STREAMING ALGORITHMS

ANALYSIS OF ASTRONOMY IMAGES & CATALOGS

10/26/2015

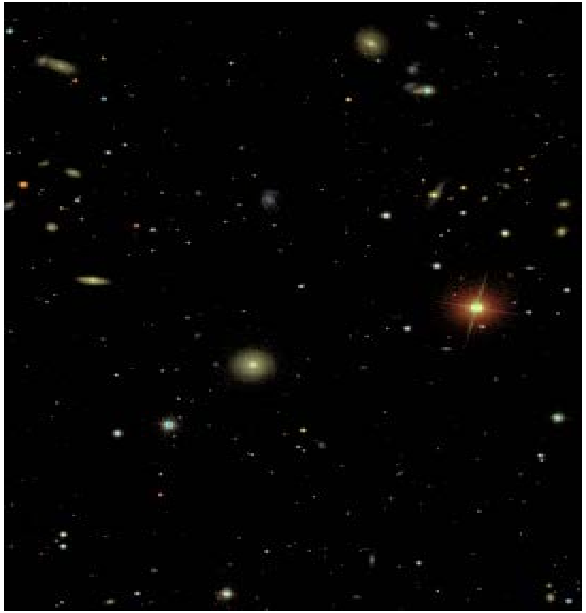
Tamás Budavári / Johns Hopkins University

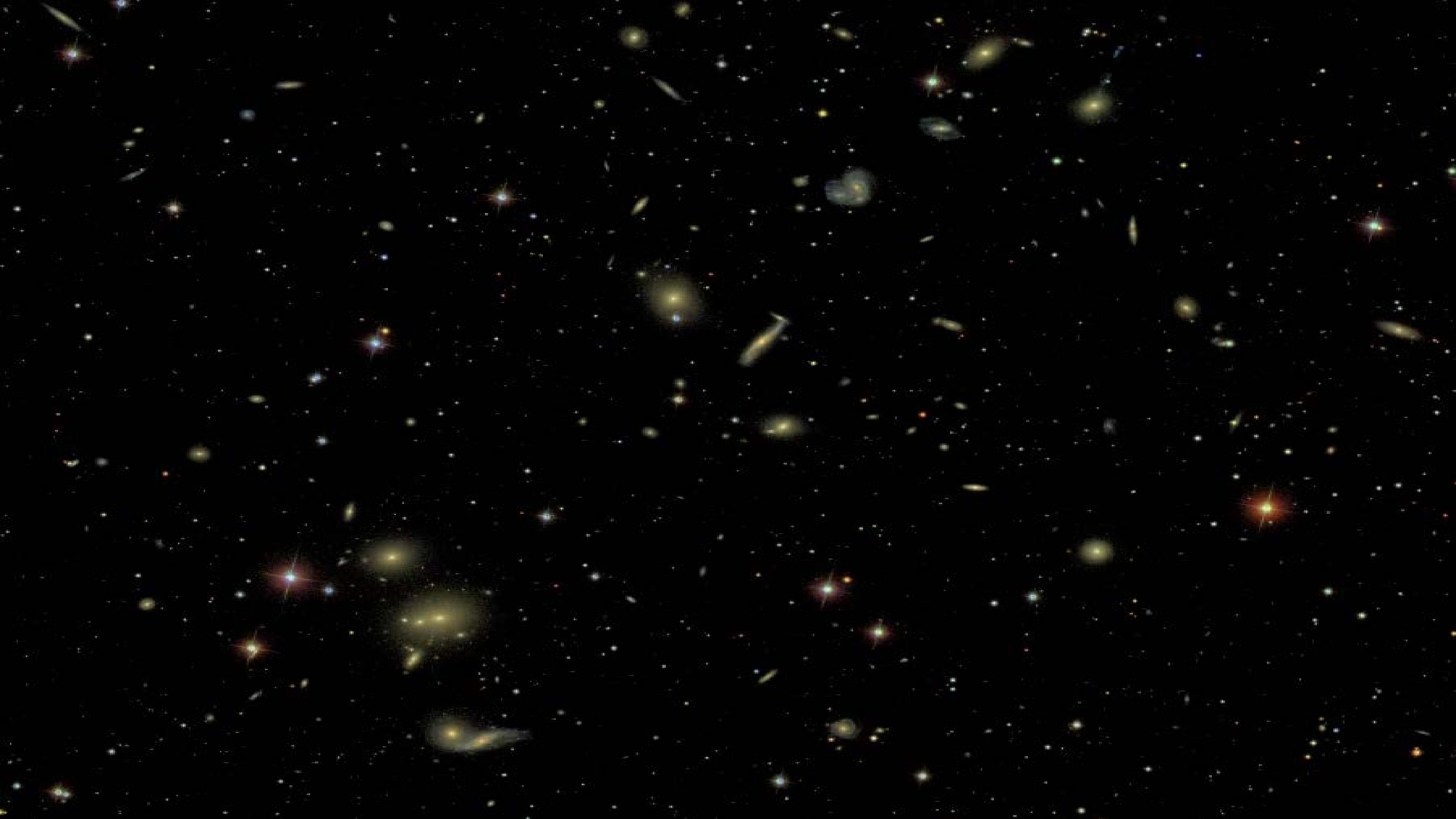
Astronomy Changed!

