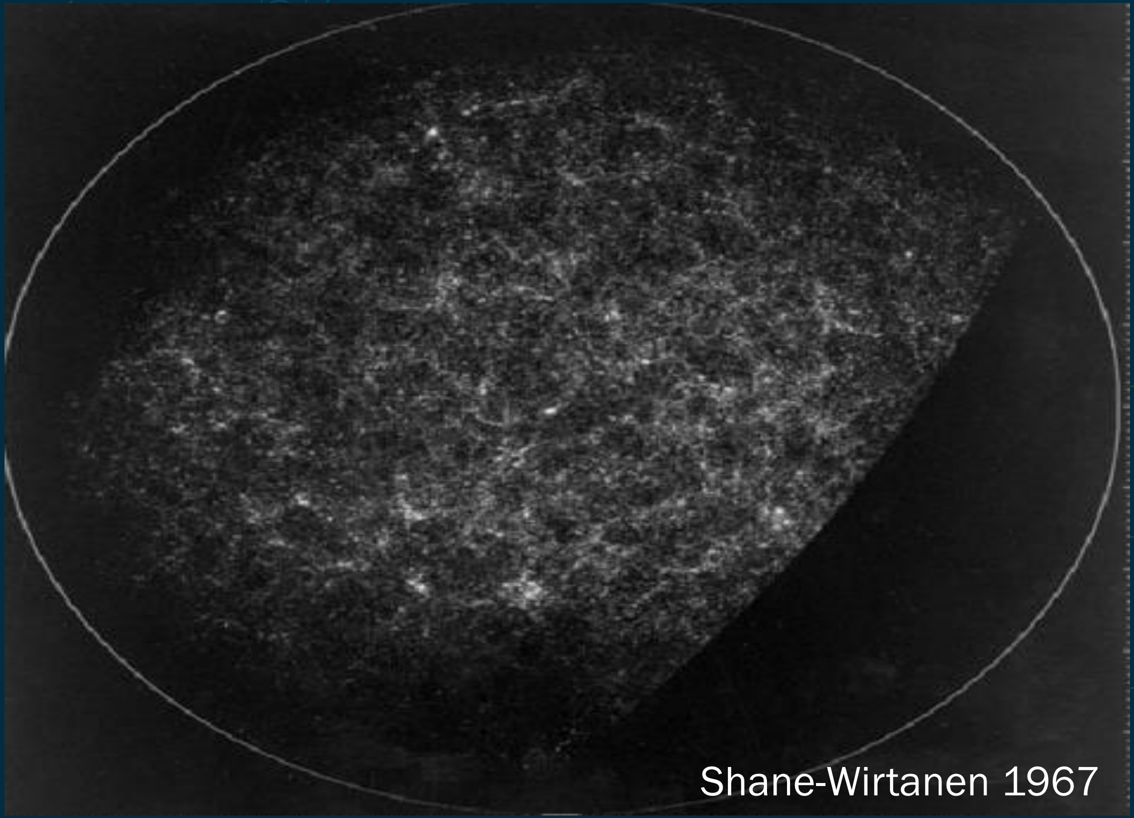
Image Processing in Astronomy: Current Practice & Challenges Going Forward

Mario Juric
University of Washington

With thanks to Andy Connolly, Robert Lupton, Ian Sullivan, David Reiss, and the LSST DM Team

Large Synoptic Survey Telescope

An analog universe (1950s)

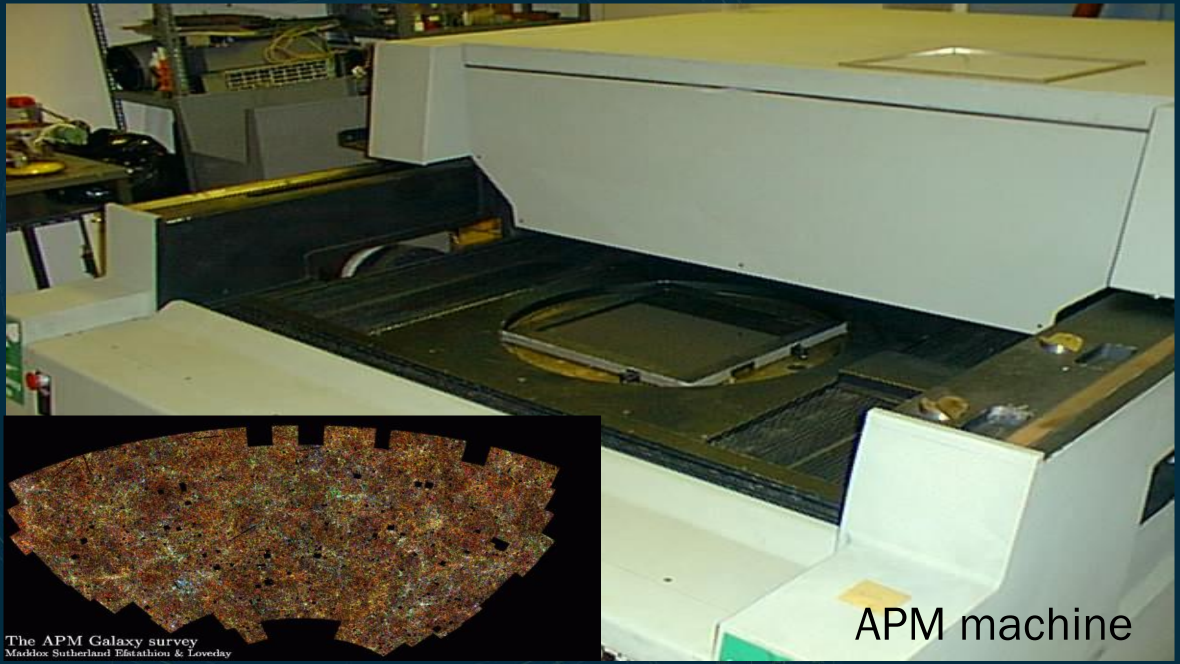


Shane-Wirtanen 1967

Visual inspection:
- 1.6 million “pixels”
- 10x10 arcmin pixels

Processing rate
- Analysis 1947 – 1954
- Published 1967

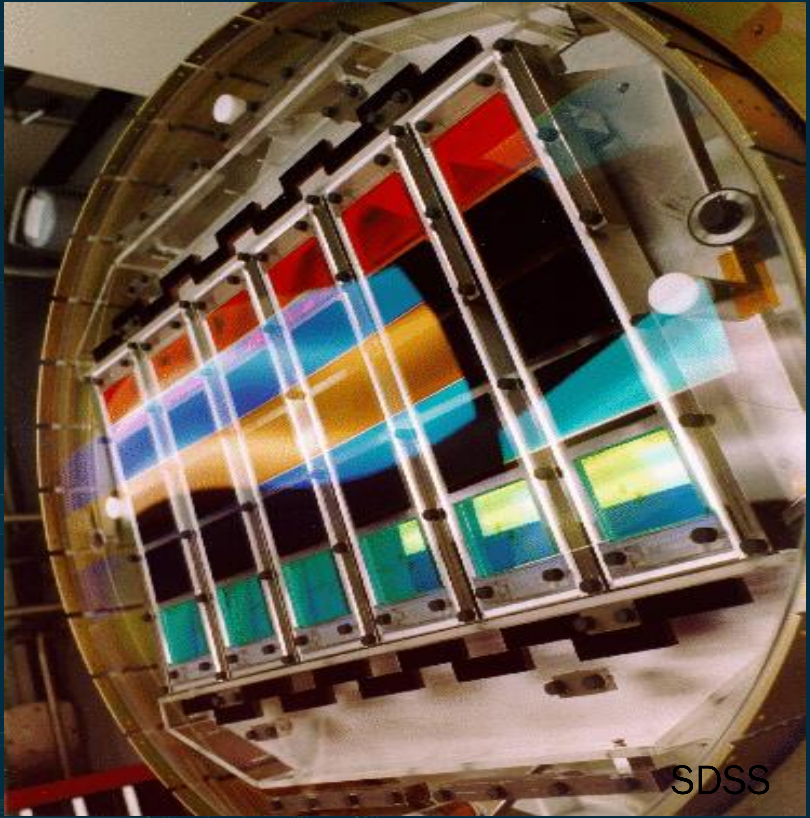
A digitized universe (1980's)



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

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

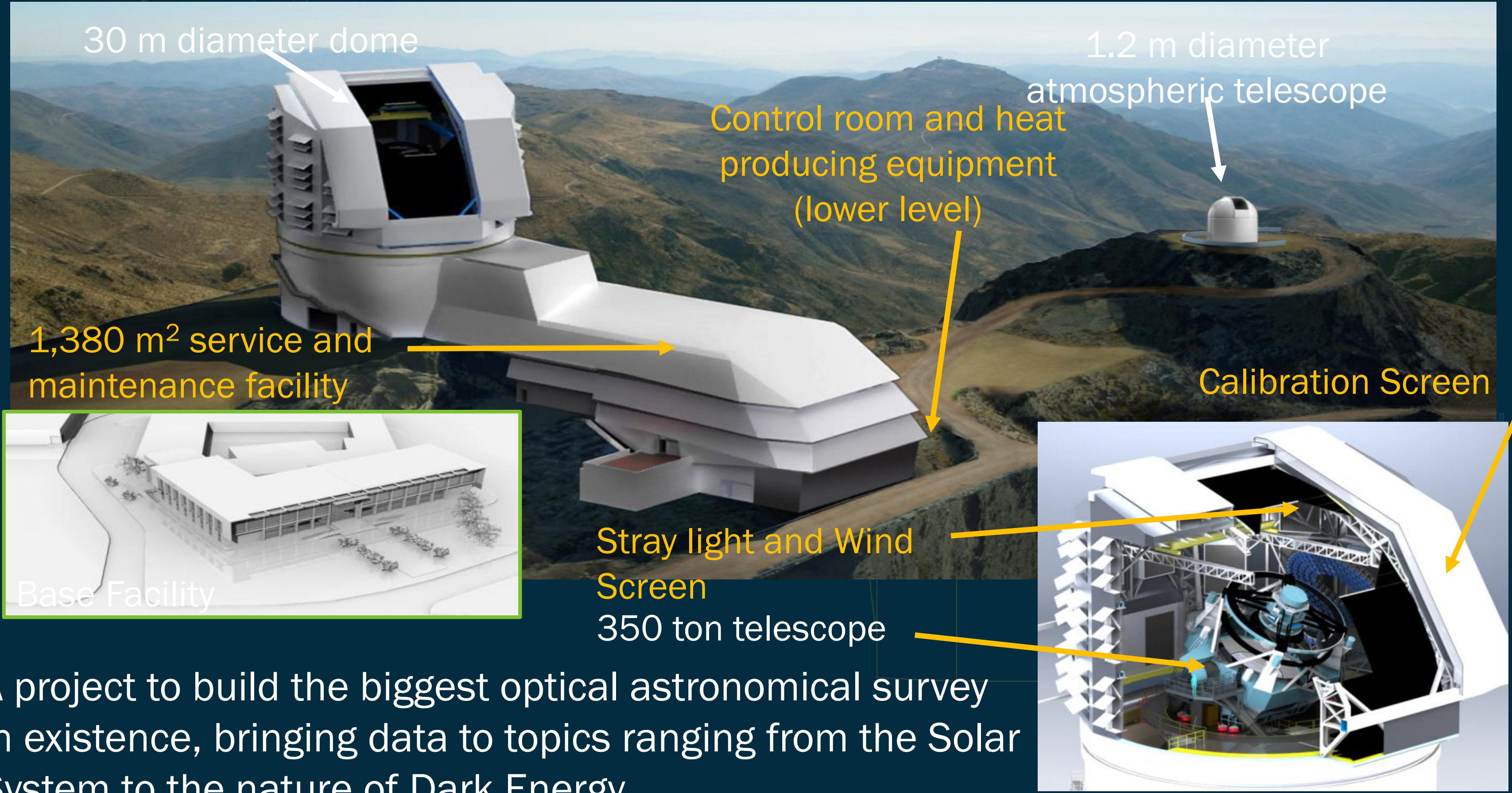
A digit<u>al</u> universe (2000s)



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

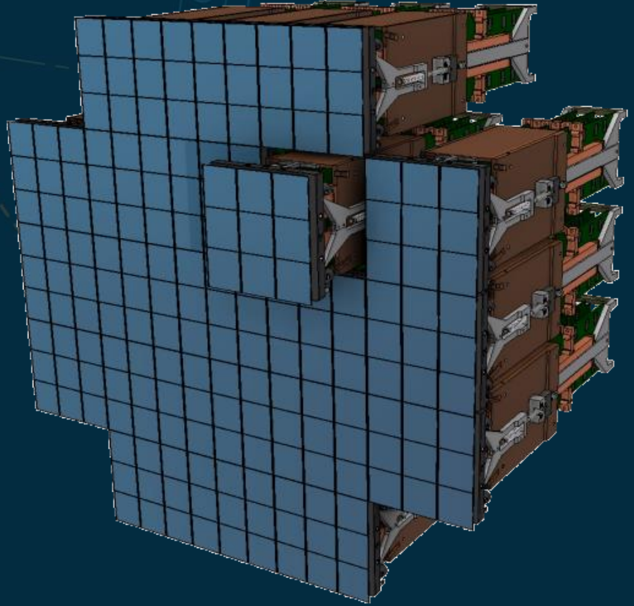
Processing rate
- 4 MB/s
- 250K lines of code

