XXX Canary Islands Winter School, Nov 5-9, 2018.

DiRAC

Challenges in Image Processing

Turning Petabytes of Data into Terabytes of Information

Prof. Mario Juric
DIRAC Institute | eScience Institute | UW Astronomy

With thanks to Andy Connolly, Robert Lupton, Ian Sullivan, David Reiss, Jake vanderPlas + LSST Project and Collaborations

Data Intensive Research in Astrophysics and Cosmology
College of Arts & Sciences | University of Washington

Context

Were used to performing analysis in catalog space, using objects and observations. But how those catalogs are created is a mystery to most astronomers.

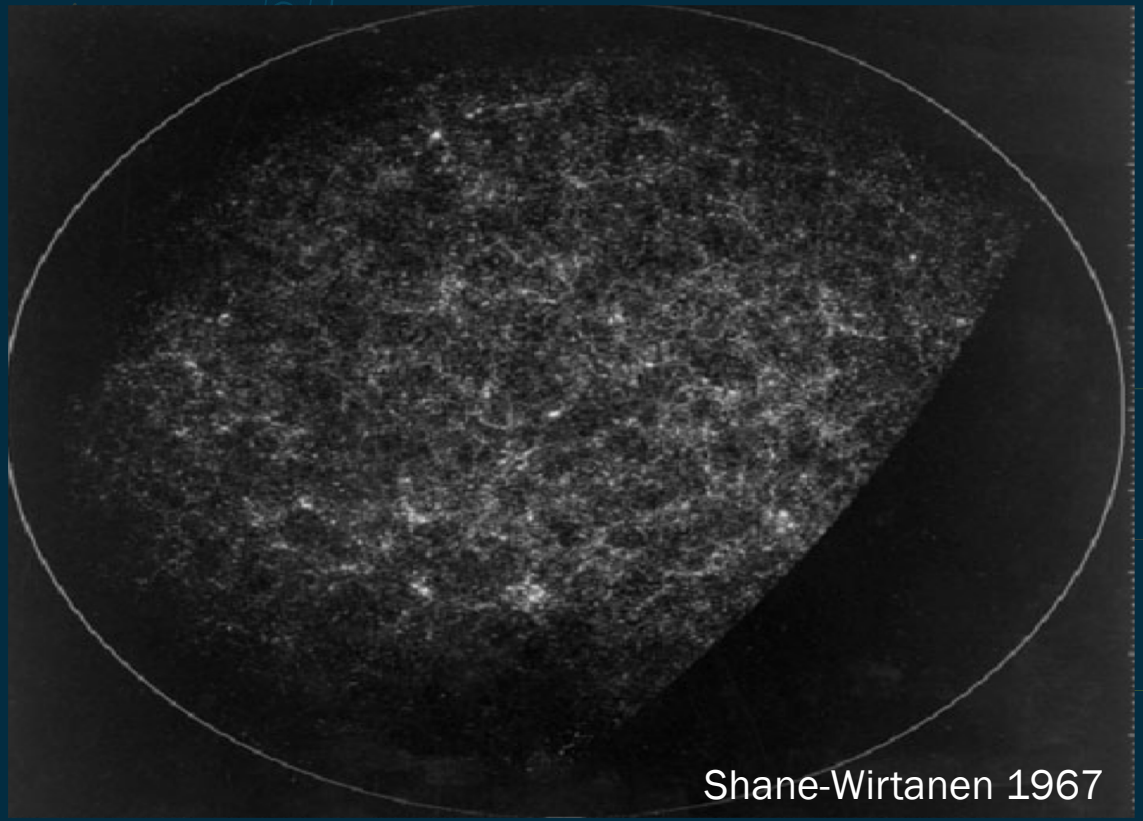
Yet science need push us towards:

- Measurements of subtle (10-6) effects
- Ever higher precision (<1% photometry, ~mas astrometry)
- Ever more data (and thus the need for automated operation)

At these precision levels it becomes critical to understand our instruments and the measurement process. Otherwise, the instrumental noise (bias) may be misinterpreted for a signal.

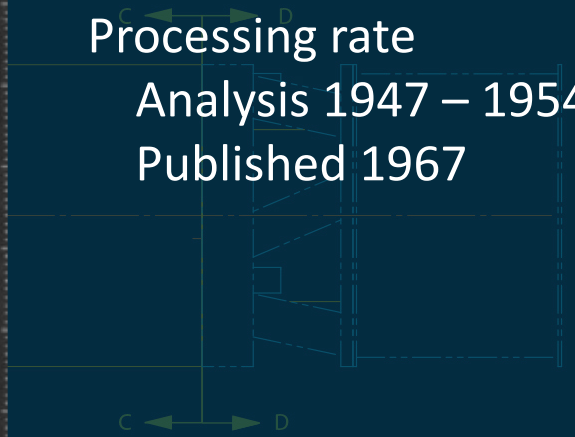
This lecture is a high-level walk-through to give you a feel for how we turn images to catalogs, and the associated challenges today and in the future.

An analog universe (1950s)



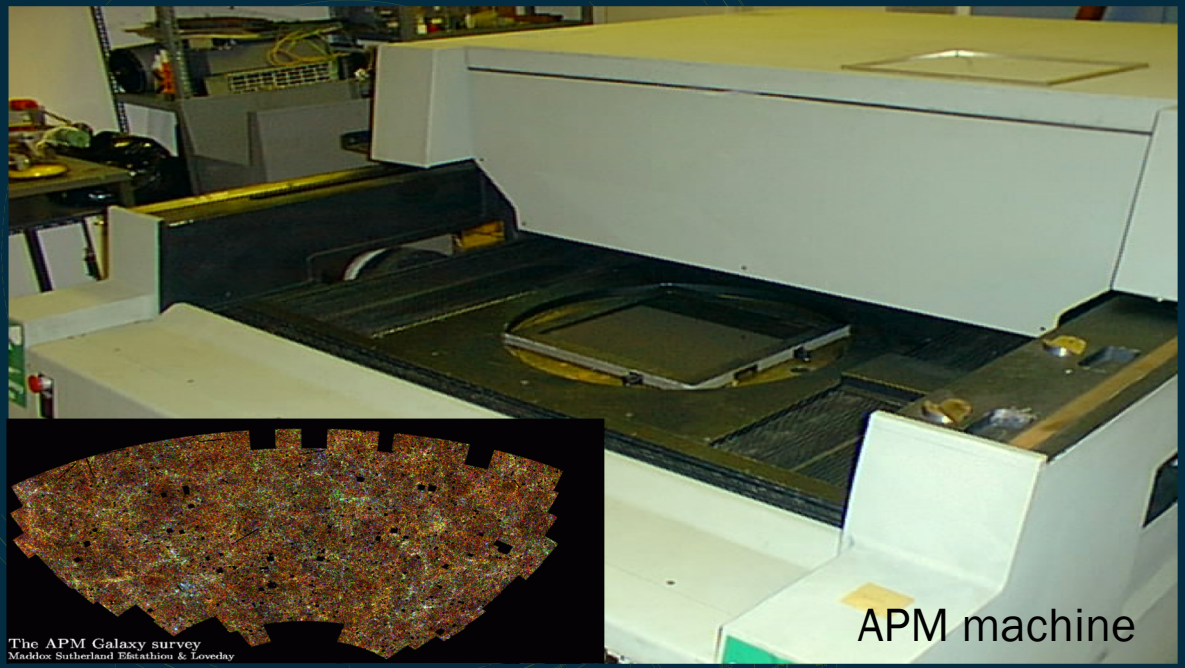
Visual inspection:
- Lick survey, 1246 plates
- 1.6 million “pixels”
- 10x10 arcmin pixels

Processing rate
- Analysis 1947 – 1954
- Published 1967



Map of galaxy counts down to V<19

A digitized universe (1980's)

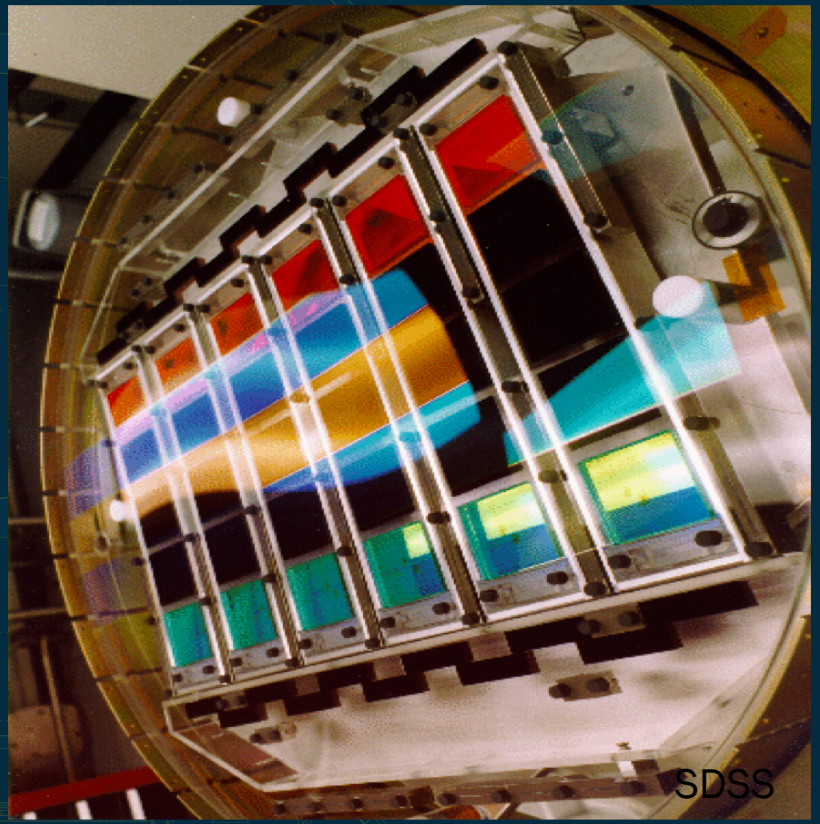


Microdensitometers:
- 10^{12} pixel sky
- 0.5 arcsec pixels

Processing rate
- 4 hours a plate
- 200 Mhz Pentium-pro

Automatic Plate Measuring facility in Cambridge, UK.

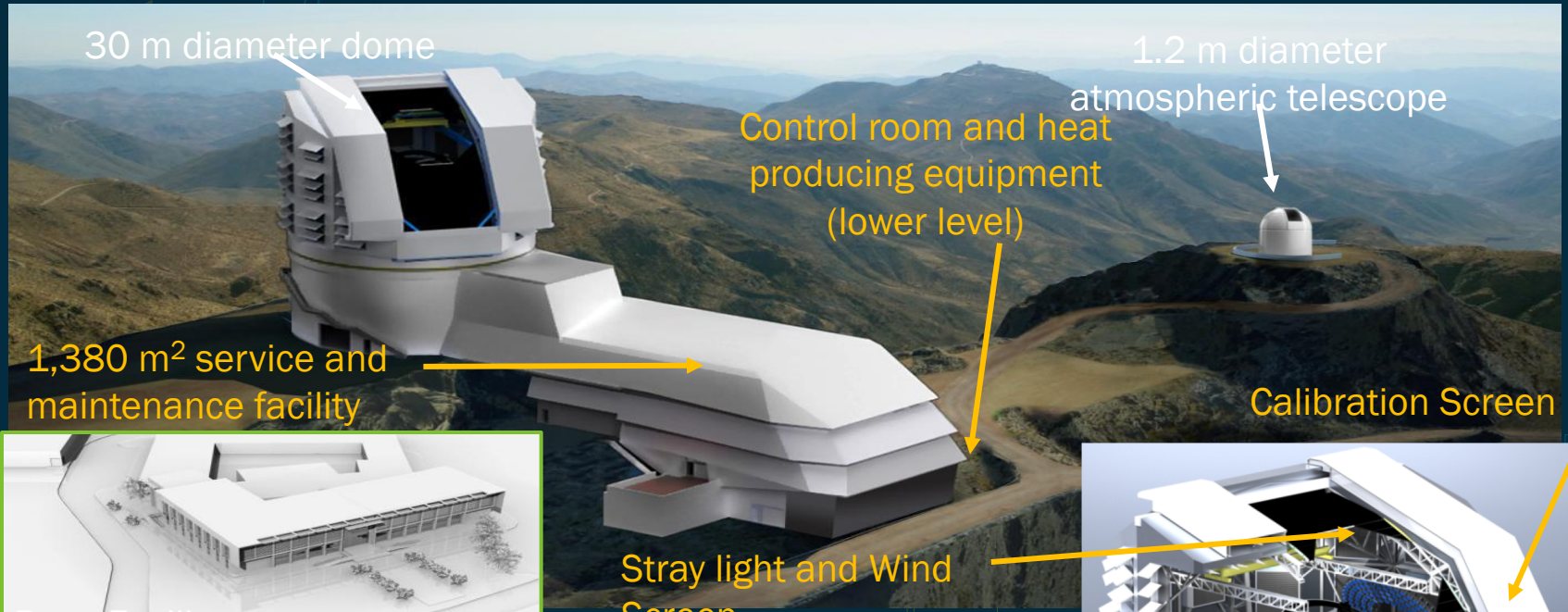
A digital universe (2000s)



Mosaic Cameras
- 10^{11} pixels sky
- 0.45 arcsec pixels

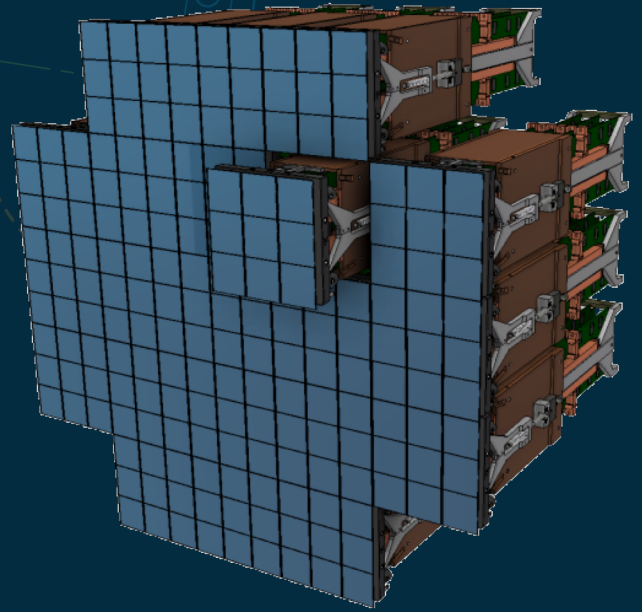
Processing rate
- 4 MB/s
- 250K lines of code

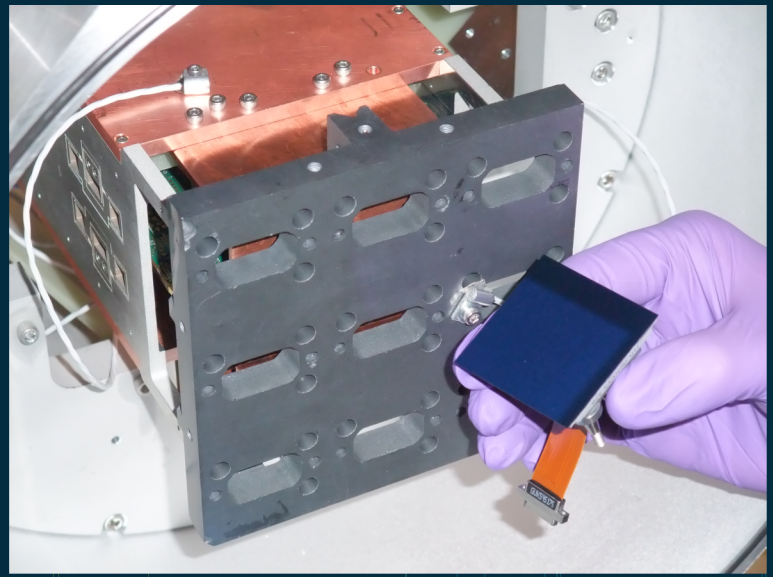
The Large Synoptic Survey Telescope



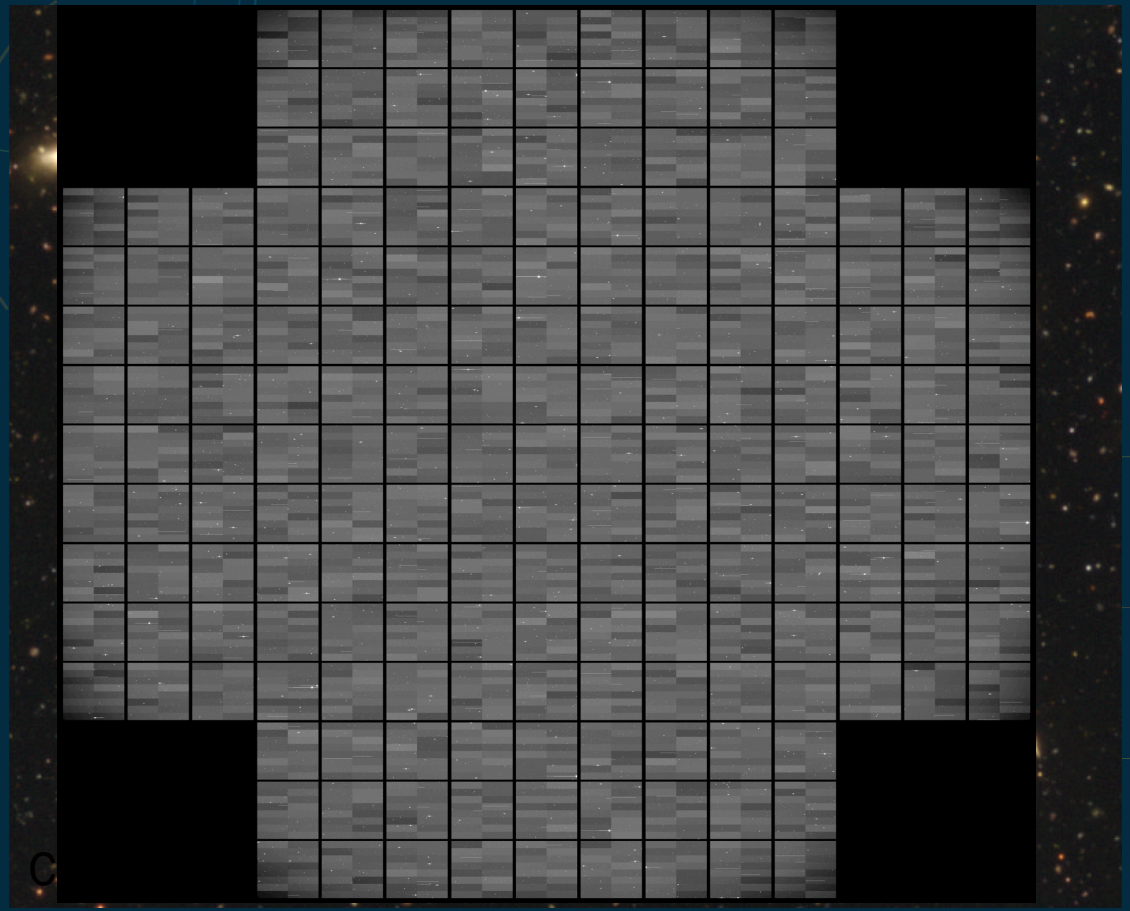
A project to build the biggest optical astronomical survey in existence, bringing data to topics ranging from the Solar System to Dark Energy.

