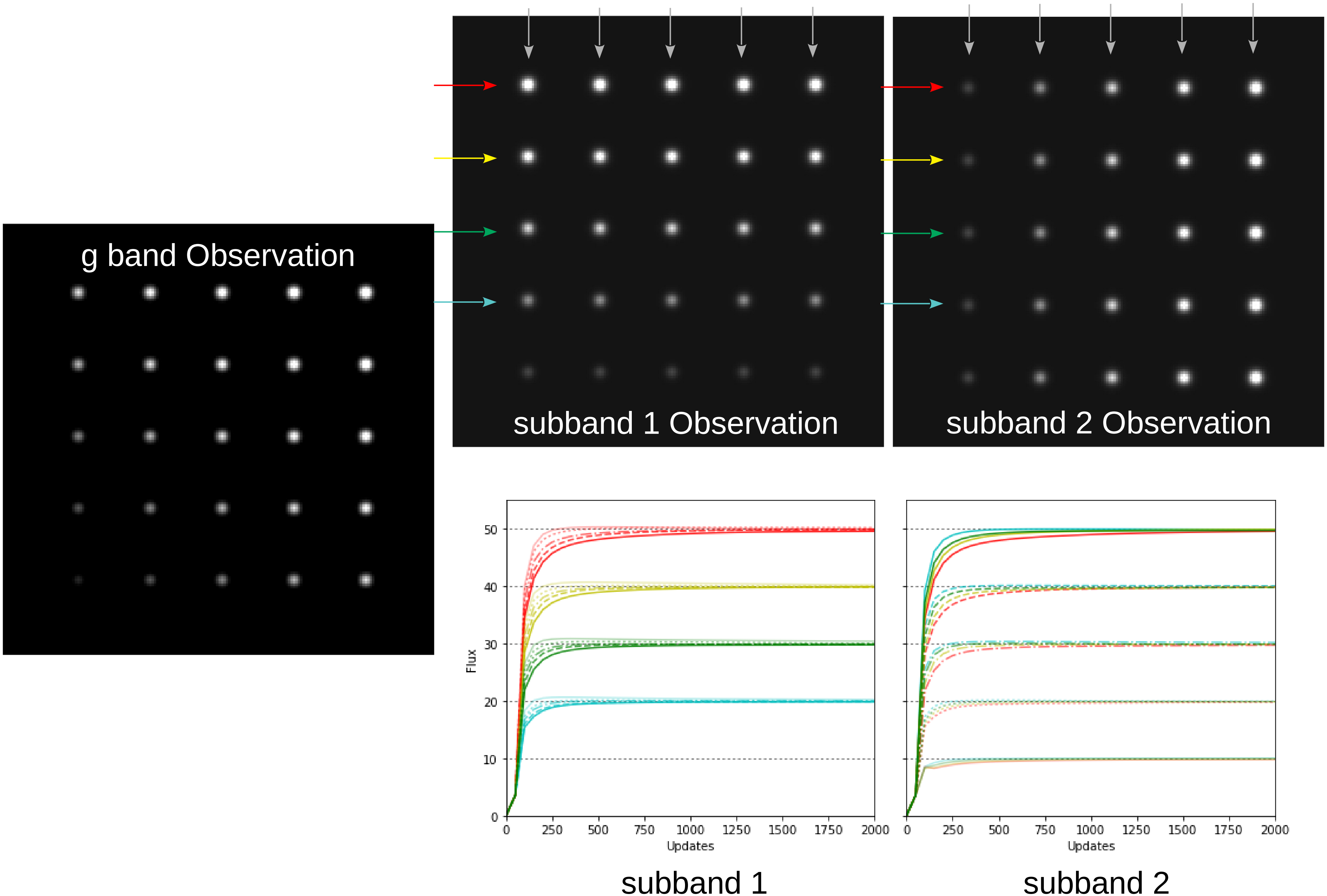


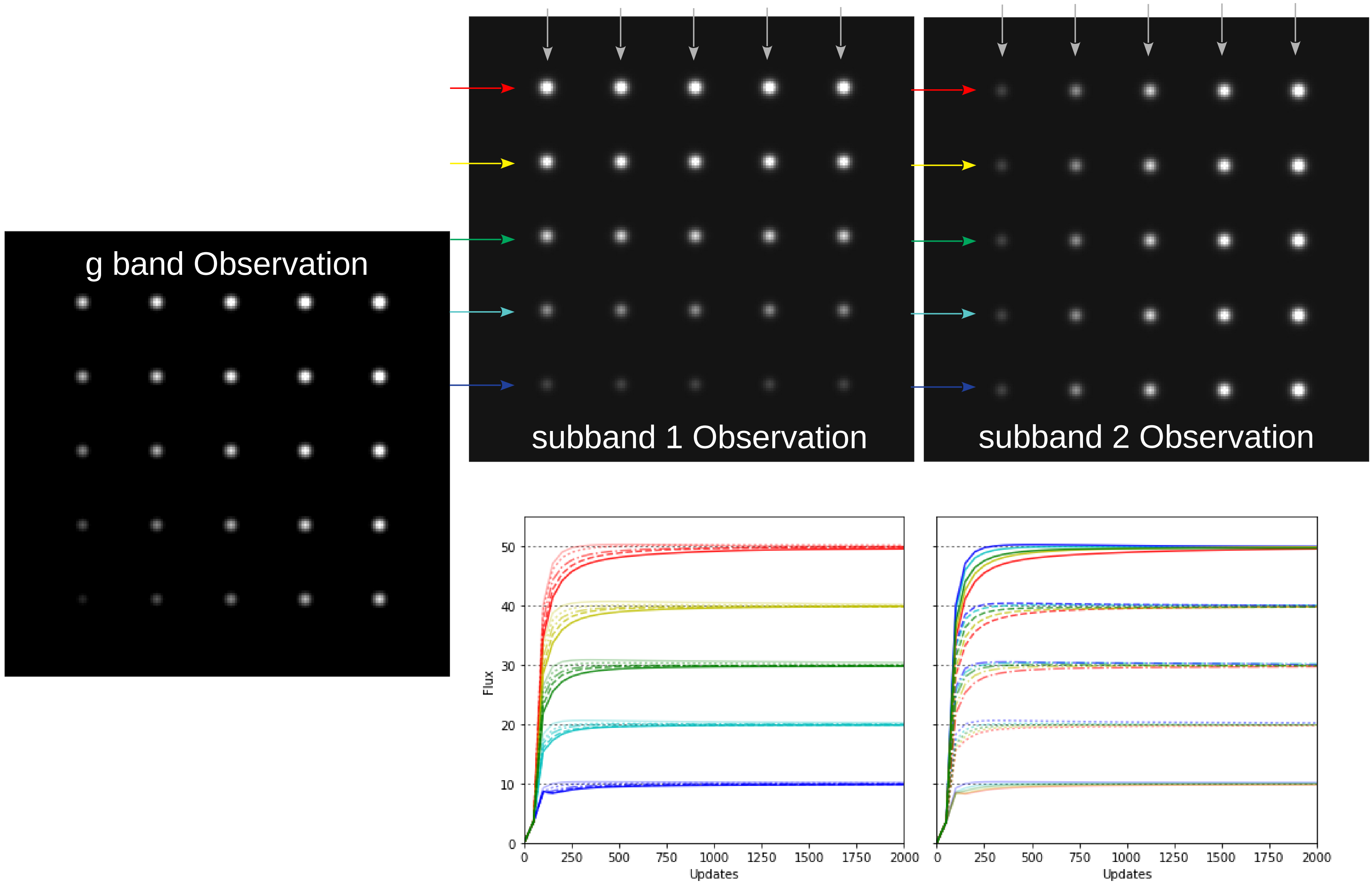
subband 1

subband 2



g band Observation
subband 1 Observation
subband 2 Observation
Flux
Updates
0
10
20
30
40
50
0
250
500
750
1000
1250
1500
1750
2000
subband 1
subband 2