- Always been data-driven

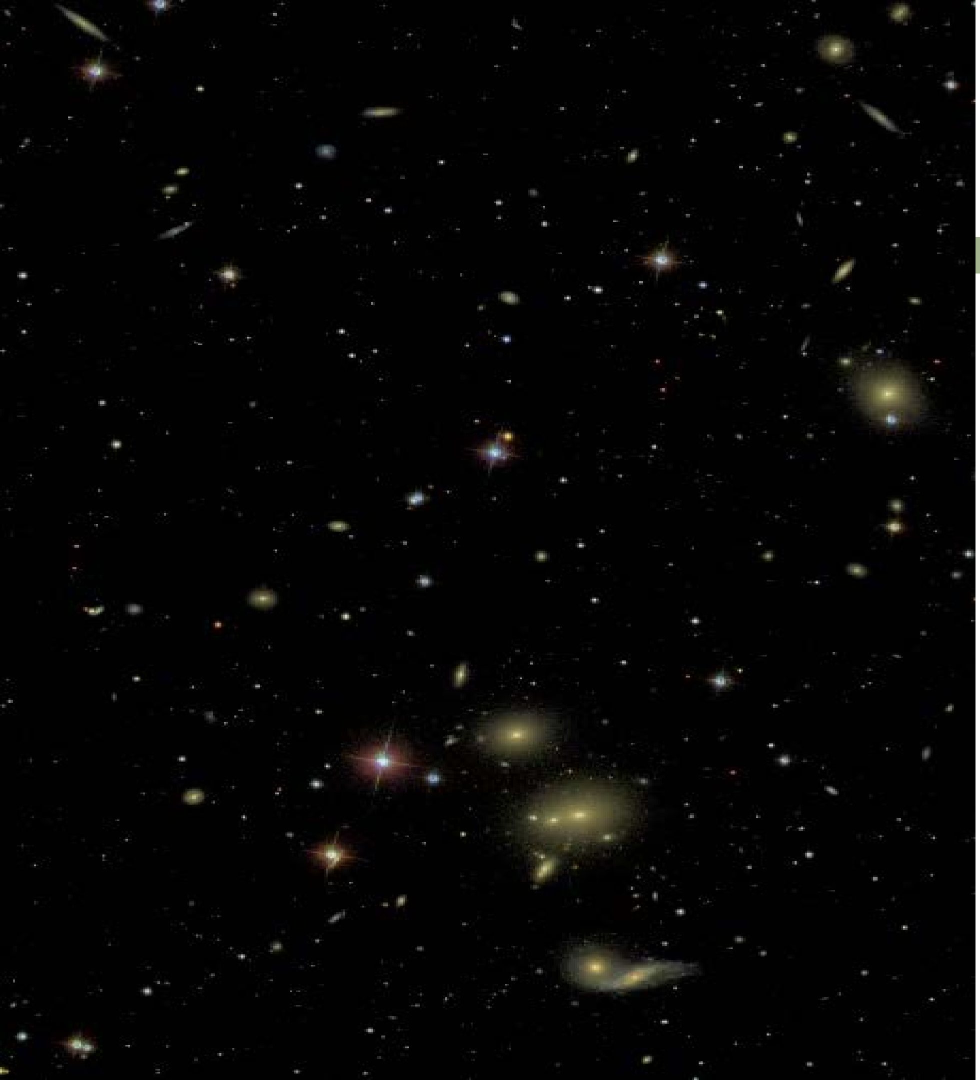
  But we used to know
  the sources by heart!

- Today large collections






Sloan Digital Sky Survey
> 200 Football Fields



Keeping Up?

- Processing pipeline
- Feature extraction
  - $O(n)$

- What is difficult?
  - $O(n \log n)$
  - $O(n^2)$, …

Worse w/ Moore's law

Sloan Digital Sky Survey

- Cosmic Genome Project 2001-2010
  - Table w/ 500M rows, 400+ cols
  - Database of 35TB
  - Software revolution in astro!
  - Astronomers learn SQL
    - Can't look at the data anymore

Science is Interactive

Too much to be accurate?

By the time you do the calculations, the answer might have changed…

Science is Interactive

- Rethink the basic methods
  - Chunks of data

$$D = \{D_1, D_2, D_3, \ldots D_N\}$$

  - Improving answers over time

Too much to be accurate?

By the time you do the calculations, the answer might have changed…

Incremental is Natural

- Bayes' rule
  - All data
  - After D_1
  - Same

$$p(\theta|D_1 D_2) = \frac{p(\theta)\, p(D_1 D_2|\theta)}{p(D_1 D_2)}$$

$$= \frac{p(\theta|D_1)\, p(D_2|\theta, D_1)}{p(D_2|D_1)}$$

$$= \frac{p(\theta)\, p(D_1|\theta)\, p(D_2|\theta, D_1)}{p(D_1)\, p(D_2|D_1)}$$

Streaming Analysis

- E.g., Mean
  - Data set

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

  - Data stream

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$$

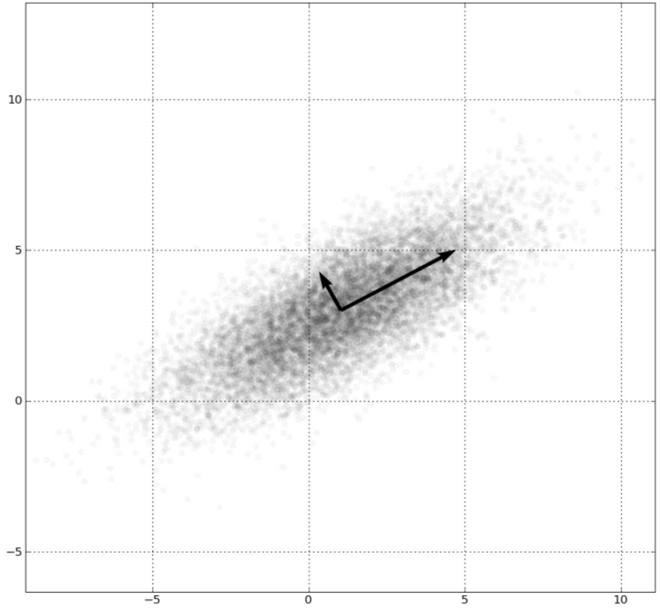
- How
  - Single pass over data
- Why
  - Low memory
  - Interactive
  - Extendable