LSST camera: A 3.2 Gigapixel camera





Modular design: 3200 Megapix = 189 x16 Megapix CCD
9 CCDs share electronics: raft (21=camera)
100 μm deep depletion devices (10 μm pixels)



Sensors at scale:

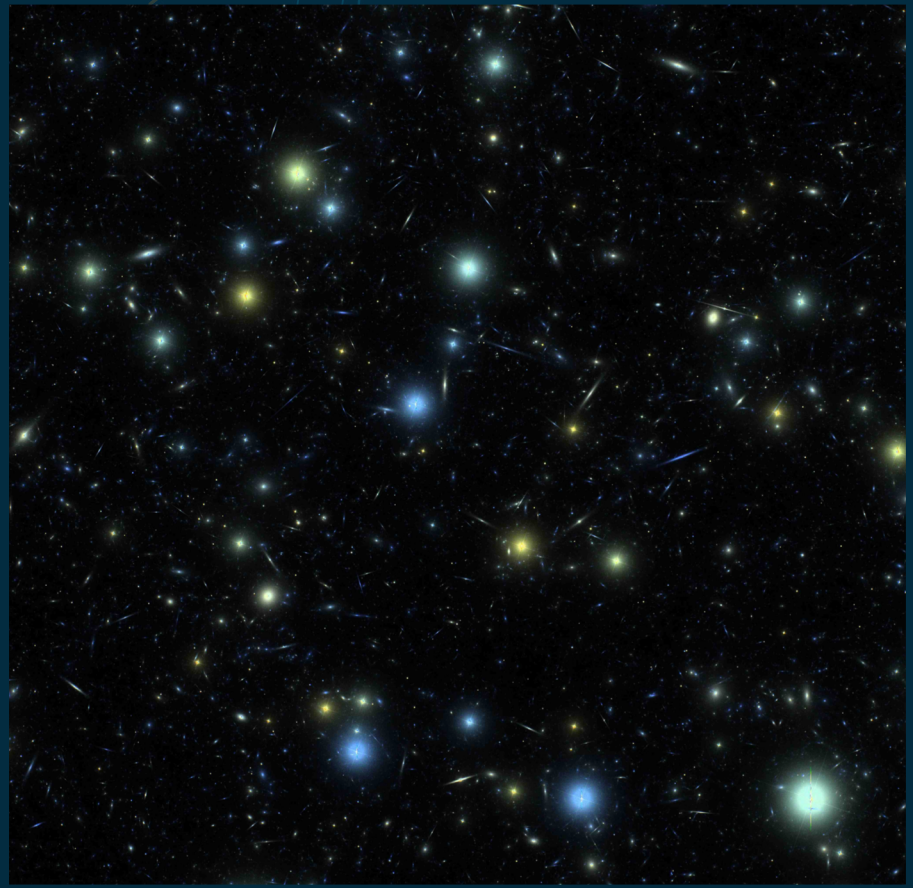
- 6.5×10^9 pixels sky
- 0.2 arcsec pixels

Processing rate

- 170 MB/s
- 1.1M lines of code

Credit: John Peterson

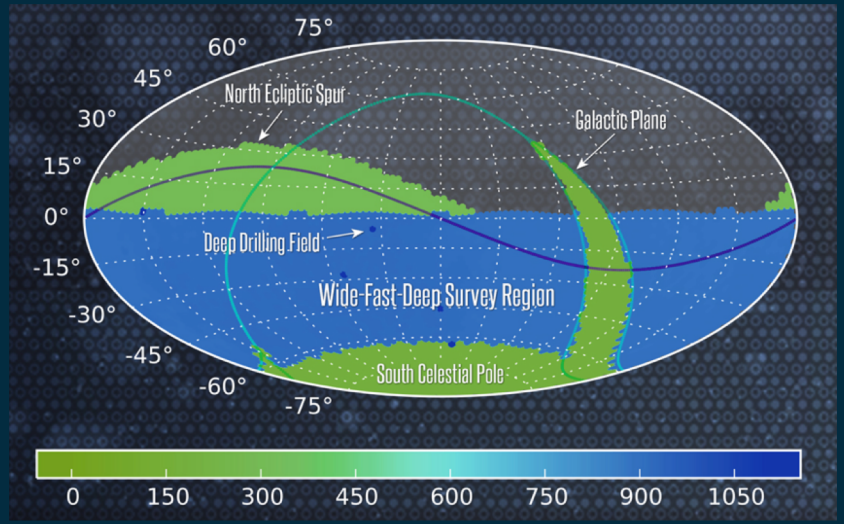
A Big Data Universe



1 chip (4kx4k, 18 bits/pixel), 0.5% of the full image.

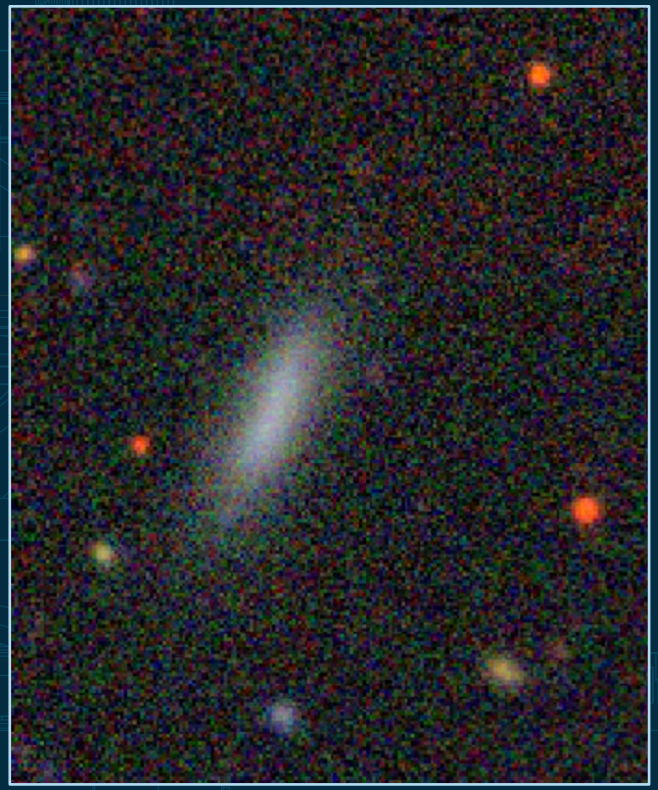
Expect ~2000 exposures per night, 300 nights a year, for 10 years.

Roughly 4 PB of raw imaging data per year.

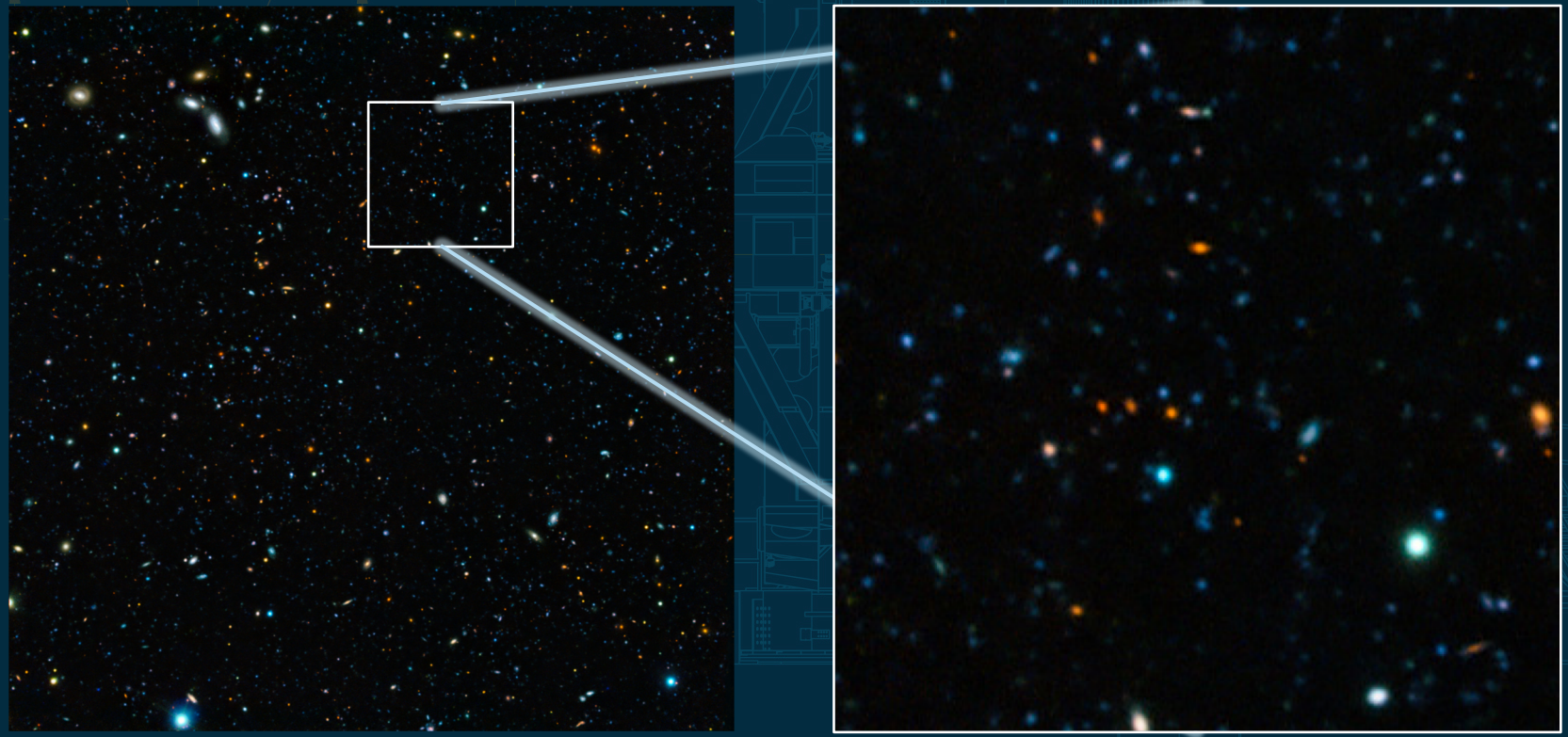


What do astronomers care about?

- What's on the image?
  - Stars (point sources)
  - Galaxies (extended objects)
- Where is it?
  - Relatively (in pixels, to ~few hundredths of a pixel)
  - Absolutely (coordinates on the sky)
- How bright is it?
- Is it changing in time?
- Is it moving?
- What is its shape?
  - Of a particular object
  - Statistically, for a class of objects



Understanding Astronomical Images



GOODS-South Field, deep VLT stack

How do astronomical images come to be?

Credit: John Peterson (Purdue) and the PhoSim Team

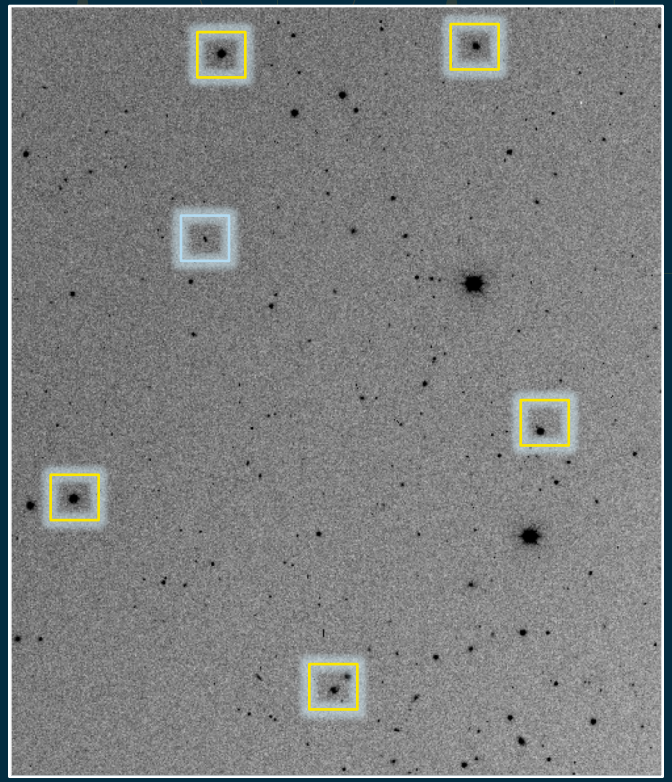

Optics
+Tracking
+Diffraction
+Detector Misalignments & Perturbations
+Lens Misalignments
+Mirror Misalignments Perturbations, & Micro-roughness
+Detector
+High Altitude Atmosphere
+Mid Altitude Atmosphere
+Low Altitude Atmosphere
+Pixelization
+Saturation & Blooming

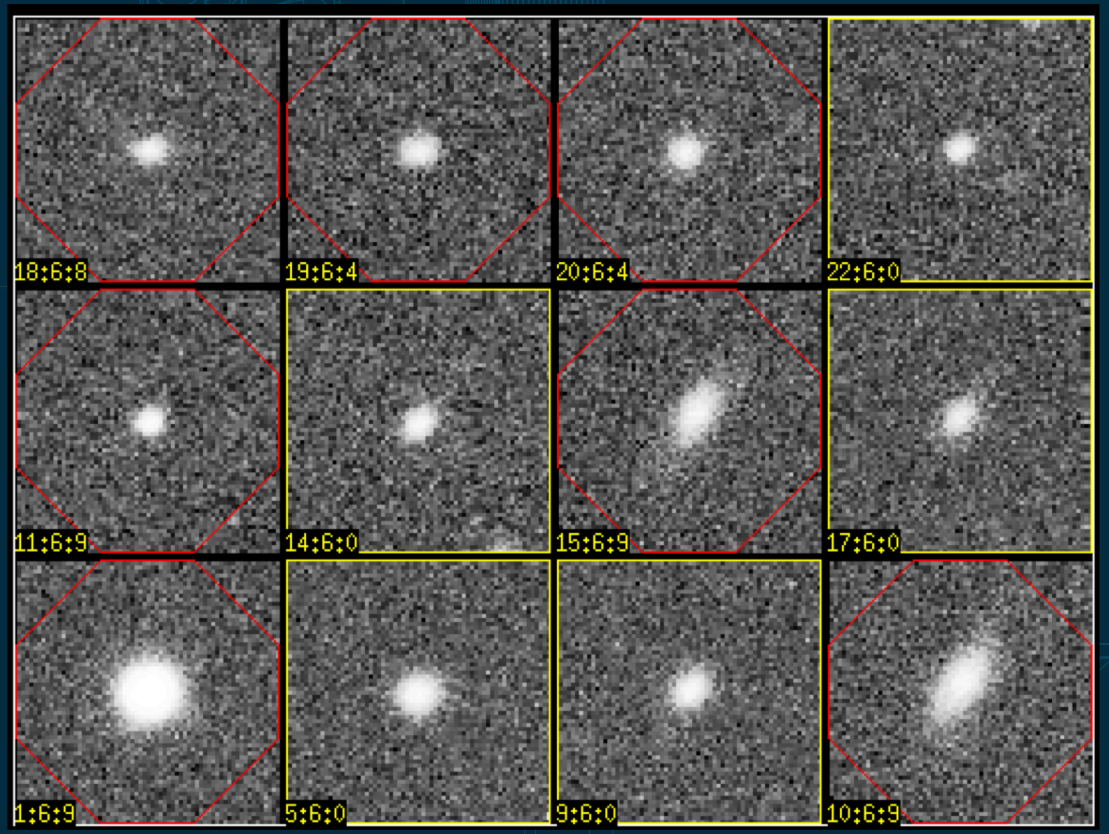
Astronomical Images (approximated)

$$\boldsymbol{I}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x}) \otimes \boldsymbol{S}(\boldsymbol{x}) + \boldsymbol{\varepsilon}(\boldsymbol{x})$$

Observed PSF Truth Noise

Point Spread Function Estimation (Today)





Sampling, modeling, interpolation

Challenge: How could we do better?

Utilize all available information.

The PSF is a combination of deterministic and stochastic elements.

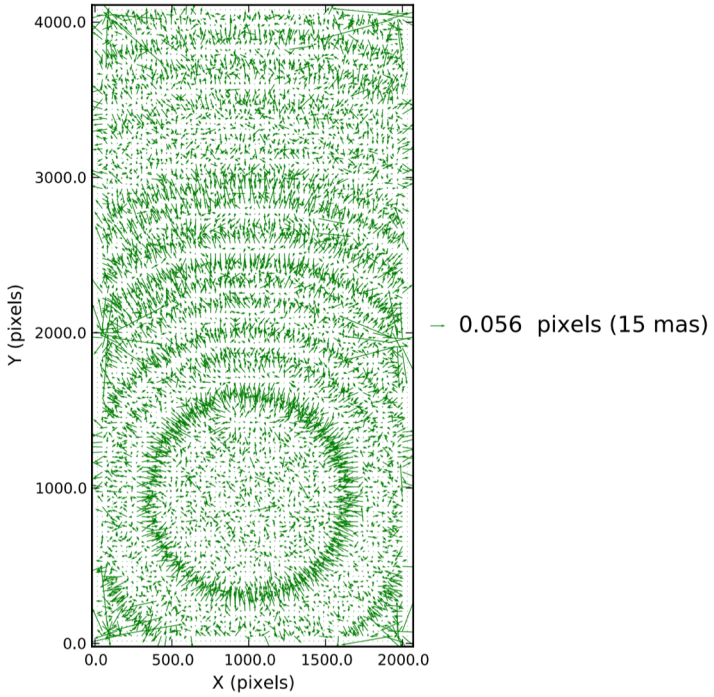
Deterministic: Camera, Optics
Stochastic: Atmospheric perturbations

Directly measure the deterministic elements.

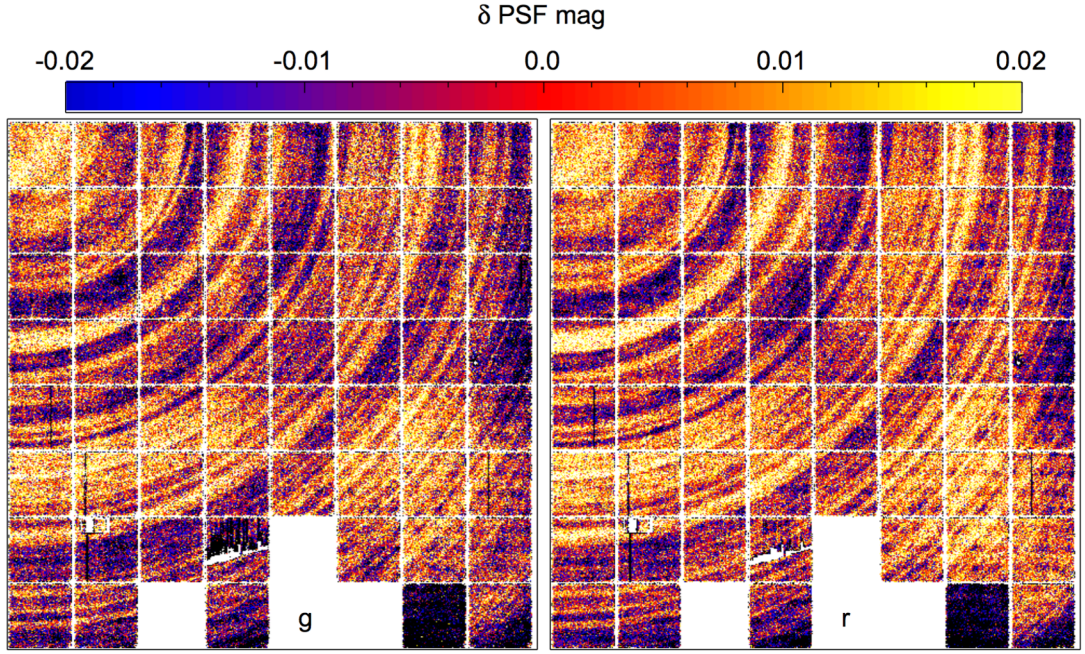
Constrain the stochastic elements (e.g., using the knowledge of the power spectrum of fluctuations in the atmosphere).