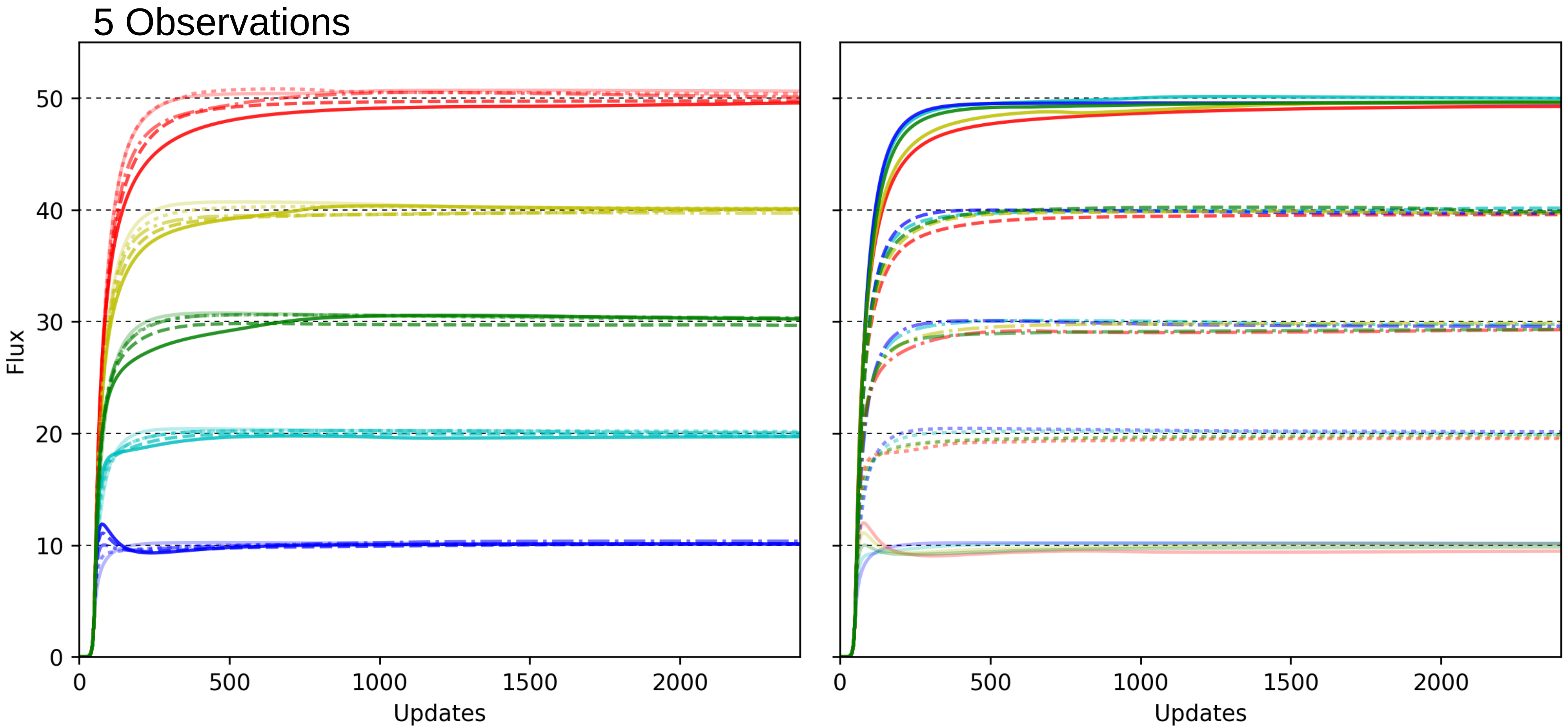

g band Observation
subband 1 Observation
subband 2 Observation
Flux
Updates
0
10
20
30
40
50
0
250
500
750
1000
1250
1500
1750
2000
subband 1
subband 2