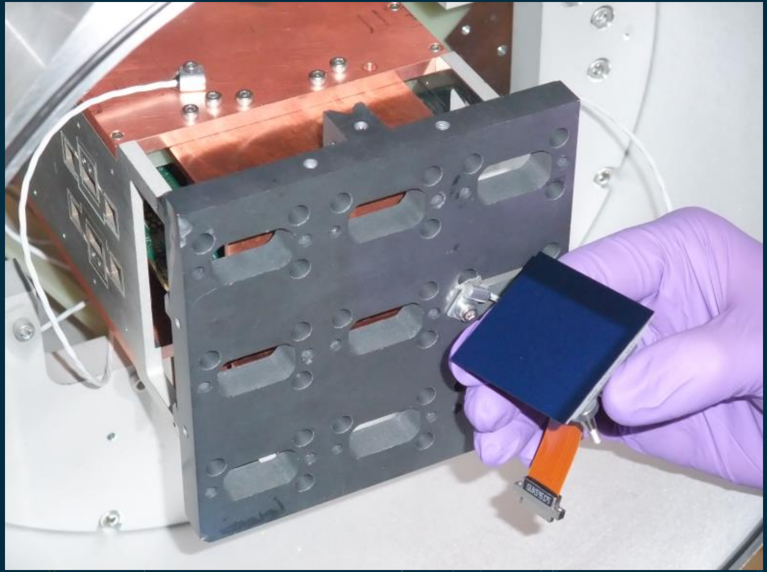
The Large Synoptic Survey Telescope



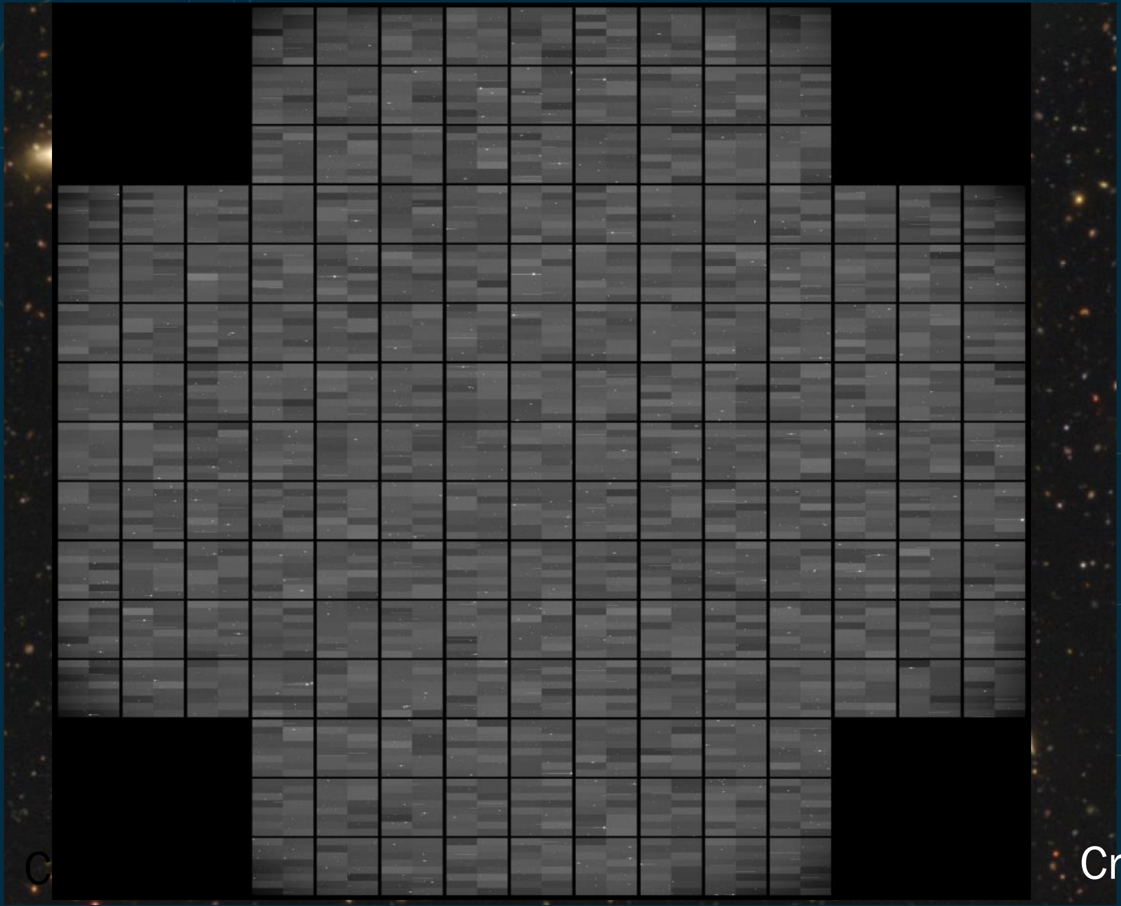
A project to build the biggest optical astronomical survey in existence, bringing data to topics ranging from the Solar System to the nature of Dark Energy.

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:

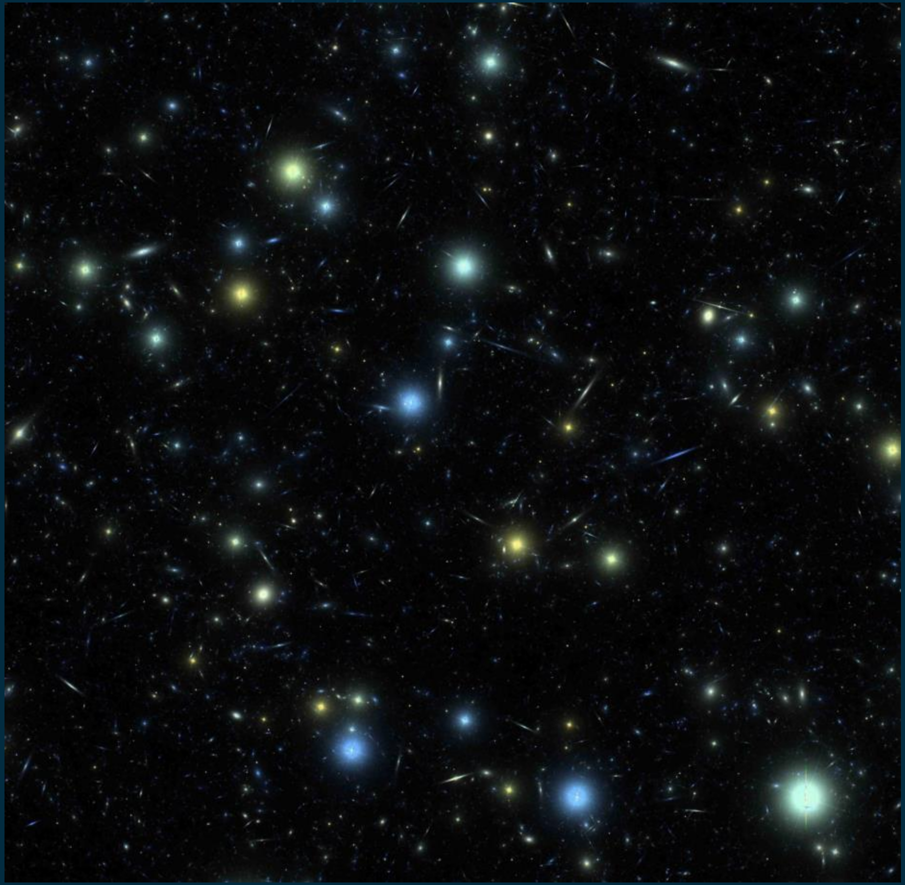
- 6.5×10^{9} pixels sky
- 0.2 arcsec pixels

Processing rate

- 170 MB/s
- 1.1M lines of code

Credit: John Peterson

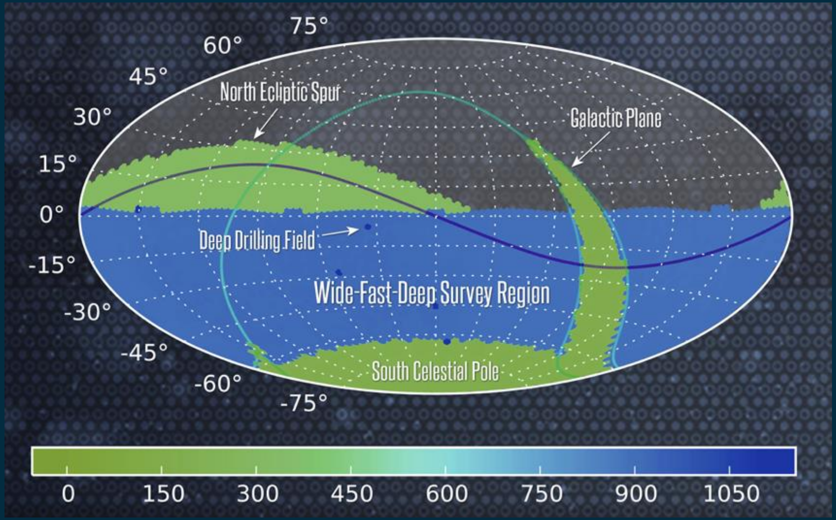
A Big Data Universe



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

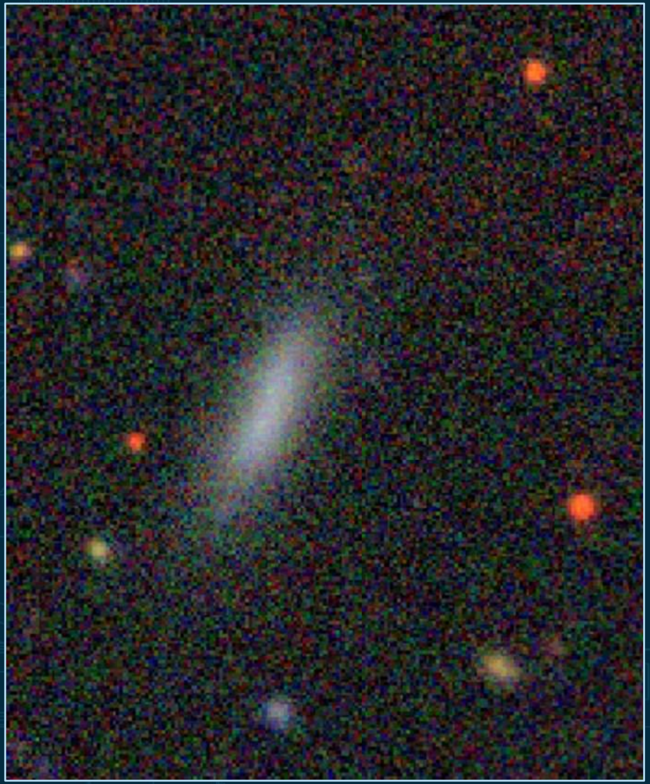
Expect ~2000 exposures per night, 300 nights a year, for 10 years.

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

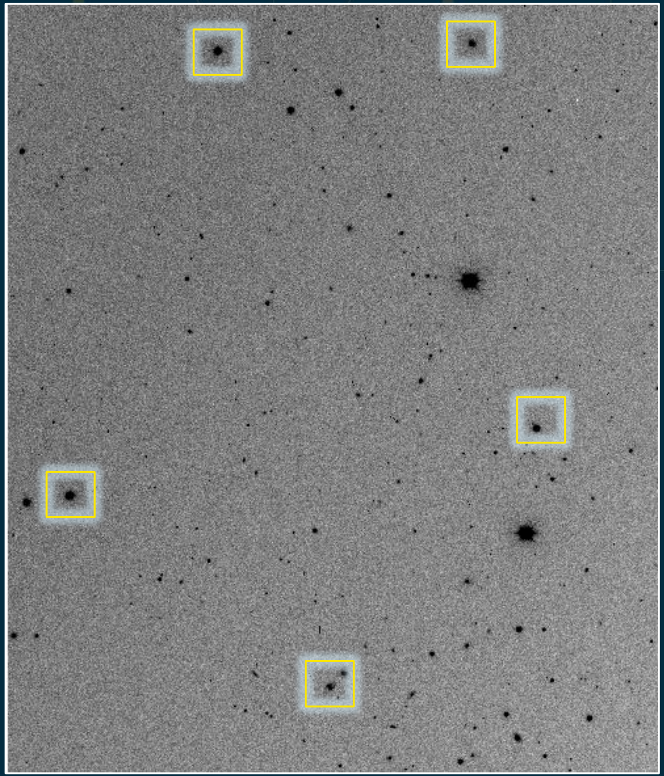


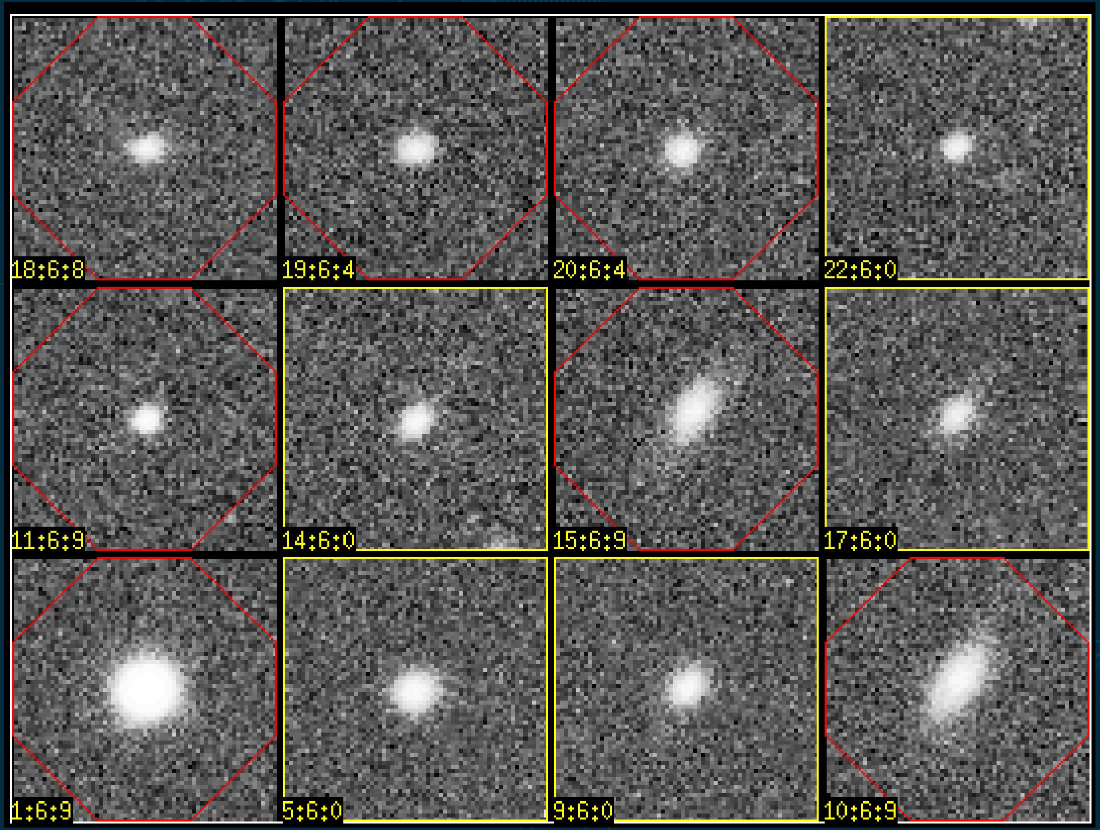
How do astronomical images come to be?