e.g., Jee & Tyson 2011

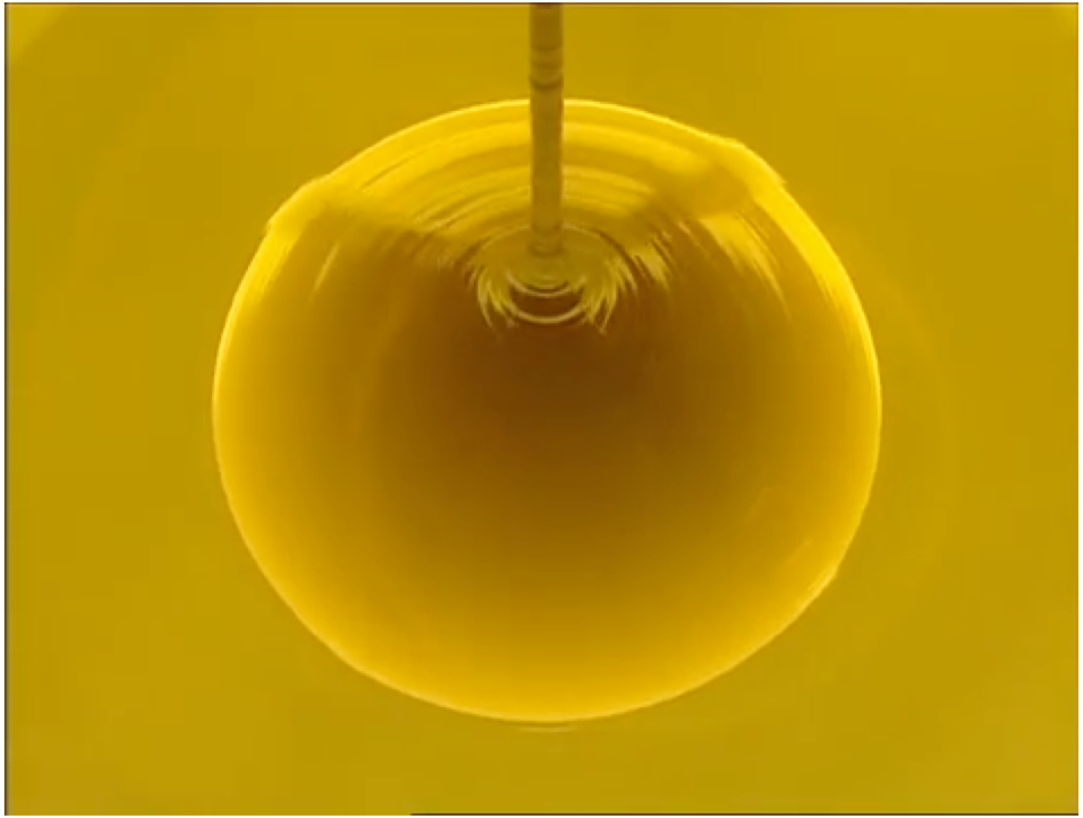
Life is not so Simple: Tree Rings



Plazas, Bernstein & Sheldon 2014



Magnier et al. 2017



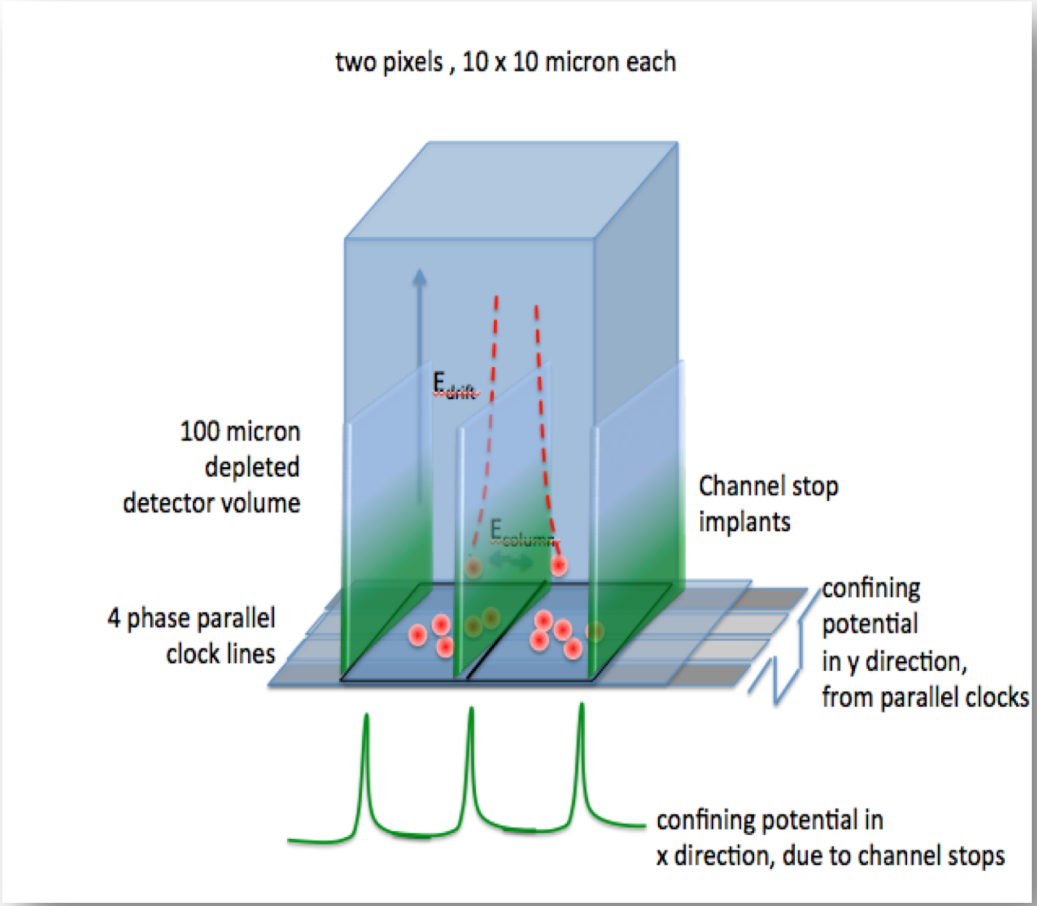
https://www.youtube.com/watch?v=jh2z-g7GJxE

Charge Diffusion

In LSST sensors, one sees tree rings variations at at a ~percent level.

Varying dopant density in silicon boules creates parasitic lateral E fields.

These DO NOT behave as QE variations. Naïve flat-fielding makes the problem WORSE by ~2x.



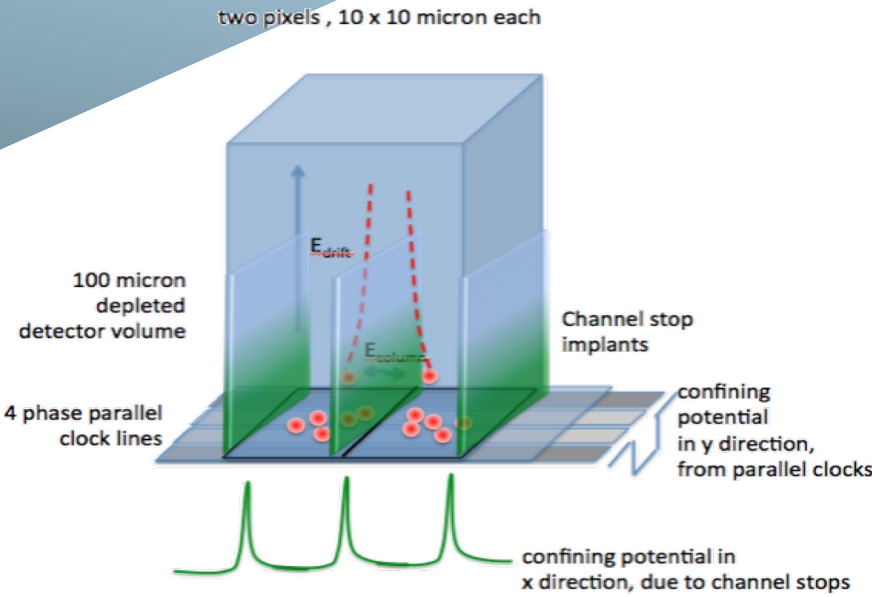
Nonlinearities: the Brighter-Fatter Effect

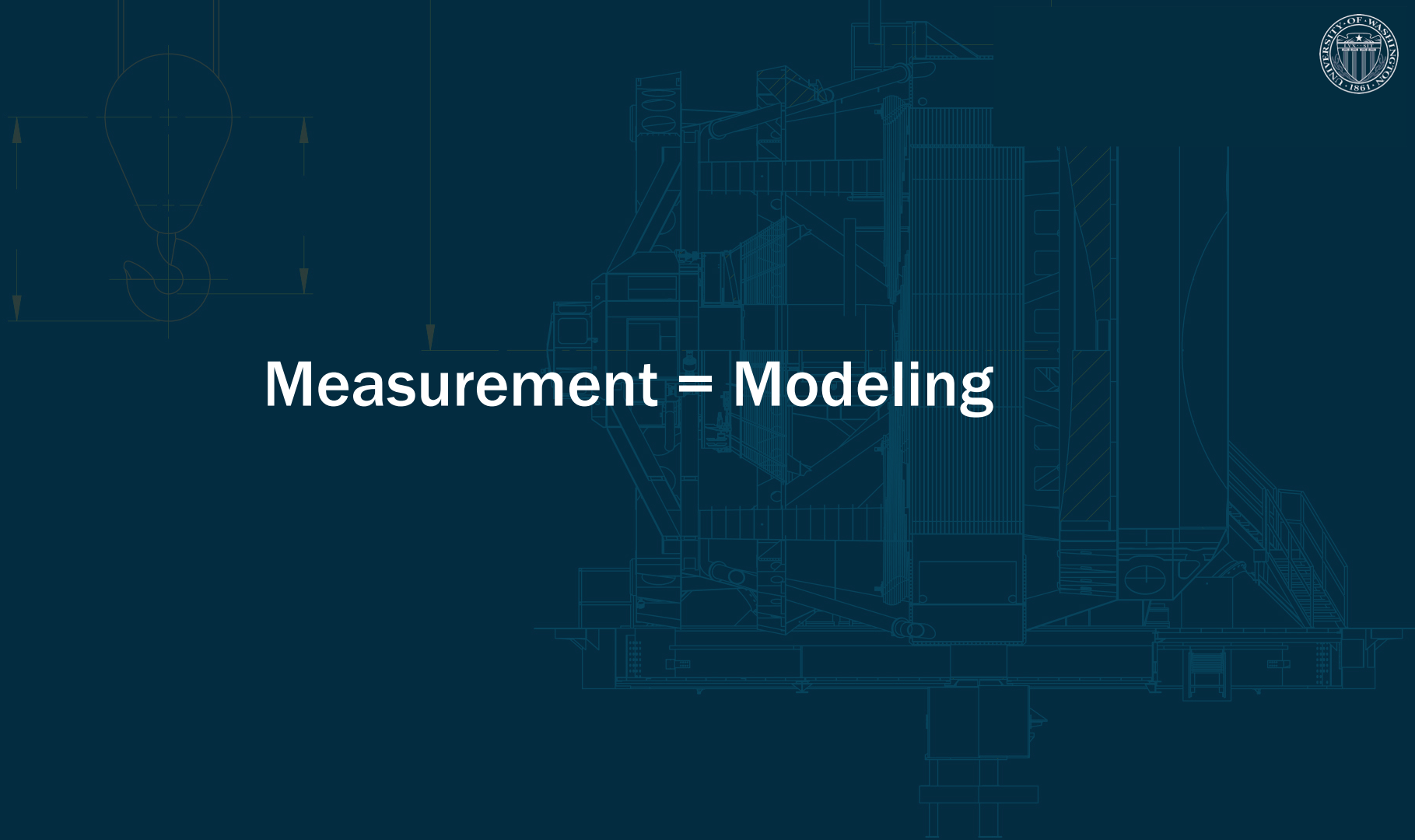
Most of today's devices (DES, HSC, LSST, GPC1) are thick. The photon converting at the top has a long way to go to reach the bottom

- Tree rings (and related effects)
- "Brighter-fatter" effect

As the potential wells fill up with electrons, the bias voltage drops making it easier for electrons to be diverted to neighboring pixels.

Correlates the values of neighboring pixels; results in an intensity-dependent PSF.





Measurement = Modeling

Deconvolve?

$$I(x) = \phi(x) \otimes S(x) + \varepsilon(x)$$

Observed PSF Truth Noise

If we can estimate the PSF, can we simply deconvolve?

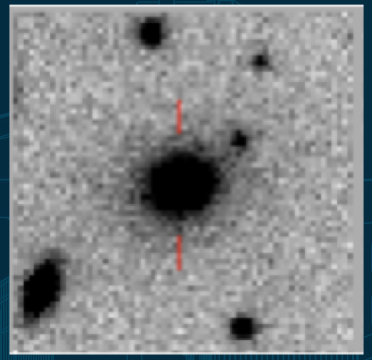
No. (at least not easily.)
Deconvolution without regularization amplifies noise.
Errors in the estimate of the PSF are further amplified in deconvolved images.

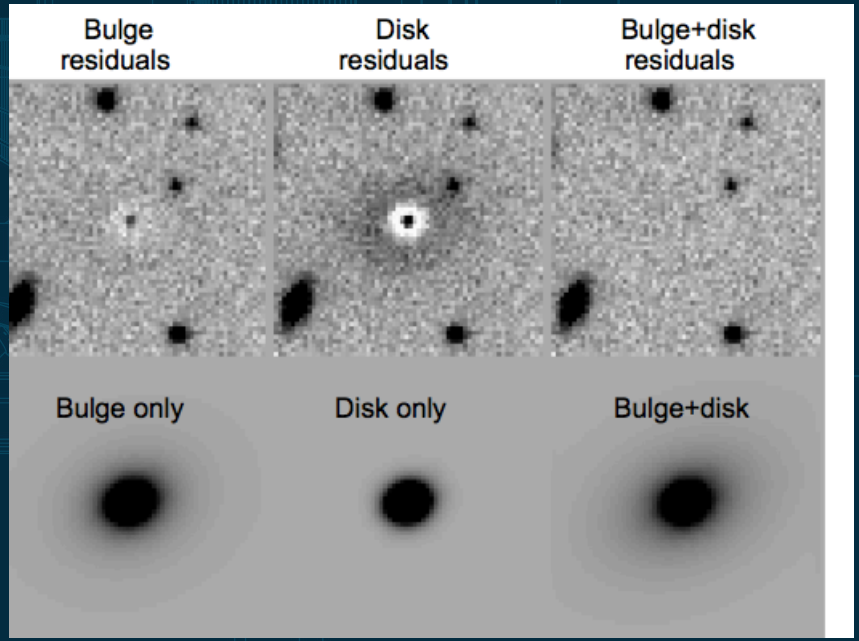
Forward Modeling

$$I(x) = \phi(x) \otimes S(x) + \varepsilon(x)$$

Observed PSF Truth Noise

$$S(x) = \sum M_i(x, y)$$

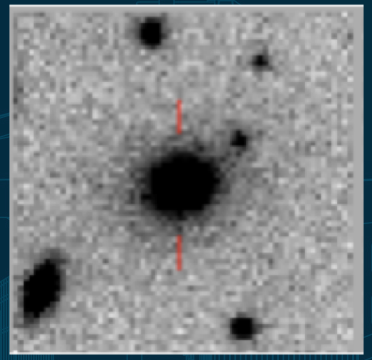




Object characterization Models:

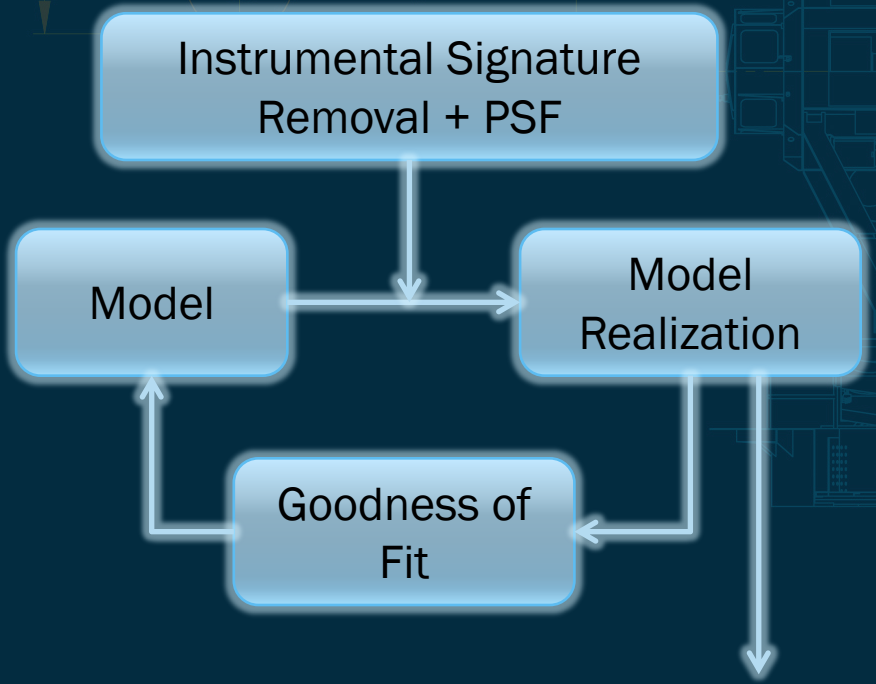
- Stars: Point Source model
- Galaxies: Double Exponential models

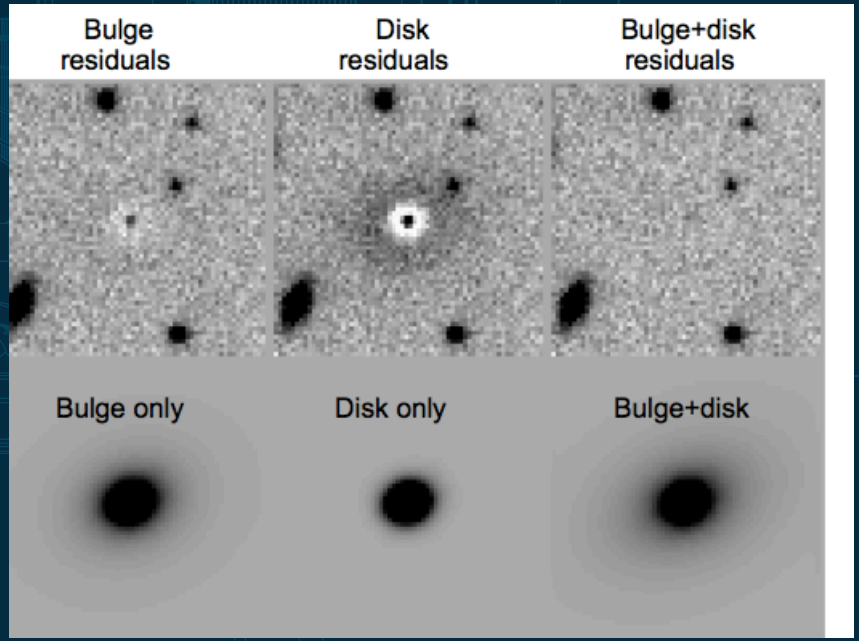
Modeling object properties



Object characterization (models):

- Stars: Point Source model
- Galaxies: Double Exponential models





1,231,051,050 rows (SDSS DR10, PhotoObjAll table)

~500 columns

Cataloging the Sky...

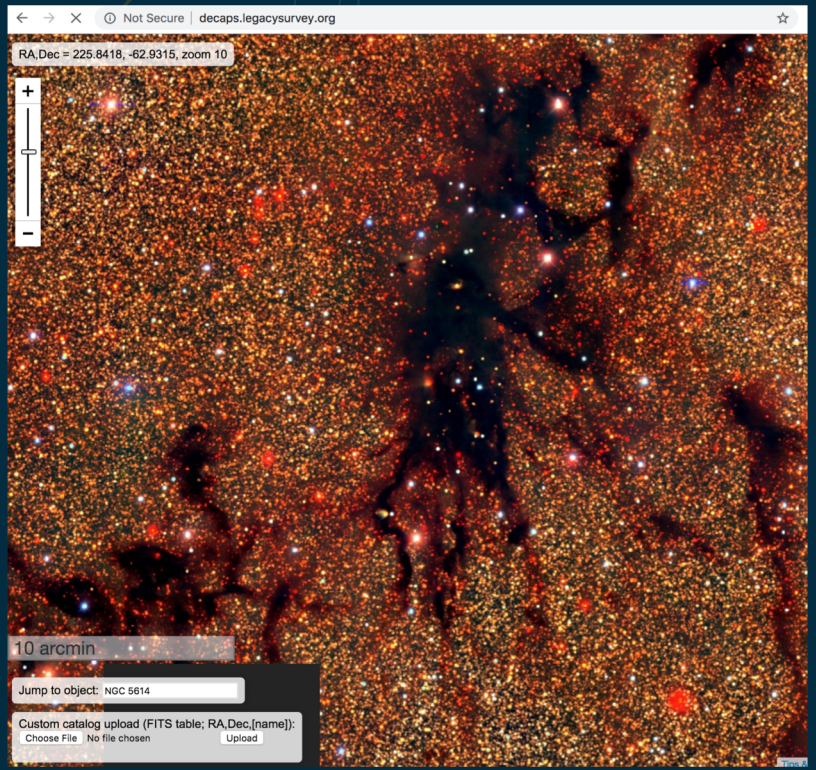
What we're doing is decomposing and modeling the sky in a way that makes physical sense.

… Compressing the Sky

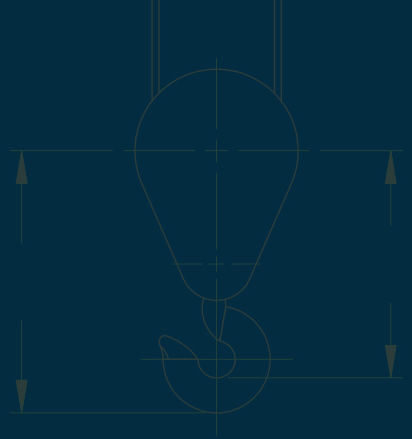
What we're doing is decomposing and modeling the sky in a way that makes physical sense.