subband 1

subband 2

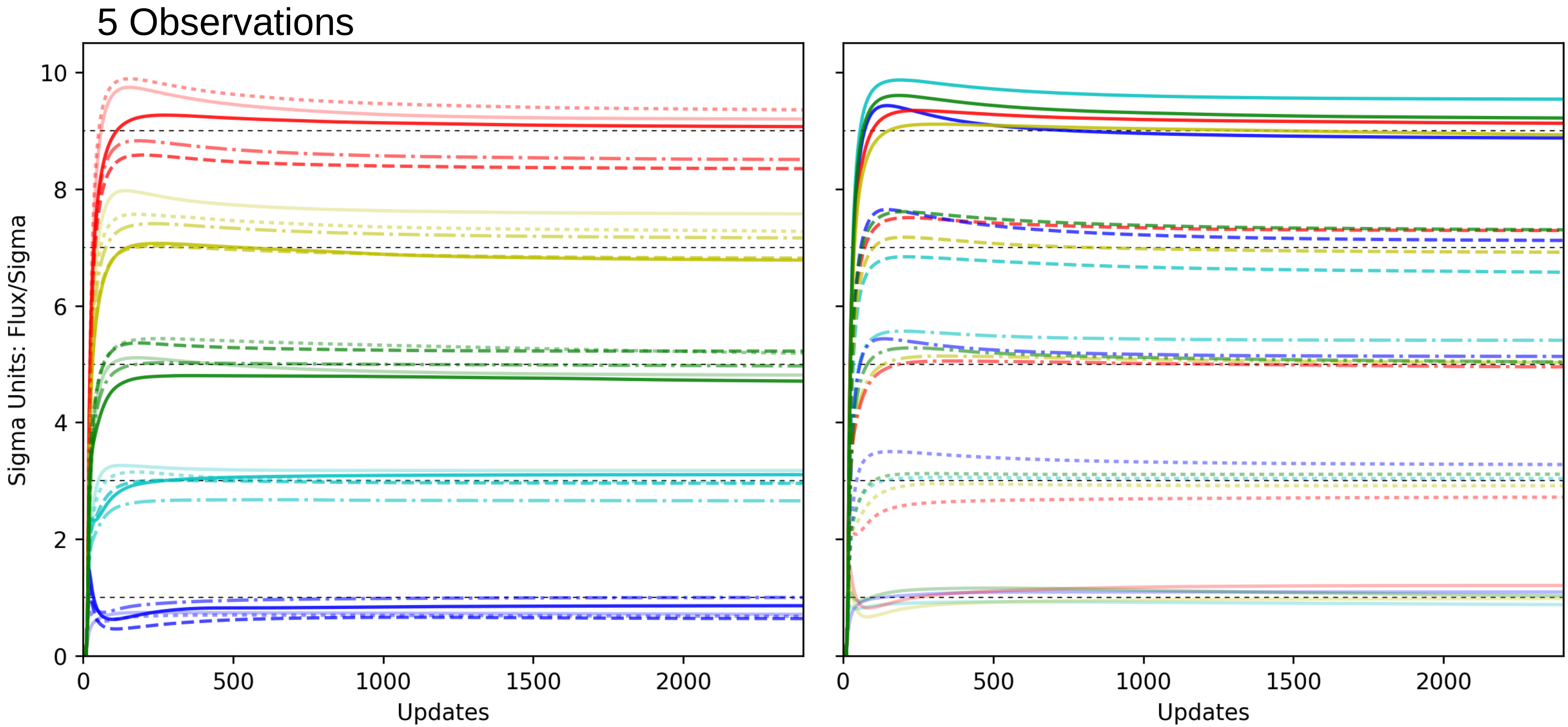
Noiseless – Flux Recovery



subband 1

subband 2

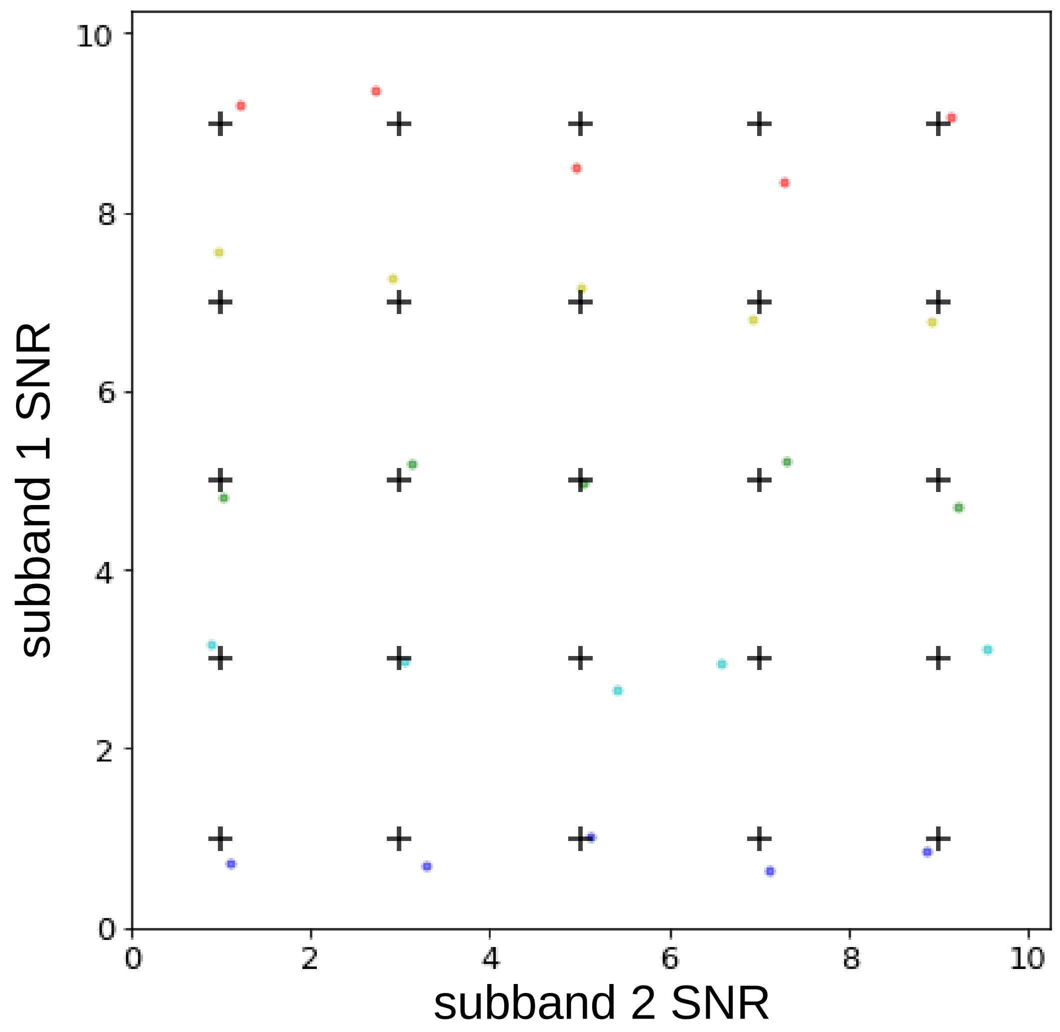
Introducing Noise



subband 1 subband 2

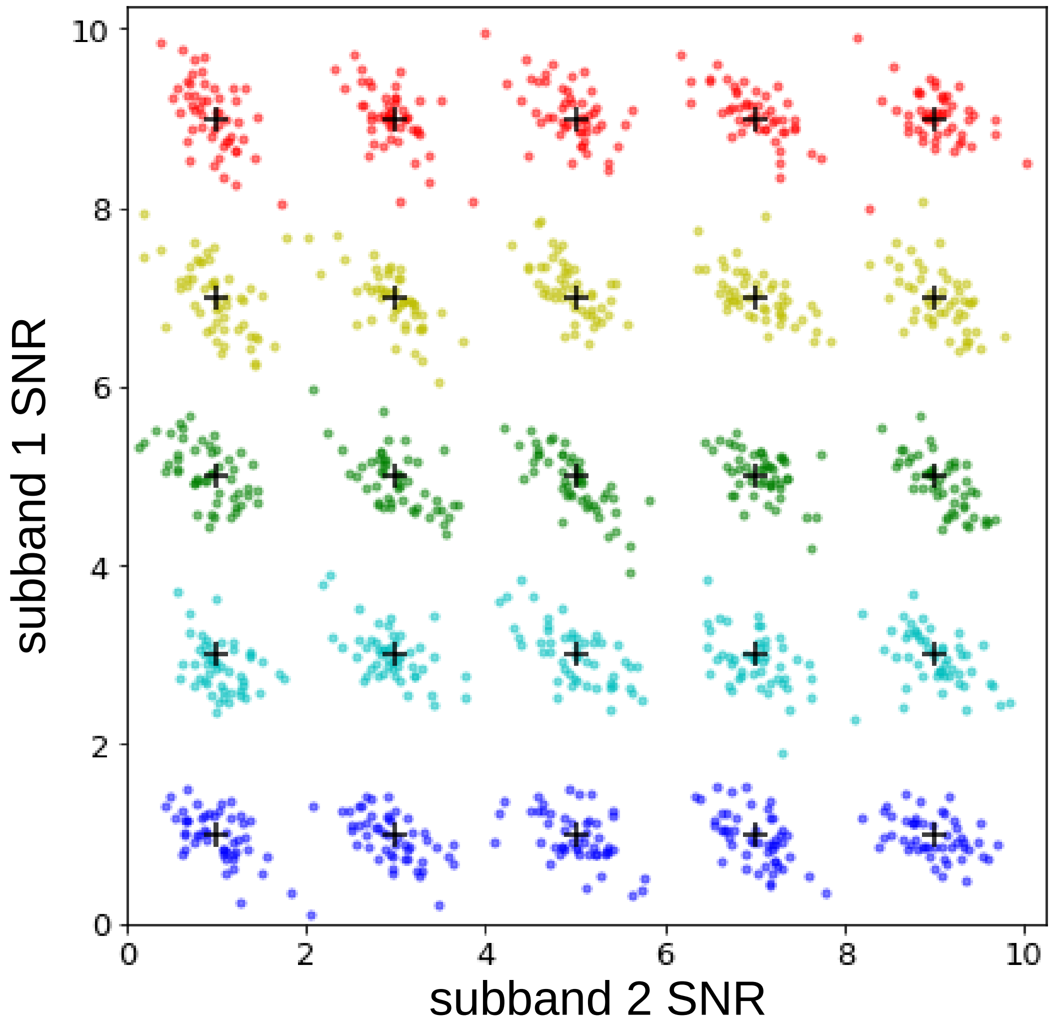
Flux Separation

- 5 Observations
- 1 Realization



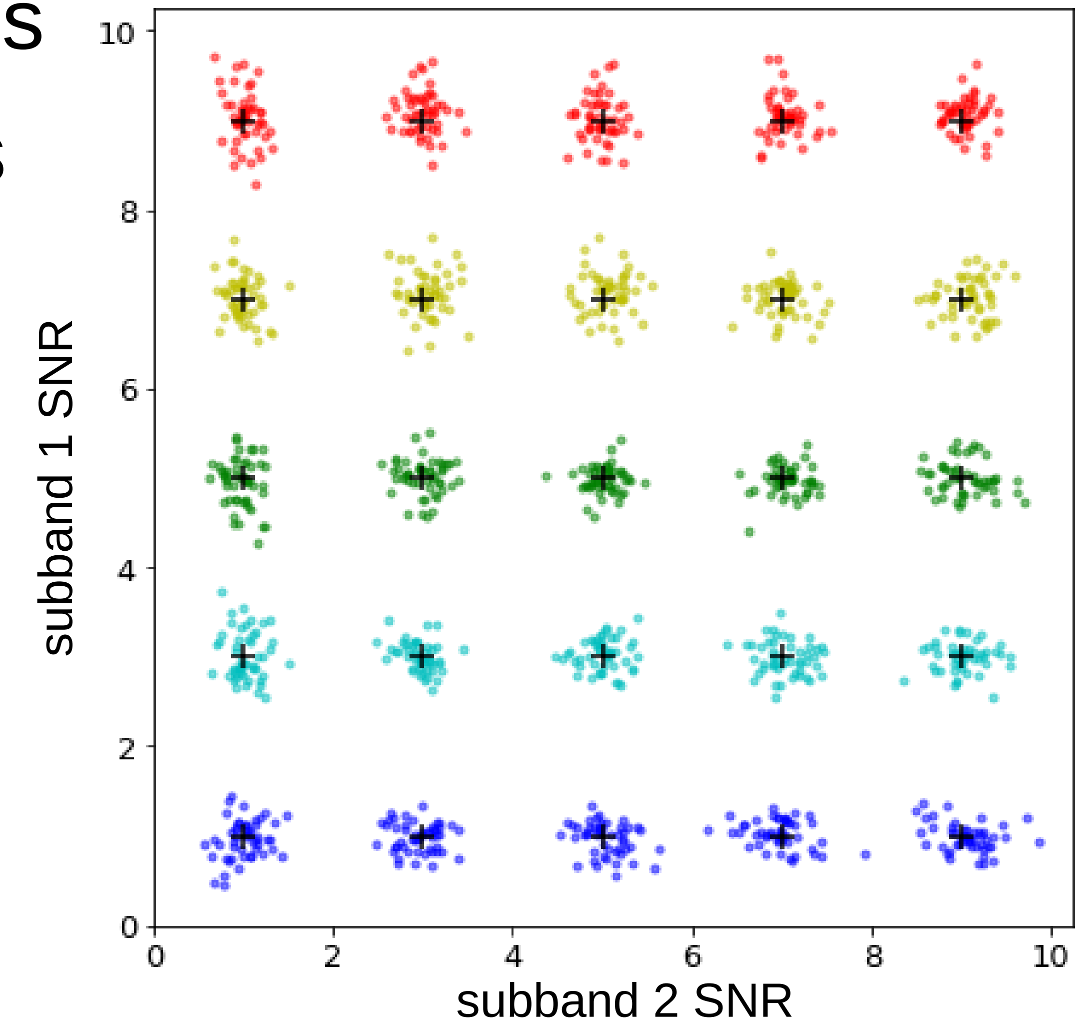