Credit: John Peterson (Purdue) and the PhoSim Team

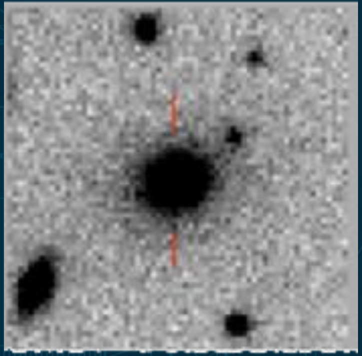

Optics
+Tracking
+Diffraction
+Detector Misalignments & Perturbations
+Lens Misalignments
+Mirror Misalignments Perturbations, & Micro-roughness
+Detector
+High Altitude Atmosphere
+Mid Altitude Atmosphere
+Low Altitude Atmosphere
+Pixelization
+Saturation & Blooming

Point Spread Function Estimation



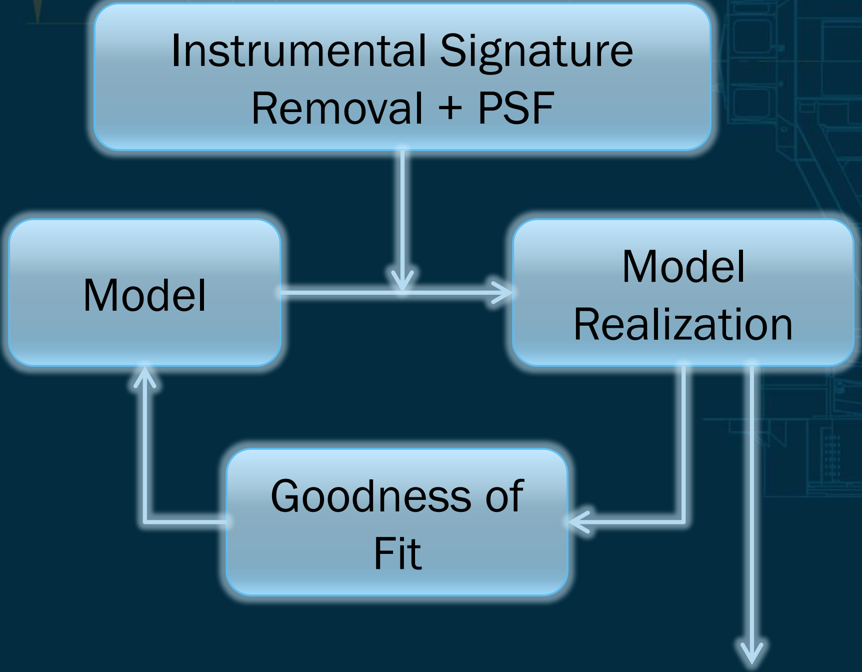


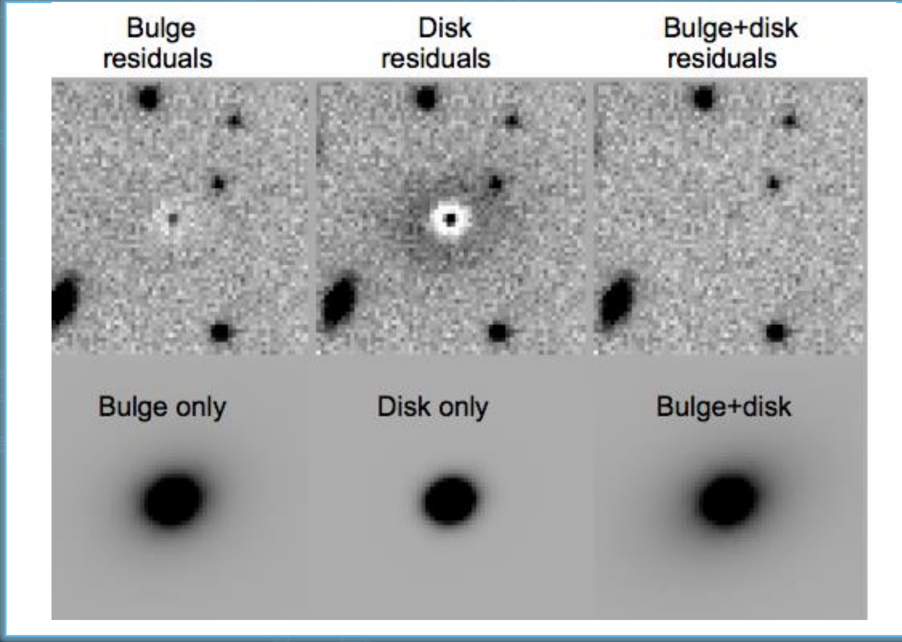
Learning by fitting models



Object characterization (models):

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

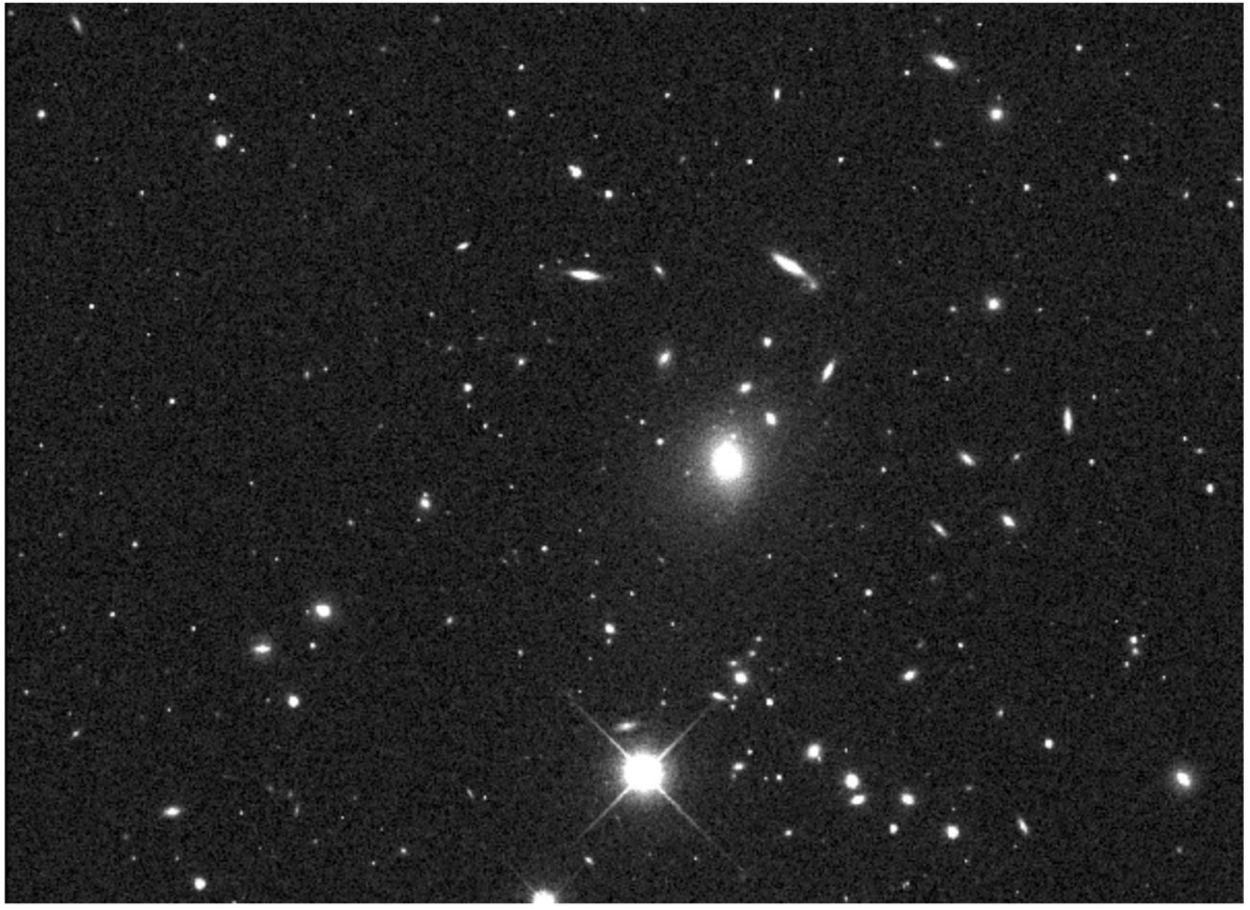




7757,301,1,74,187,6,8.12783867556709,26.627245975921,17.37402,17.92875,0.02894481,0.02568013
7757,301,1,74,188,3,8.1273232252419 2,26.6251199416623,20.1466,21.35297,0.3003744,0.3302762
4288,301,1,39,682,3,24.5161170422305,-1.16579446393527,22.97032,24.3259,0.2672399,0.5240437
4288,301,1,39,683,3,24.5179406515354,-1.1792069022485,22.62052,25.09109,0.1850479,0.6585805
4288,301,1,39,684,6,24.5189463293148,-1.15915086108891,21.4247,23.04125,0.06608655,0.1968172
4136,301,1,61,935,6,36.4715922759092,-1.06093938828308,22.71782,23.14112,0.158014,0.1799687
4136,301,1,61,936,3,36.4717583013136,-1.1378448207726,22.81683,23.88123,0.1742272,0.3260605
4136,301,1,61,937,3,36.4717582434391,-1.13784497192974,22.81147,23.87586,0.1734457,0.3247895
4288,301,1,40,311,3,24.6839203338022,-1.23631696217547,21.2002,21.67521,0.05694564,0.06316777
4288,301,1,40,312,3,24.6840602692246,-1.21784918362007,20.30287,21.04976,0.02972161,0.04032911
4288,301,1,40,313,6,24.6840216690377,-1.08292772289886,24.92263,25.72778,0.68427,0.5471938
5598,301,1,61,792,3,351.787950113407,6.14573538435867,22.43574,23.83793,0.1125768,0.2867745
5598,301,1,61,793,3,351.787950113434,6.14573538316393,22.43573,23.77753,0.1174541,0.2741905
5598,301,1,61,794,6,351.787349107439,6.14481612145222,24.6701,24.8507,0.4675894,0.4849904
2699,301,1,48,527,3,12.0760019016408,-3.32677418219699,22.18116,23.27577,0.126546,0.2270369
2699,301,1,48,528,6,12.0770027529666,-3.32913243320258,22.12757,23.79366,0.1215217,0.3403472
2699,301,1,48,529,3,12.0832728187538,-3.52539818226738,22.29741,23.32008,0.1377919,0.2253349
94,301,1,38,279,6,340.524768659138,-0.843090883870374,20.75028,21.13888,0.04839022,0.04644739
94,301,1,38,280,6,340.52579365662 8,-0.965210498356983,24.23321,26.07633,0.8058653,0.7702702
94,301,1,38,281,6,340.53257887691,-1.02365035629542,21.10608,21.15248,0.06605724,0.04672106
4288,301,1,76,766,3,30.0899167300738,-1.25189466355601,22.57054,22.91144,0.1807321,0.1984752
4288,301,1,76,767,3,30.0899962733195,-1.14314301954175,21.99419,23.78533,0.1094794,0.4162939
4288,301,1,76,768,6,30.089915624703 5,-1.19236549812758,22.64196,22.91776,0.1927483,0.2022232
7937,301,1,84,354,3,5.59132226470183,26.6120638856974,22.63659,23.76799,0.2950225,0.5986399
7937,301,1,84,355,3,5.58984744314193,26.6157353187021,21.27009,22.13992,0.08808753,0.1490507
7937,301,1,84,356,6,5.59007089167375,26.5703384153732,24.52393,26.33231,1.060107,0.6528299
3996,301,1,81,615,6,216.223187662076,11.947
3996,301,1,81,616,3,216.23835672109,12.05342
3996,301,1,81,617,3,216.219379232008,11.9171
6354,301,1,29,2997,6,332.302798261878,41.217
6354,301,1,29,2998,6,332.355200585822,41.244

1,231,051,050 rows (SDSS DR10, PhotoObjAll table)
~500 columns

Cataloging the Sky...