Streaming Analysis

- Complex Analyses
  - Catalogs
  - Spectra
  - Images
- How
  - Single pass over data
- Why
  - Low memory
  - Interactive
  - Extendable

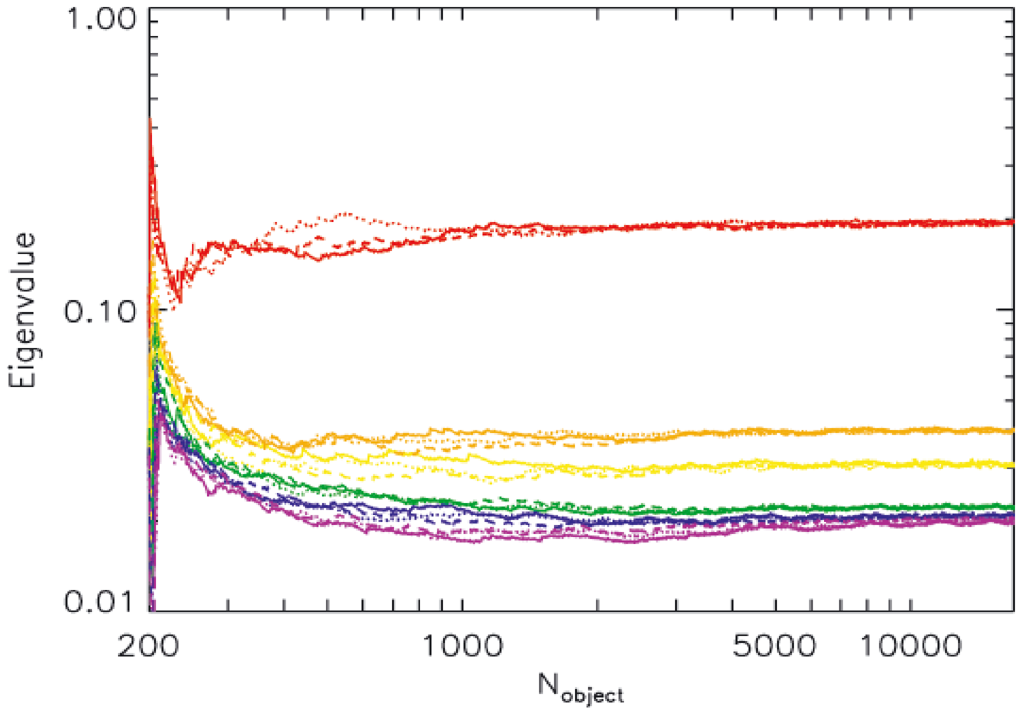
Principal Component Analysis

- Clear meaning & method
  - Directions of largest variations
  - Eigenproblem of covariances
- Issues
  - Needs lots of memory
  - Very sensitive to outliers



Monitoring Convergence

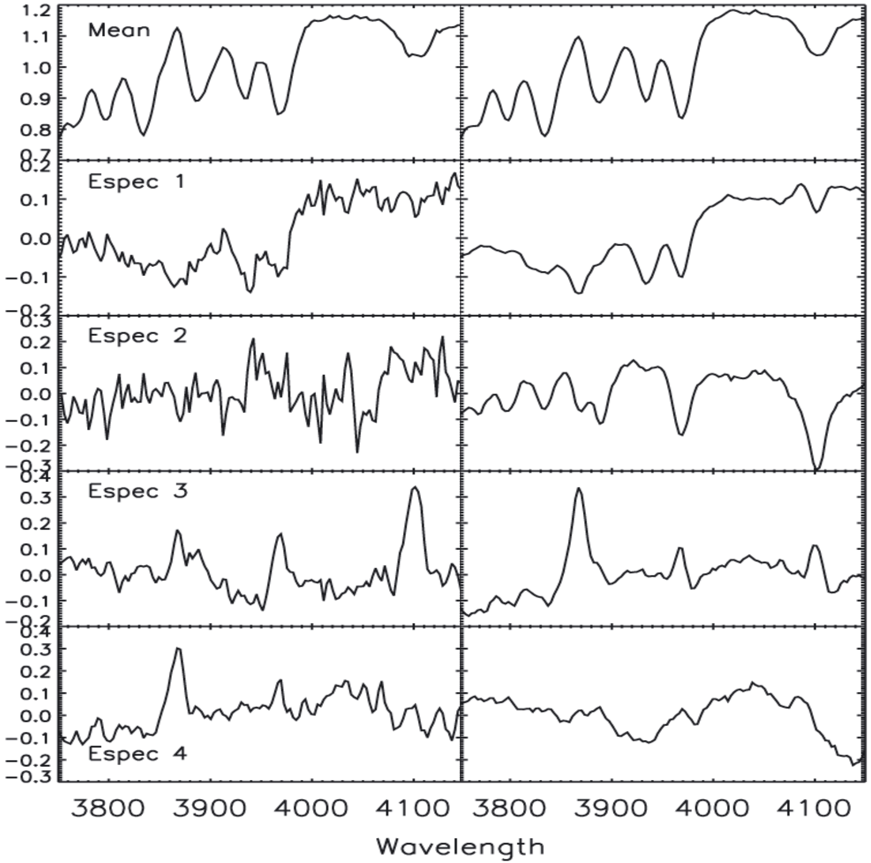
- Visual feedback
- Automatic checks
- Sublinear time
- Robust statistics



