Wide Field Imaging Surveys

The Challenges of Turning Photons into Science

Bryan Scott

CIERA/Northwestern

Adapted from talks by Adam Miller and Mario Juric

What is a survey?

What is a survey?

Proposed definition:

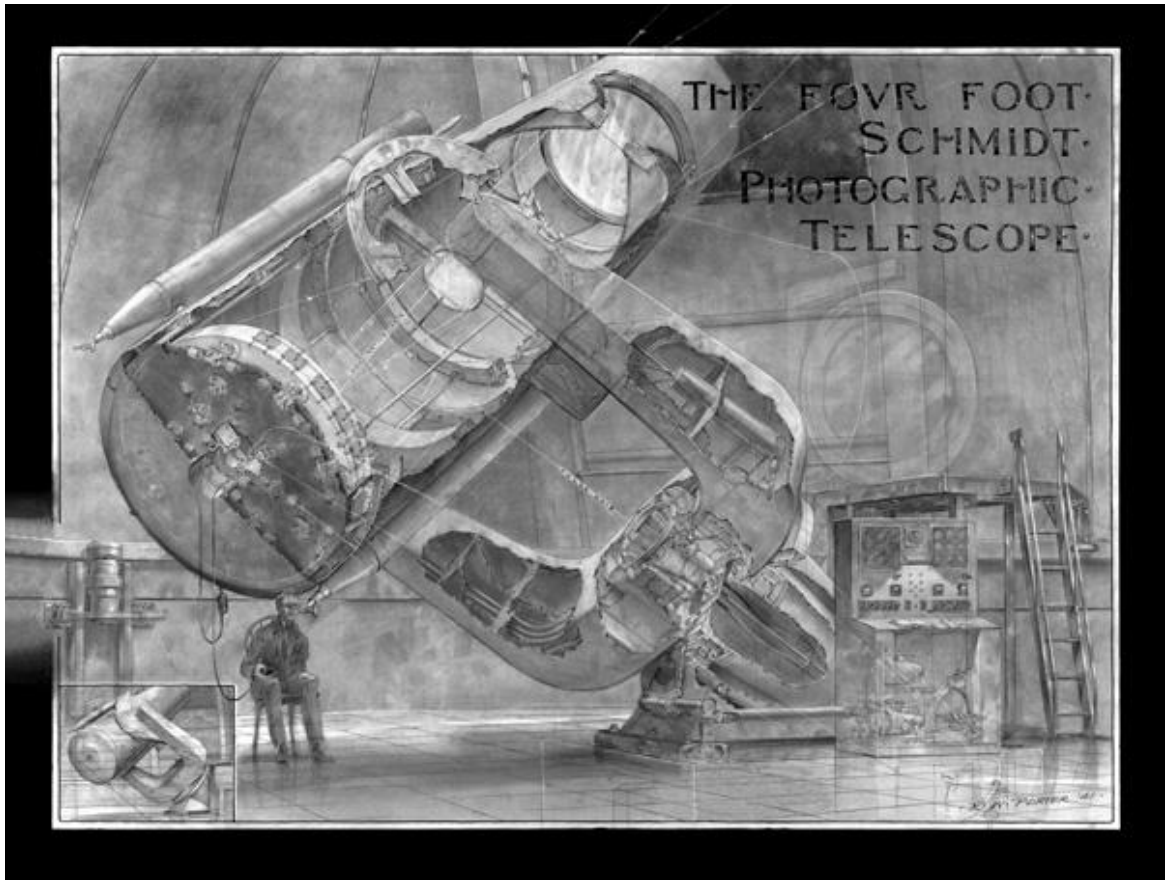
Survey: (noun) A system for turning photons into science.

A timeline of Wide Field Imaging Surveys

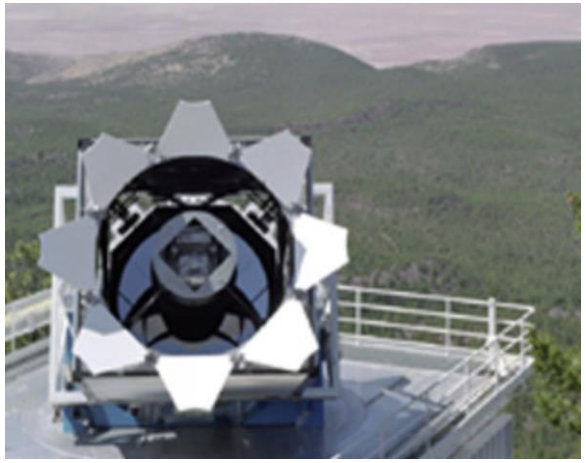
First sky surveys were done in deep antiquity (Hipparchos, ~140 BCE)

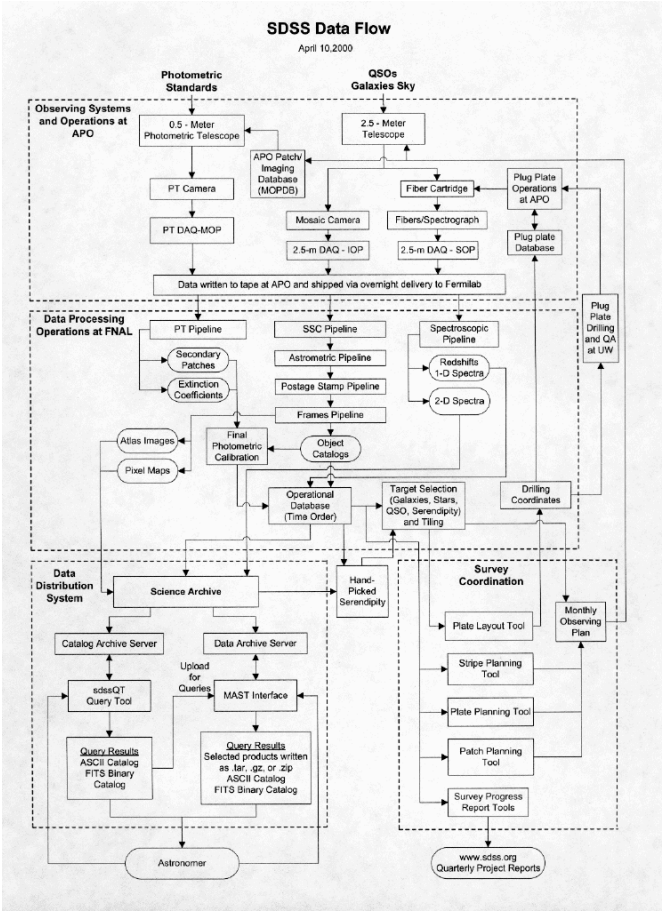
Modern sky surveys *really* start with Palomar Sky Survey I (POSS I) in 1940s-1950s and II in 1980s.

The first "digitized sky survey" in 1990s composed of several photographic plate surveys.



Sloan Digital Sky Survey





Catalogs with ~3 billion rows.
>100 GB SQL database in early data releases.

Survey Design Considerations

Designing a survey typically means comparing and optimizing a "figure of merit" or a metric. These are typically (but not always) science case or domain specific.

Entendue is a general figure of merit for imaging surveys. It is defined as:

$$G = \mathbf{A}\,\Omega$$

Volumetric Survey Speed is related to the Entendue but accounts for the limiting magnitude and volume (related to event rates for transients). See Bellm, 2016.