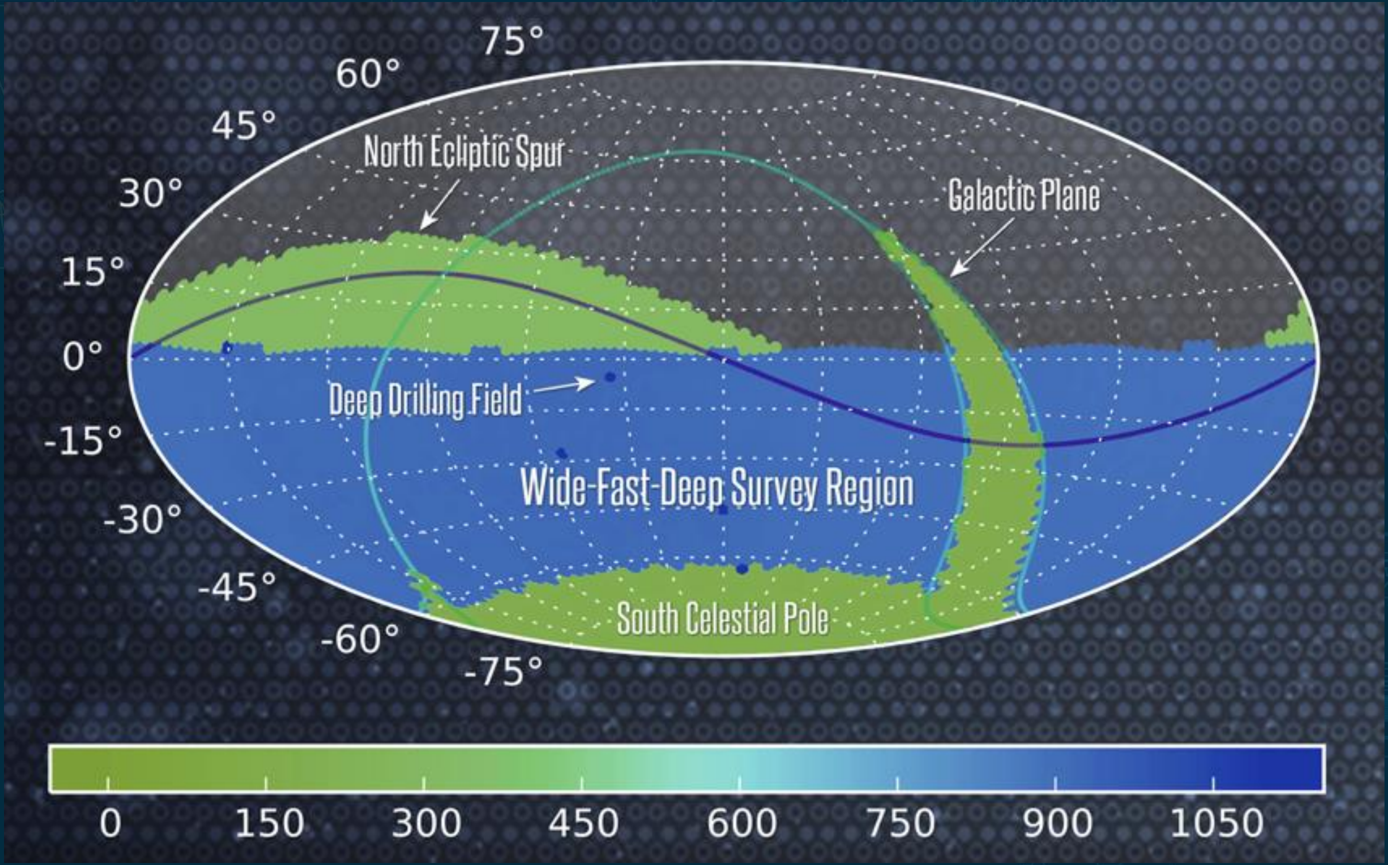
What we're doing is decomposing and modeling the sky in a way that makes physical sense.

… 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 fancy lossy compression technique.

Beyond a Single Image



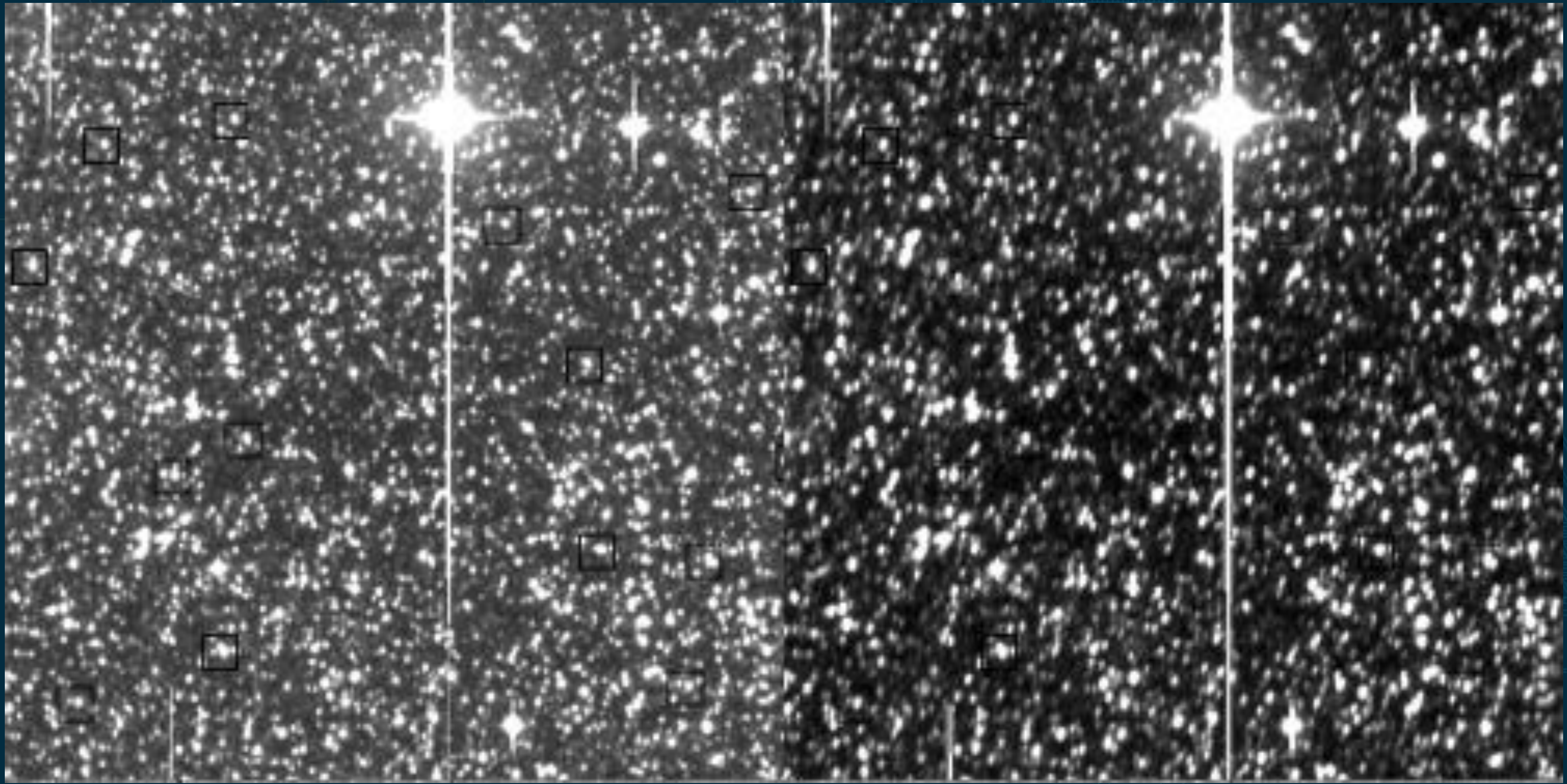
Beyond a Single Image

Going Deep: Coaddition

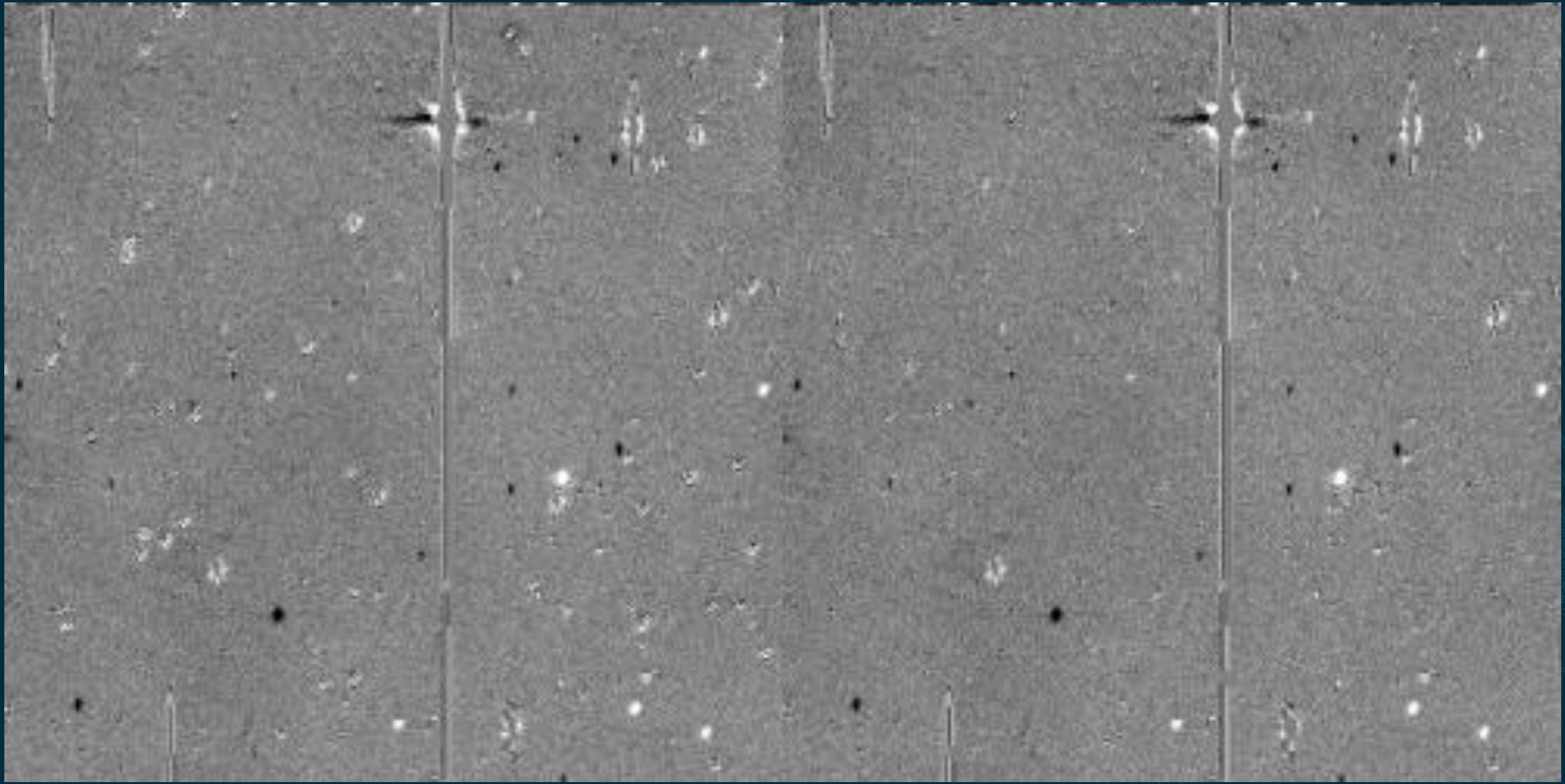
Changes in Time: Image Differencing

Image Differencing

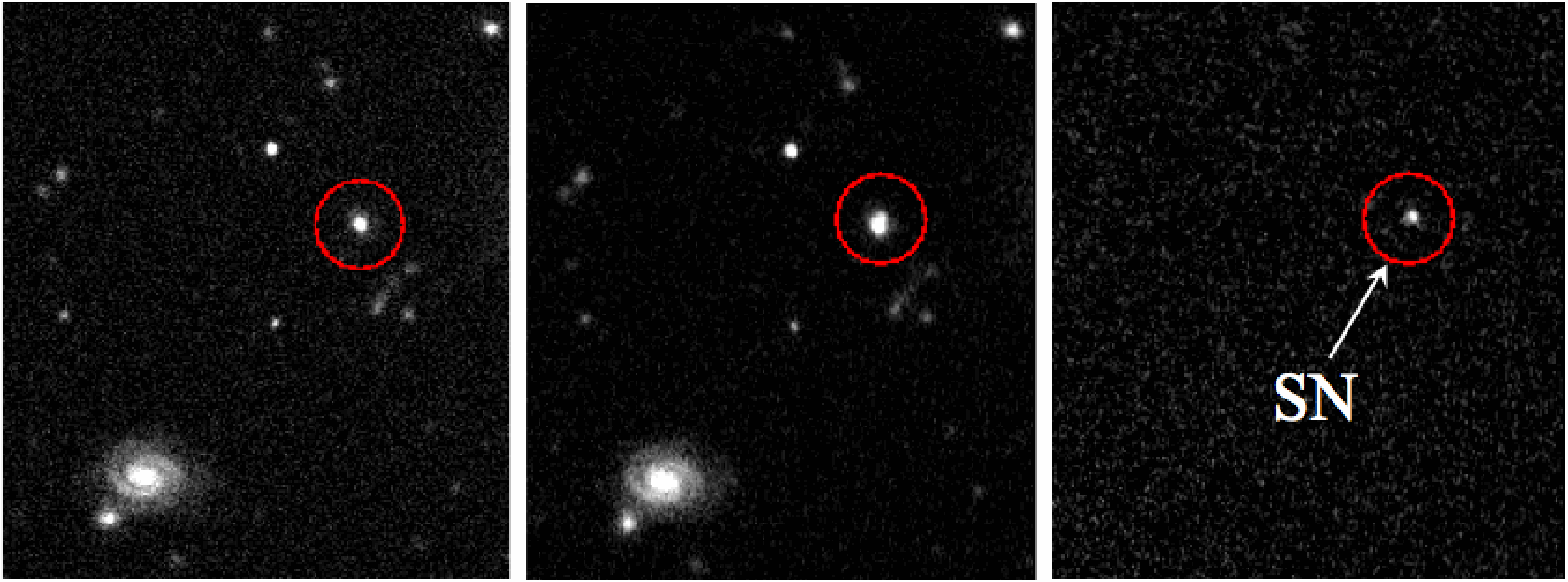
Why Difference?