Galaxy Spectra

TB, Wild+ (2008)

- High SNR eigenfunctions
  - Sign of robustness
- Speedup on SDSS
  - From 3 days on a large-memory machine
  - To 15 mins on desktop



Source

Conference Paper: **Real time change point detection by incremental PCA in large scale sensor data**

Dmitry Mishin · Kieran Brantner-Magee · Ferenc Czako · Alexander S Szalay

[Hide abstract]

ABSTRACT: The article describes our work with the deployment of a 600-piece temperature sensor network, data harvesting framework, and real time analysis system in a Data Center (hereinafter DC) at the Johns Hopkins University. Sensor data streams were processed by robust incremental PCA and K-means clustering algorithms to identify outlier and changepoint events. The output of the signal processing system allows us to better understand the temperature patterns of the DataCenter's inner space and make possible the online detection of unusual transient and changepoint events, thus preventing hardware breakdown, optimizing the temperature control efficiency, and monitoring hardware workloads.

2014 IEEE High Performance Extreme Computing Conference, Waltham, MA USA; 09/2014

Streaming Algorithms for Halo Finders

Zaoxing Liu* [1], Nikita Ivkin* [1], Lin F. Yang† [2], Mark Neyrinck† [3], Gerard Lemson† [3],
Alexander S. Szalay† [3], Vladimir Braverman* [4], Tamas Budavari†, Randal Burns*, Xin Wang† [3]

*Department of Computer Science

†Department of Physics & Astronomy

Johns Hopkins University

Baltimore, MD 21218, USA

***Abstract*—Cosmological N-body simulations are essential for studies of the large-scale distribution of matter and galaxies in the Universe. This analysis often involves finding clusters of particles and retrieving their properties. Detecting such "halos" among a very large set of particles is a computationally intensive problem, usually executed on the same super-computers that produced the simulations, requiring huge amounts of memory.**

Recently, a new area of computer science emerged. This area, called streaming algorithms, provides new theoretical methods to compute data analytics in a scalable way using only a single pass over a data sets and logarithmic memory.

I. Introduction

The goal of astrophysics is to explain the observed properties of the universe we live in. In cosmology in particular, one tries to understand how matter is distributed on the largest scales we can observe. In this effort, advanced computer simulations play an ever more important role. Simulations are currently the only way to accurately understand the nonlinear processes that produce cosmic structures such as galaxies and patterns of galaxies. Hence a large amount of effort is spent on running simulations modelling representa-

Streaming Algorithms for Halo Finders

Zaoxing Liu* [1], Nikita Ivkin* [1], Lin F. Yang† [2], Mark Neyrinck† [3], Gerard Lemson† [3], Alexander S. Szalay† [3], Vladimir Braverman* [4], Tamas Budavari†, Randal Burns*, Xin Wang† [3]

**Department of Computer Science*

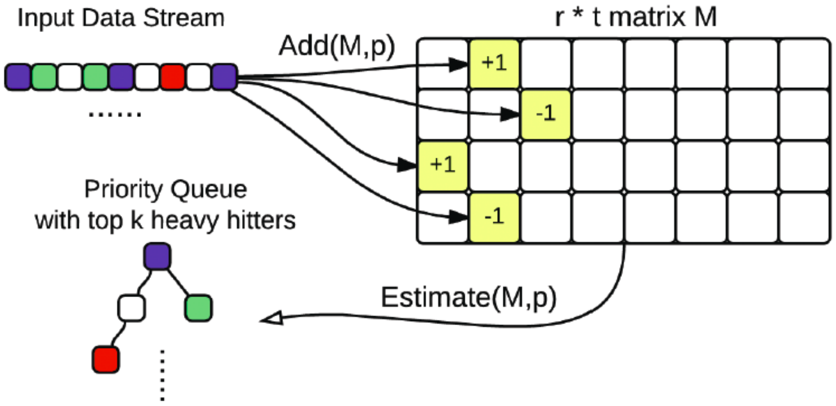
†Department of Physics & Astronomy

Johns Hopkins University

Baltimore, MD 21218, USA

***Abstract*—Cosmological N-body simulations are essential for studies of the large-scale distribution of matter and galaxies in the Universe. This analysis often involves finding clusters of particles and retrieving their properties. Detecting such "halos" among a very large set of particles is a computationally intensive problem, usually executed on the same super-computers that produced the simulations, requiring huge amounts of memory.**

Recently, a new area of computer science emerged. This area, called streaming algorithms, provides new theoretical methods to compute data analytics in a scalable way using only a single pass over a data sets and logarithmic memory.