Survey Design Considerations

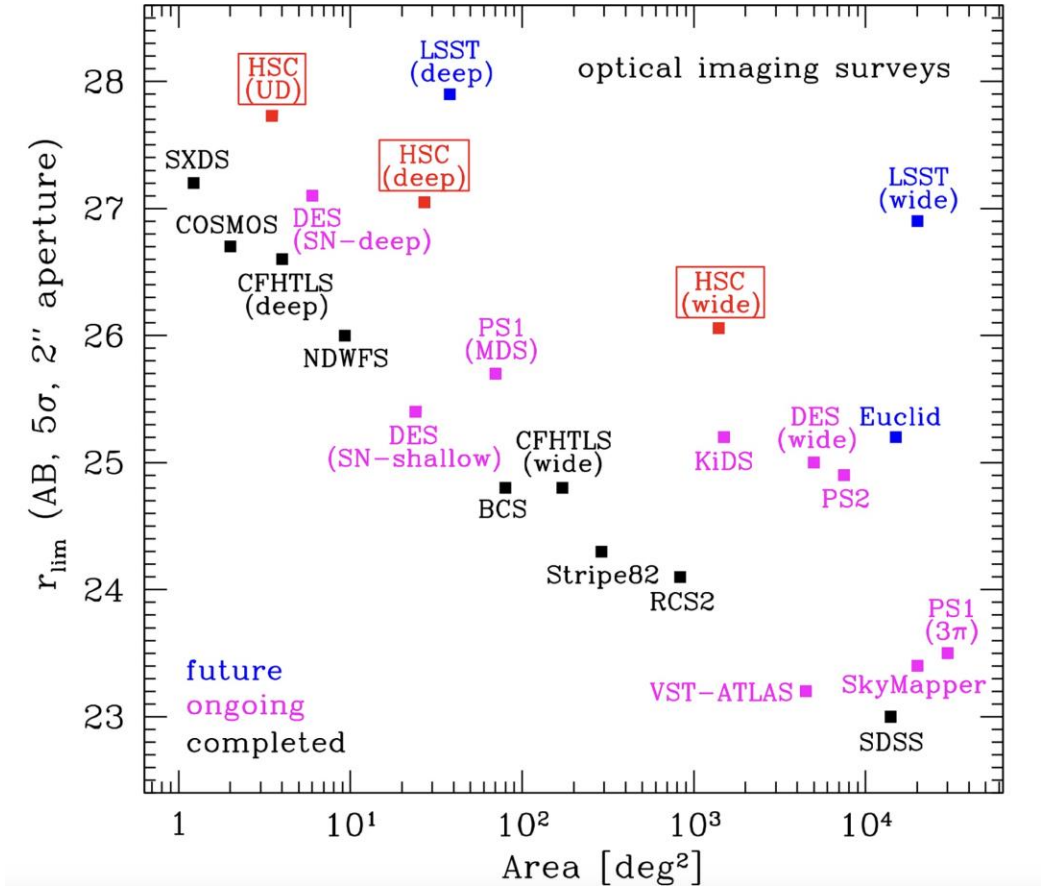
Figures of merit for survey optimization may also be science case specific.

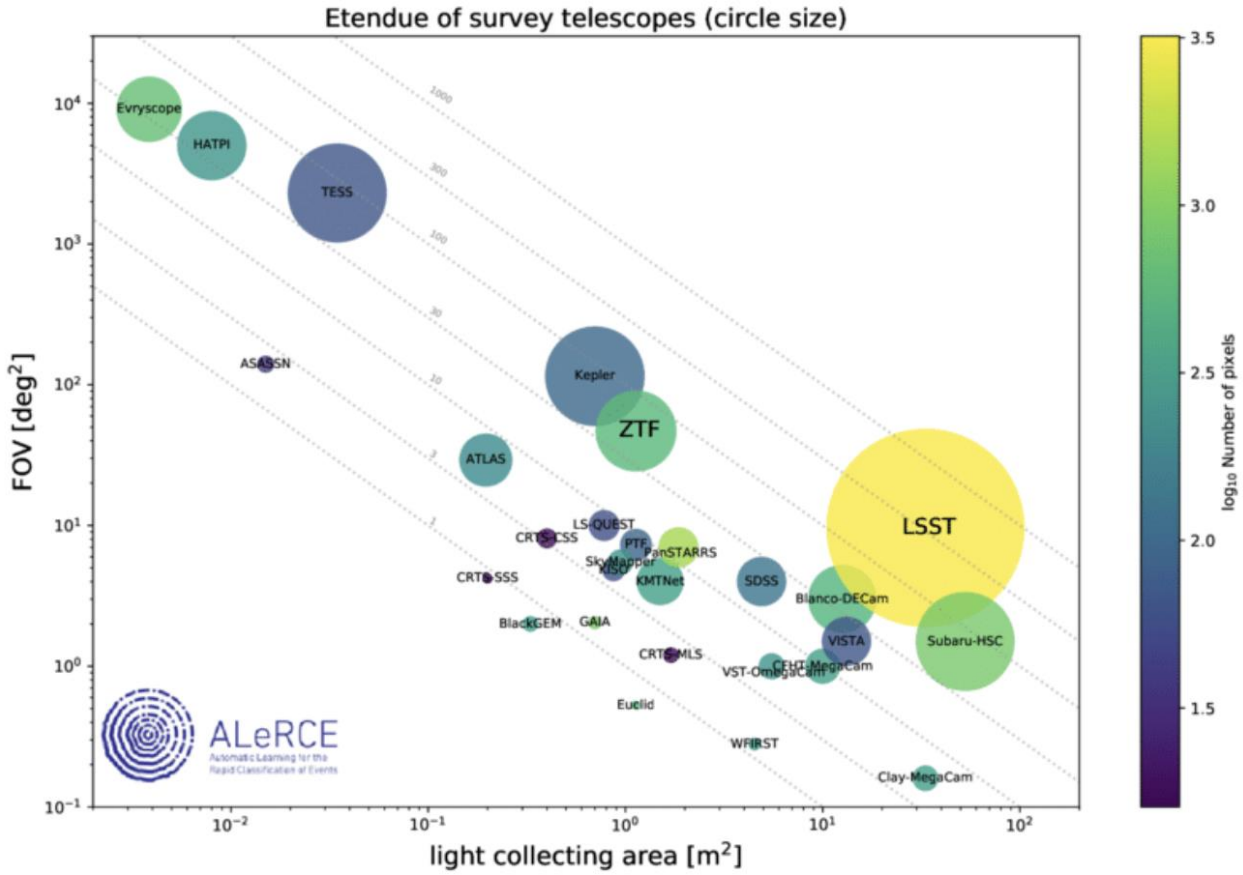
An example of this is the Dark Energy Task Force FOM:

$$C = \begin{pmatrix} \sigma^2_{w_0 w_0} & \sigma^2_{w_0 w_a} \\ \sigma^2_{w_0 w_a} & \sigma^2_{w_a w_a} \end{pmatrix} \longrightarrow \mathrm{DETF\ FoM} = (\det C)^{-1/2}$$