Why Difference?

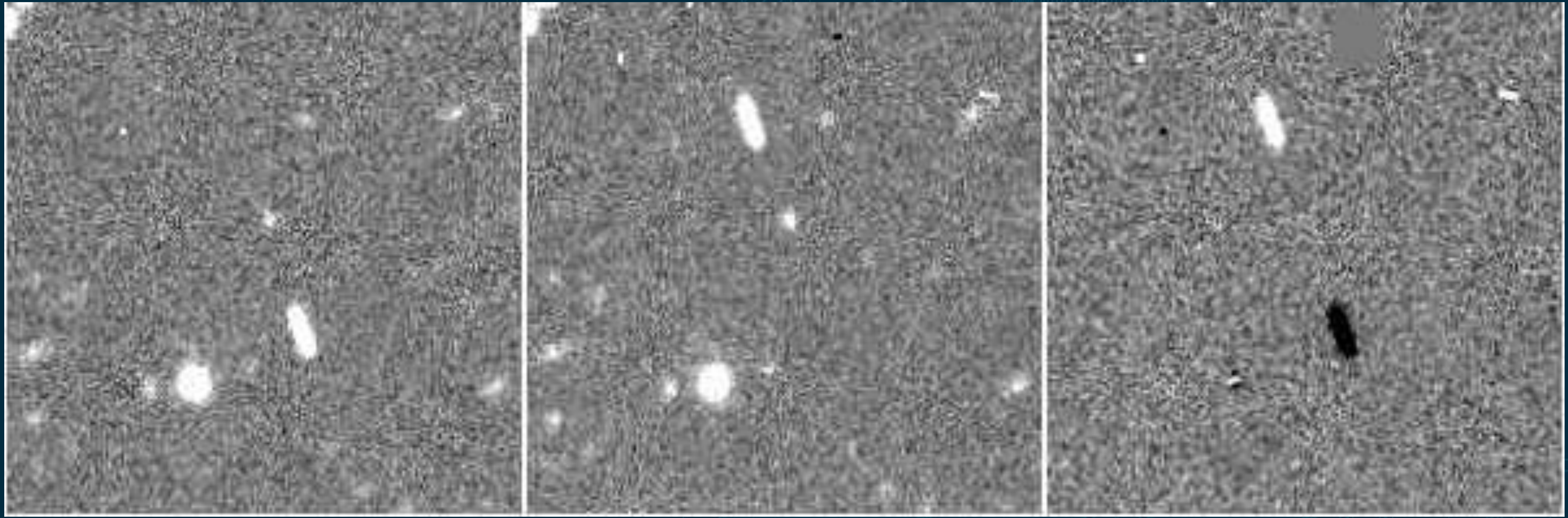


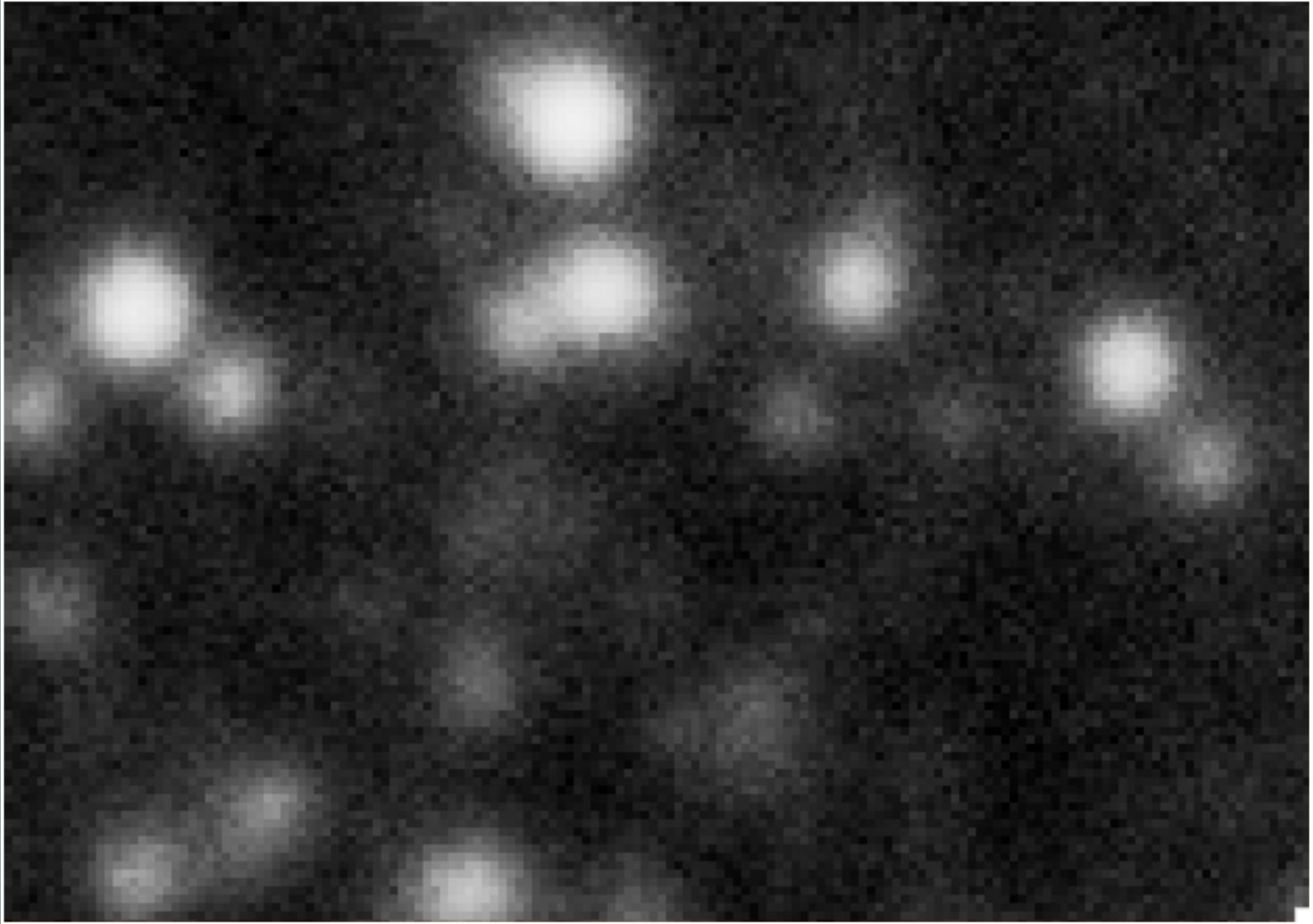
Why Difference?

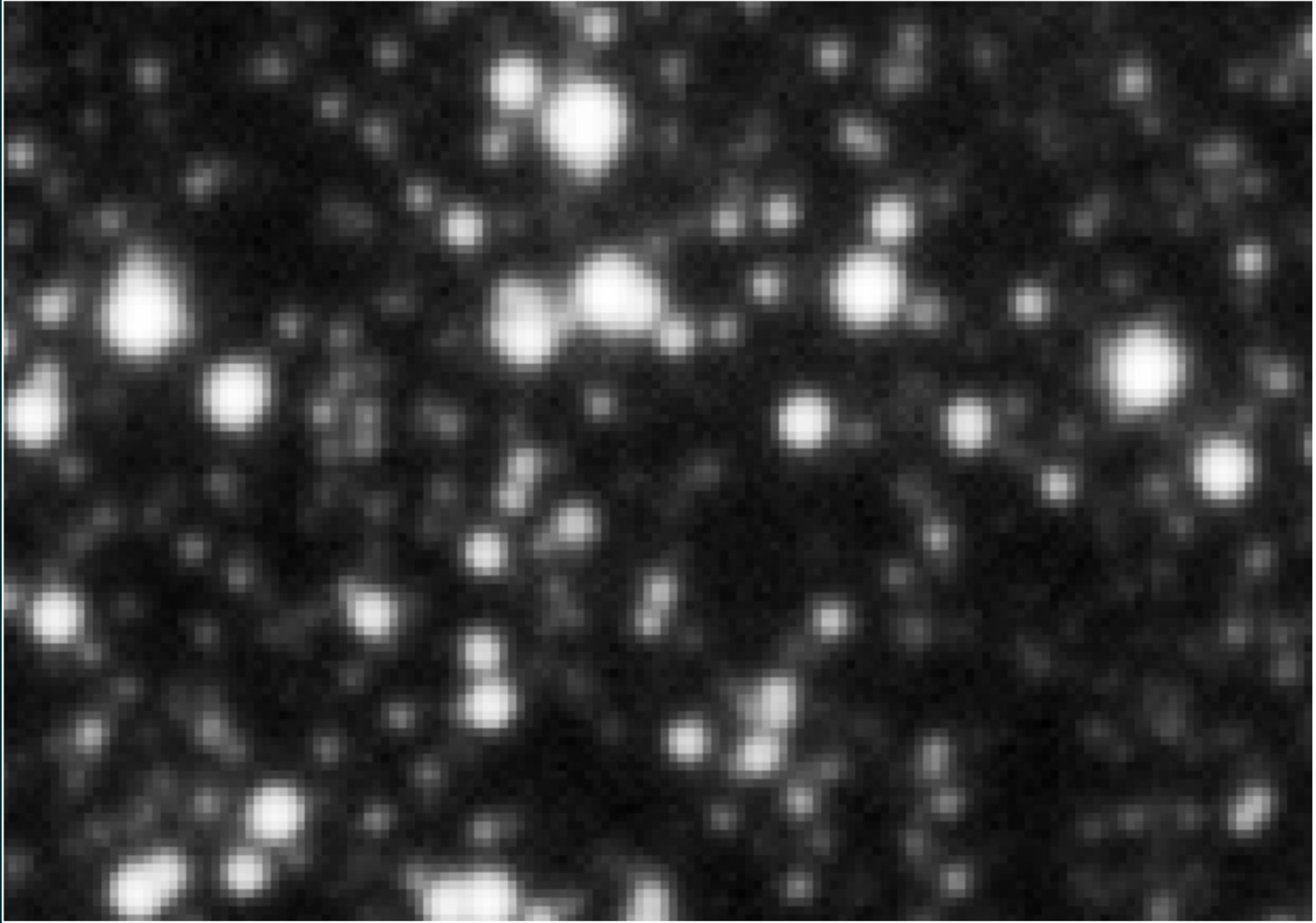


First image Second image Difference image

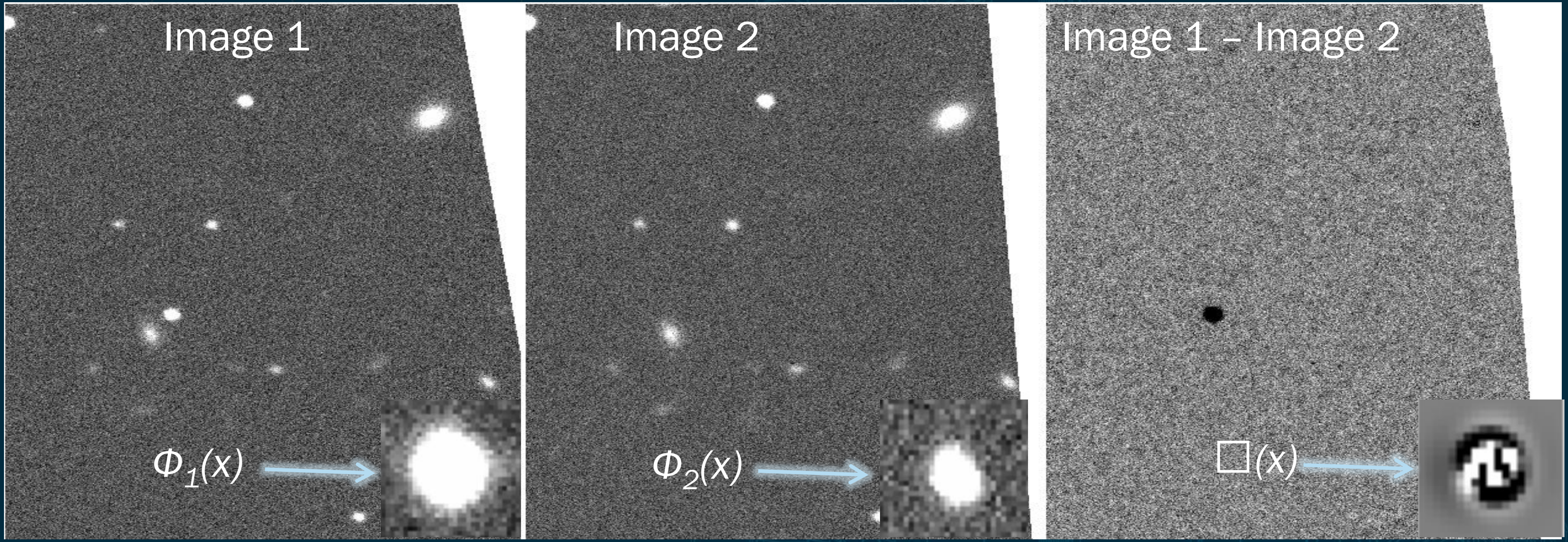
Align, Subtract, Profit! Right?...