The Count-Sketch Algorithm

Optimal Image Coaddition

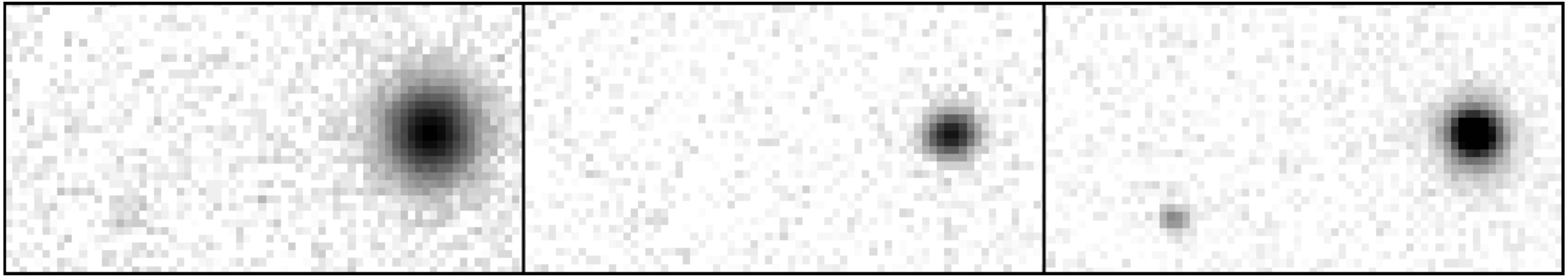
Matthias Lee

Multiple Exposures

- Each observation
  - Low Signal-to-Noise
  - Blury
  - Variable Quality

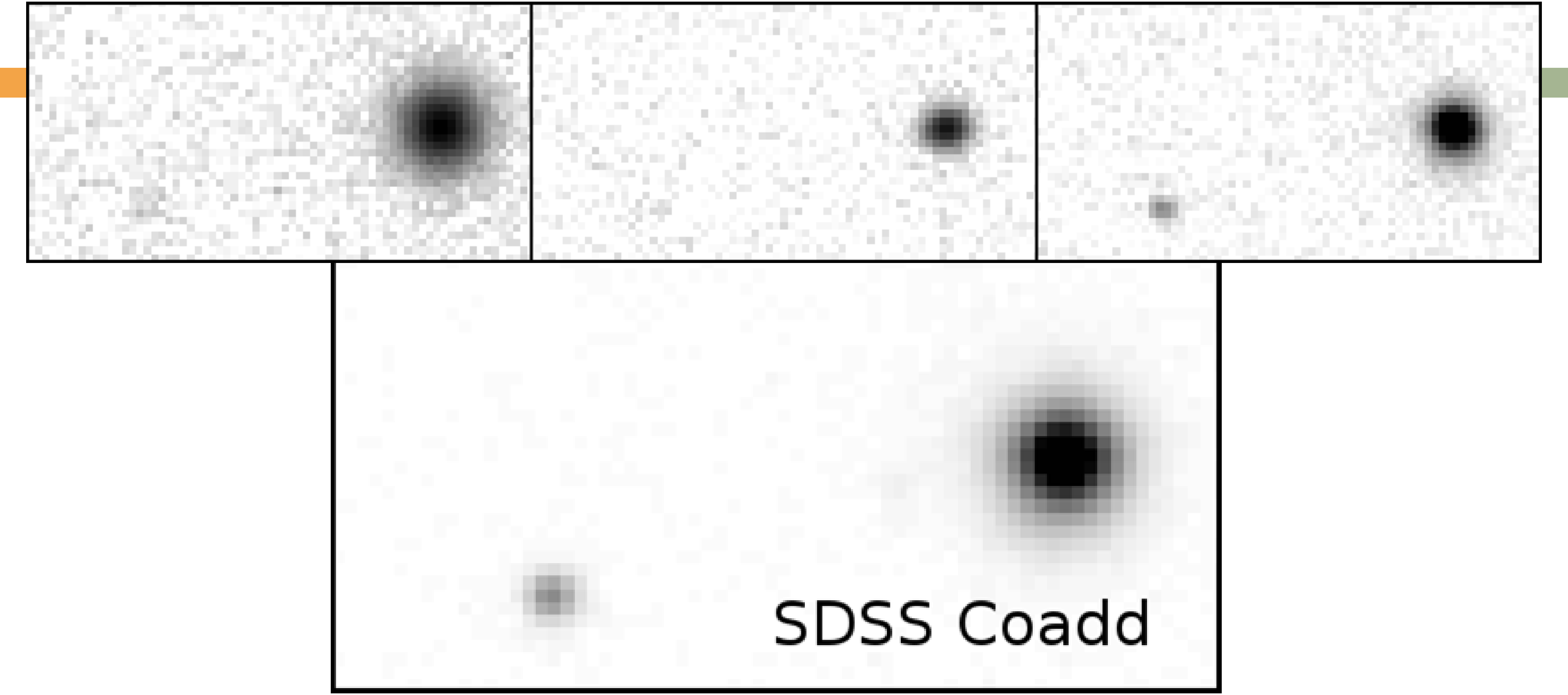
BIG DATA !

SDSS FRAMES



Traditional Solutions

- Lucky Imaging
  - Keep only the best/sharpest images
  - Discard majority of exposures
- Coadding
  - Higher Signal-to-Noise Ratio
  - Worst acceptable PSF


SDSS Coadd

Our Goal

- Improved quality
  - Best signal-to-noise ratio
  - Sharper & deeper images
  - Even higher resolution

Computational Optics

Single Frame

- Correcting Hubble optics; Richardson-Lucy deconvolution
- White (1994), Starck+ (1994), Fruchter+ (1997), Fish+ (1995)
- Degeneracies due to limited information

Multiple Frames

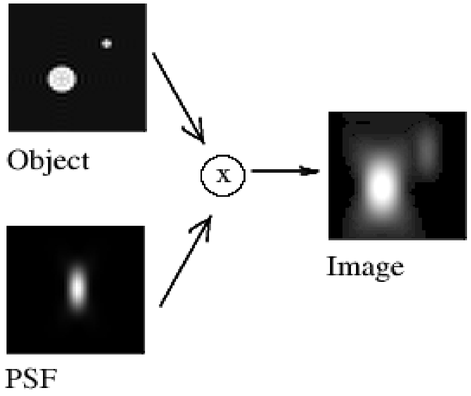
- Harmeling+ (2009, 2010)
- More data for inference

Simple Model for Exposures

- Latent "true" image convolved with unknown point-spread functions
- Plus noise

- Simultaneous solution?