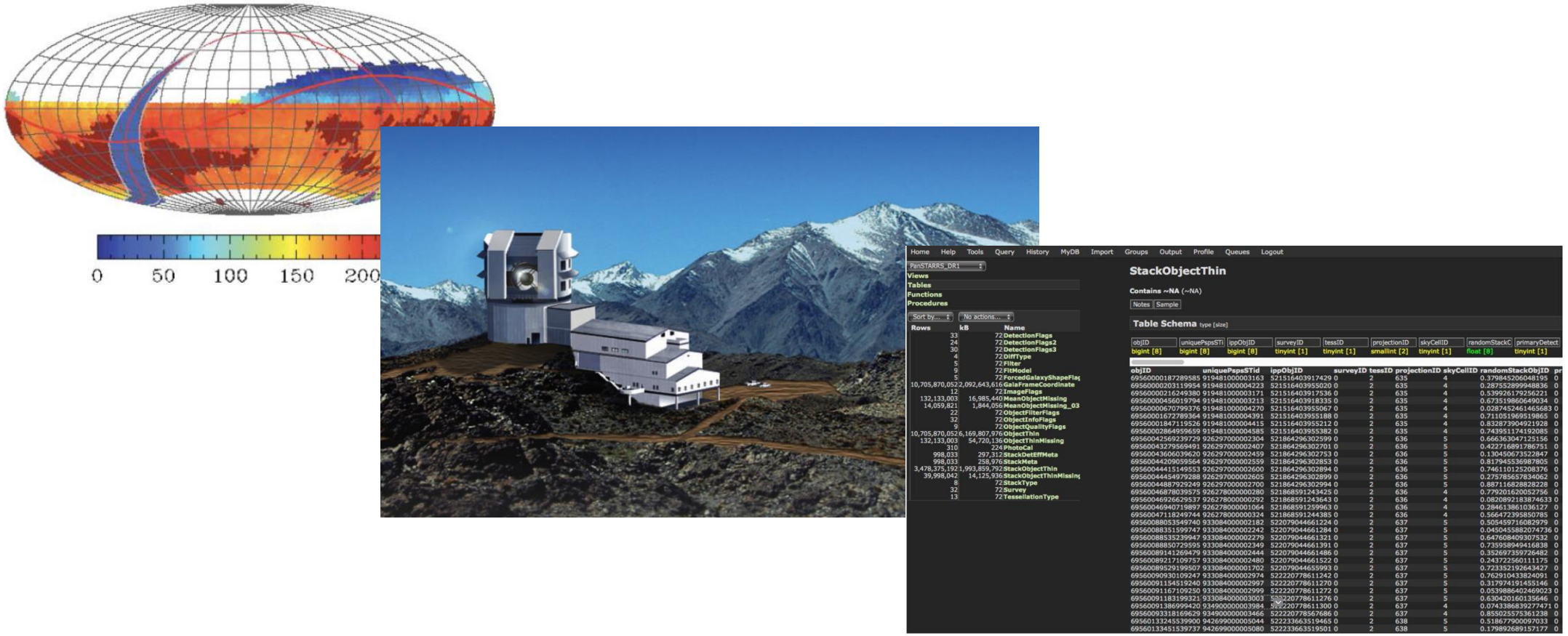
Which defines what is meant by a "Stage III/IV/V" experiment. A *lower (a smaller area of the confidence ellipse)* FOM means a later stage experiment.

Current generation surveys

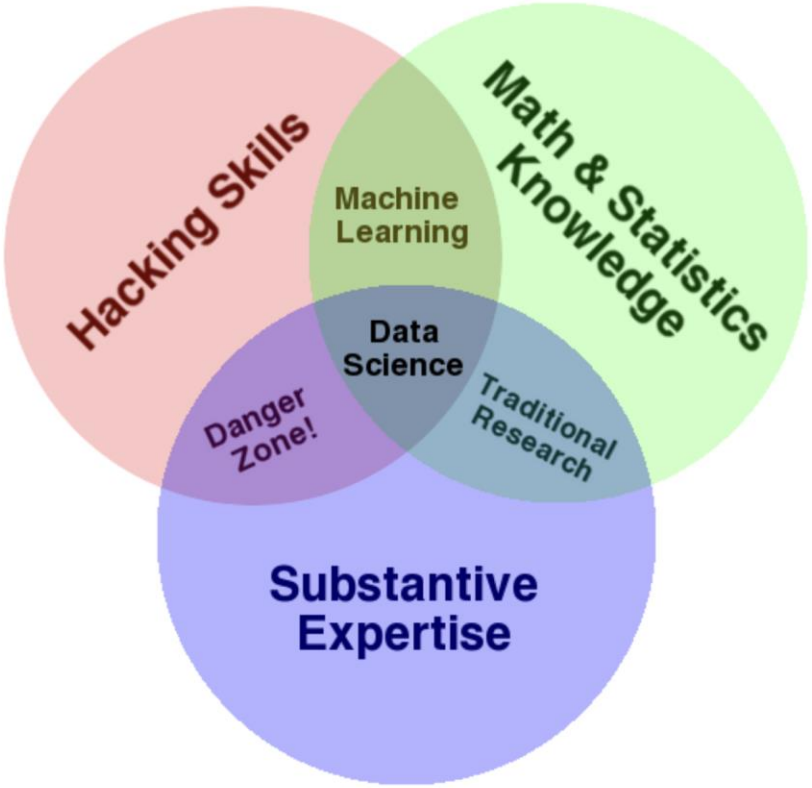




Rubin Observatory and the Legacy Survey of Space & Time



Skill Landscape of Astronomy in ~2030



Domain Knowledge is an essential ingredient for the data science practitioner.

When domain knowledge matters…

There are no galaxies fainter
than i≈27.5mag.[Perhaps this signals the edge of the universe...]

When domain knowledge matters…

There are no galaxies fainter
than i≈27.5mag.[Perhaps this signals the edge of the universe...]

the inverse-square law:

flux$\propto r^{-2}$ and the sensitivity limit of the LSST detector.

When domain knowledge matters…

There are no galaxies fainter than i≈27.5mag.[Perhaps this signals the edge of the universe...]

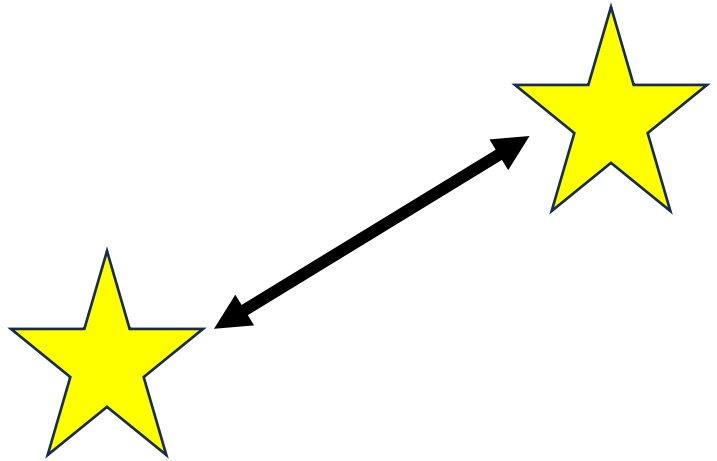
the inverse-square law:

flux$\propto r^{-2}$ and the sensitivity limit of the LSST detector.

We know fainter galaxies do exist, but they are either too distant or intrinsically dim to be detected by LSST.