Flux Separation

- 5 Observations
- 50 Realizations



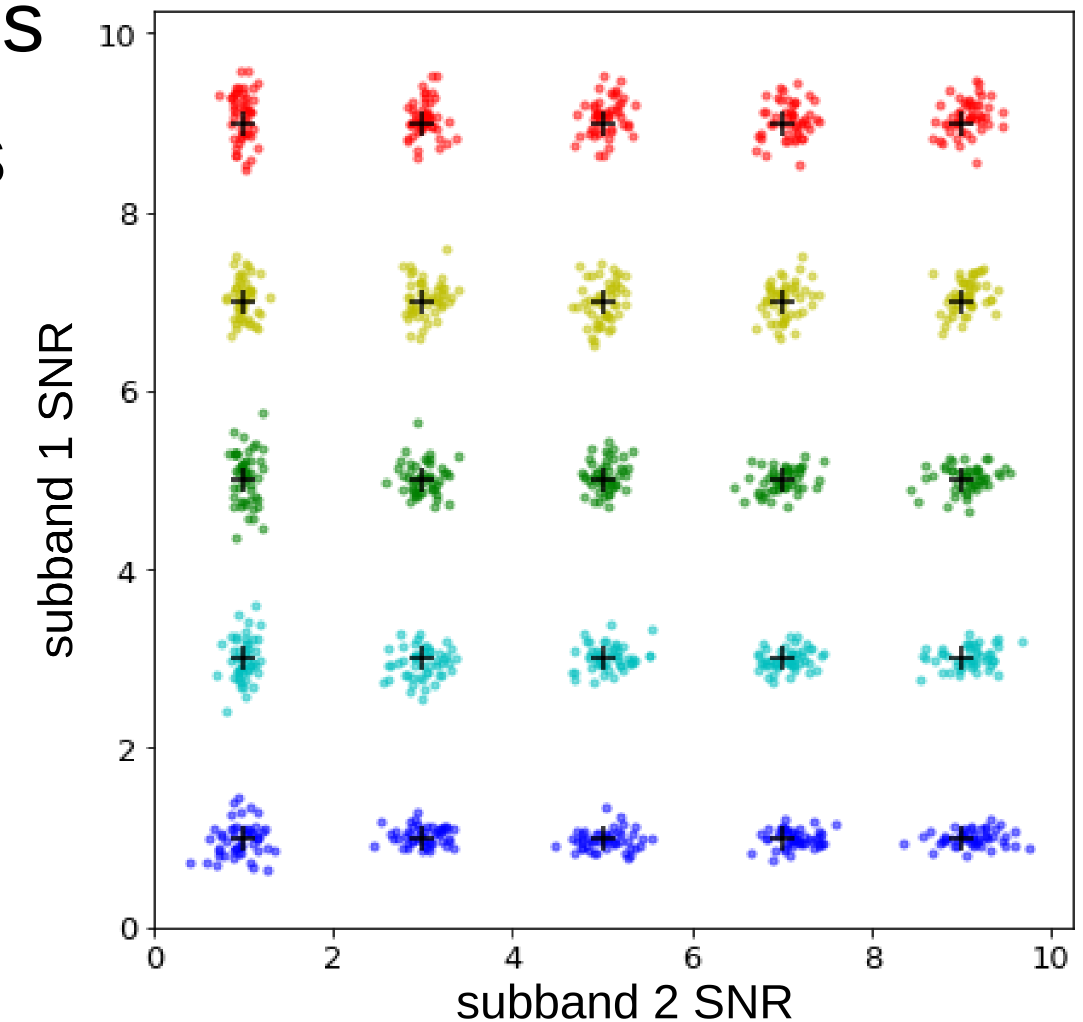
Flux Separation

- 15 Observations
- 50 Realizations



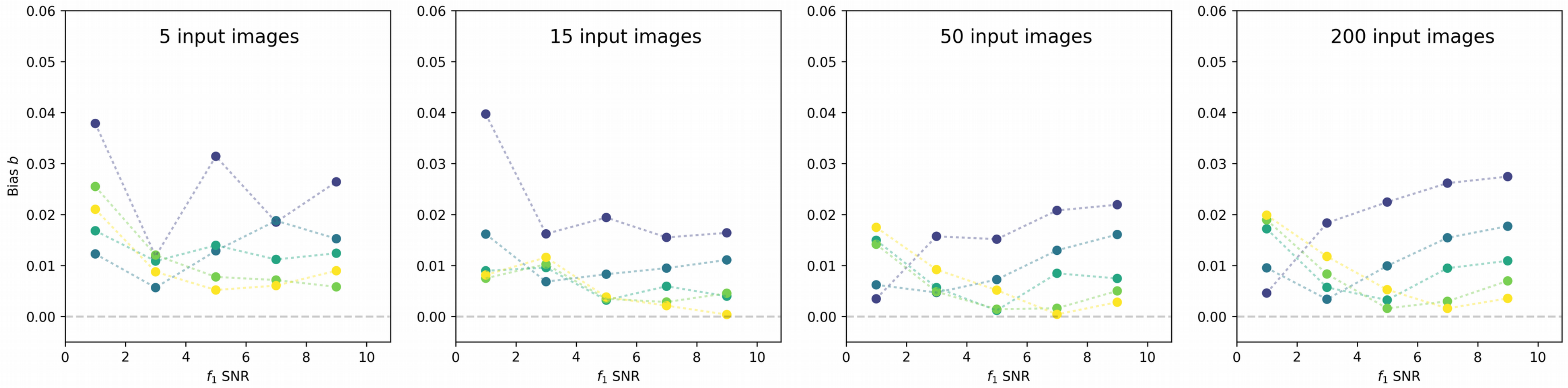
Flux Separation

- 50 Observations
- 50 Realizations



Positional Accuracy

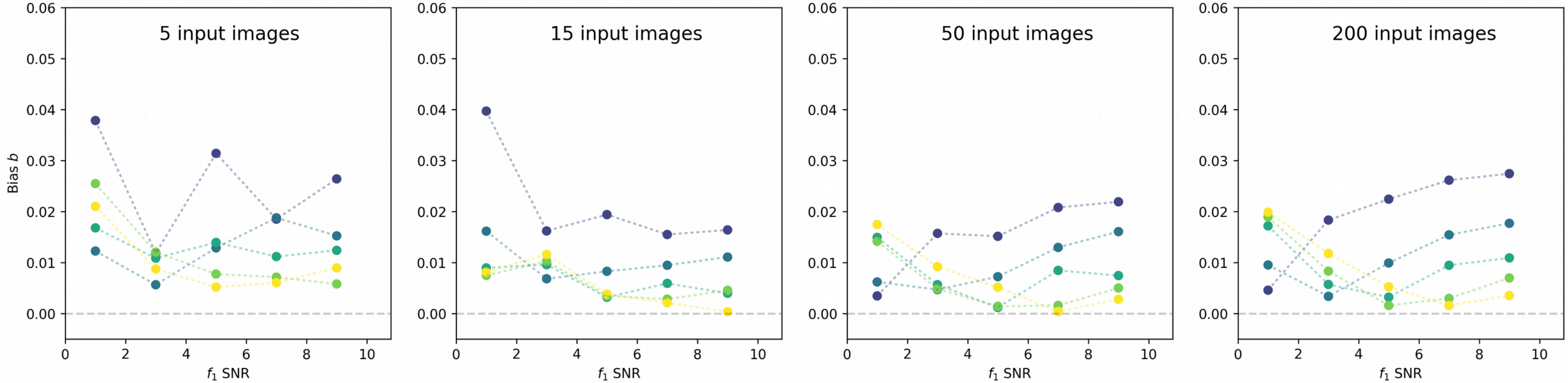