Alard-Lupton '97 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

Works well but…

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

Assumes template PSF is smaller than image PSF

- Convolve with a PSF when detecting sources so maybe not a problem

Assumes the “template” has no noise

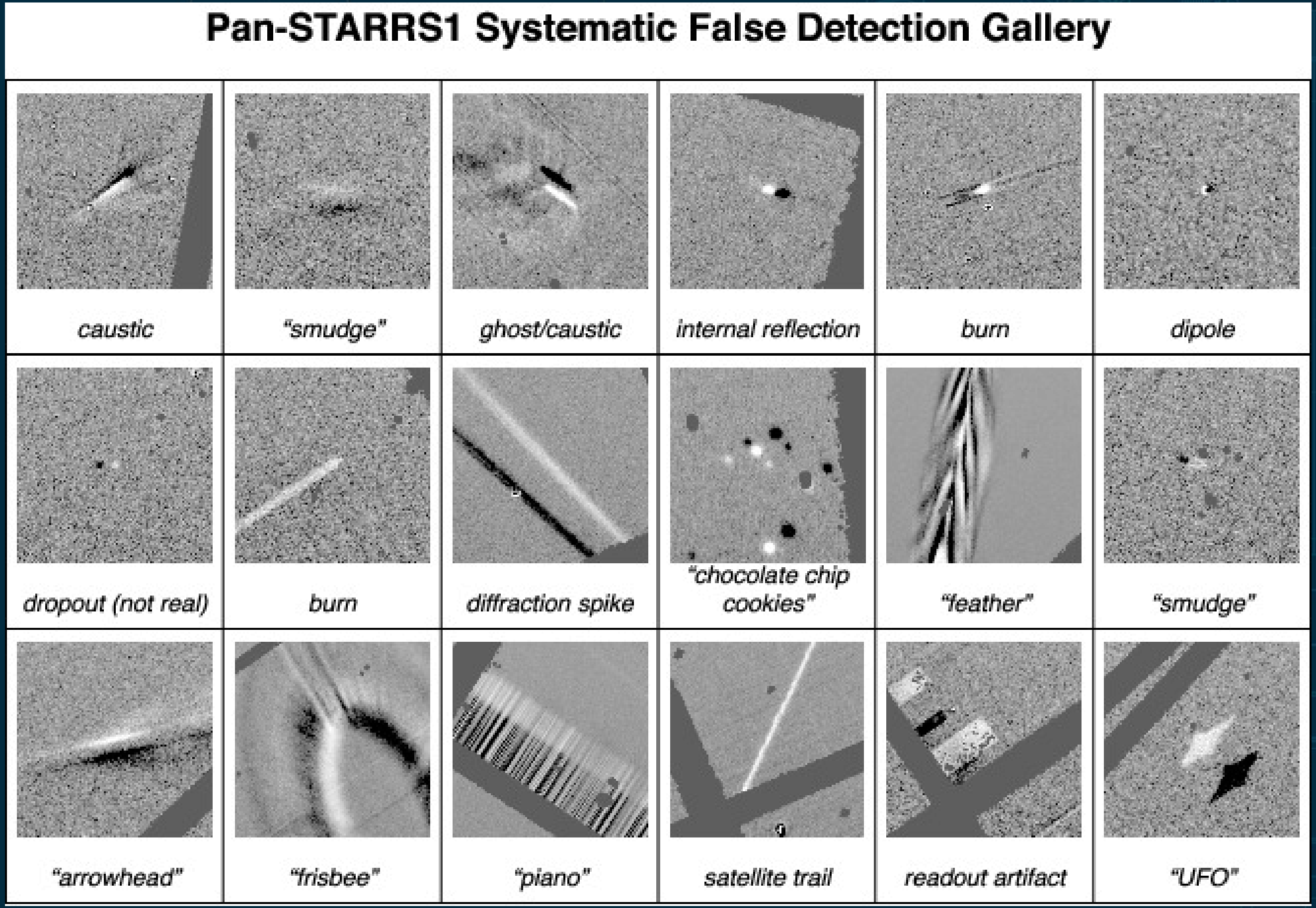
- Derived from a previous set of observations
- Recent work has shown that for Gaussian, heteroschedastic noise we can take a Fourier Transform, and compute the log-likelihood and prewhiten the images

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

Zackay et al 2016

David Reiss @ UW developing production code for LSST

Real world is more complicated...

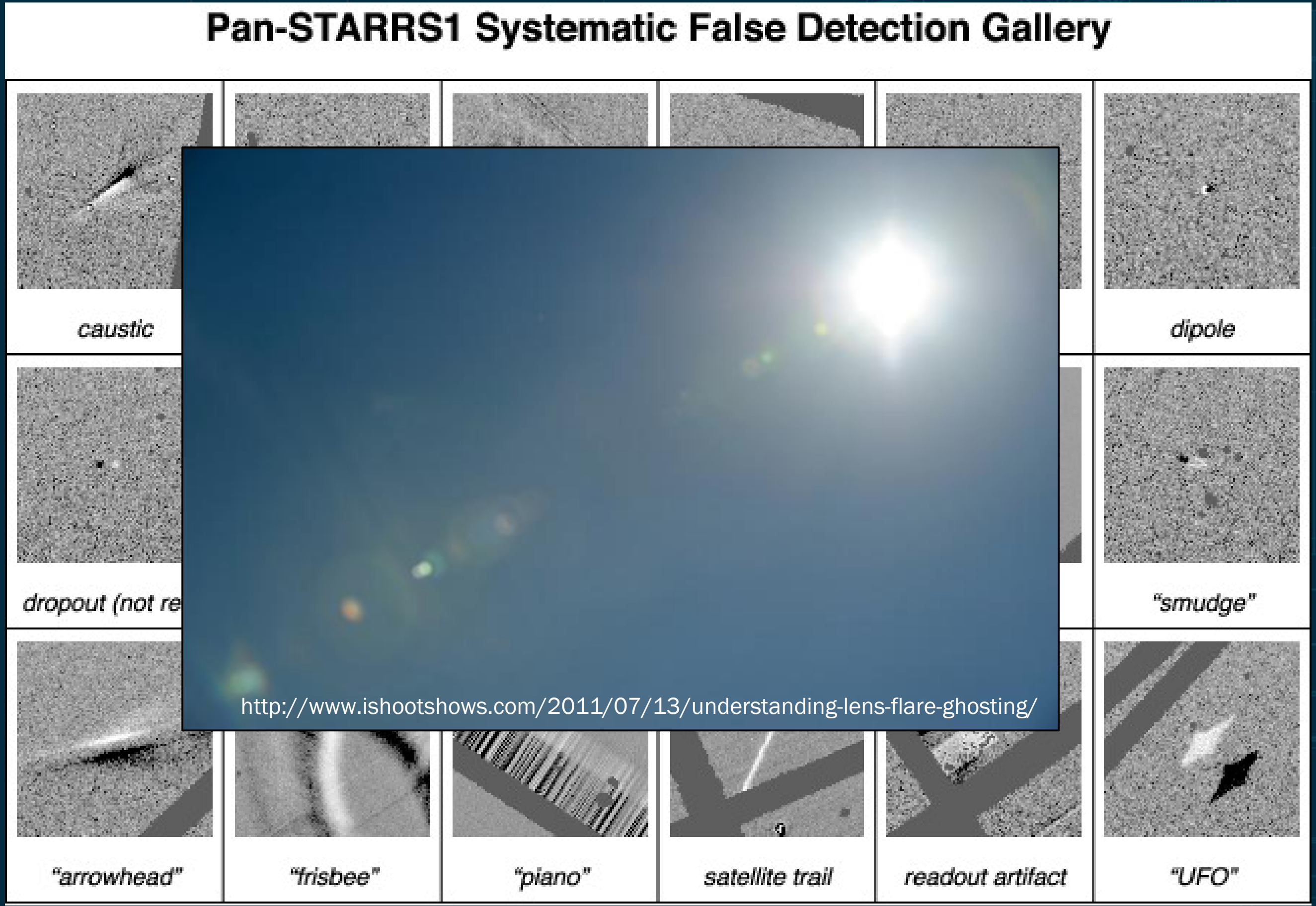


In real-world applications, differencing two images is never perfect. In fact, it's far from perfect!

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

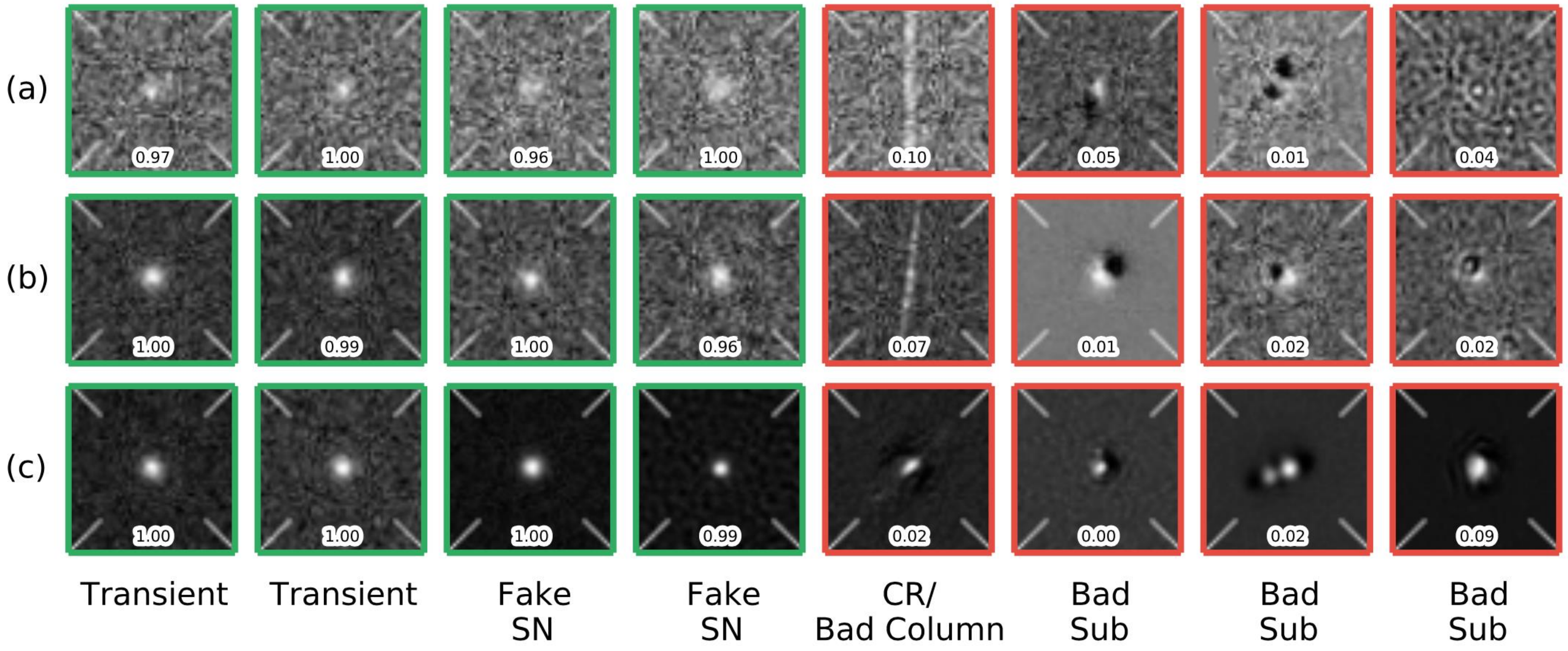
Typically, we see 100:1 to 10:1 false-to-true detection ratios (that is not a typo!).

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).

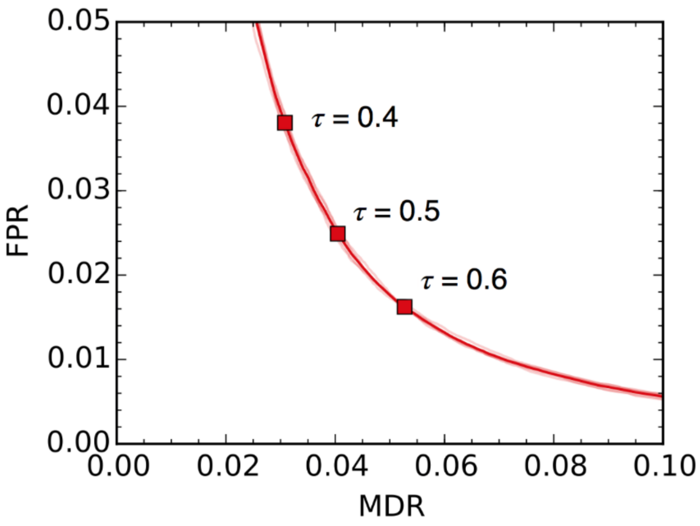


Fig. 7.— 5-fold cross-validated receiver operating characteristics of the best-performing classifier from §3.5. Six visually indistinguishable curves are plotted: one translucent curve for each round of cross-validation, and one opaque curve representing the mean. Points on the mean ROC corresponding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

Fake SN

Bad Column

Sub

Sub

Sub

TABLE 4

…CAN ON REPROCESSED DES Y1 TRANSIENT CANDIDATE …

	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

…se detections, teach the computer to recognize the

…ssign a "score", τ, to each (τ=1 -> real, $\tau = 0$ -> false).