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



Blind Deconvolution

- We solve for the true image & all PSFs
  - Gaussian likelihood function yields quadratic minimization

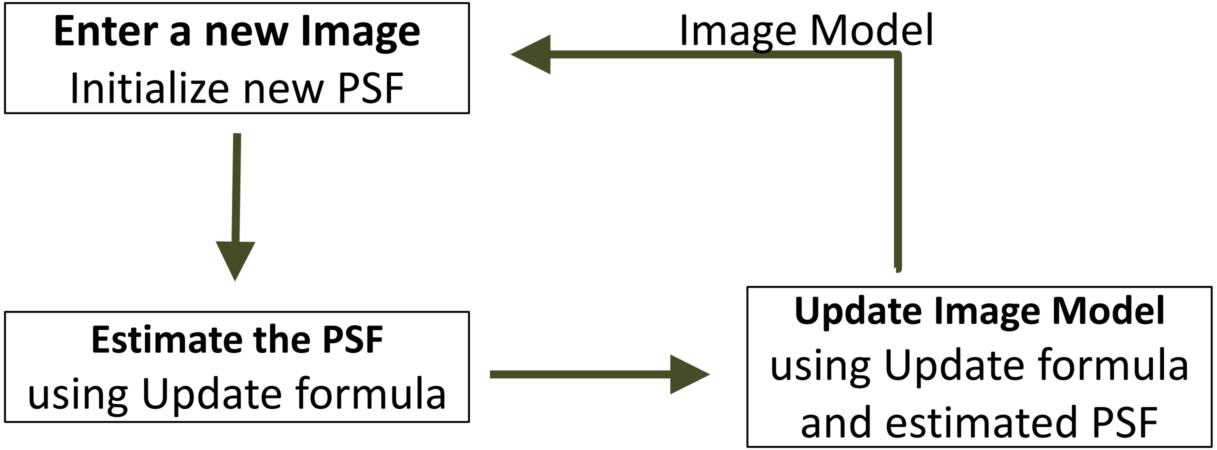
$$\left|y_t - Fx\right|^2$$

  - Multiplicative updates
    cf. Richardson-Lucy

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

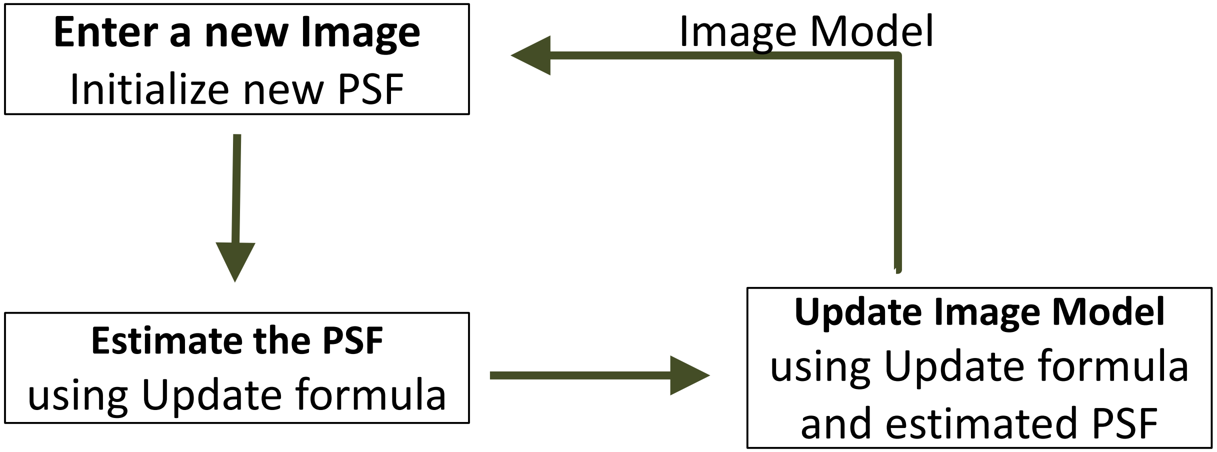
Multi-Frame Blind Deconvolution

- General iterative approach:



Multi-Frame Blind Deconvolution

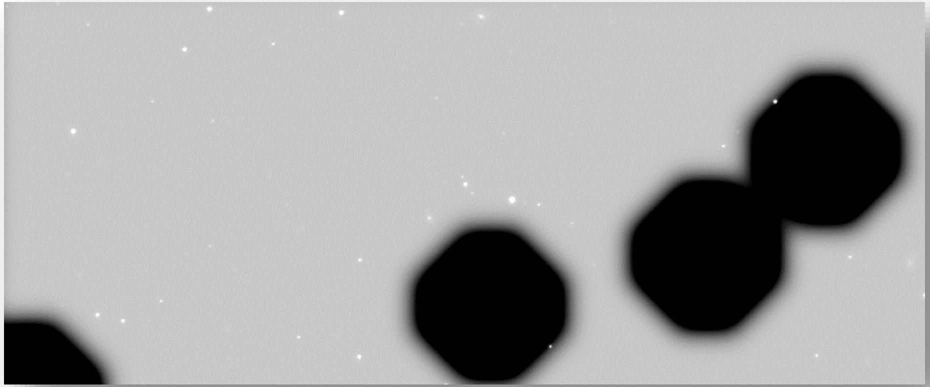
- General iterative approach:



The devil is in the details!