Position Bias of Coadd



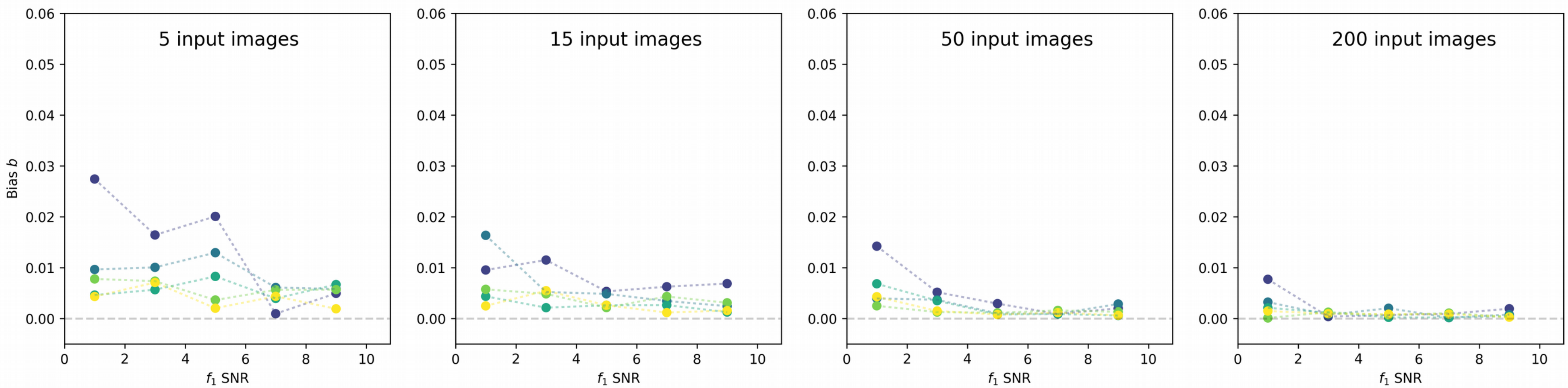
Thesis Work

Positional Accuracy

Position Bias of Coadd

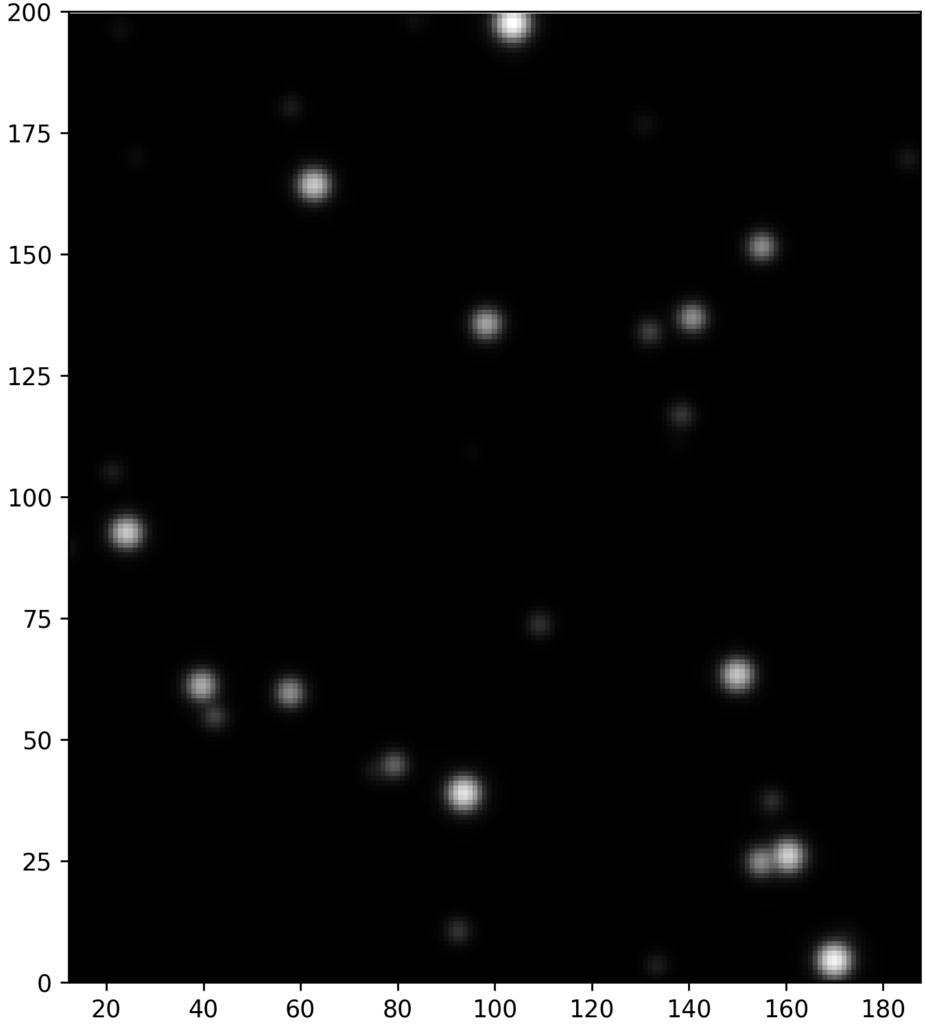


Position Bias of DCR Result

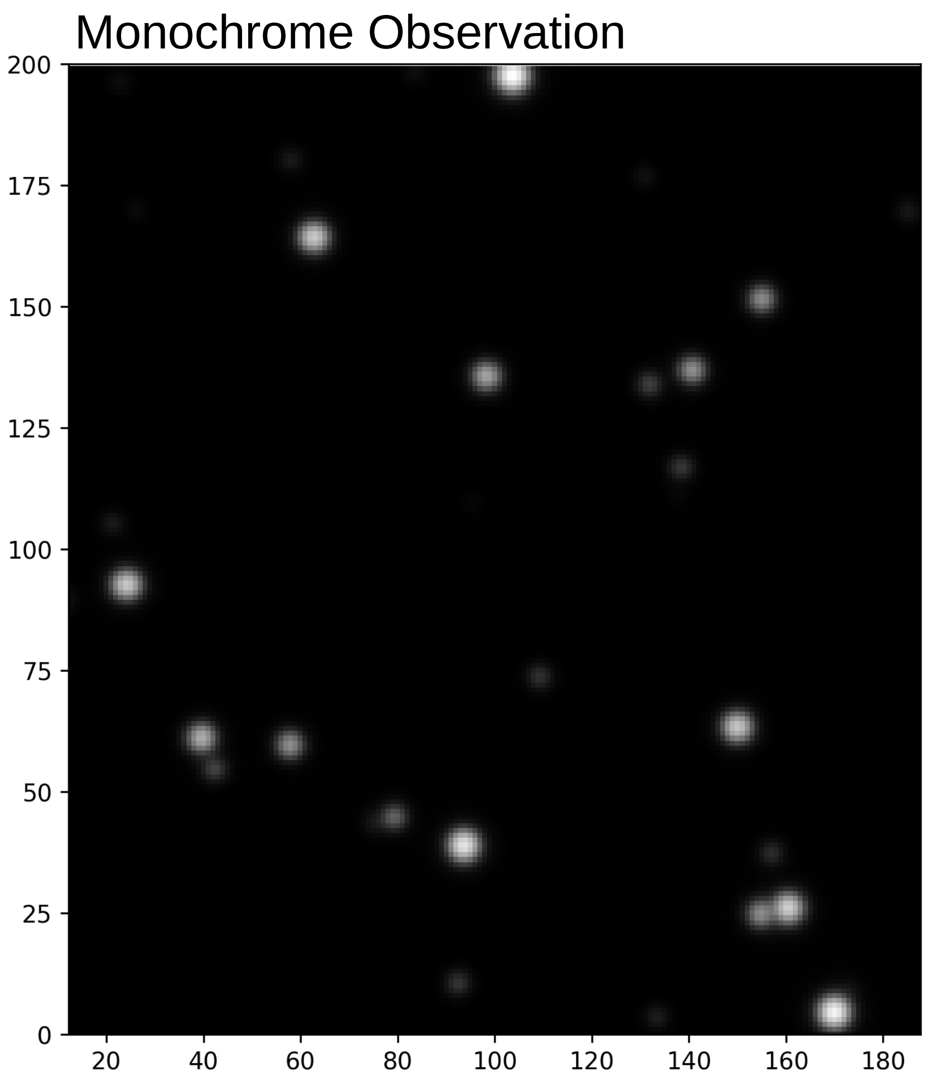


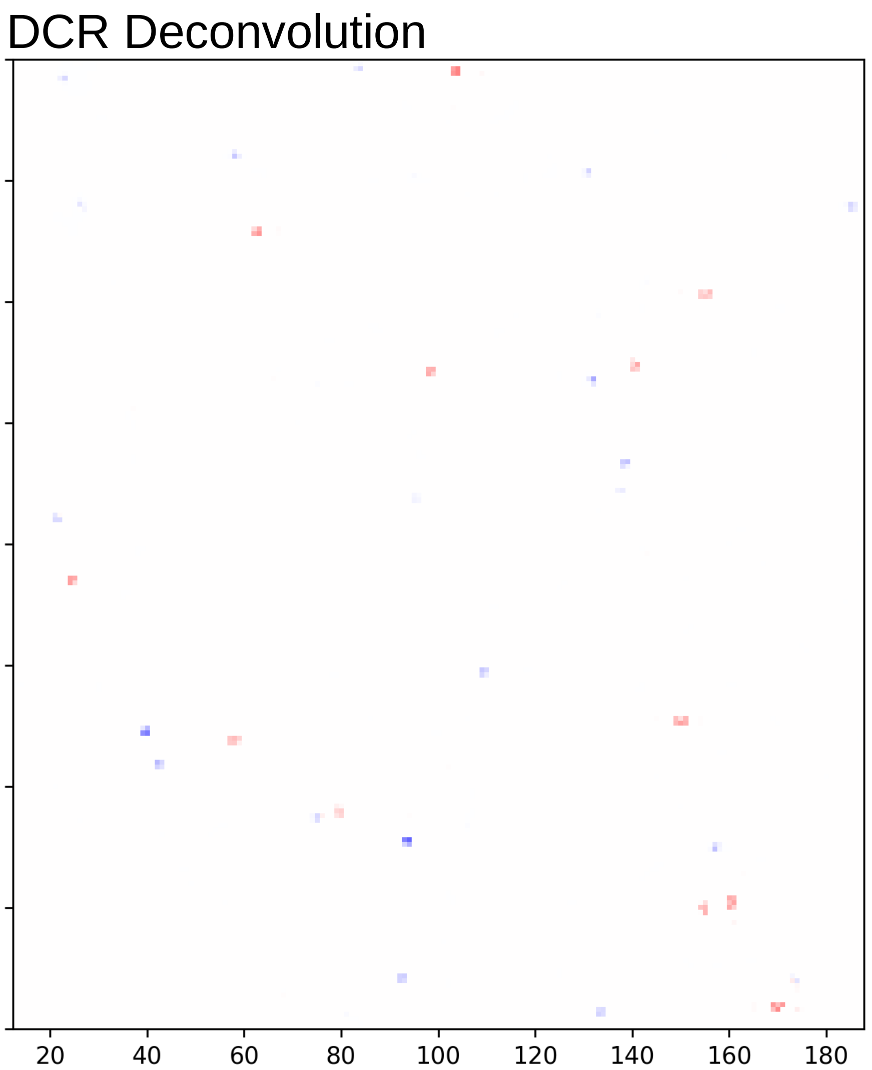
Realistic Simulation, 2 Sub-Bands