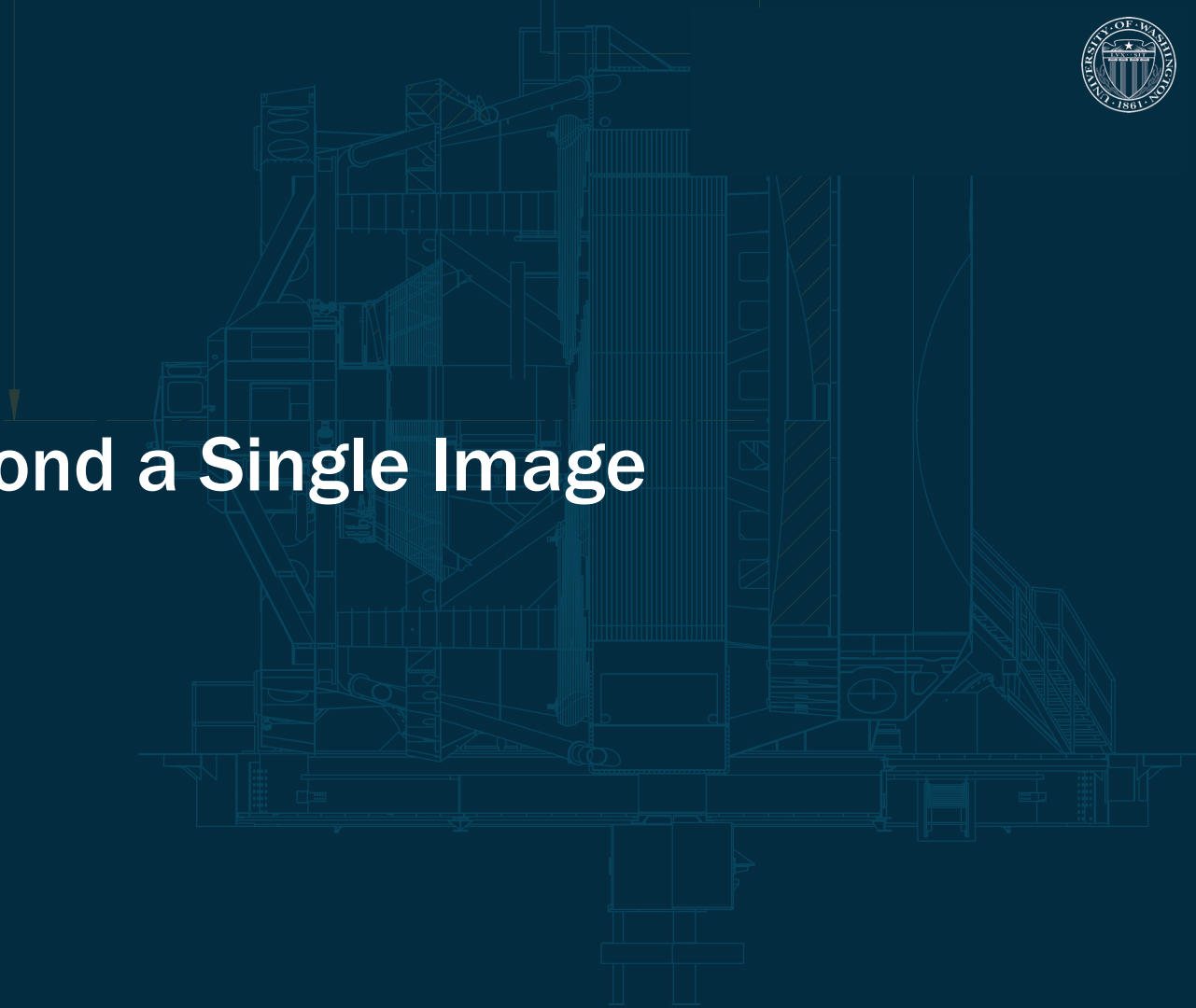
But you may also think of this as developing a very astronomy-specific lossy compression technique.

Challenges: Incomplete model space

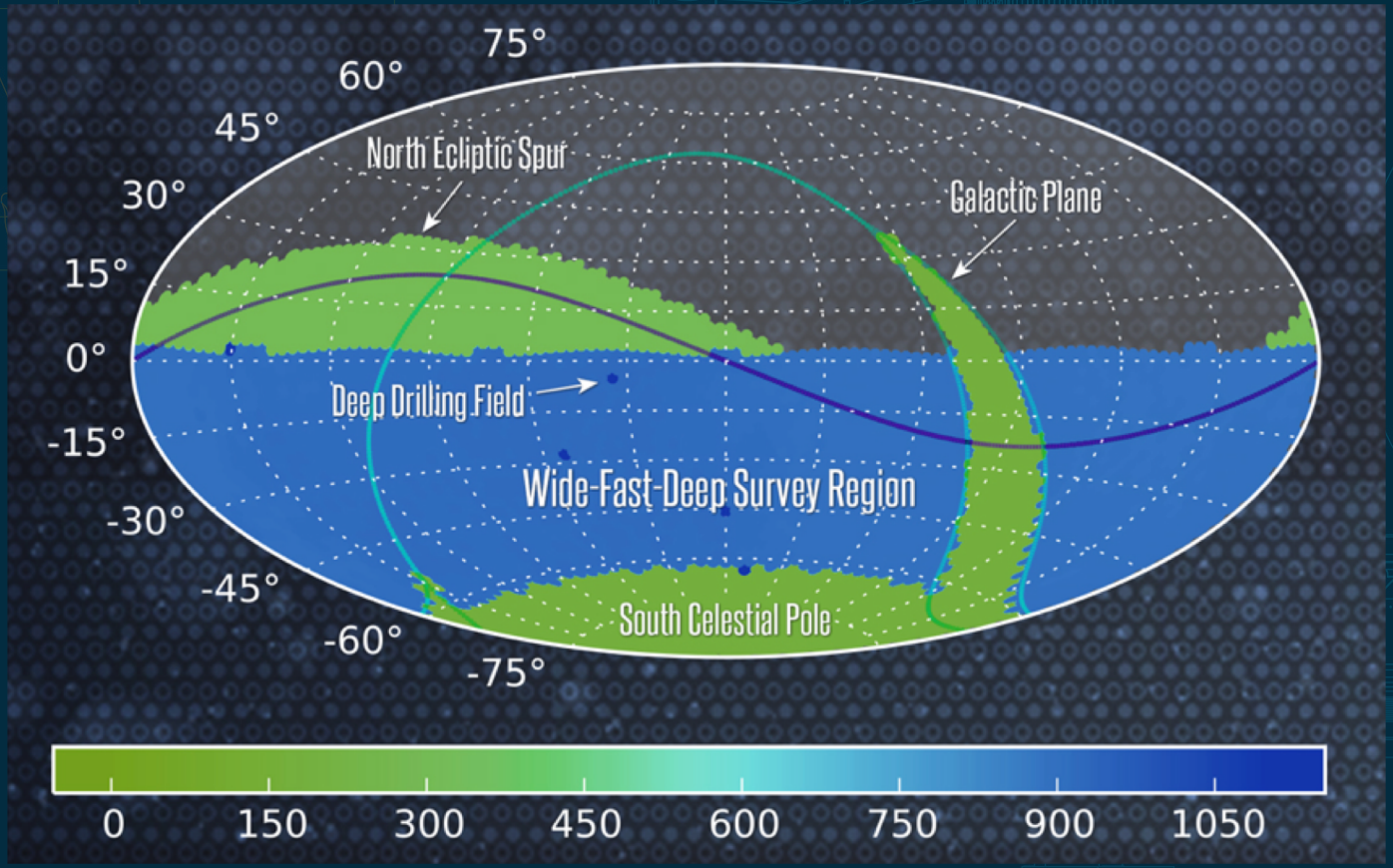


Beyond a Single Image

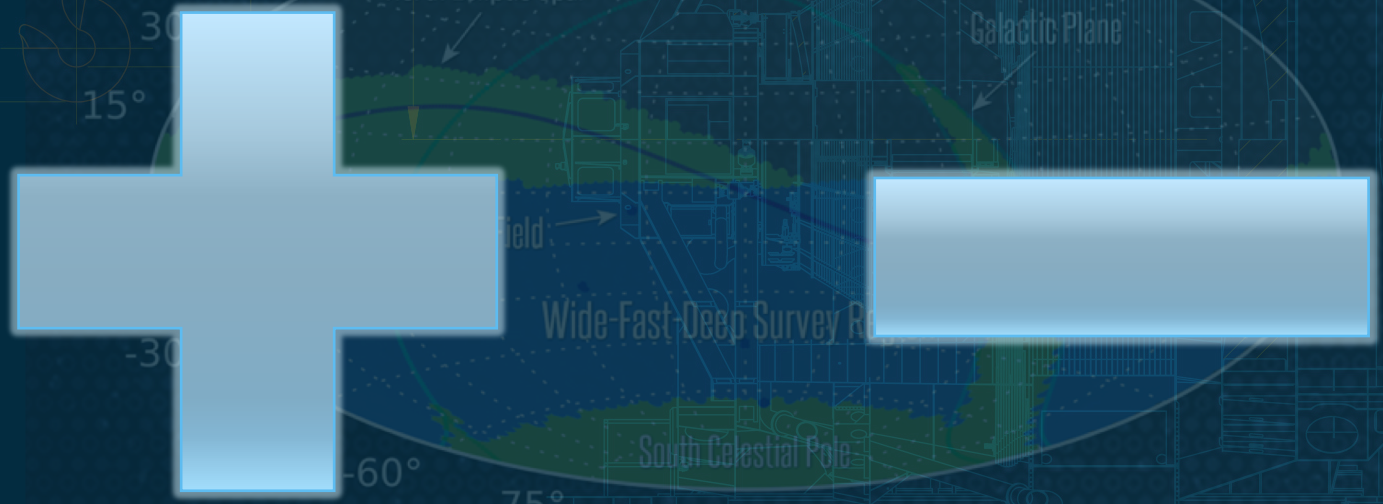




Beyond a Single Image

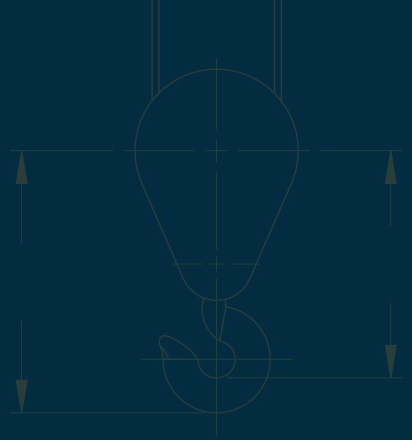


Beyond a Single Image



Going Deep: Coaddition

Changes in Time: Image Differencing



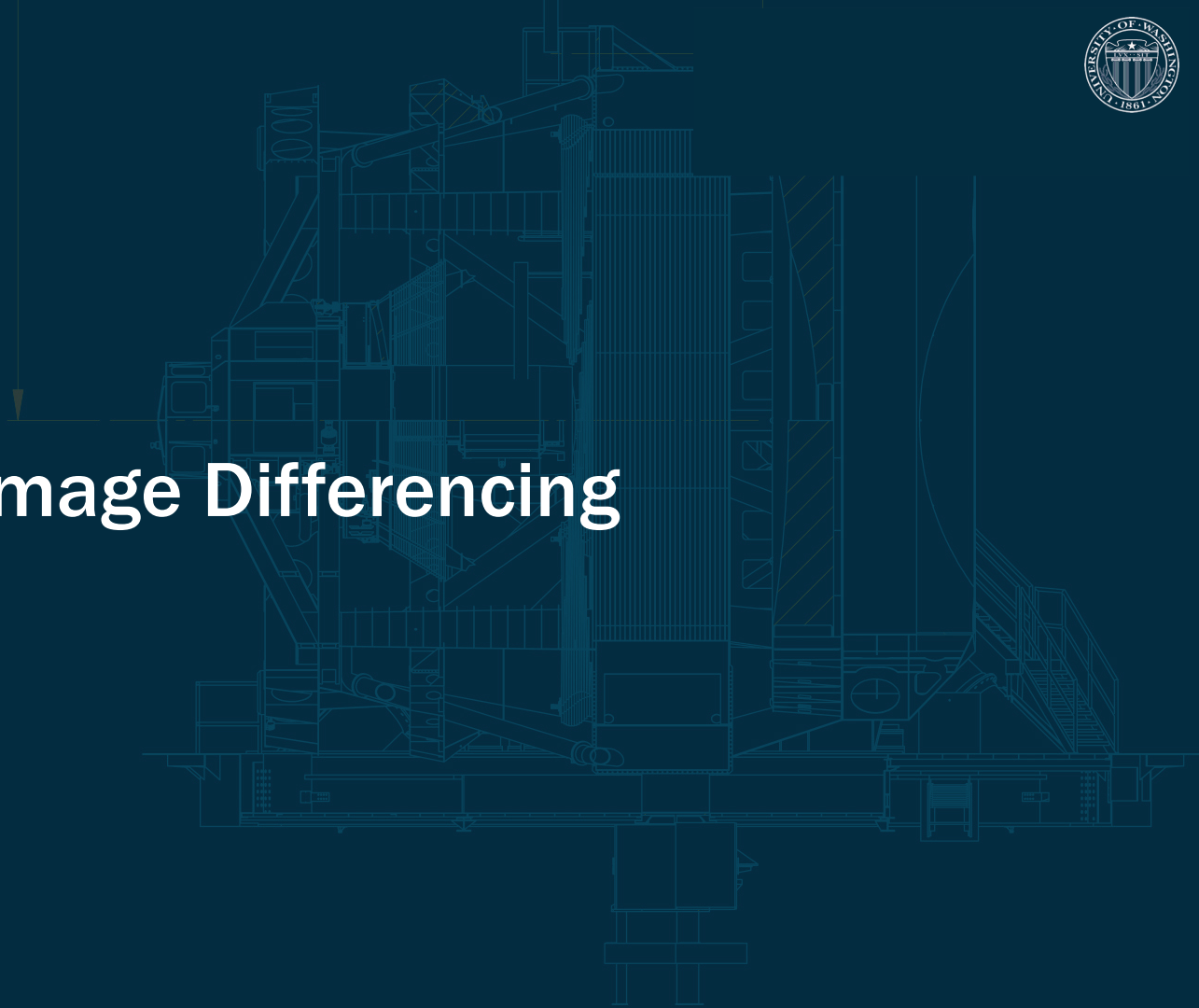
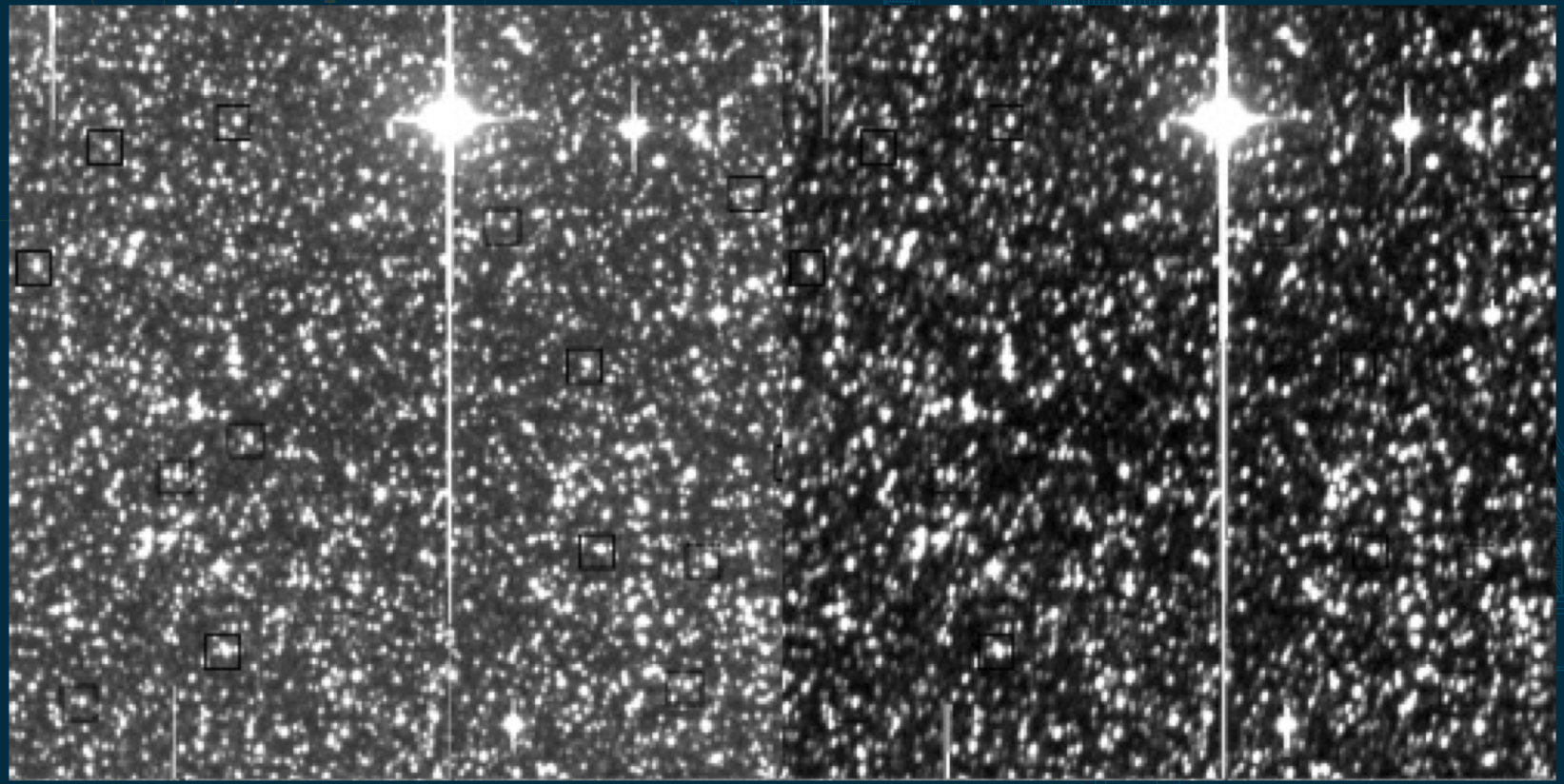
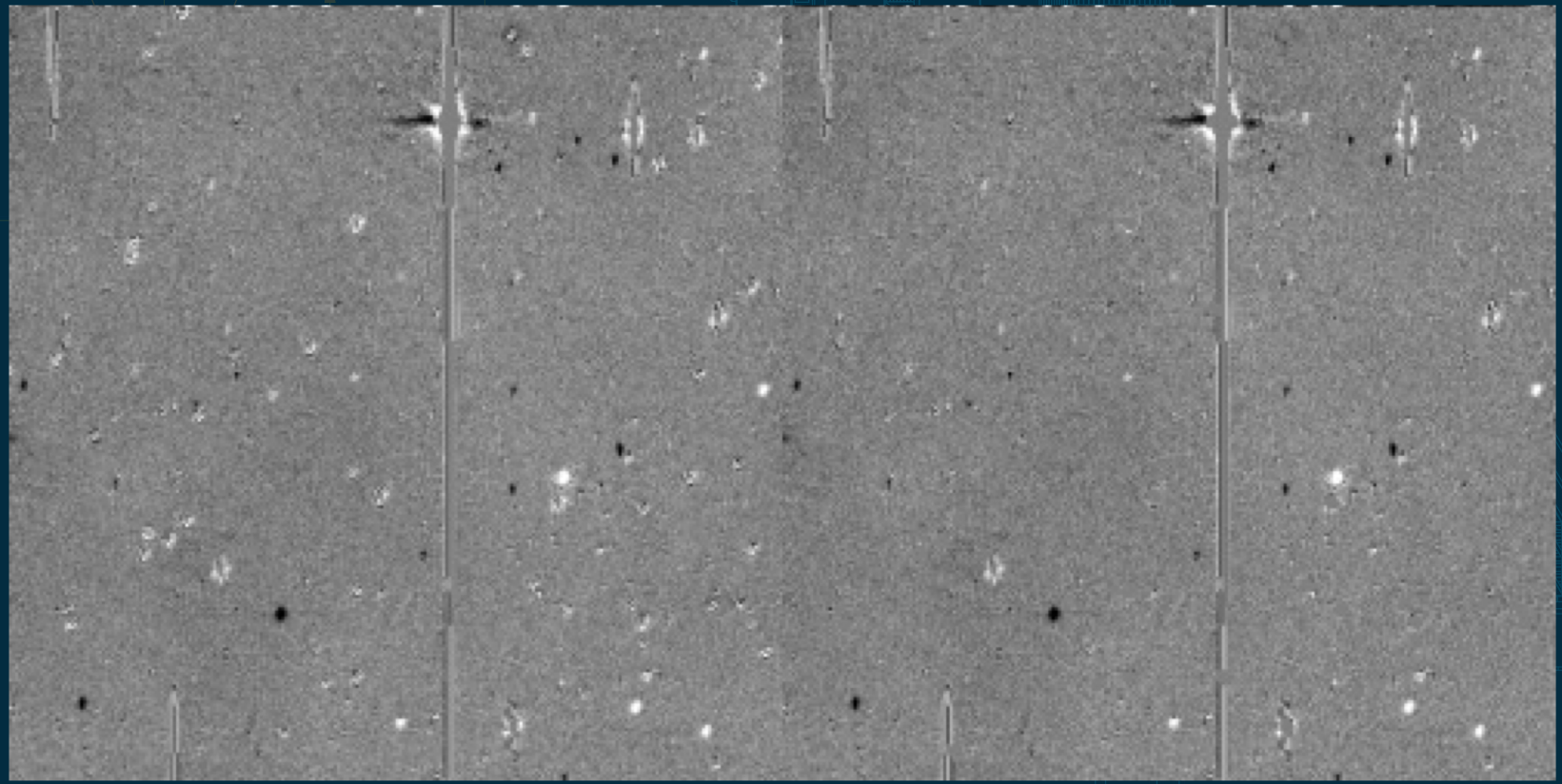


Image Differencing

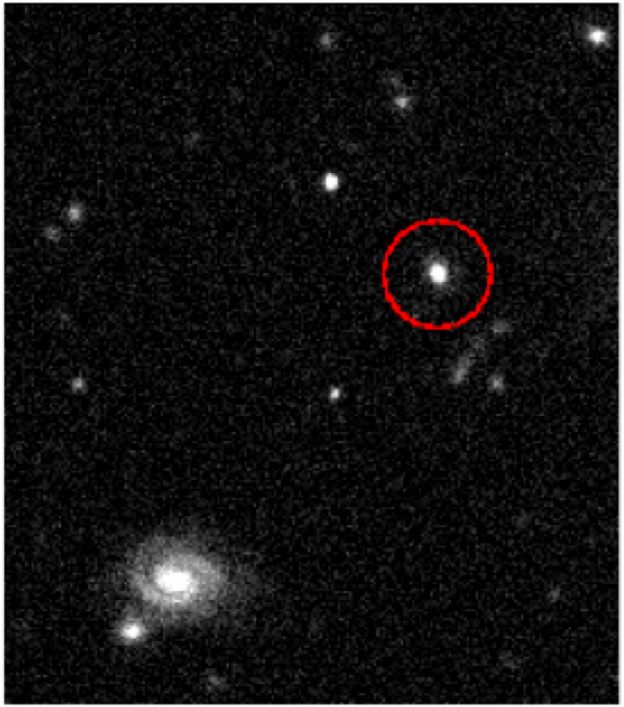
Why Difference?