(a)

(b)

(c)

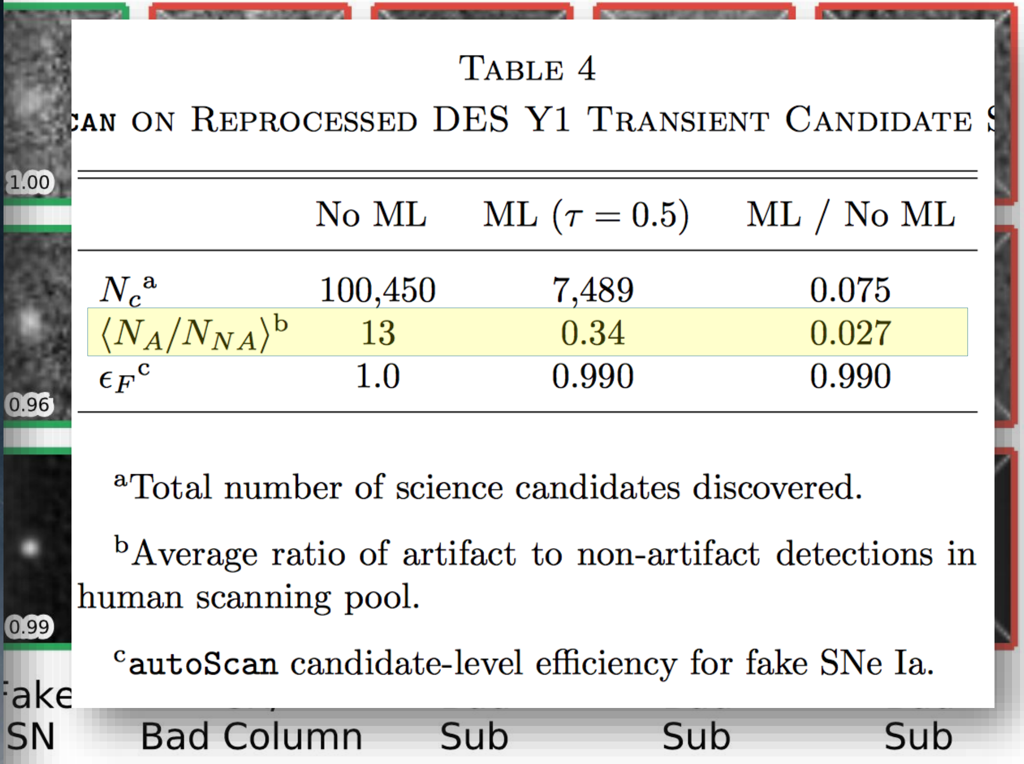


TABLE 4

AN 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.

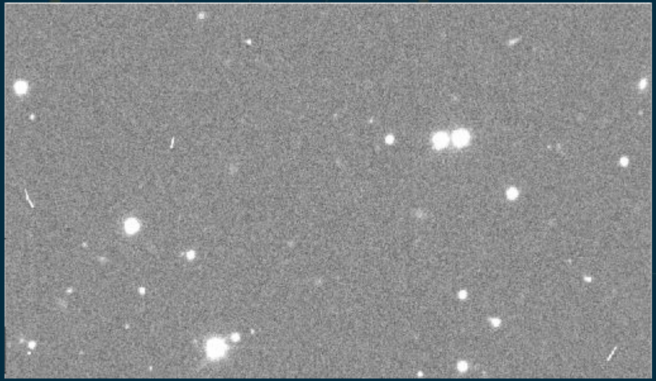
[c] `autoScan` candidate-level efficiency for fake SNe Ia.

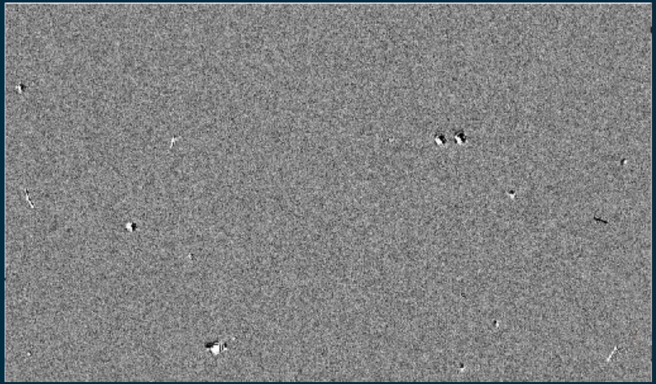
Above: Figure 1, Goldstein al. (2015), AJ, 150, 82

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

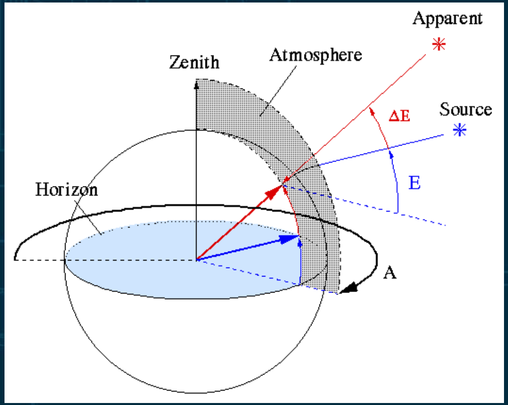




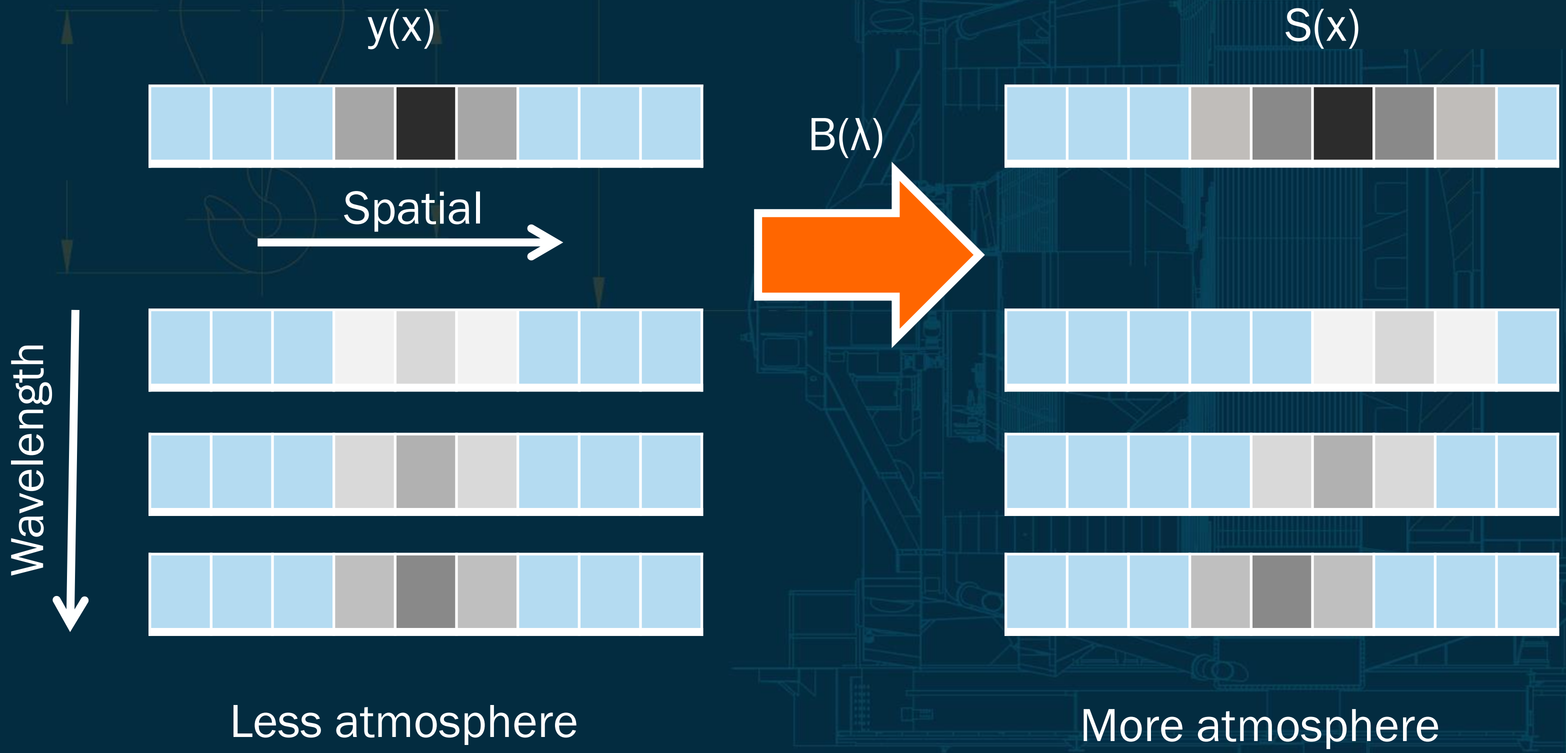
Astrometric misalignment introduces dipoles in the images, misalignment of >2% of a pixels will dominate the number of false positives

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)



Example: differential refraction

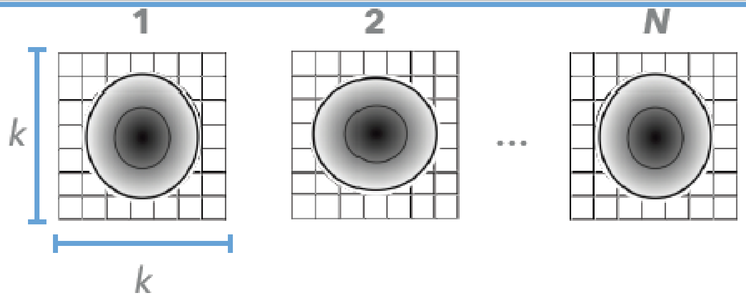


$$S(x) = B(x, \lambda) y_\lambda(x)$$

DCR-corrected Templates: work by Ian Sullivan, UW LSST Group

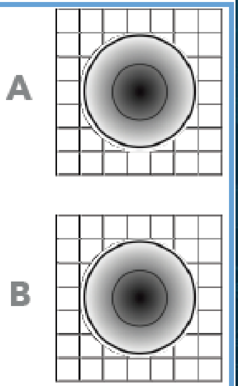
All pixels within a *k x k* size kernel, from all *N* images : $\overrightarrow{s}$

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



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

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



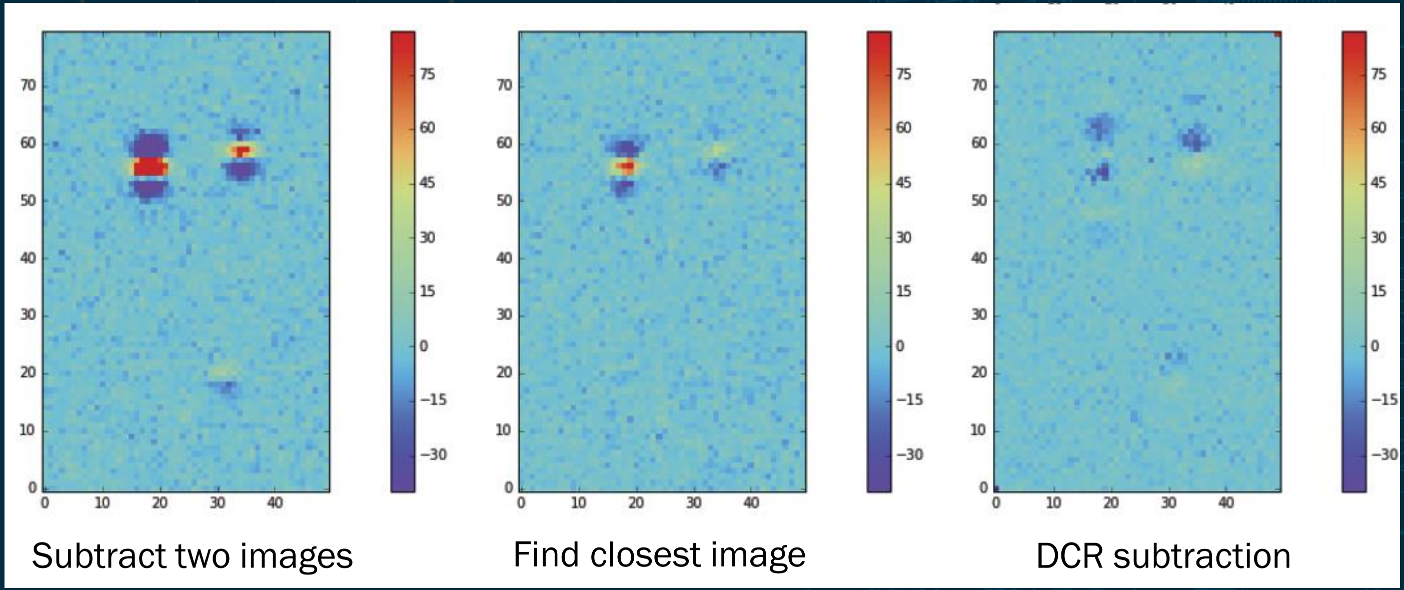
$$QB\overrightarrow{y} = P\overrightarrow{s}$$

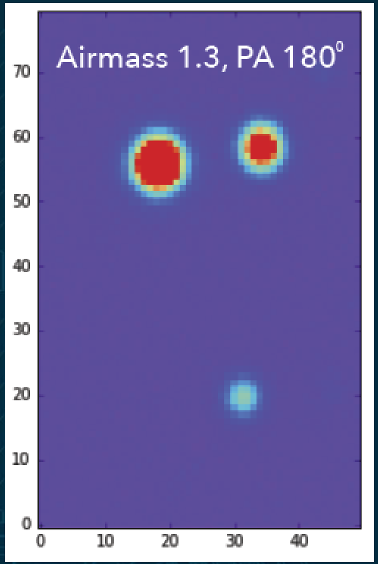
$$\| \vec{y} \| = (B^T B)^{-1} B^T (Q^T Q)^{-1} Q^T P \vec{s}$$

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

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

DCR-corrected Templates: work by Ian Sullivan, UW LSST Group





A non-negative LS algorithm works but is slow, requires small (50x50 patch solutions)
Regularized solution is about 700 hrs (for full LSST image)

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

Why co-add?