Monochrome Observation

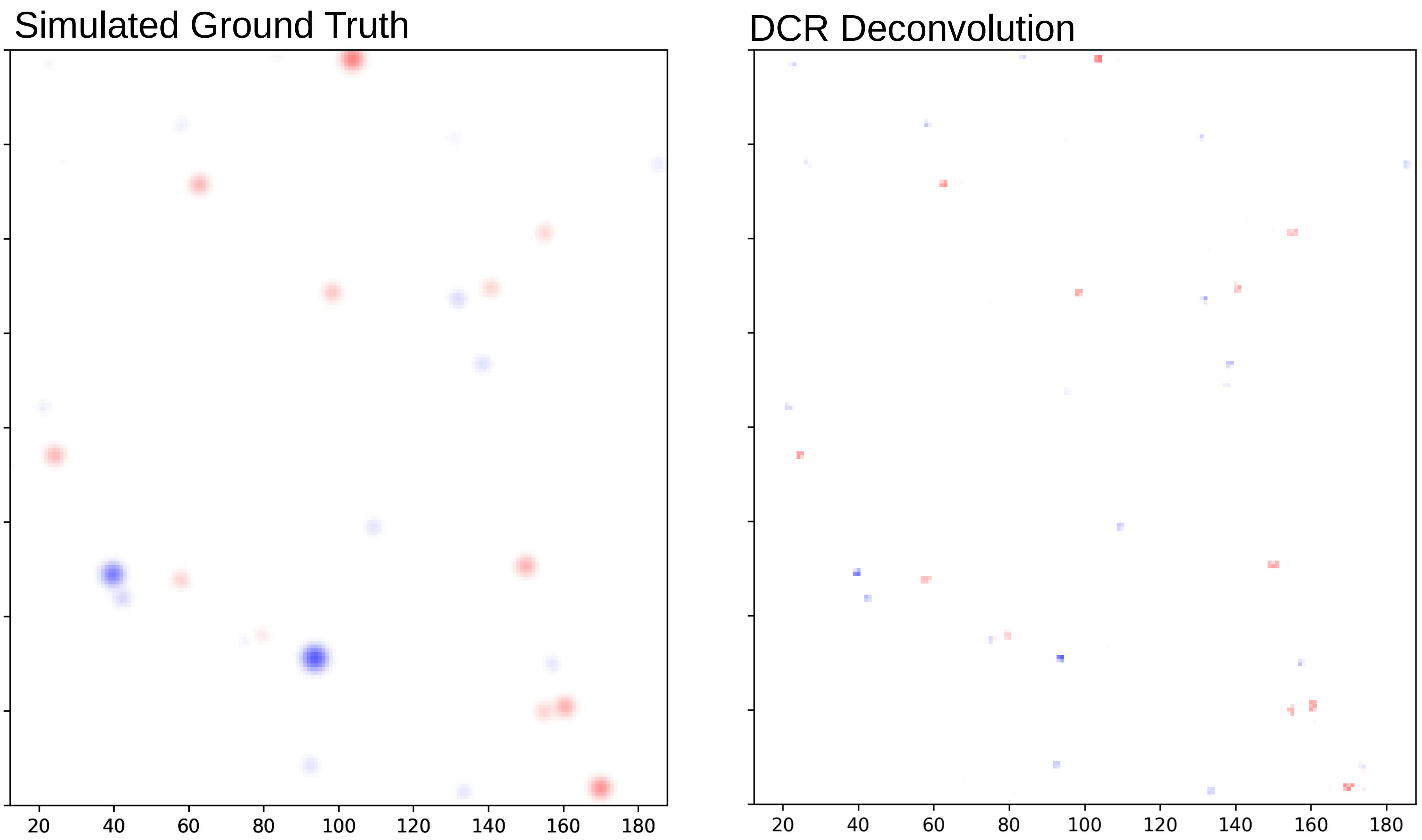


Realistic Simulation, 2 Sub-Bands



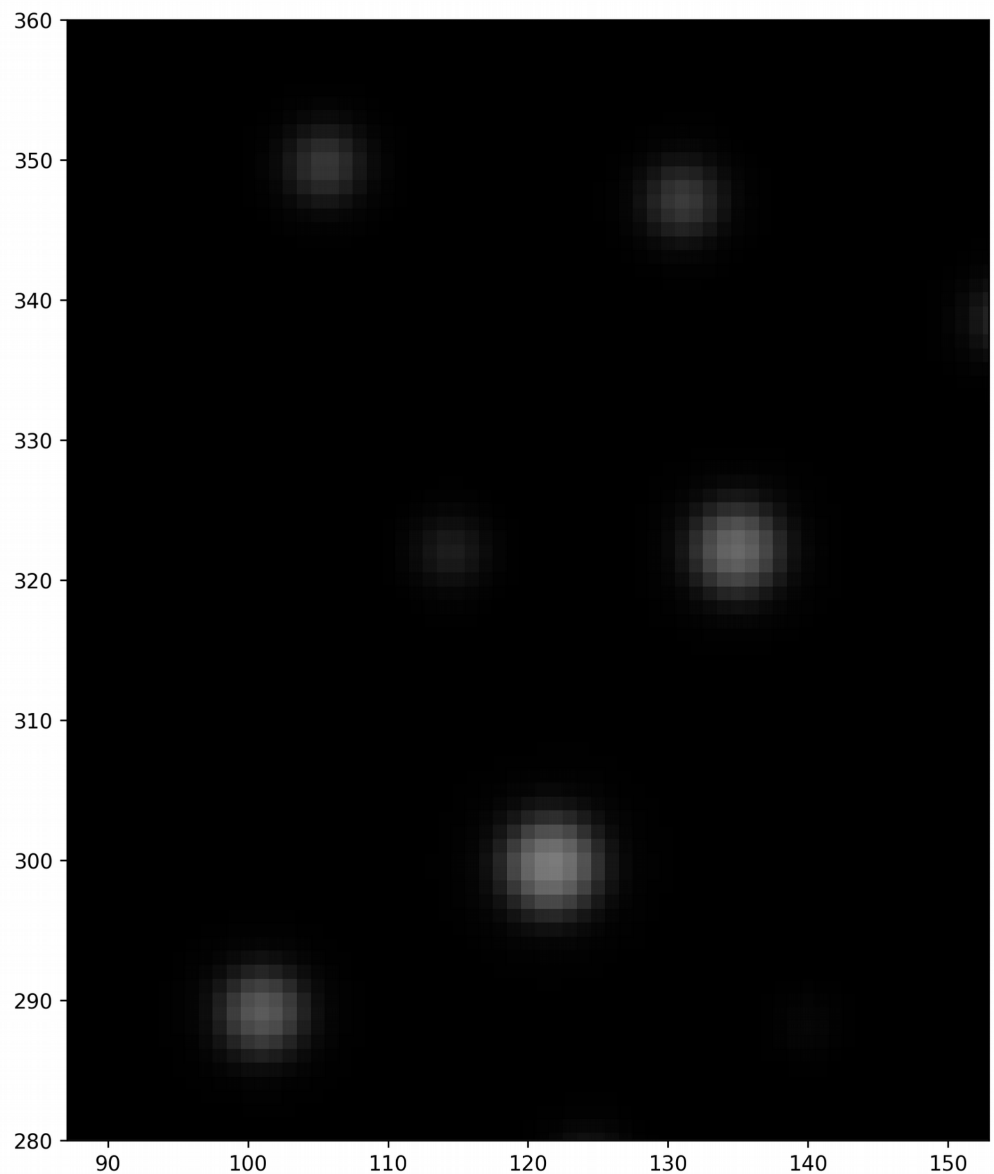


Realistic Simulation, 2 Sub-Bands

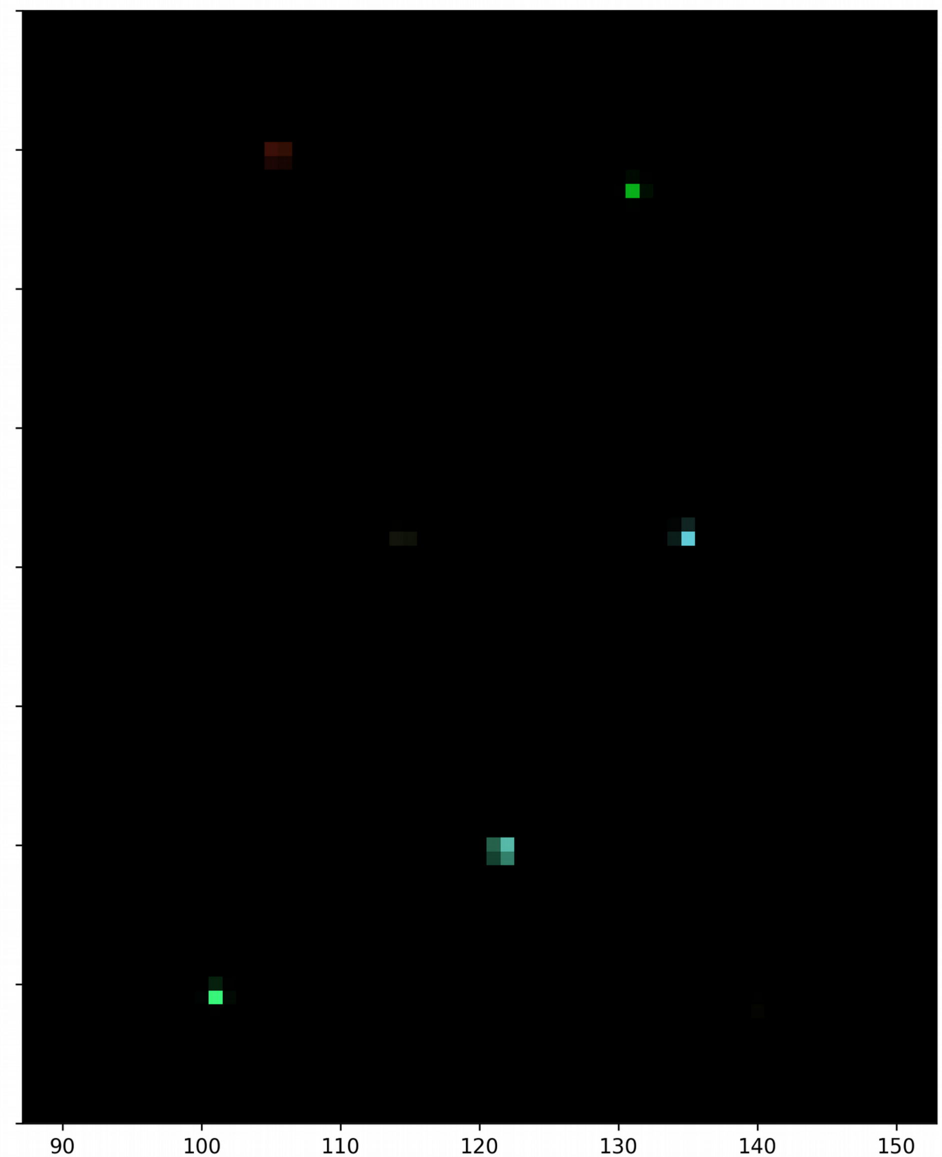


Realistic Simulation, 3 Sub-Bands

Monochrome Observation

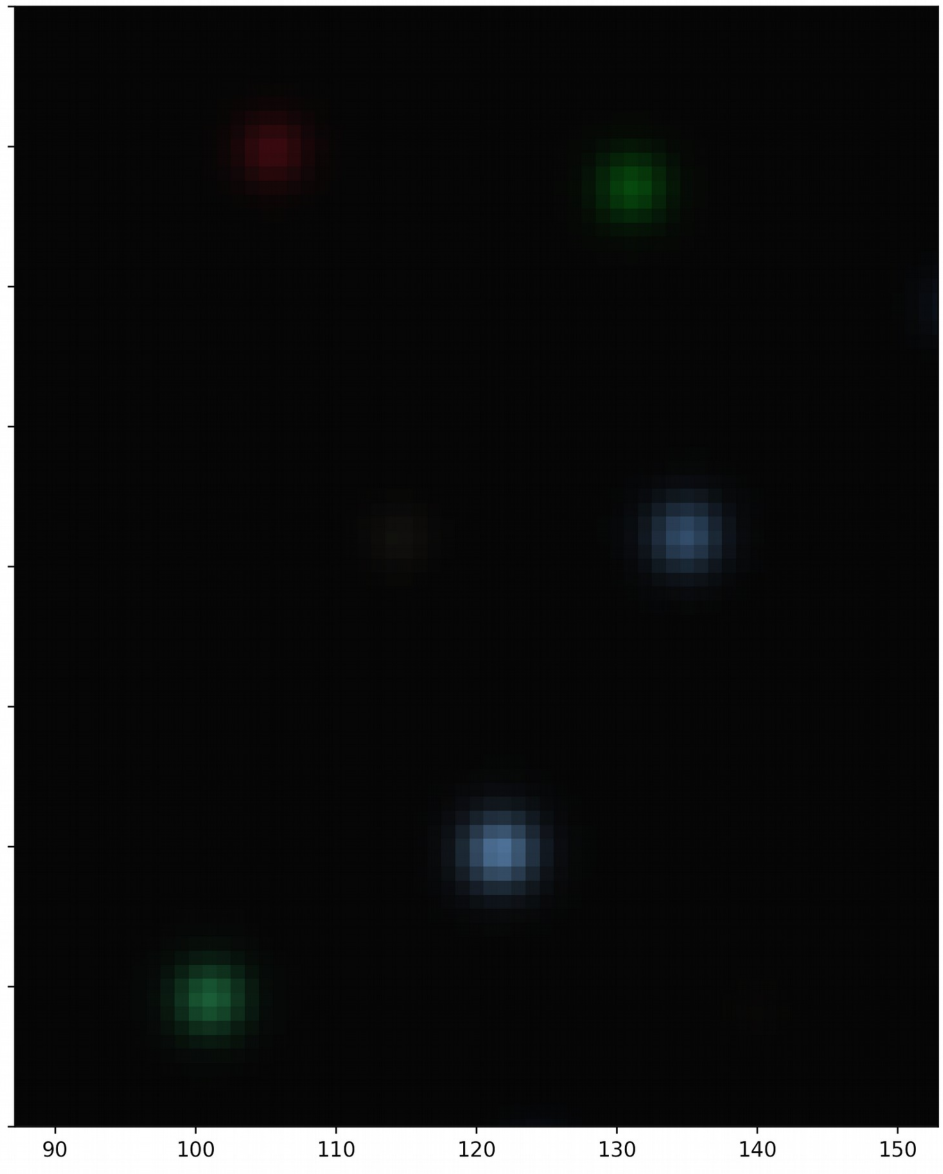