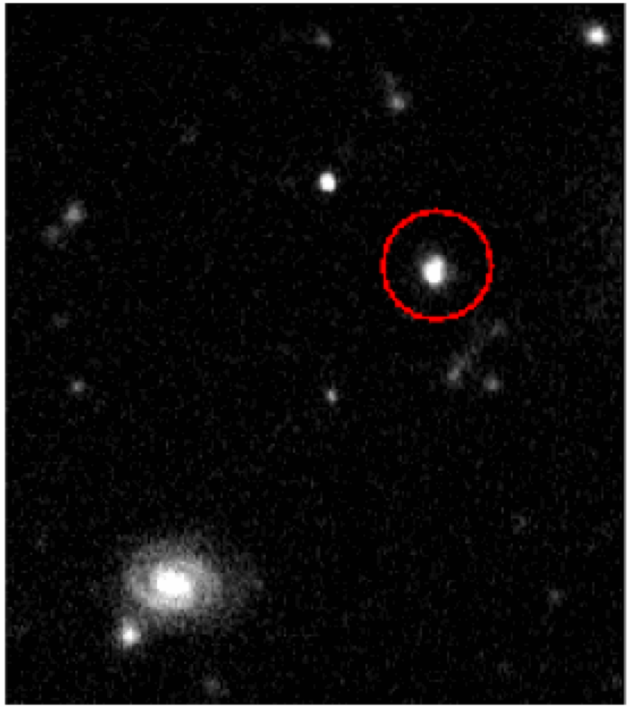
Why Difference?



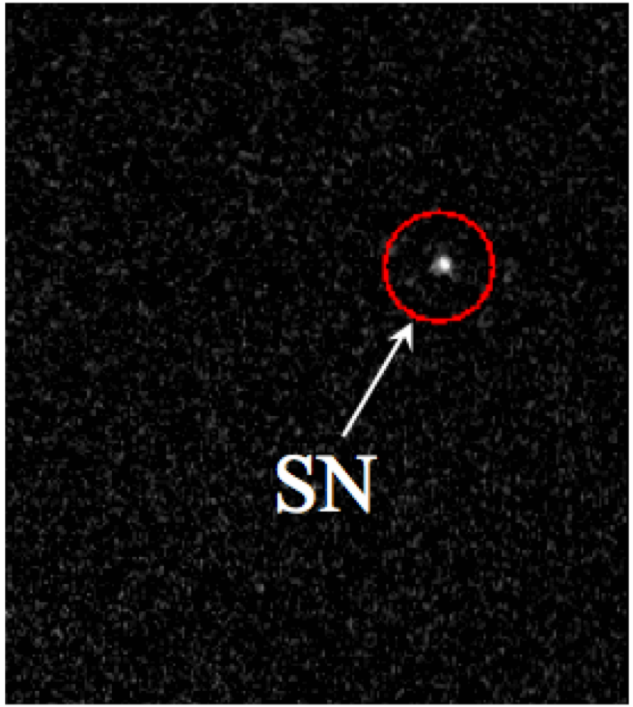
Why Difference?



First image

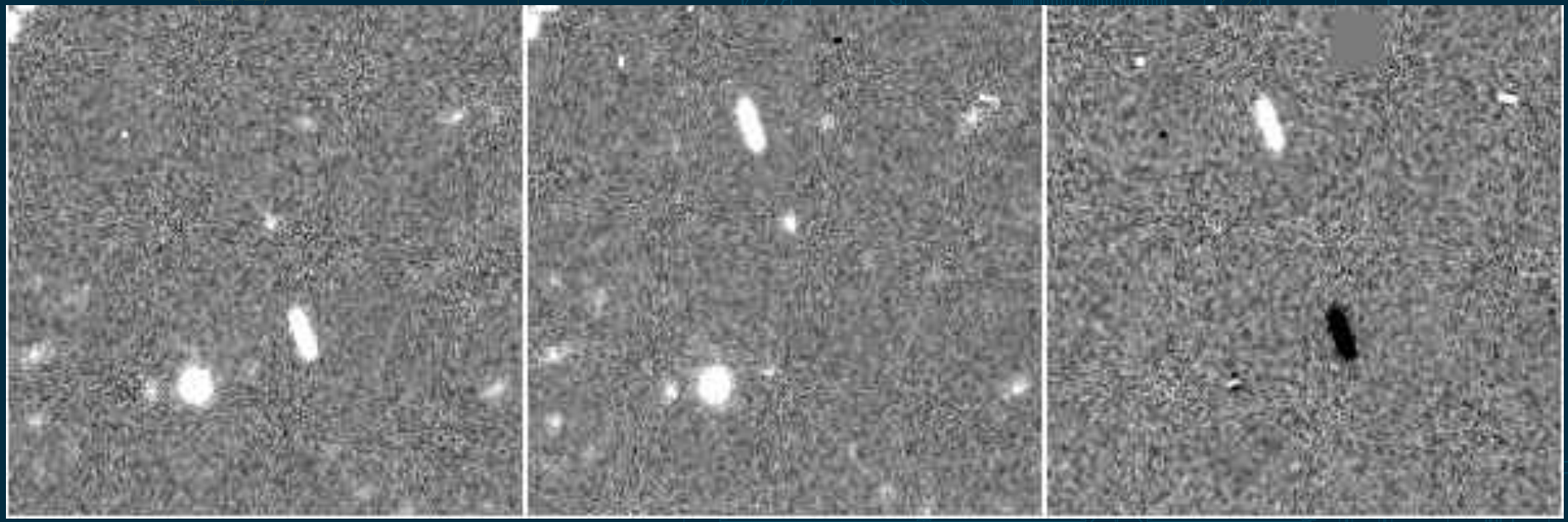


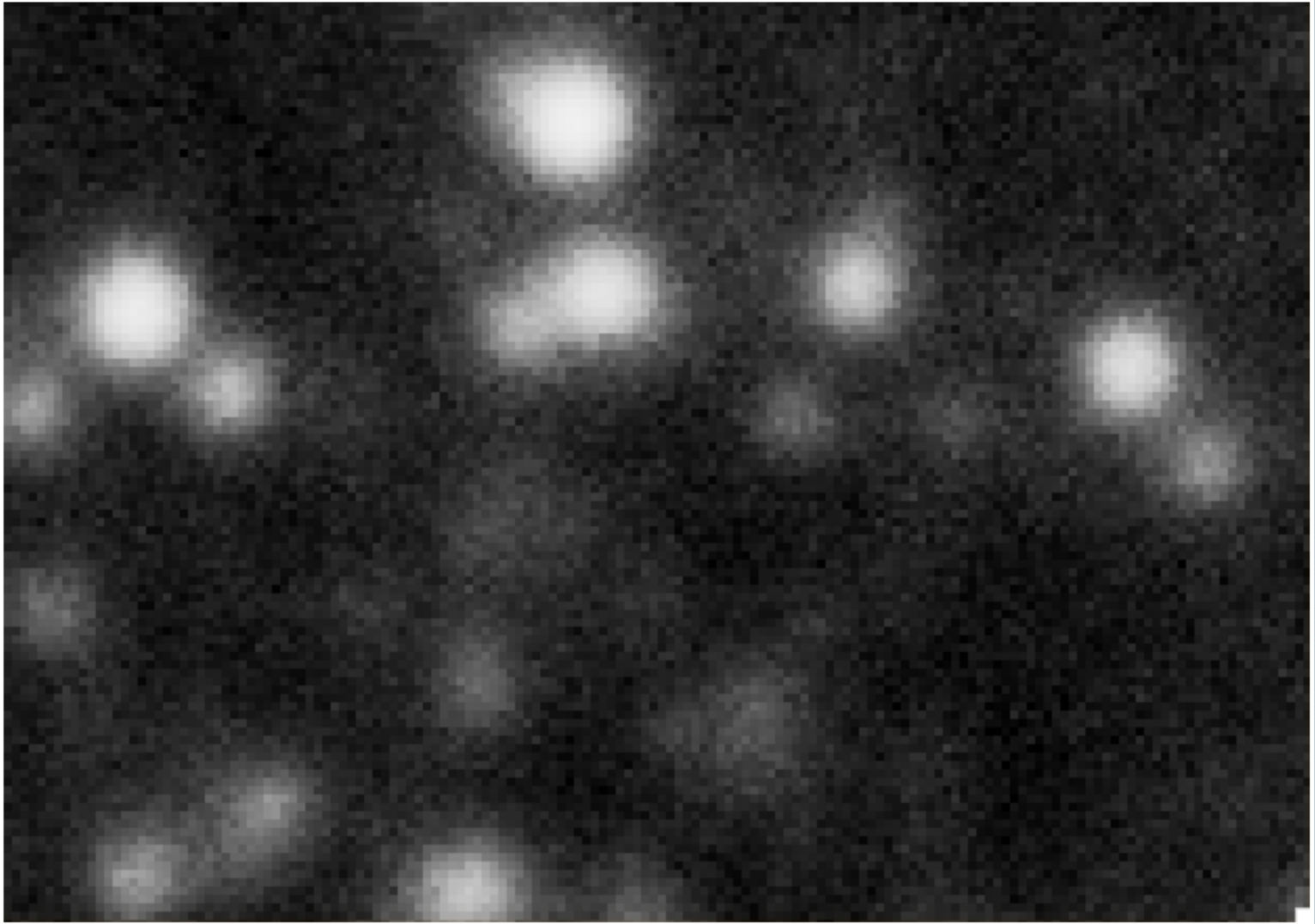
Second image

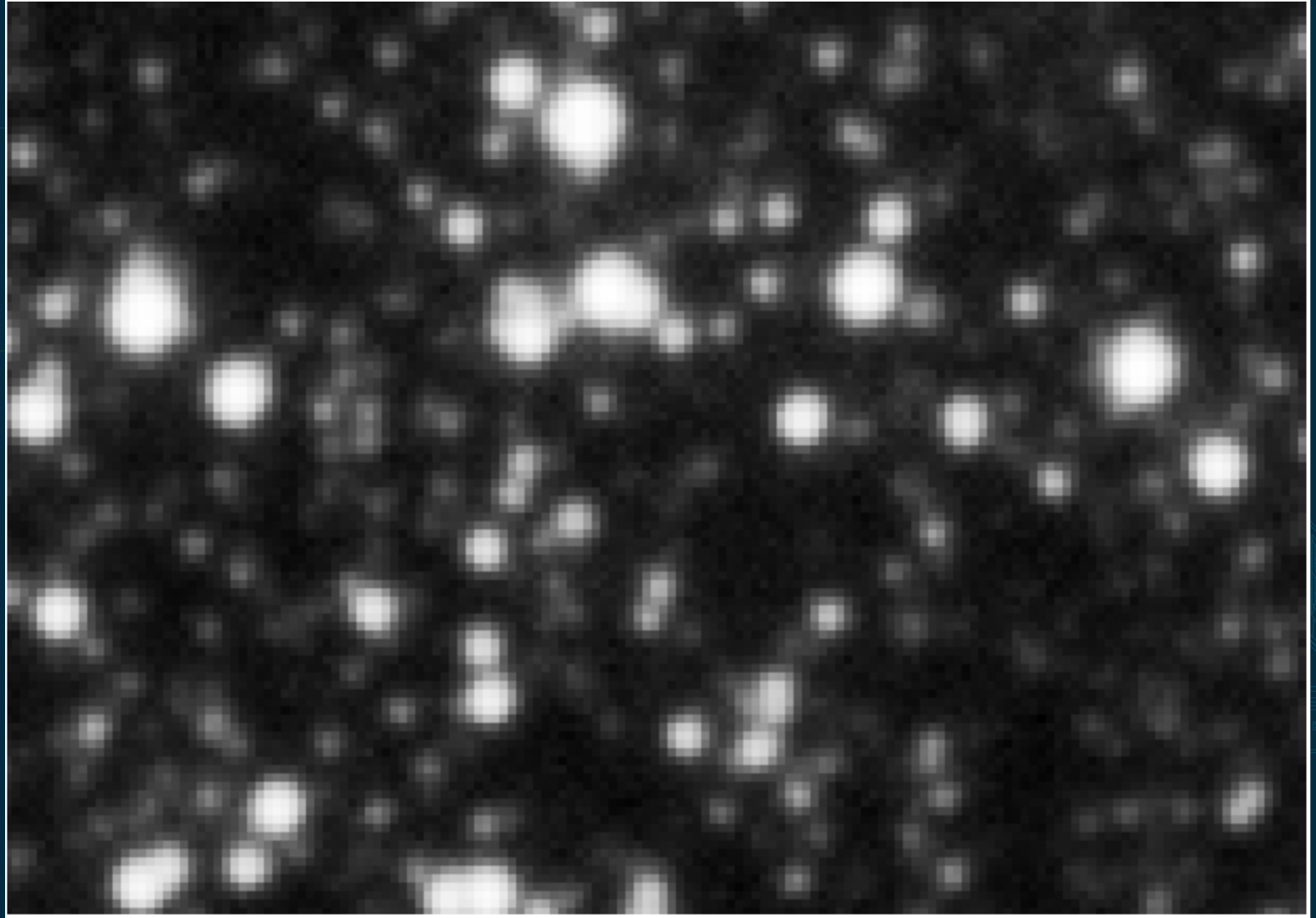


Difference image

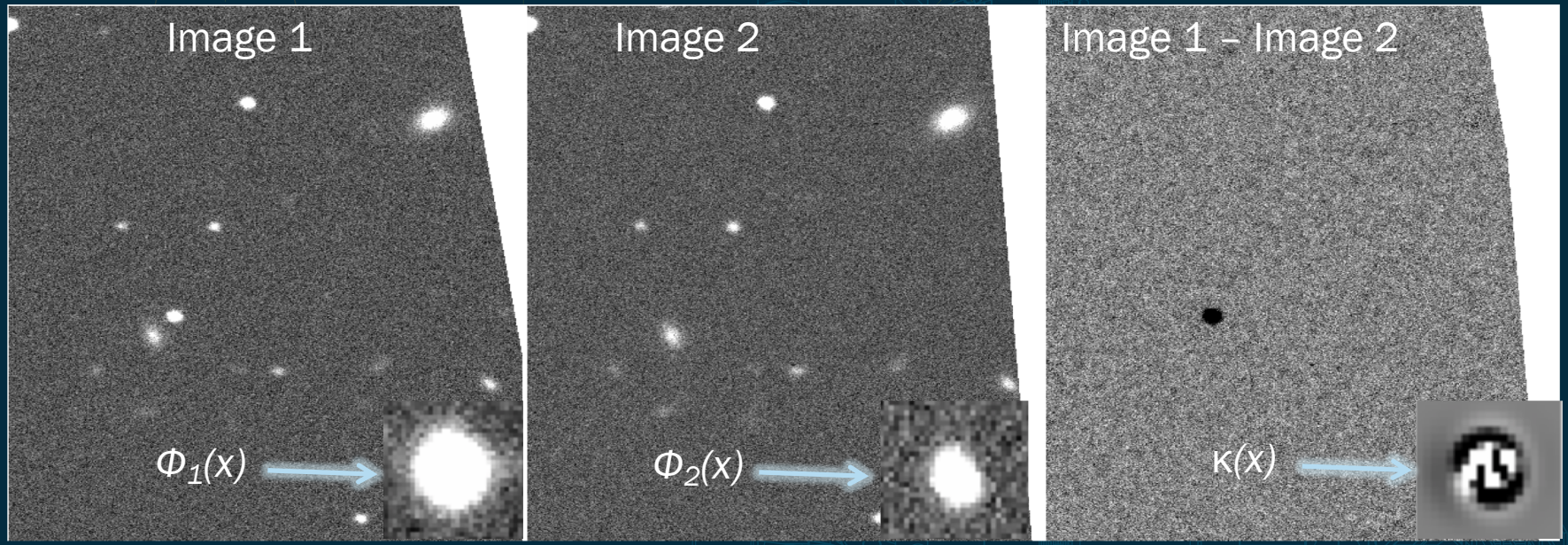
Align, Subtract, Profit! Right?...







Alard-Lupton Algorithm



$$I(x) = \phi(x) \otimes S(x) + \varepsilon(x)$$

PSF Truth Noise

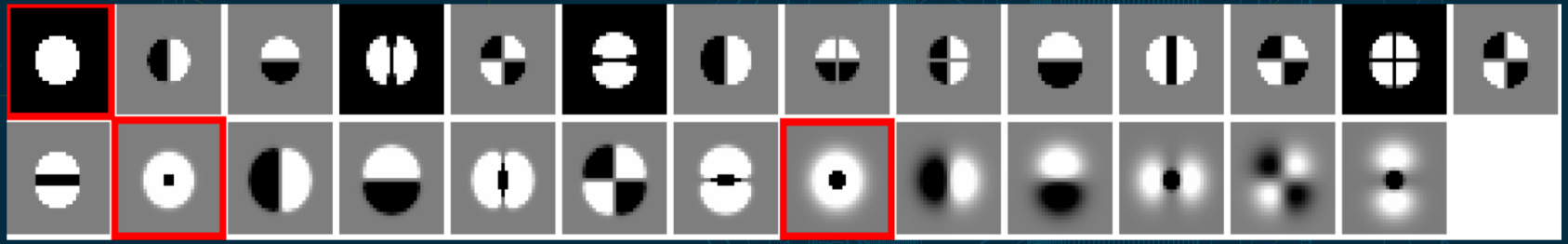
Image

$$D(x) = I_1(x) - \kappa(x) \otimes I_2(x)$$

Image Difference

How to match image quality

Difference imaging requires solving for the mapping kernel



$$k(x) = \sum_i a_i B_i(x) = \sum_{n,p,q} a_i e^{-\frac{u^2+v^2}{2\sigma_n^2}} u^p v^q$$

$$\left| \frac{I_1 - \sum_i a_i B_i(x) \otimes I_2}{\sigma} \right|^2$$

Alard and Lupton 97…

Works well but...

$$D(x) = I_1(x) - \kappa(x) \otimes I_2(x)$$

Assumes template PSF is smaller than image PSF

- Deconvolution otherwise -> BAD!