When domain knowledge matters…

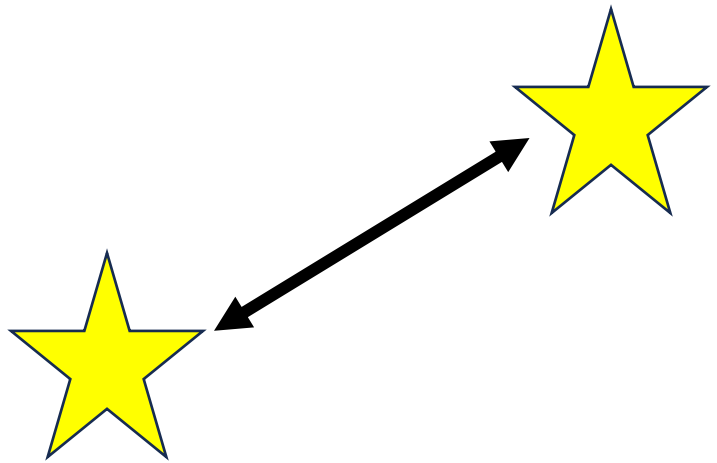
Two stars cannot be closer than ~0.35 arcsec in the sky.



When domain knowledge matters…

Two stars cannot be closer than ~0.35 arcsec in the sky.

[Perhaps there is some repulsive force between stars that keeps them separated...]

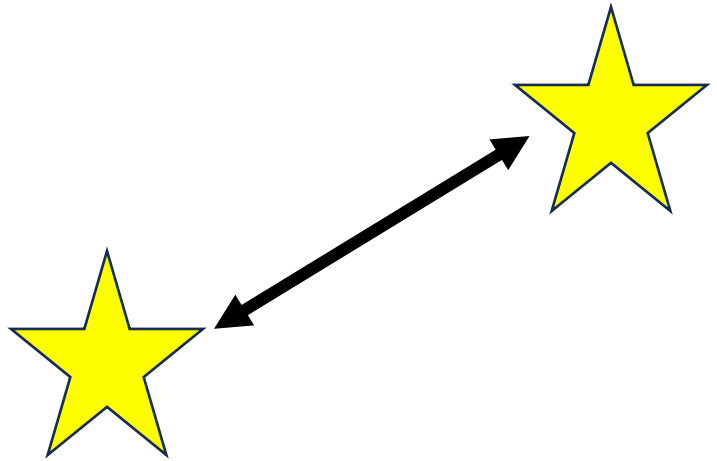


When domain knowledge matters…

Two stars cannot be closer than ~0.35 arcsec in the sky.

[Perhaps there is some repulsive force between stars that keeps them separated...]

This apparent conclusion reflects the typical seeing at Cerro Pachon (~0.7 arcsec). Very nearby stars (θ<0.3 arcsec), cannot be resolved by LSST.



When domain knowledge matters…

The Universe emits more light in the r-band than the y-band.

When domain knowledge matters…

The Universe emits more light in the r-band than the y-band.

Blue sources naturally emit more light in the r-band than the y-band, but this imbalance should be countered by red sources (due to reddening and redshift there *should* be a lot more sources with observed red colors). Many red sources ($m_r - m_y > 0$) will only be detected in the r-band, however, due to the relative sensitivity in each filter.

When domain knowledge matters…

The Universe emits more light in the r-band than the y-band.

Blue sources naturally emit more light in the r-band than the y-band, but this imbalance should be countered by red sources (due to reddening and redshift there *should* be a lot more sources with observed red colors). Many red sources ($m_r - m_y > 0$) will only be detected in the r-band, however, due to the relative sensitivity in each filter.

This apparent conclusion is a bit more subtle than the previous two, and there are multiple factors contributing to this incorrect assertion. LSST will be far more sensitive in the r-band than the y-band (lower sky backgrounds and higher detector efficiency are the primary reasons).

Pushing the Boundaries: The 3 Vs of LSST

Volume: LSST will produce data volumes 1-2 orders of magnitude larger than any existing experiment.

Variety: LSST will (potentially) detect phenomena that have not been seen before.

Velocity: LSST will produce alerts every 60s (about 20x faster than existing surveys).

Success in this era will require substantial working knowledge of both "hacking" and "stats/mathematical analysis", but progress will be impeded without a corresponding expertise in how the data were acquired and why the Universe produced those data in the first place.

Volume

	ZTF	LSST
Number of detections	1 trillion	7 trillion
Number of objects	1 billion	37 billion

Velocity

	ZTF	LSST
Number of detections	1 trillion	7 trillion
Number of objects	1 billion	37 billion
Nightly alert rate	1 million	10 million
Nightly data rate	1.4 TB	15 TB
Alert latency	< 20 minutes	60 seconds

Variety

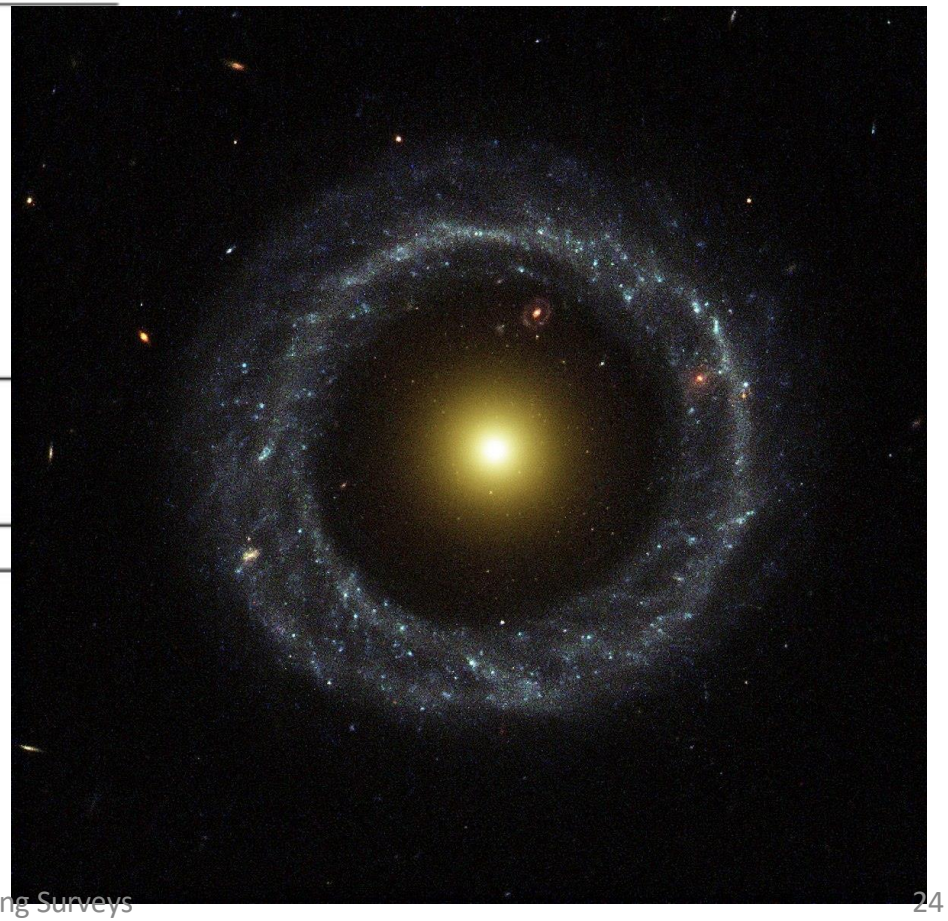
Table 1
Summary of Transient and Variable Models for PLAsTiCC