SDSS Southern Coadd 9

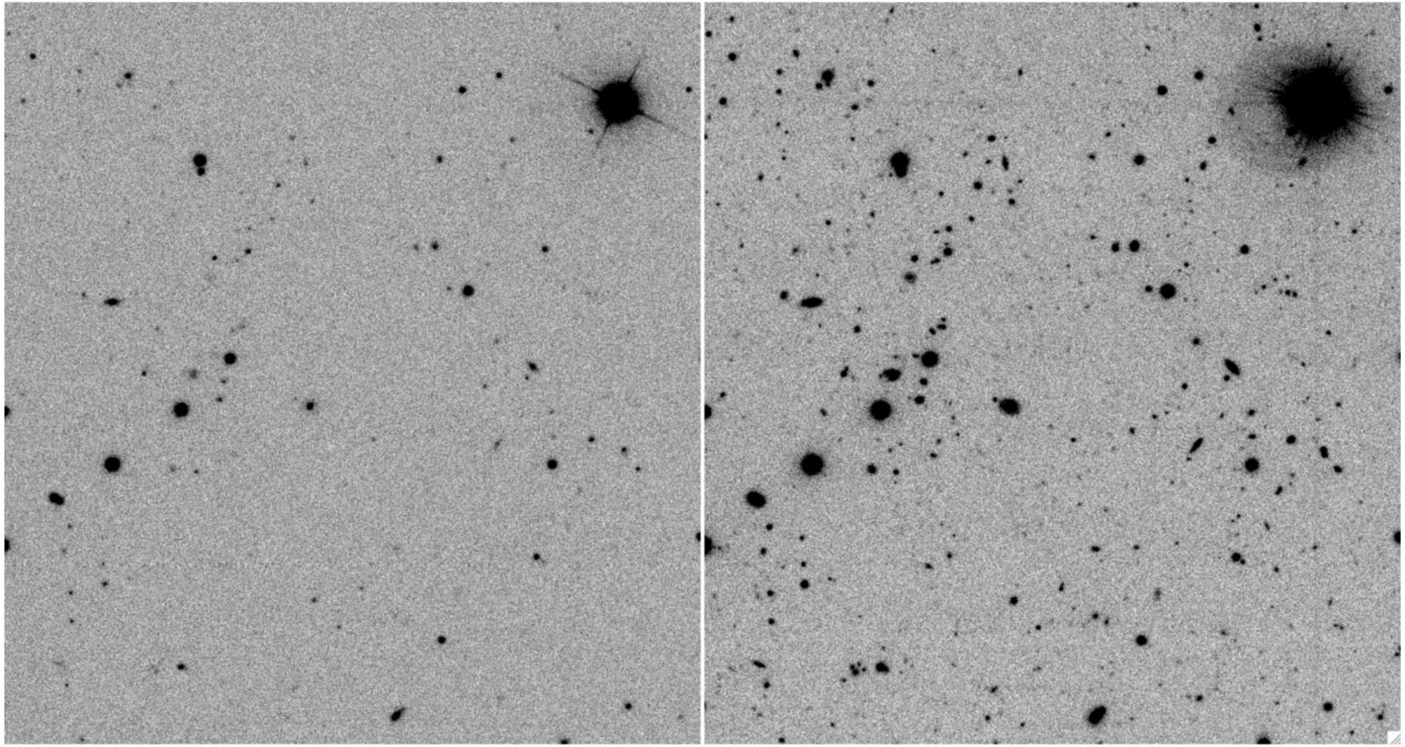


FIG. 2.— Comparison between single pass (left) and coadd (right) images in r-band for run 206, camcol 3, field 505, RA=15, Dec=0. Images are shown with the same scale, contrast and stretch. The single pass counterpart (run 5800, camcol 3, field 505) is one out of 28 images used in the coaddition of this 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.

<u>Pros</u>:

See fainter objects!

Computationally inexpensive

Left: Annis et al. 2011

Why co-add?

SDSS Southern Coadd 9

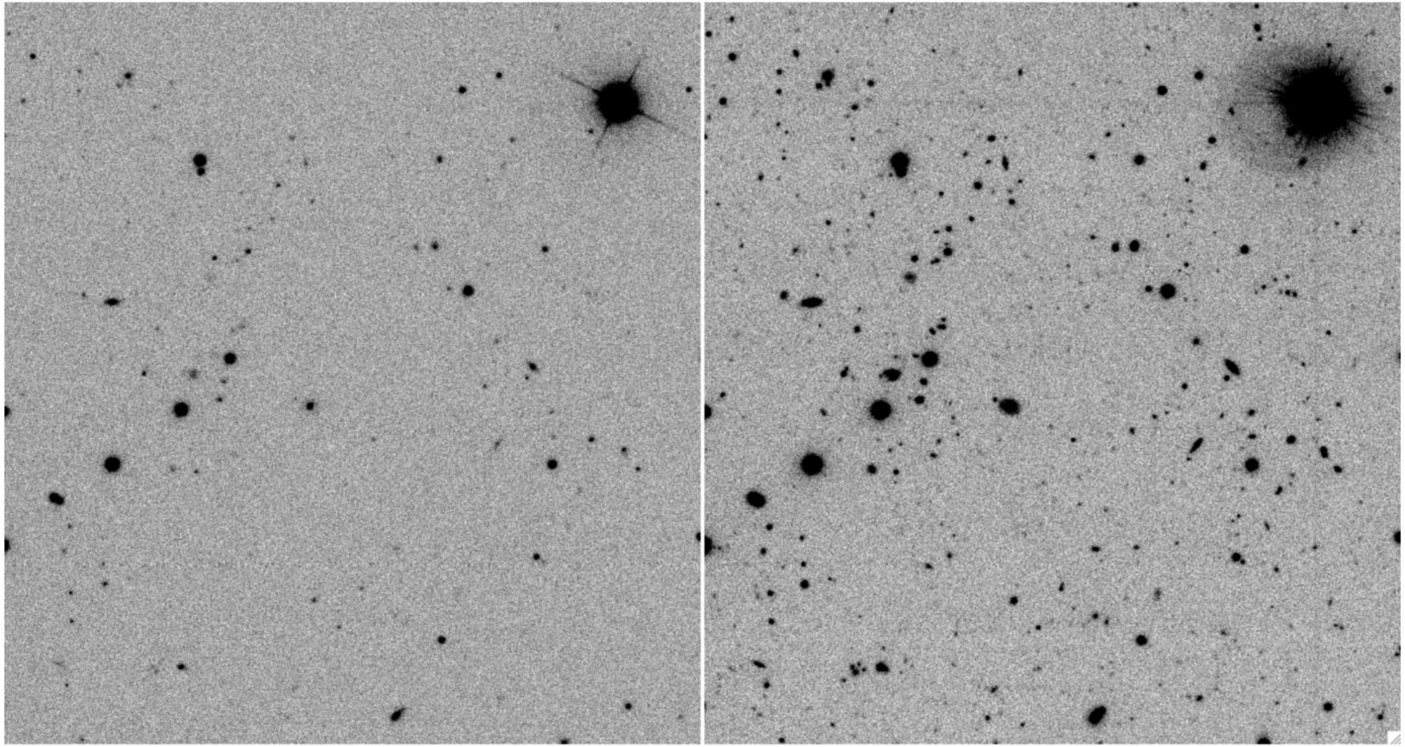


FIG. 2.— Comparison between single pass (left) and coadd (right) images in r-band for run 206, camcol 3, field 505, RA=15, Dec=0. Images are shown with the same scale, contrast and stretch. The single pass counterpart (run 5800, camcol 3, field 505) is one out of 28 images used in the coaddition of this 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.

<u>Pros</u>:

See fainter objects!

Computationally inexpensive

<u>Cons</u>:

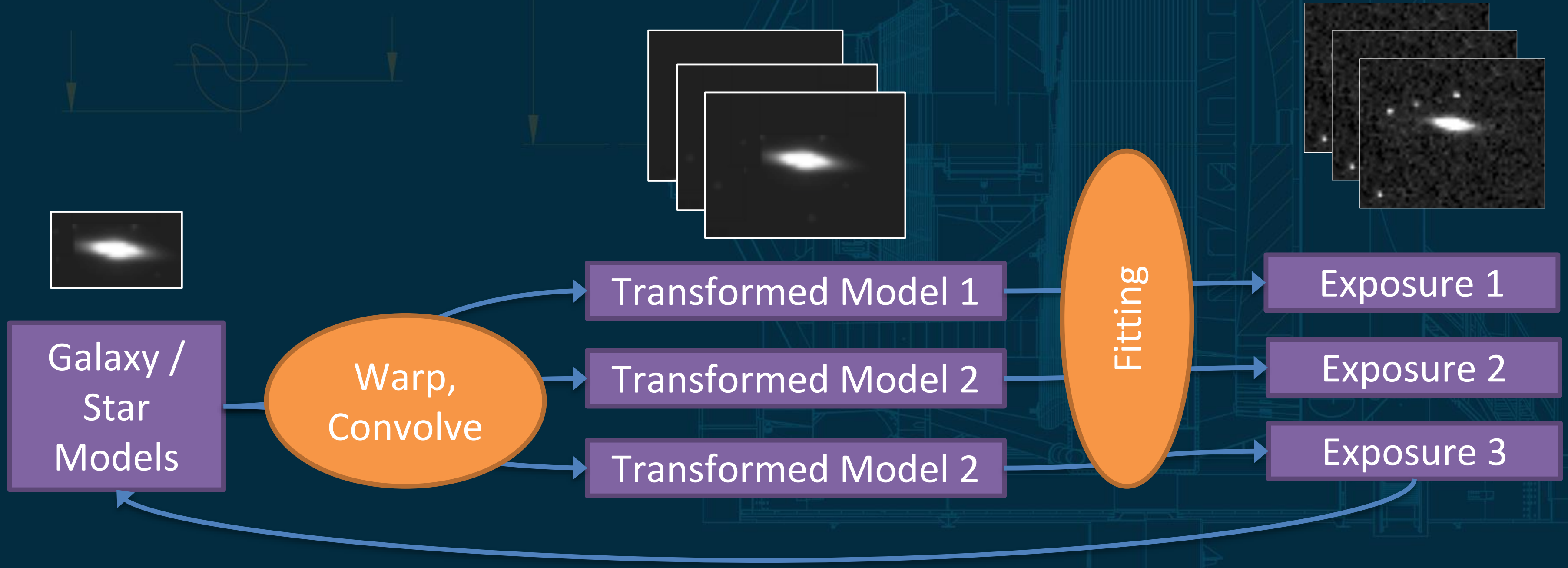
- Complicated PSF
- Correlated noise
- Loss of motion and time variability information
- Loss of information.

Left:

Annis et al. 2011

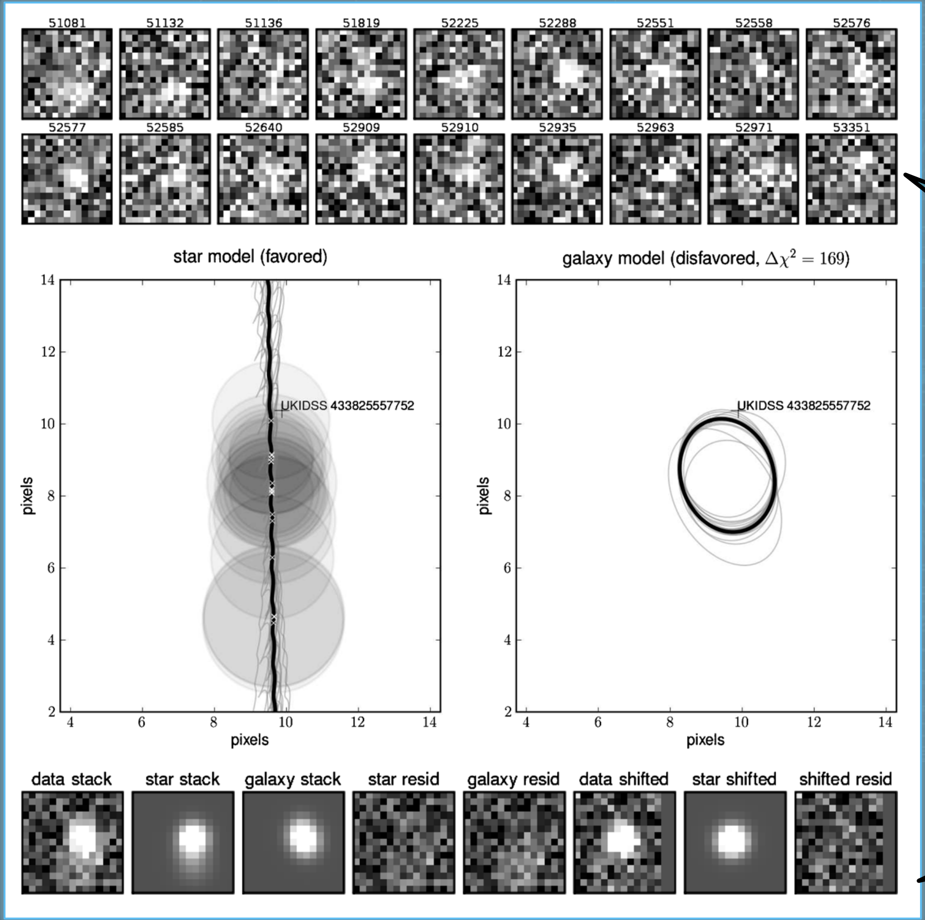
Better than Coaddition: Multi-Epoch Fitting (MultiFit)

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)

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 super 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!)

We're building a ~2 PFLOP machine to do this (cca ~2025). Still, this is all cheaper than building a bigger telescope (and/or launching it into space)!

Where we (know we) could use some help

- Detection and characterization of diffuse sources
- Rejection of instrumental artifacts; how do we get better at removing ghosts, glints, flares, etc.
- Multi-epoch image processing (scene reconstruction)
  - We fit models of individual sources, but not the whole scene. Can we do better?
  - Caveat: We deeply care about the measured object quantities (e.g., fluxes, shapes)
- Efficient implementations of existing algorithms

- Probably many other places that we don't know about!