Masking Pixels

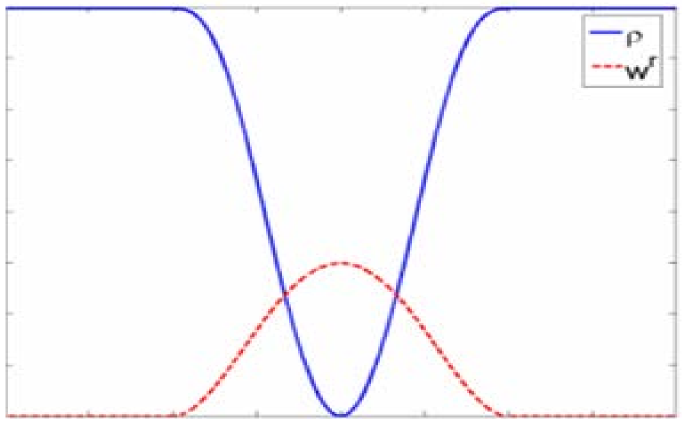
- Ignore gaps as well as bad & saturated areas



- But we solve for missing parts, too!

Robust Statistics

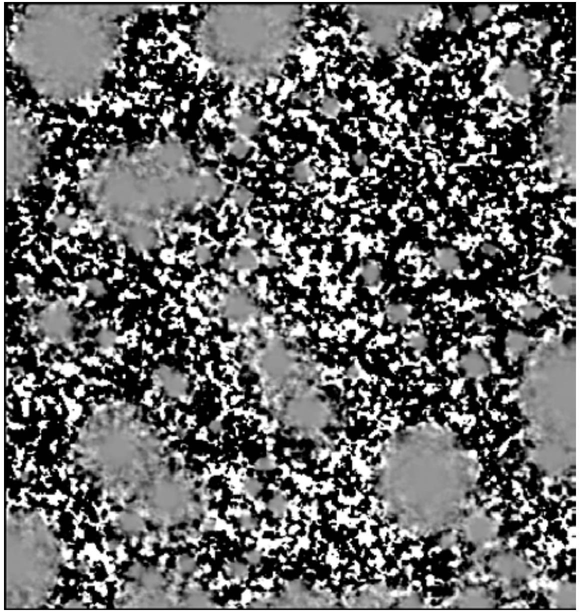
- Quadratic cost function is dominated by bad pixels
  - Bad convergence across images
- Robust $\rho(r)$ instead of r^2
  - Quadratic for small residuals
  - Limited where bad
- Simple weighting
  - Integrates with streaming



Careful Updates

- Artifacts from nowhere
  - Large updates of small values
- Limit the influence of updates
  - Say, no more than 2x

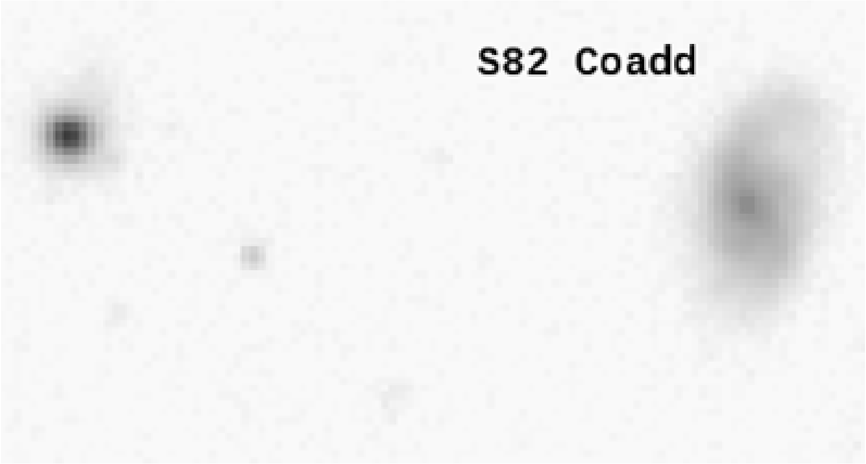
$$u'_t = \max\left\{1/d,\ \min(d,\ u_t)\right\}$$



Coadded & Reconstructed

- Coadding
  - Brings out faint sources
  - But blurs the images
- We deconvolve
  - Sharper
  - Deeper