Model Class Num[a]: Name	Model Description	Contributor(s)[b]	N_{event} Gen[c]	N_{event} Train[d]	N_{event} Test[e]	Redshift Range[f]
90: SN Ia	WD detonation, SN Ia	RK	16,353,270	2313	1,659,831	<1.6
67: SN Ia-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	<0.9
52: SN Iax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	<1.3
42: SN II	Core collapse, SN II	SG,LG:RK,JRP:VAV	59,198,660	1193	1,000,150	<2.0
62: SN Ibc	Core collapse, SN Ibc	VAV:RK,JRP	22,599,840	484	175,094	<1.3
95: SLSN-I	Super-lum. SN (magnetar)	VAV	90,640	175	35,782	<3.4
15: TDE	Tidal disruption event	VAV	58,550	495	13,555	<2.6
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	<0.3
88: AGN	Active galactic nuclei	SD	175,500	370	101,424	<3.4
92: RRL	RR Lyrae	SD	200,200	239	197,155	0
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	0
16: EB	Eclipsing binary stars	AP	220,200	924	96,572	0
53: Mira	Pulsating variable stars	RH	1490	30	1453	0
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2820	151	1303	0
991: μLens-Binary	μ-lens from binary lens	RD,AA	1010	0	533	0
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1702	<0.4
993: CaRT	Calcium-rich transient	VAV	2,834,500	0	9680	<0.9
994: PISN	Pair-instability SN	VAV	5650	0	1172	<1.9
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	0
Total	Sum of all models		117,128,700	7846	3,492,888	...

Variety

Table 1
Summary of Transient and Variable Models for PLAsTiCC

Model Class Num[a]: Name	Model Description	Contributor(s)[b]	N_{event} Gen[c]	N_{event} Train[d]	N_{event} Test[e]	Redshift Range[f]
90: SN Ia	WD detonation, SN Ia	RK	16,353,270	2313	1,659,831	
67: SN Ia-91bg	Peculiar type Ia: 91bg	SG,LG	1,329,510	208	40,193	
52: SN Iax	Peculiar SNIax	SJ,MD	8,660,920	183	63,664	
42: SN II	Core collapse, SN II	SG,LG:RK,JRP:VAV	59,198,660	1193	1,000,150	
62: SN Ibc	Core collapse, SN Ibc	VAV:RK,JRP	22,599,840	484	175,094	
95: SLSN-I	Super-lum. SN (magnetar)	VAV	90,640	175	35,782	
15: TDE	Tidal disruption event	VAV	58,550	495	13,555	
64: KN	Kilonova (NS-NS merger)	DK,GN	43,150	100	131	
88: AGN	Active galactic nuclei	SD	175,500	370	101,424	
92: RRL	RR Lyrae	SD	200,200	239	197,155	
65: M-dwarf	M-dwarf stellar flare	SD	800,800	981	93,494	
16: EB	Eclipsing binary stars	AP	220,200	924	96,572	
53: Mira	Pulsating variable stars	RH	1490	30	1453	
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN	2820	151	1303	
991: μLens-Binary	μ-lens from binary lens	RD,AA	1010	0	533	
992: ILOT	Intermed. Lum. Optical Trans.	VAV	4,521,970	0	1702	
993: CaRT	Calcium-rich transient	VAV	2,834,500	0	9680	
994: PISN	Pair-instability SN	VAV	5650	0	1172	
995: μLens-String	μ-lens from cosmic strings	DC	30,020	0	0	
Total	Sum of all models		117,128,700	7846	3,492,888	



Consider: How long would it take to perform basic processing of all of LSST on your laptop?

The bare minimum for image processing includes bias (subtraction) and flat-field (division) corrections. Assume your laptop has a single 3 GHz processor that requires 1 tick to perform a single addition operation and 4 ticks to perform a single multiplication operation.

Consider: How long would it take to perform basic processing of all of LSST on your laptop?

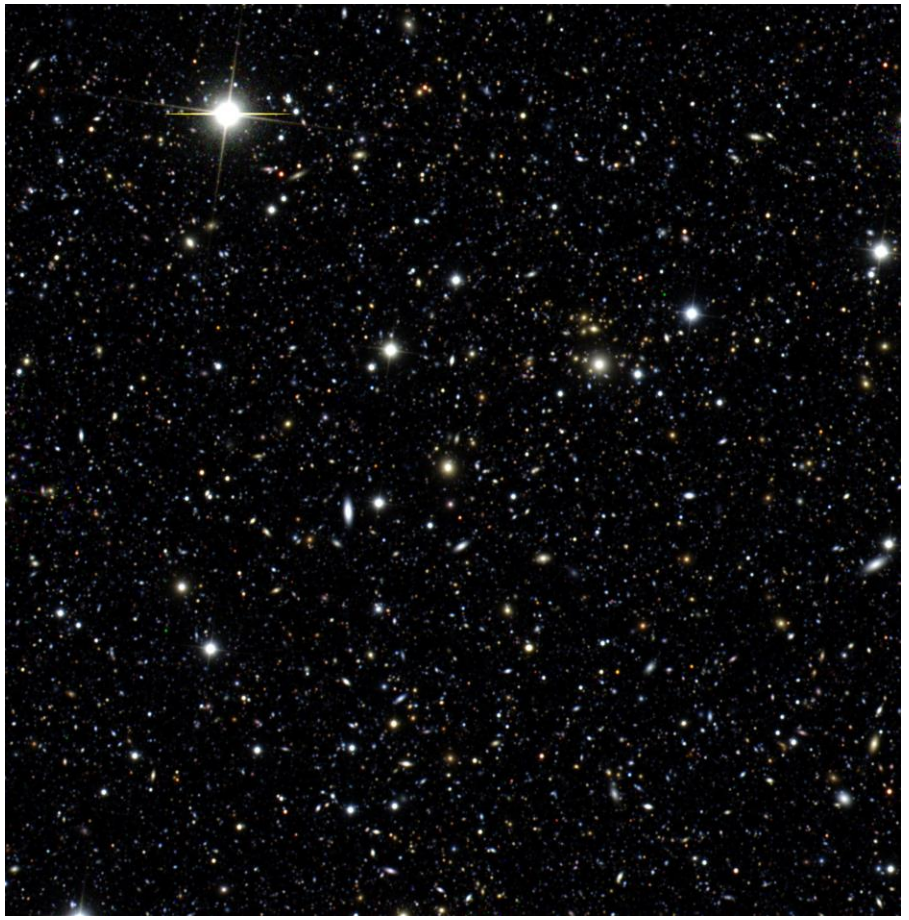
The bare minimum for image processing includes bias (subtraction) and flat-field (division) corrections. Assume your laptop has a single 3 GHz processor that requires 1 tick to perform a single addition operation and 4 ticks to perform a single multiplication operation.

$$\frac{3.2 \times 10^9\,\mathrm{pix}}{\mathrm{field\,obs}} \times \frac{\mathrm{field}}{10\,\mathrm{deg}^2} \times 20{,}000\,\mathrm{deg}^2 \times \frac{5\,\mathrm{ticks}}{\mathrm{pix}} \times \frac{\mathrm{s}}{3 \times 10^9\,\mathrm{ticks}} \times 1000\,\mathrm{obs} \approx 4\,\mathrm{months}$$