DCR Deconvolution

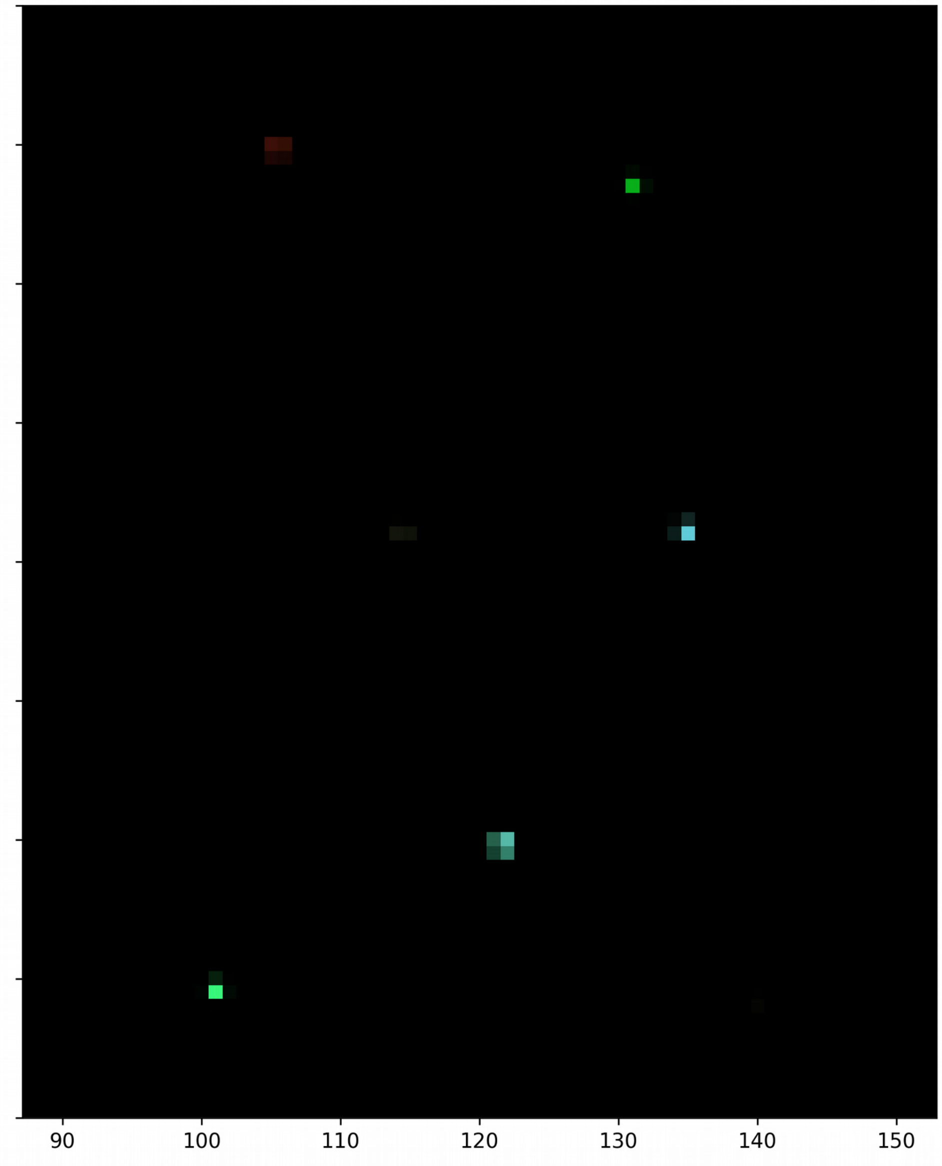


Realistic Simulation, 3 Sub-Bands

Simulated Ground Truth



DCR Deconvolution



Thesis Work

My Contributions

- Through:
  - Developing and implementing new algorithms
  - Advanced software and hardware

- Results:
  - Achieve high performance