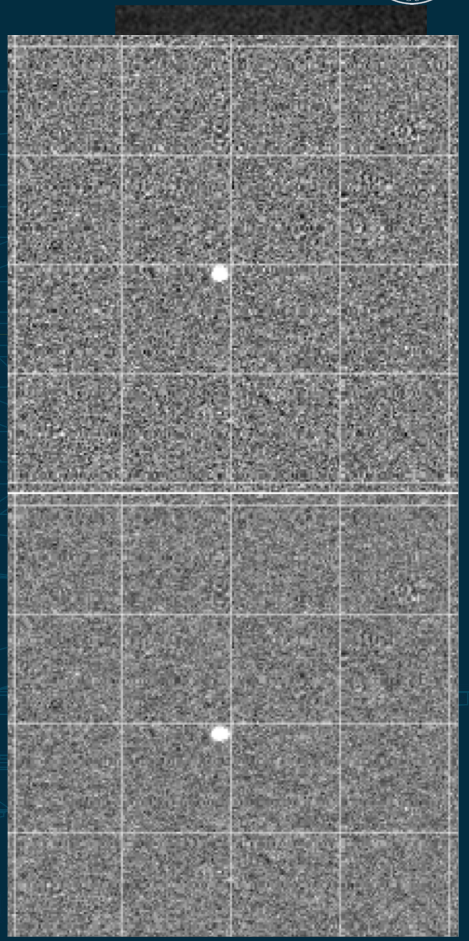
Assumes the “template” has no noise

- Derived from a previous set of observations
- Recent work has shown that for Gaussian, heteroschedastic noise we can take a Fourier Transform, and compute the log-likelihood and prewhiten the images

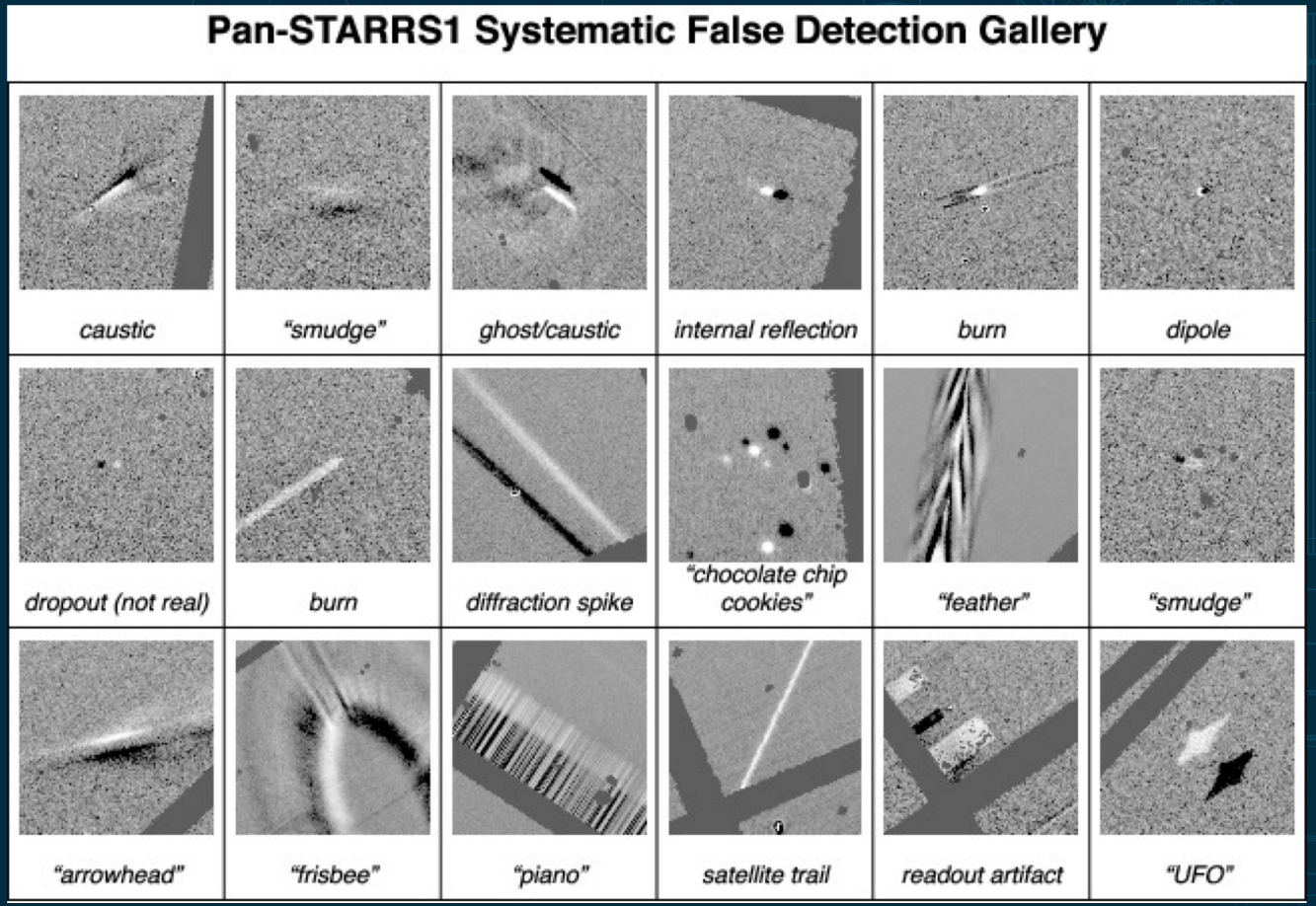
$$\hat{D}(k) = (I_1(k) - \kappa(k) I_2(k)) \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \kappa^2(k)\sigma_2^2}}$$

Zackay et al 2016

This is running on ZTF!



Real world is more complicated…

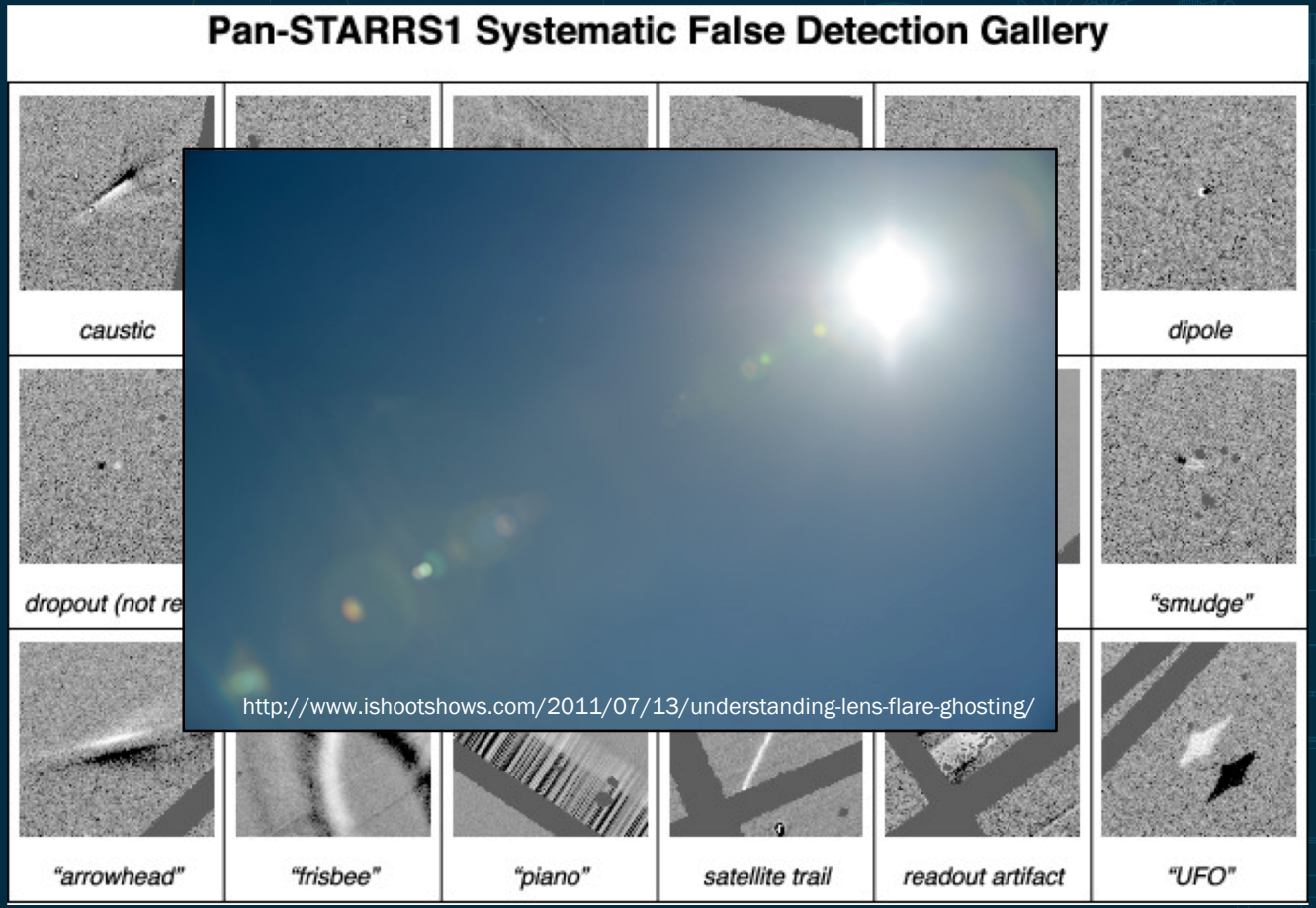


In real-world applications, differencing two images is never perfect. In fact, it's far from perfect!

Typically, the observed difference images are littered with artifacts, *false positives*, that make astronomers sad and unhappy.

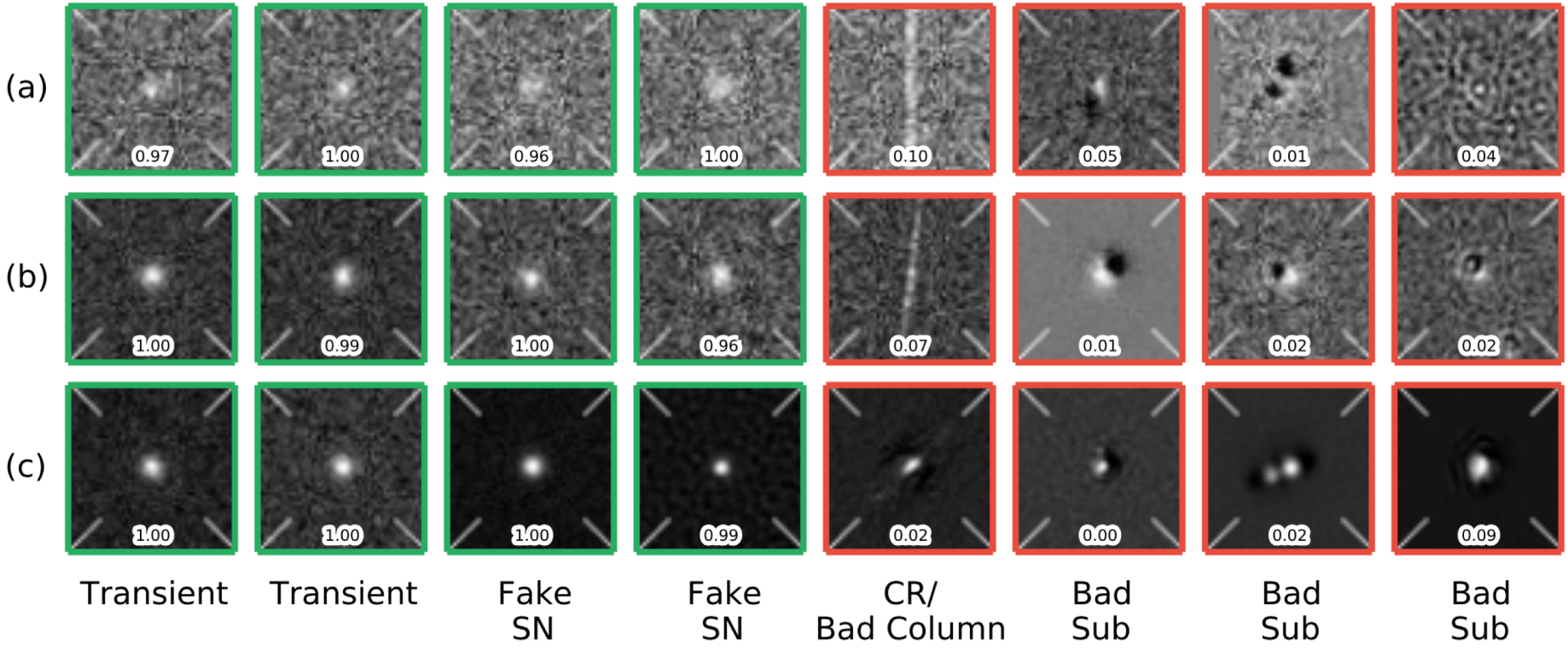
Typically, we see 100:1 to 10:1 false-to-true detection ratios (that is not a typo!).

Real world is more complicated...



Where do these things come from?

- Image misalignments
- Imperfect knowledge of PSF variation along the image
- CCD defects
- Readout electronics artifacts
- Optical ghosts and glints
- ...



Above: Figure 1, Goldstein al. (2015), AJ, 150, 82

Machine Learning to the Rescue!

Given a sample of real detections and false detections, teach the computer to recognize the difference between the two and judge assign a "score", τ, to each (τ=1 -> real, $\tau = 0$ -> false).

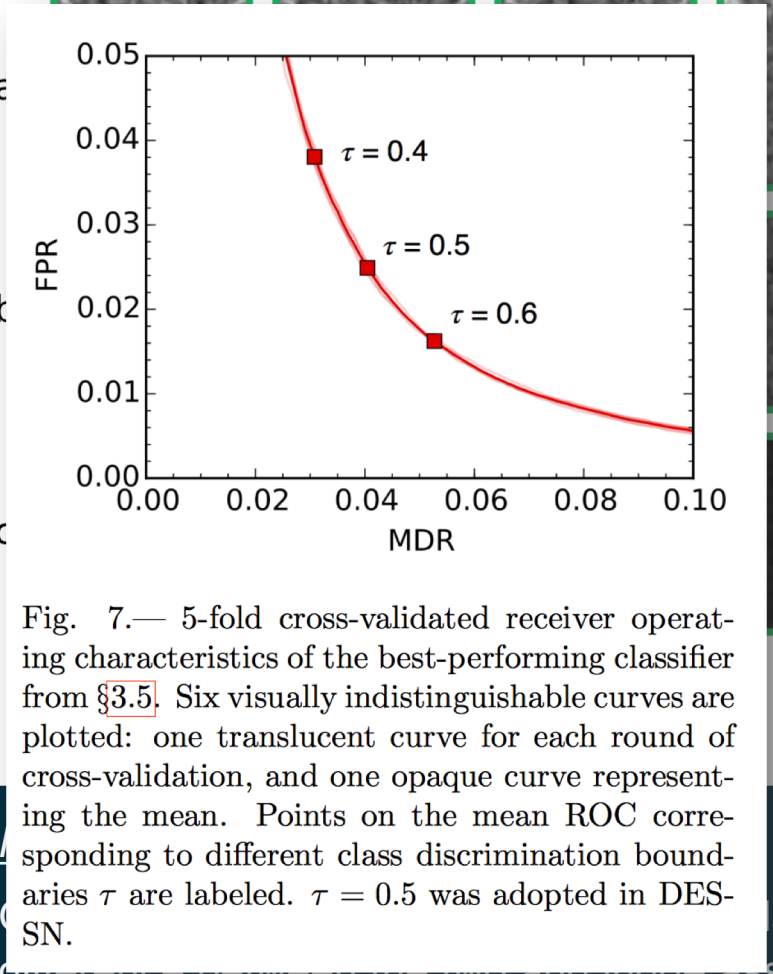


Fig. 7.— 5-fold cross-validated receiver operating characteristics of the best-performing classifier from §3.5. Six visually indistinguishable curves are plotted: one translucent curve for each round of cross-validation, and one opaque curve representing the mean. Points on the mean ROC corresponding to different class discrimination boundaries τ are labeled. $\tau = 0.5$ was adopted in DES-SN.

TABLE 4

CAN ON REPROCESSED DES Y1 TRANSIENT CANDIDATE S

	No ML	ML ($\tau = 0.5$)	ML / No ML
N_c[a]	100,450	7,489	0.075
$\langle N_A/N_{NA}\rangle$[b]	13	0.34	0.027
ϵ_F[c]	1.0	0.990	0.990

[a]Total number of science candidates discovered.

[b]Average ratio of artifact to non-artifact detections in human scanning pool.

[c]`autoScan` candidate-level efficiency for fake SNe Ia.

1.00

0.96

0.99

Fake SN

Bad Column

Sub

Sub

Sub

Above: Figure 1, Goldstein al. (2015), AJ, 150, 82

se detections, teach the computer to recognize the difference between the two and judge assign a "score", τ, to each (τ=1 -> real, τ = 0 -> false).

(a)

(b)

(c)

1.00

0.96

0.99

Fake

SN

Bad Column

Sub

Sub

Sub

TABLE 4

CAN ON REPROCESSED DES Y1 TRANSIENT CANDIDATE S

	No ML	ML ($\tau = 0.5$)	ML / No ML
N_c[a]	100,450	7,489	0.075
$\langle N_A/N_{NA}\rangle$[b]	13	0.34	0.027
ϵ_F[c]	1.0	0.990	0.990

[a]Total number of science candidates discovered.

[b]Average ratio of artifact to non-artifact detections in human scanning pool.

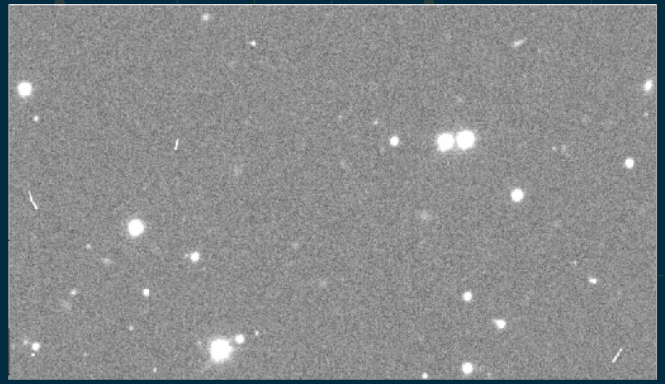
[c]`autoScan` candidate-level efficiency for fake SNe Ia.

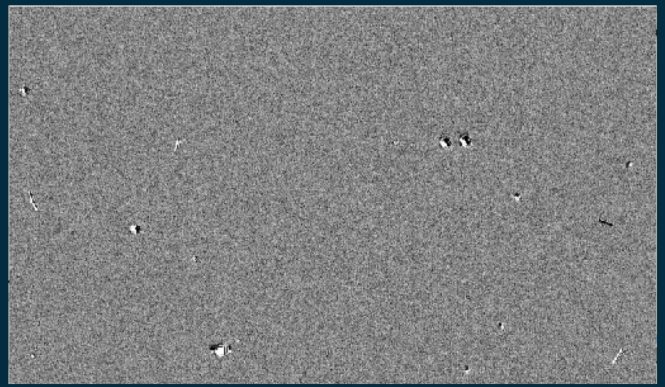
Above: Figure 1, Goldstein al. (2015), AJ, 150, 82

Machine Learning to the Rescue!

Given a sample of real detections and false detections, teach the computer to recognize the difference between the two and judge assign a “score”, τ, to each (τ=1 -> real, $\tau = 0$ -> false).