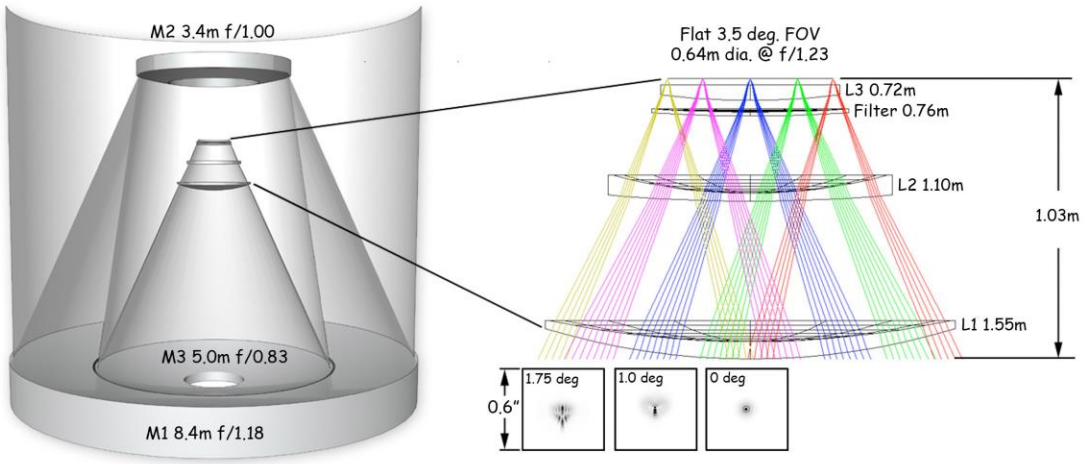
What do we measure?

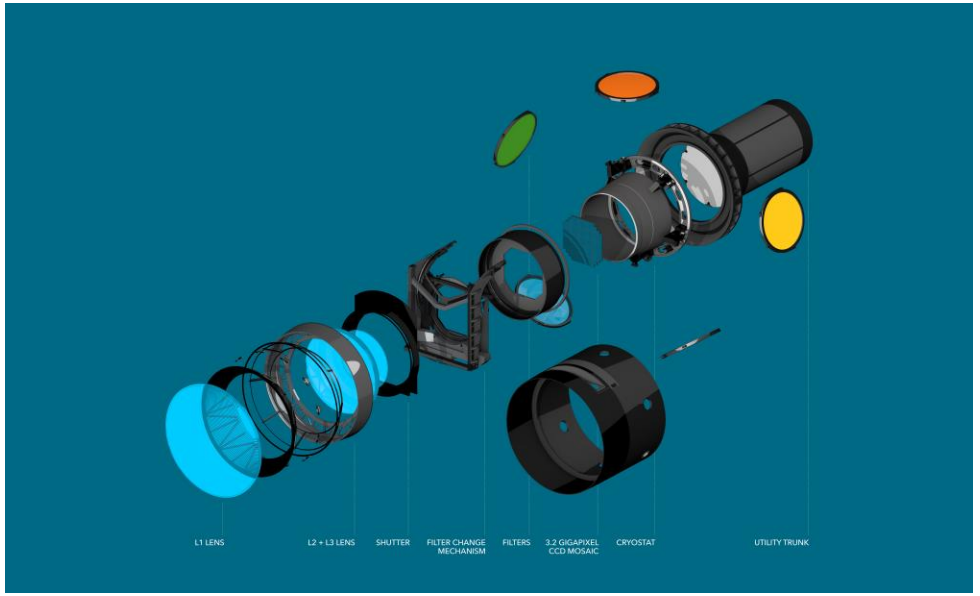
Fundamentally, the thing we care about is measuring fluxes (and positions - though these two are related).

In principle, flux measurements are straight forward: count the number of photons per unit energy per unit time.



Not so simple…



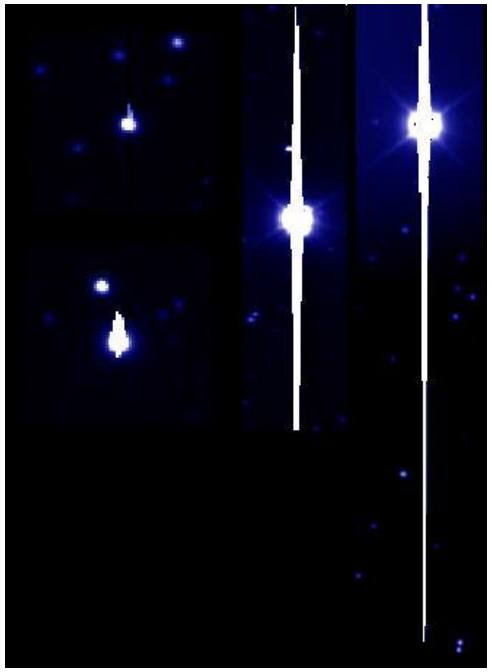


In practice, things are not this simple:

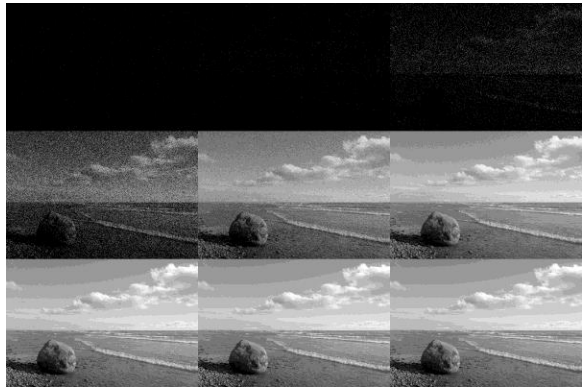
telescope's optical elements are not 100% efficient
(we *can* measure inefficiencies and correct them → complicates the uncertainties beyond Poisson)

our detectors introduce noise to our measurements

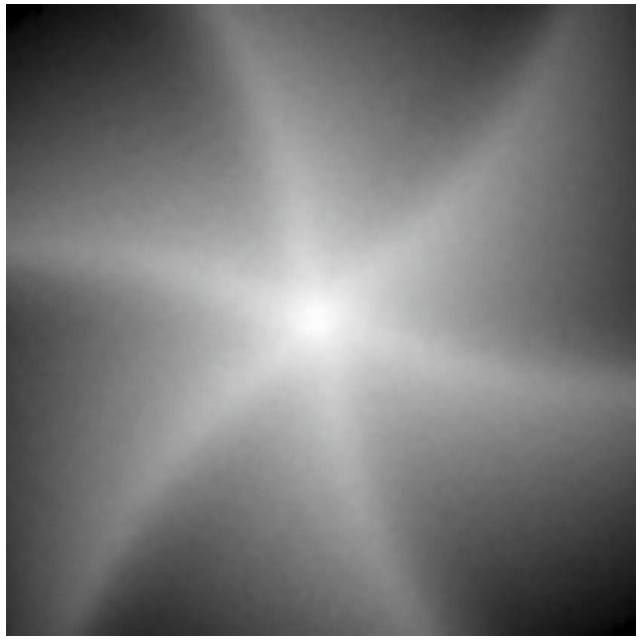
Sources of Uncertainty



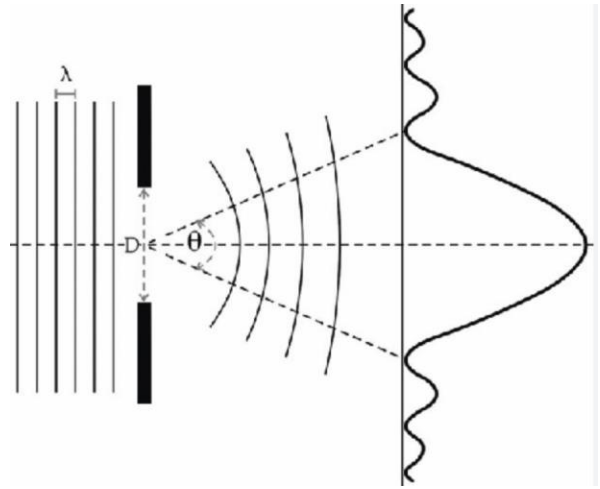
Saturation of the detector (flux)