Coadded Image

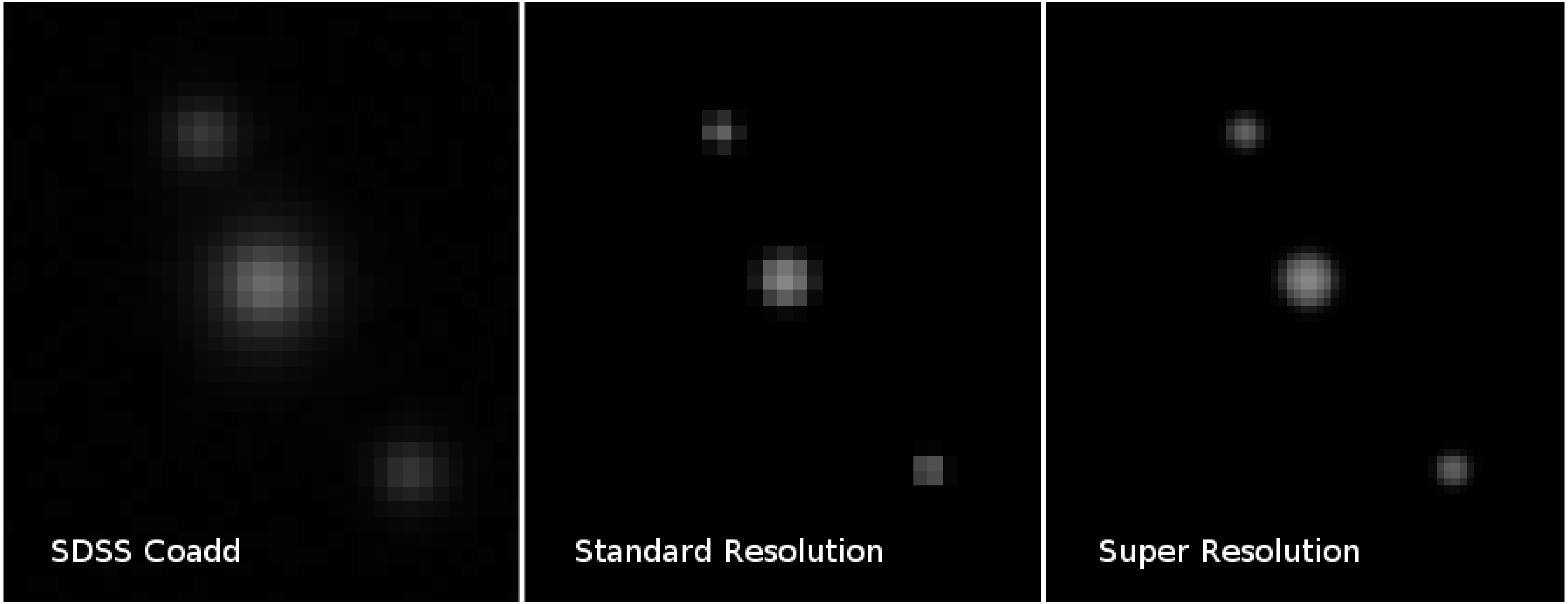


Coadded & Reconstructed

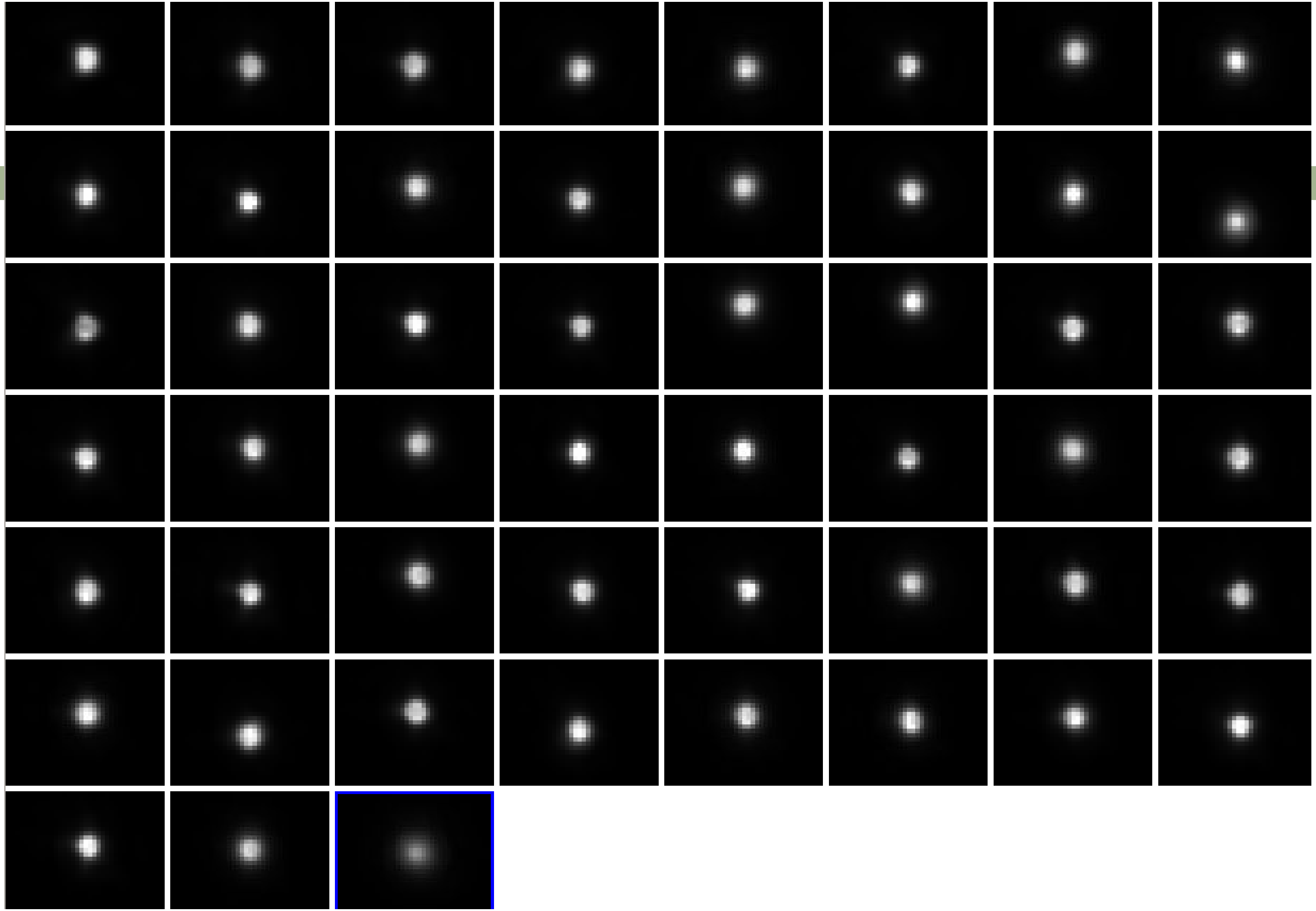
- Coadding
  - Brings out faint sources
  - But blurs the images
- We deconvolve
  - Sharper
  - Deeper

Deconvolved Image

Super Resolution



PSFs



Next Steps

- Works great on GPUs
  - 140 images (2k × 2k) under 5 mins
  - 140 images (4k × 4k) in 10 mins
- Pipeline for real surveys
- Fit for sky background

Summary

- Streaming and randomized algorithms can help
  - Reduce memory requirements of big analyses
  - Provide the best solution within the given time
  - Integrate with intuitive improvements
- Promising applications – ready for next-gen?

KEEP ASTRONOMY INTERACTIVE!