Option #2: Understand the root cause(s)

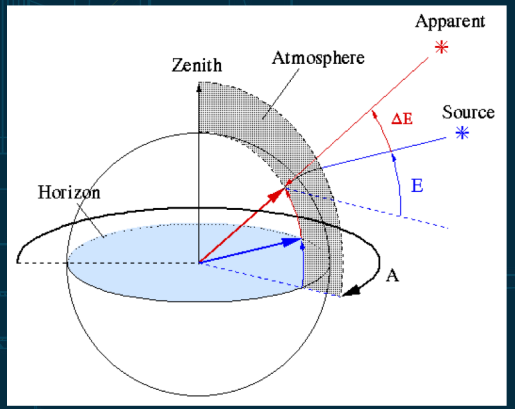




Example: Astrometric misalignment introduces dipoles in the images, misalignment of >2% of a pixels will dominate the number of false positives

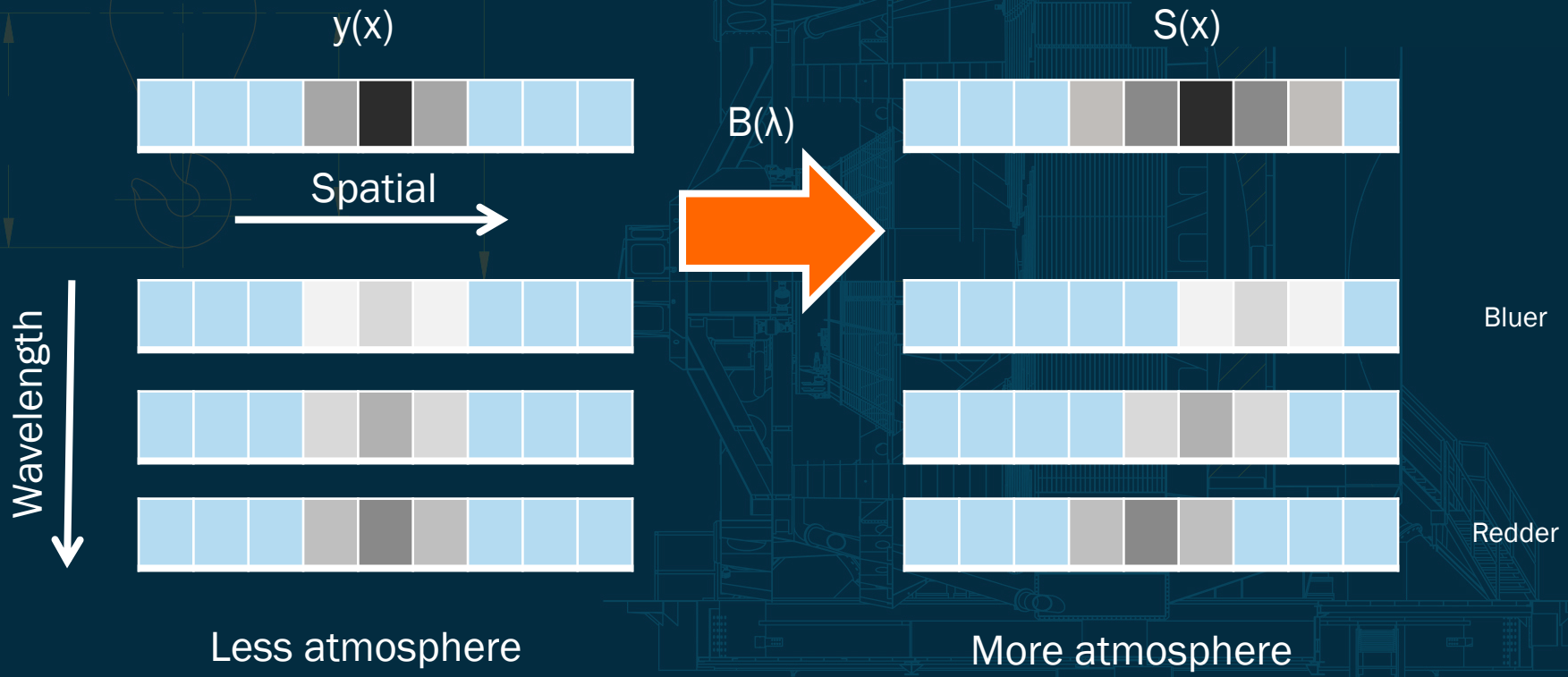
Major source is Differential Chromatic Refraction

Atmosphere refracts (shifts) a source more for blue light than red (even for light measured through the same filter)



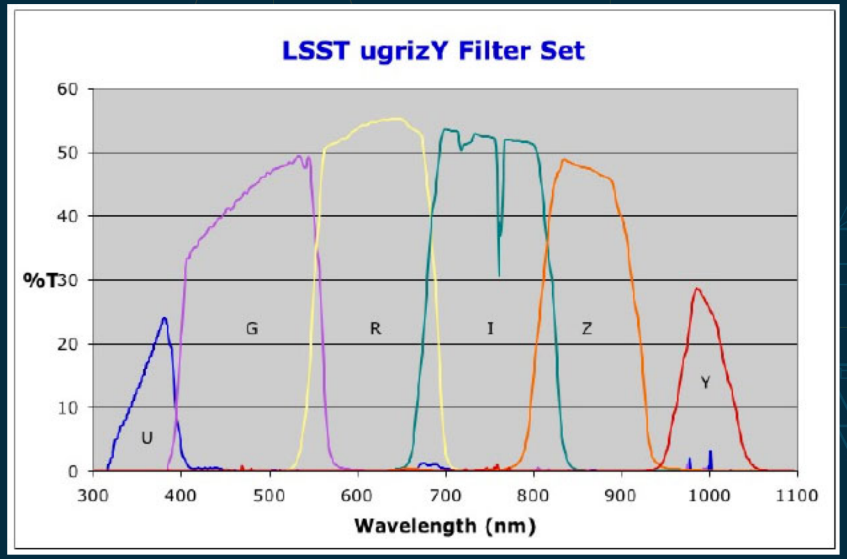
A bandpass is not a delta function.

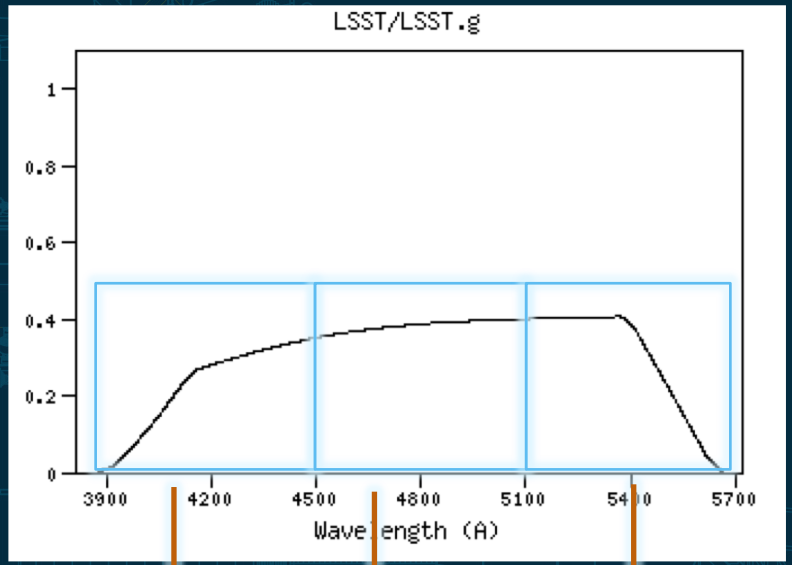
How differential refraction works

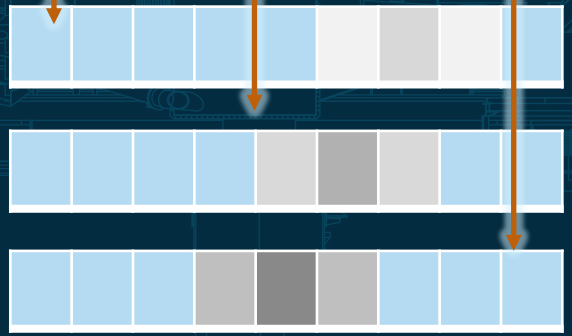


$$S(x) = B(x, \lambda) y_\lambda(x)$$

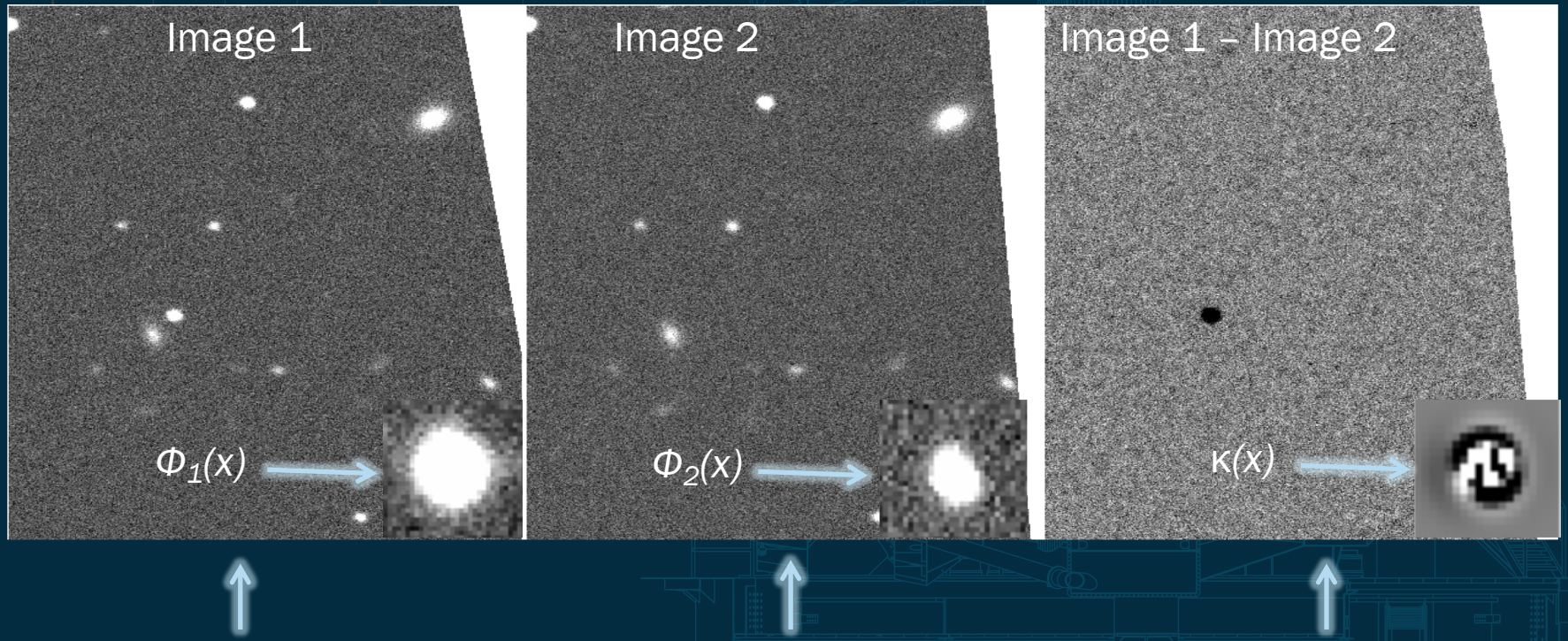
A bandpass is not infinitely narrow







A few more details…



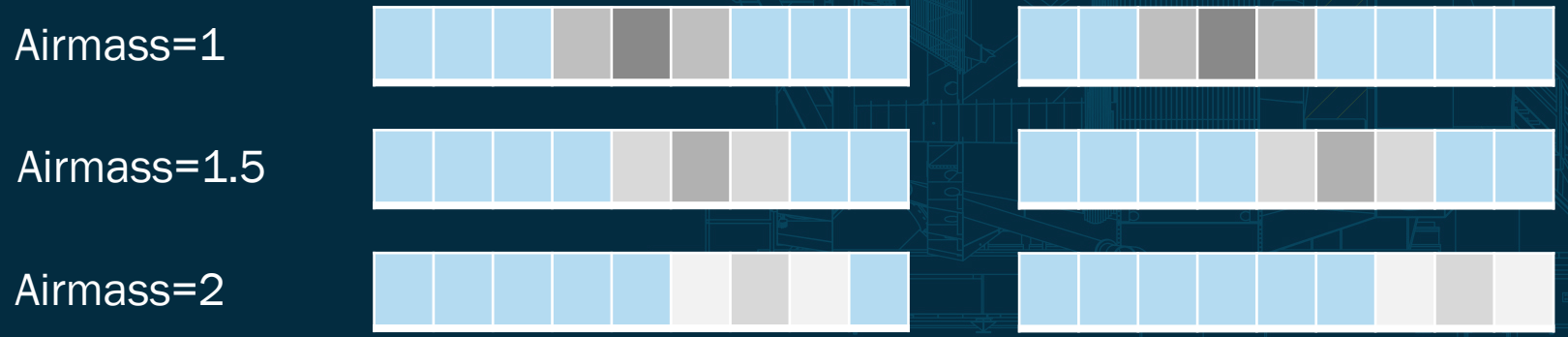
"Template Image"
(often a *coadd*)

"Science Image"

Image Difference

A solution

Instead of having just one template, infer the template as a function of wavelength based on many observations at different airmasses.



A solution

Instead of having just one template, infer the template as a function of wavelength based on many observations at different airmasses!

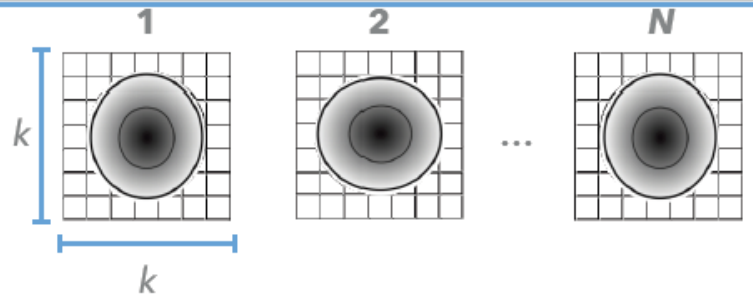
Effectively, given N exposures at different airmassess, infer the images $I_T(\lambda=\text{blue})$, $I_T(\lambda=\text{center})$, $I_T(\lambda=\text{red})$ for the template(s).

Then, for each science image, we can construct the template exactly for that airmass, before doing the differencing.

DCR-corrected Templates: Ian Sullivan et al.

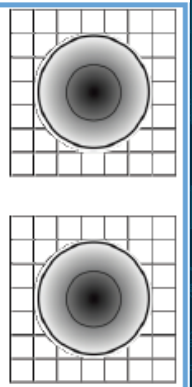
All pixels within a *k x k* size kernel, from all *N* images : $\vec{s}$

Each image has PSF $Q_1, Q_2, \ldots, Q_N$: Q



All pixels within a *k x k* size kernel, from all *M* sub-bands : $\vec{y}$

Each sub-band image has PSF $P_A, P_B, \ldots$: P



$$QB\vec{y} = P\vec{s}$$

$$\| \vec{y} \| = (B^T B)^{-1} B^T (Q^T Q)^{-1} Q^T P \vec{s}$$

$$\| \vec{s} \| = (P^T P)^{-1} P^T Q B \vec{y}$$

Assume a model for the image, y, that is made up of series of images each of a different wavelength (ie a “hyperspectral” cube)

Byproduct: measuring the intra-band spectrum!

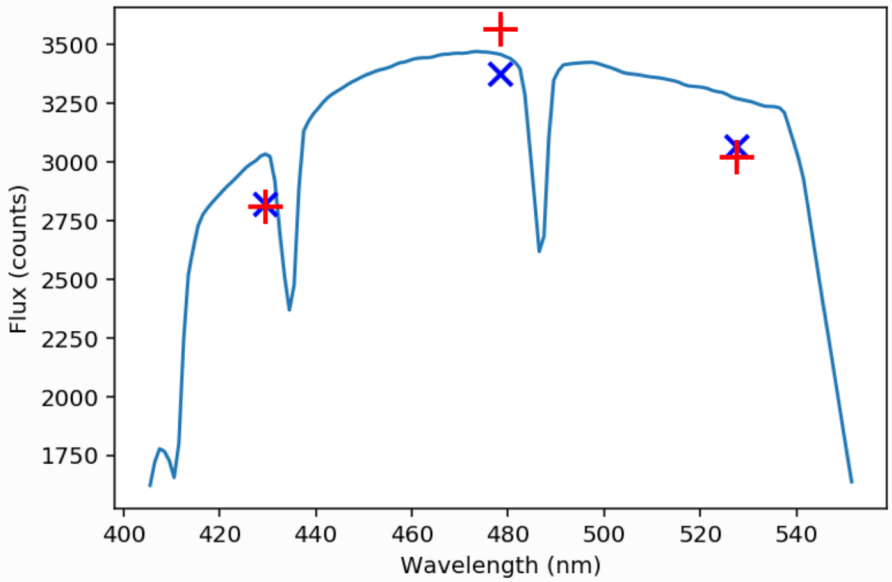


Figure 15 Example input spectrum for a type F star with surface temperature ~7130K (solid blue). The flux measured in each sub-band is marked with a with red '+', and the average values of the simulated spectrum across each subfilter is marked with a blue 'x' for comparison.

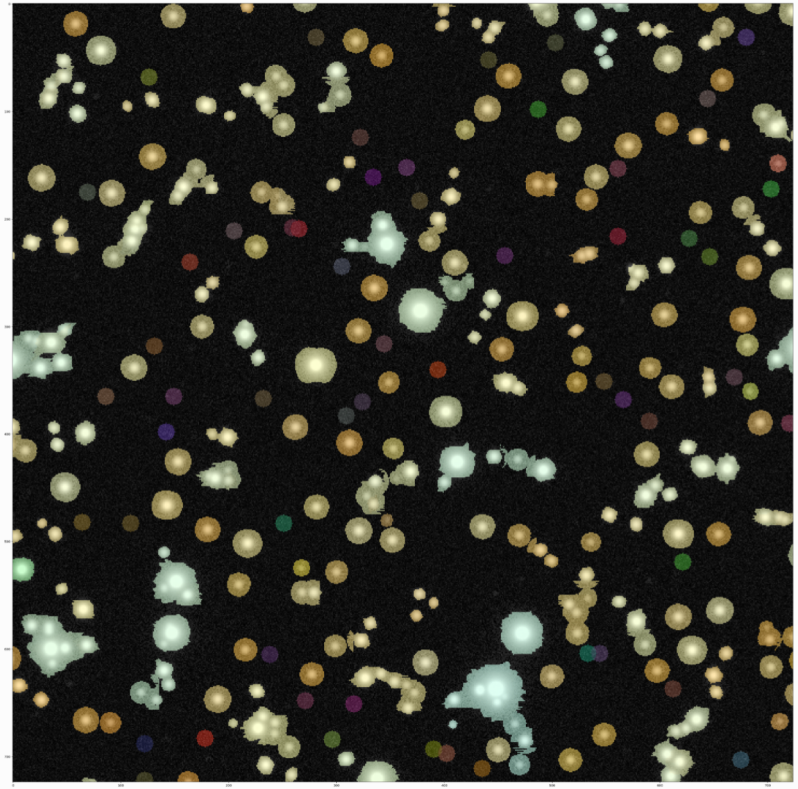
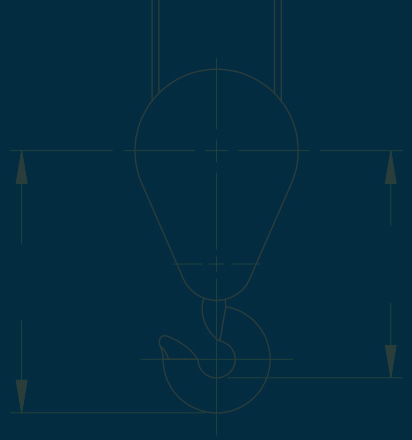


Figure 14 Source measurements in three sub-bands of the DCR sky model are converted to RGB values and used to fill the footprints of detected sources. The combined full-band model is displayed behind the footprint overlay.

This is another example of utilizing all available information and understanding of image generation processes to extract additional information.



Going Deeper: Coaddition (a.k.a. “astronomer’s HDR”)

Why co-add?