(Quantum) Shot Noise (flux)

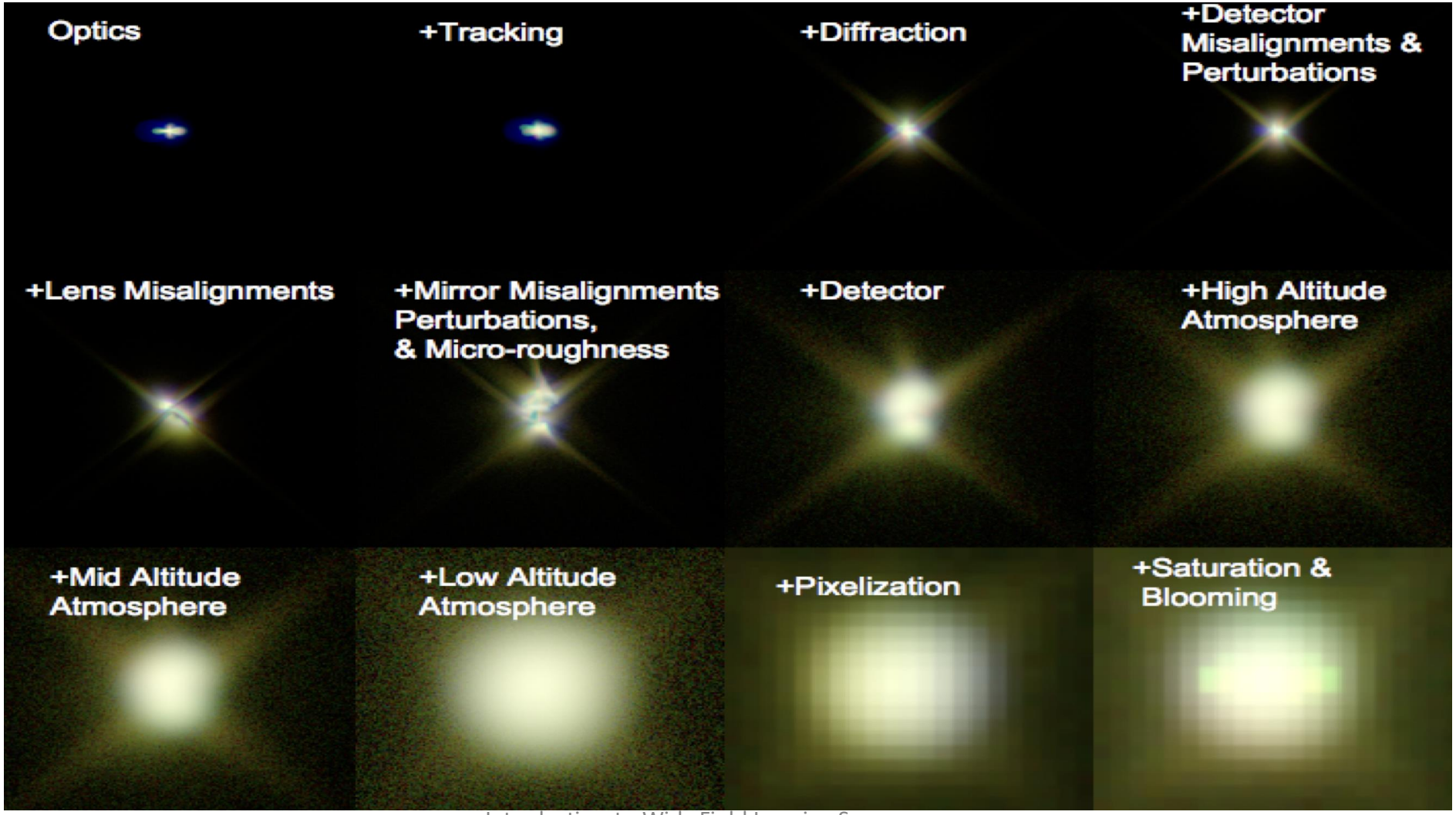


Uneven exposure due to shutter (flux)



Diffraction by telescope optics (position + flux)

Not so simple…



Takeaway

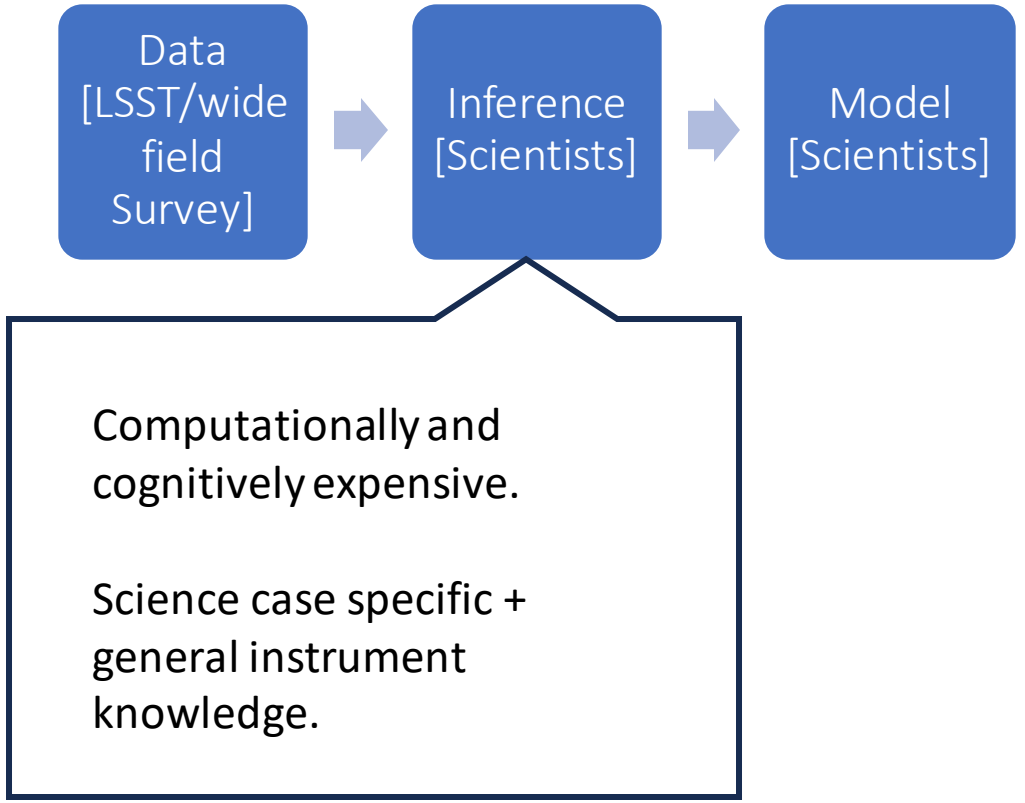
In summary, while our basic task — counting — is in principle quite simple, measuring the flux/position of an astronomical source is very complicated.