  - Efficiently tackle big-data problems
  - Data fusion at scale!

Acknowledgments

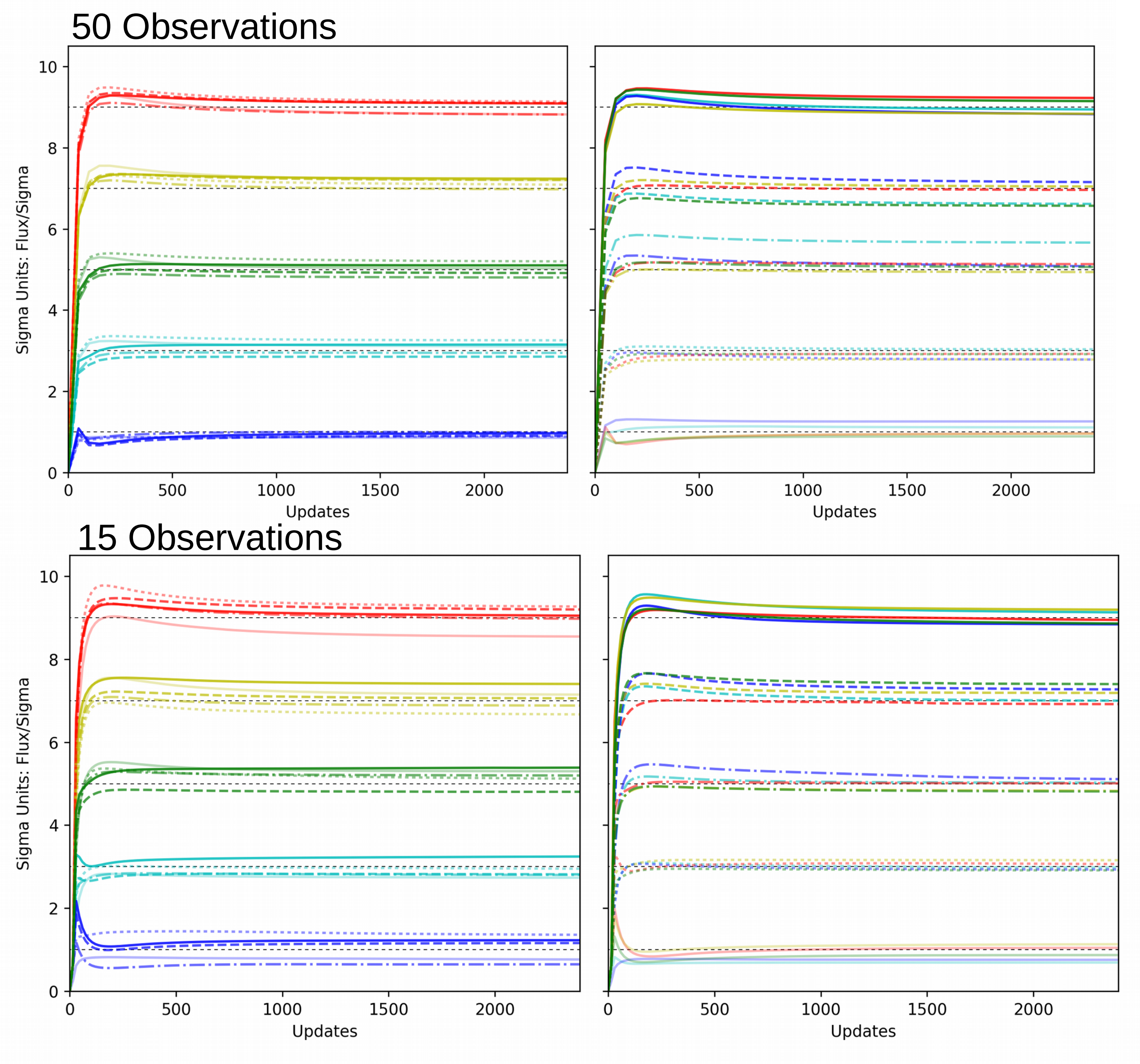
Tamás Budavári

Alex Szalay

Randal Burns

Thank You

Any Questions?



50 Observations
Sigma Units: Flux/Sigma
Updates
Updates
15 Observations
Sigma Units: Flux/Sigma
Updates
Updates

Zenith Angles