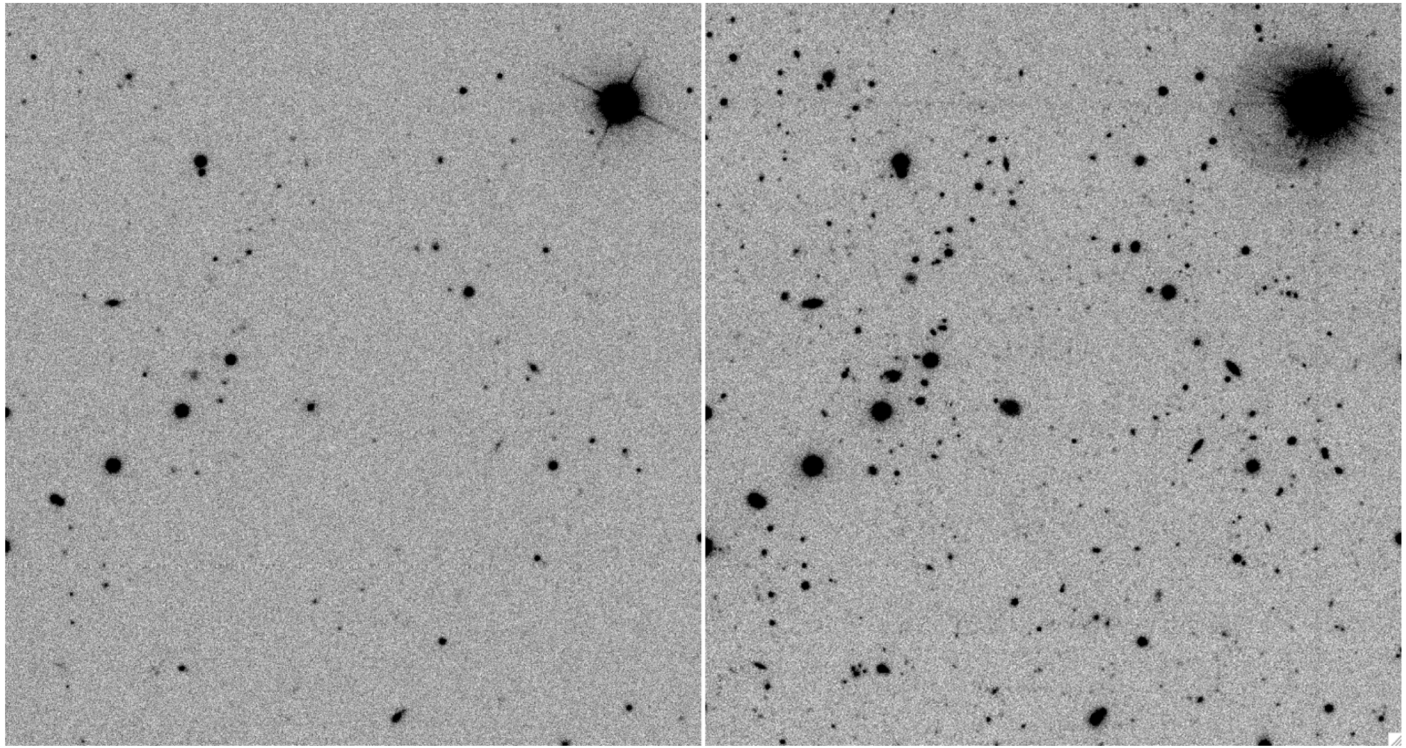


FIG. 2.— Comparison between single pass (left) and coadd (right) images in r-band for run 206, camcol 3, field 505, RA=15, Dec=0. Images are shown with the same scale, contrast and stretch. The single pass counterpart (run 5800, camcol 3, field 505) is one out of 28 images used in the coaddition of this particular image. This example illustrates the fact that a large number of objects bellow the detection threshold of each image can be well detected and measured in the coadd.

Pros:
See fainter objects!
Computationally inexpensive

Left:
Annis et al. 2011

Why co-add?

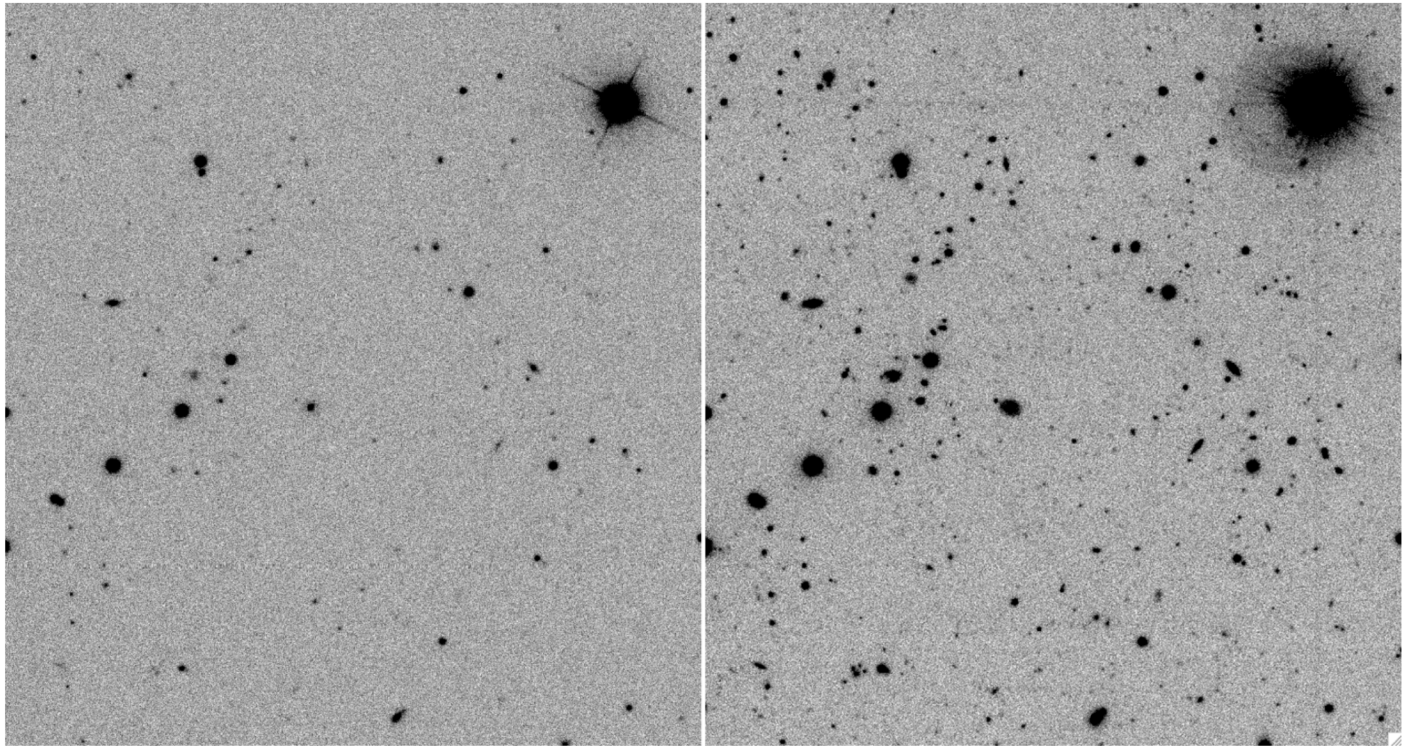


FIG. 2.— Comparison between single pass (left) and coadd (right) images in r-band for run 206, camcol 3, field 505, RA=15, Dec=0. Images are shown with the same scale, contrast and stretch. The single pass counterpart (run 5800, camcol 3, field 505) is one out of 28 images used in the coaddition of this particular image. This example illustrates the fact that a large number of objects bellow the detection threshold of each image can be well detected and measured in the coadd.

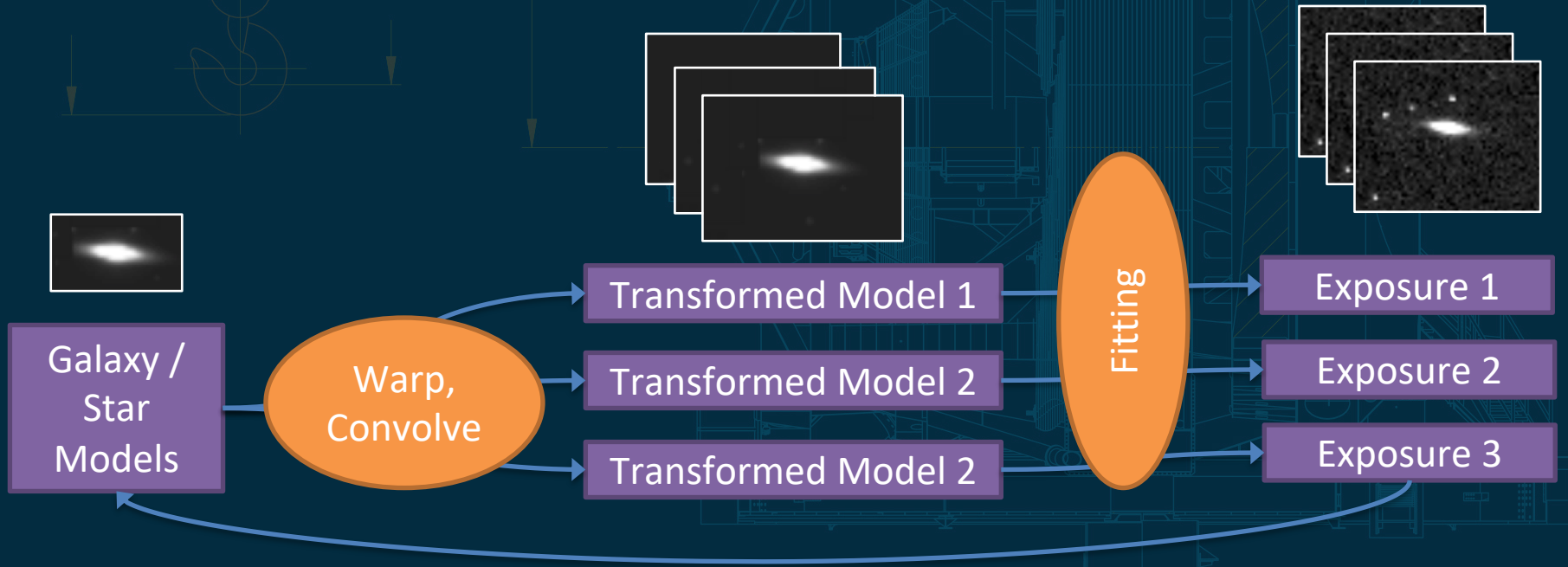
Pros:
See fainter objects!
Computationally inexpensive

Cons:
- Complicated PSF
- Correlated noise
- Loss of motion and time variability information
- Loss of information.

Left:
Annis et al. 2011

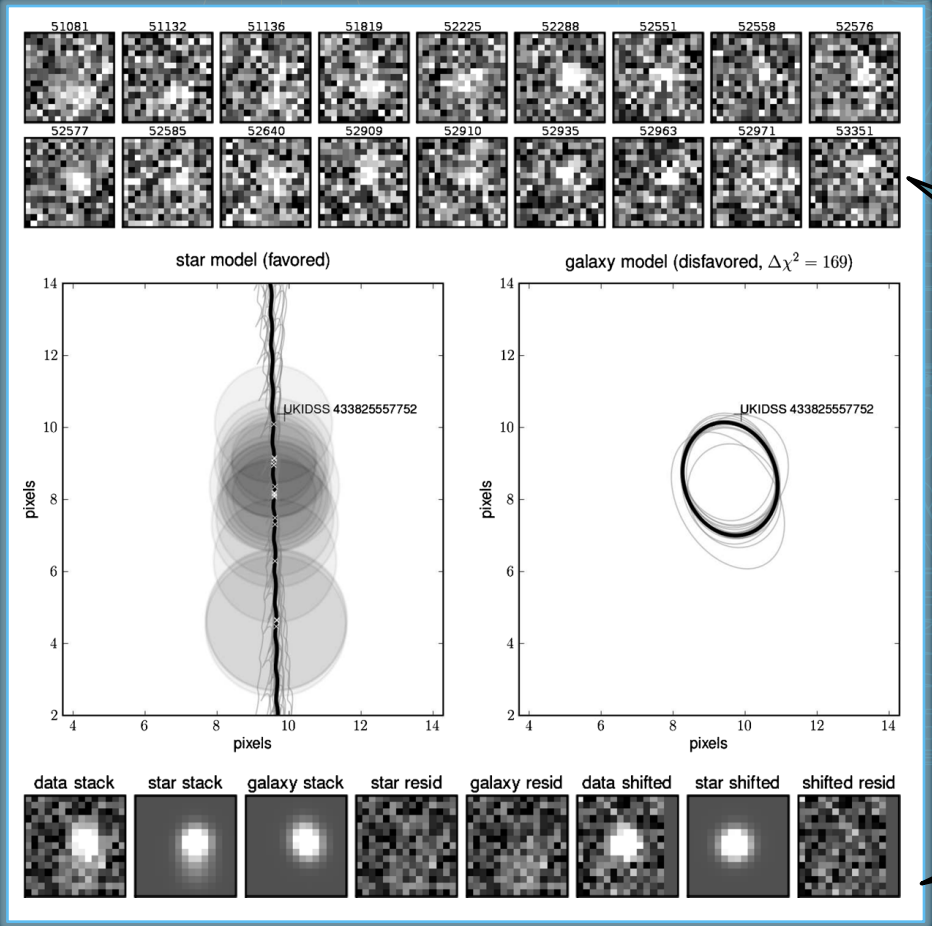
Better than Coaddition: Multi-Epoch Fitting (MultiFit)

A very simple idea: instead of co-adding pixels of individual observations, and then fitting the model to the result, why not fit the model directly to each individual observation?



MultiFit (Simultaneous Multi-Epoch Fitting)

Opportunity: Recovering motion from the noise



Lang (2009):

Fit moving source models to suspected moving stars in SDSS Stripe 82 survey.

Individual exposures: objects are undetected or marginally detected

Moving point-source and galaxy models are indistinguishable on the coadd

Downsides

Computationally extremely intensive. Scales with the number of epochs (so ~100-1000x more computationally expensive for modern surveys like LSST).

Worse, we're greedy: the physics we're trying to study is so sensitive to biases that ML estimators are not enough; we want posteriors PDFs for parameters of each observed galaxy! (20-200x more output storage!)

But computing is getting cheaper. LSST is building a ~2 PFLOP machine to do this (cca ~2025). Still cheaper than building a bigger telescope (and/or launching it into space)!

Putting it all together

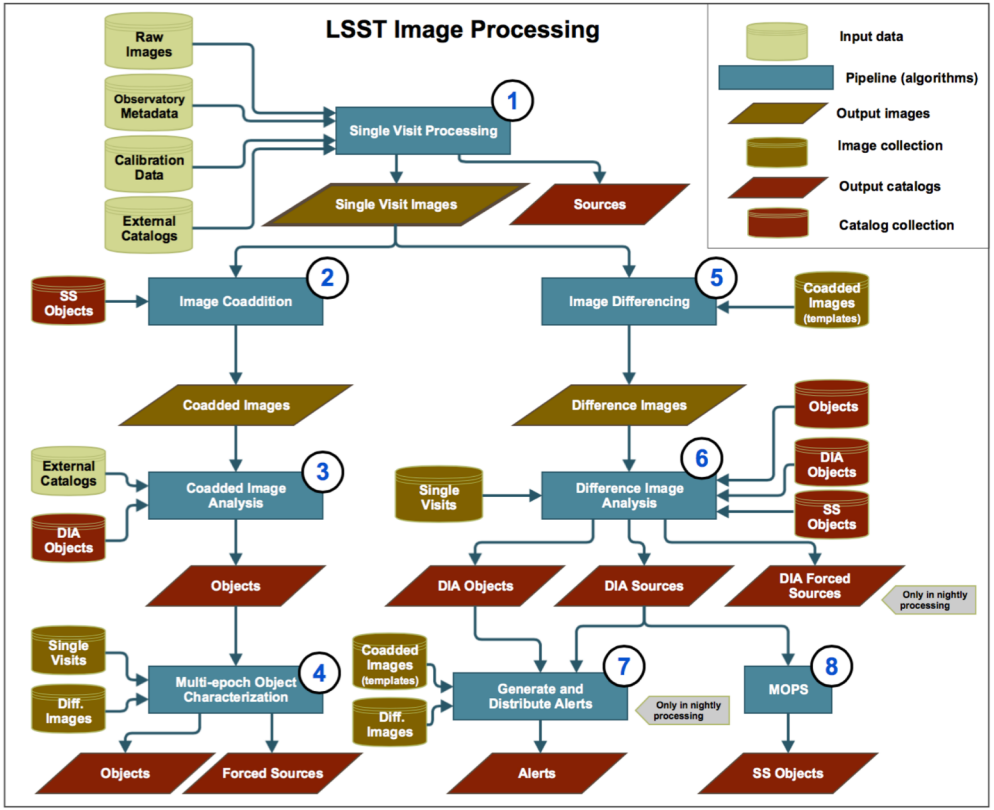
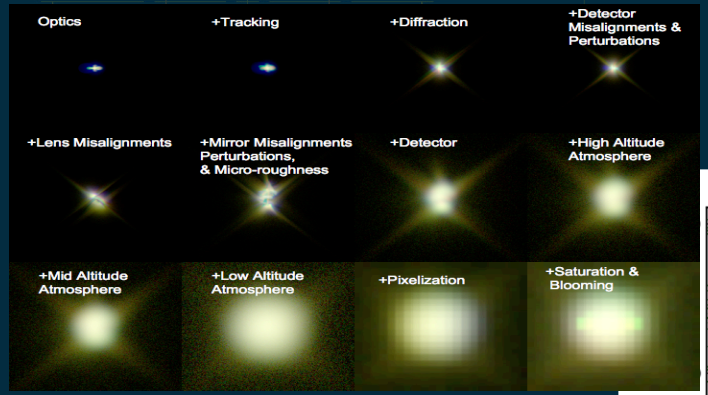


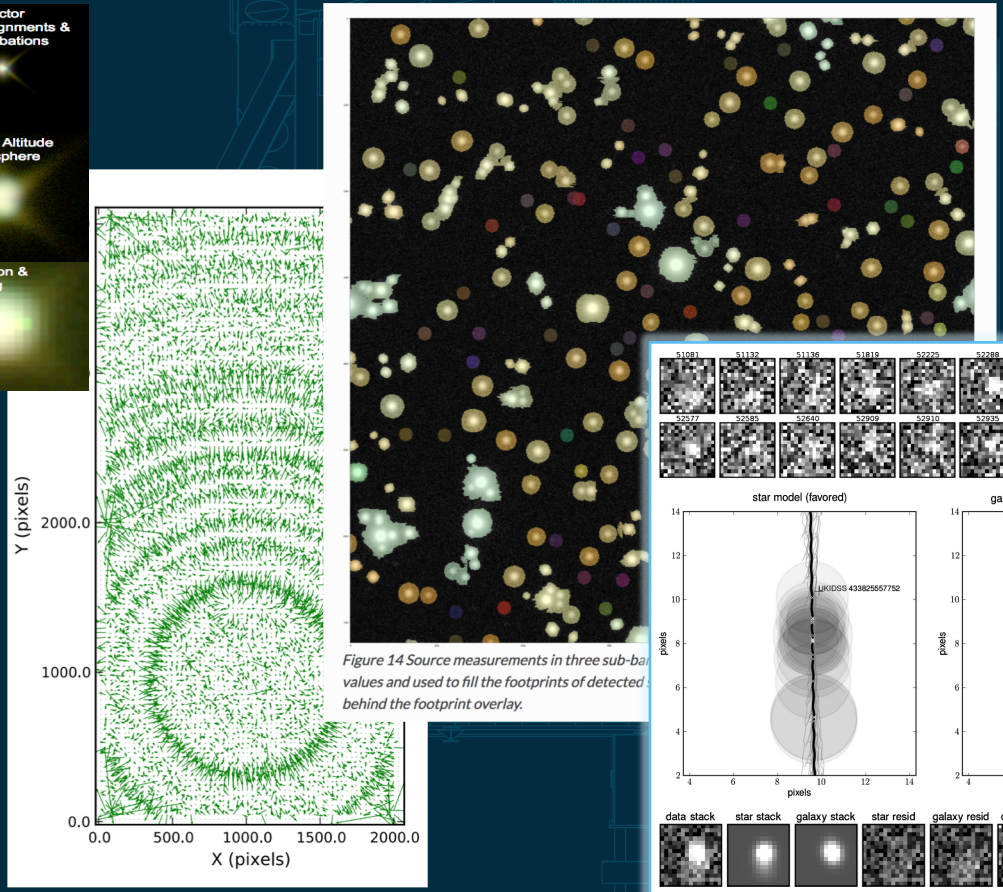
Figure 2: Illustration of the conceptual design of LSST science pipelines for imaging processing.

A Common Theme: Understanding -> Information



Much of the information “loss” comes not from the physical instrument, but the data processing method.

By improving the processing (algorithms, software), we can do a factors of few better than with the exact same dataset.



The “next gen” survey may be just software!

SDSS DR5

Photometric calibration	*r*	*u-g*	*g-r*	*r-i*	*i-z*
	2%	3%	2%	2%	3%

=>

SDSS DR12

Relative photometric calibration accuracy (RMS) (Padmanabhan et al. 2008)

u 1.3%

g 0.8%

r 0.8%

i 0.7%

z 0.8%

Summary

- Measurement used to be simple: “CCDs are linear”, “pixels are independent photon buckets”, ….
- We’re in an era of high precision astronomy; this requires we drop these simplifications
  - Understand the physics of the instrument
  - Properly perform inference (measurement)
- Thinking about these problems can reveal interesting new opportunities.
  - Many instrumental effects are information preserving (or revealing!)
  - It is our poor measurement techniques that erase information.
  - We now have enough computing power to do things right.