We control all the elements of the system, however, and a variety of different measurements can correct for these issues (though this results in more challenging uncertainty estimates).

Given all these complications, how can one make any (informed) inferences about the universe?

Inference for Large Surveys

Inference Model 1:



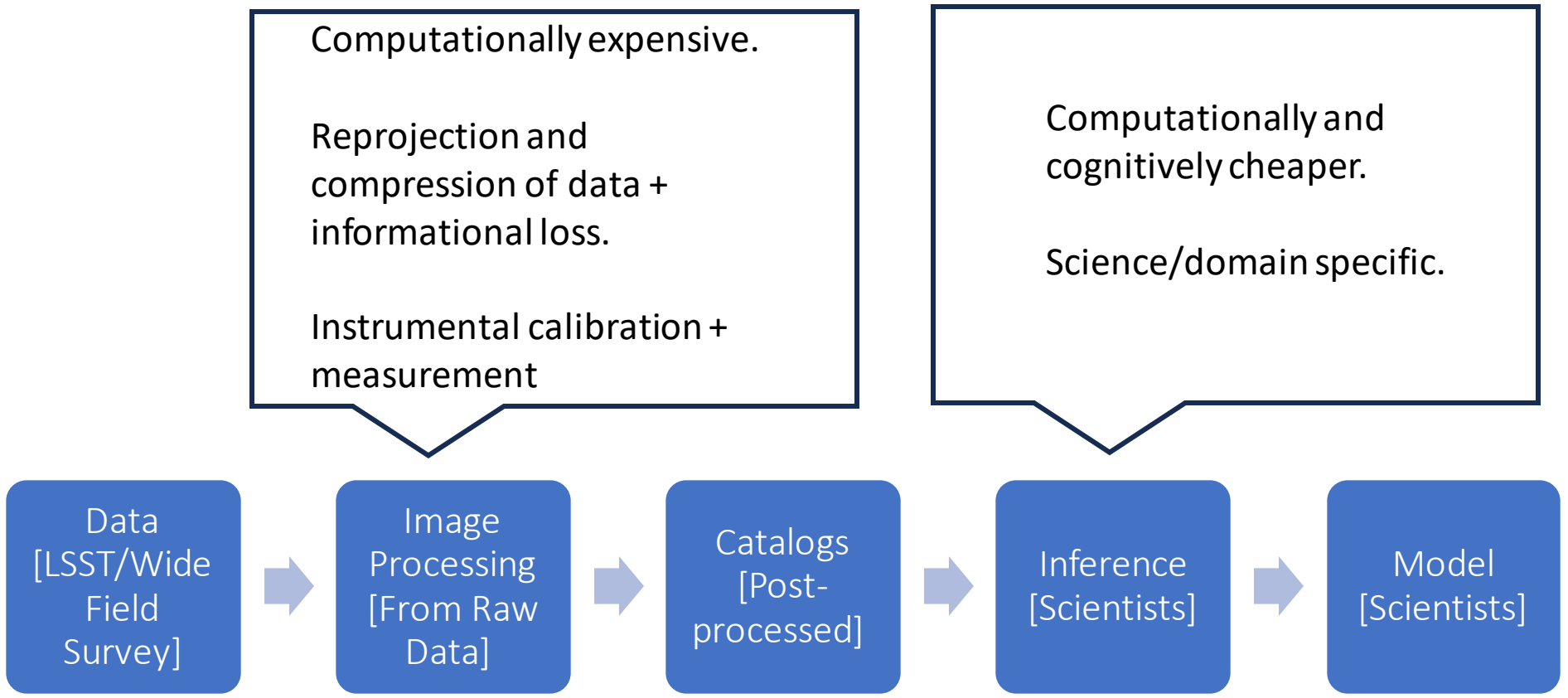
Inference for Large Surveys

Inference Model 1:

Inference Model 2:

Inference for Large Surveys

Inference Model 2:



Inference for Large Surveys

Inference Model 1:

Inference Model 2:

Astronomical Image Transformations

"True" Image

$$I(x) = \phi(x) \circledast S(x) + \epsilon(x)$$

Point Spread Function
Captures the deterministic part of an image transform

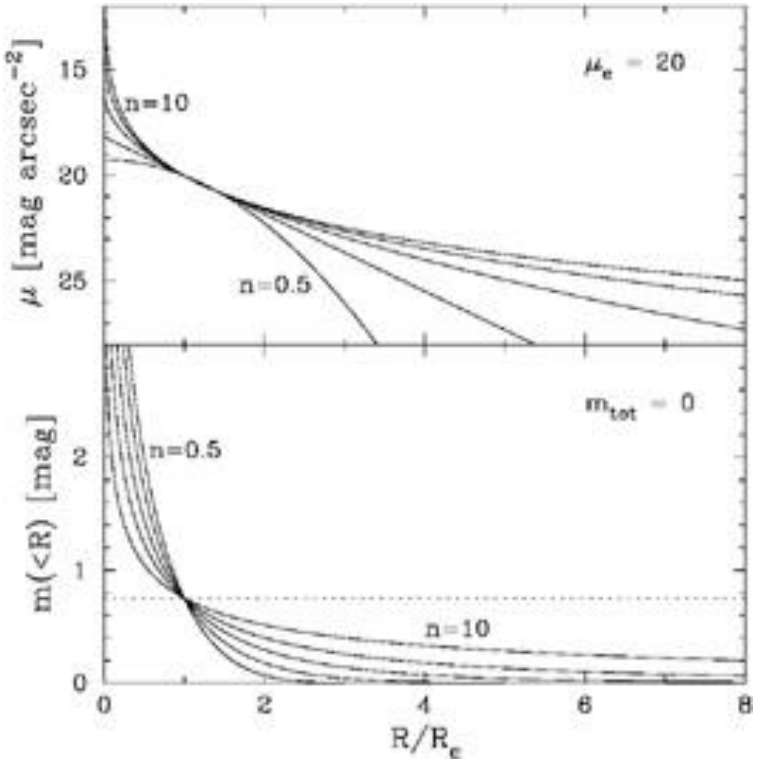
Additive Noise
Captures the Stochastic Part of the Image Transform

Modeling and Compressing the Sky

In addition to a model for $\phi(x)$, we also need a model for the 'true' image, S(x). S(x) is usually a sum over many terms, for example,

Stars – point source model

Galaxies – double exponential (Sersic profile)



Two ways of looking at this: Decomposition of the sky into component physical models –or- An Astronomy specific way of compressing the information in an image.

Co-addition and Image Differencing

Objects at the faint limit: Sums over many images, called *co-added images*, will be covered by Yusra on Wednesday.

Transient detection: *Image Differencing,* what has changed in the images as a function of time?

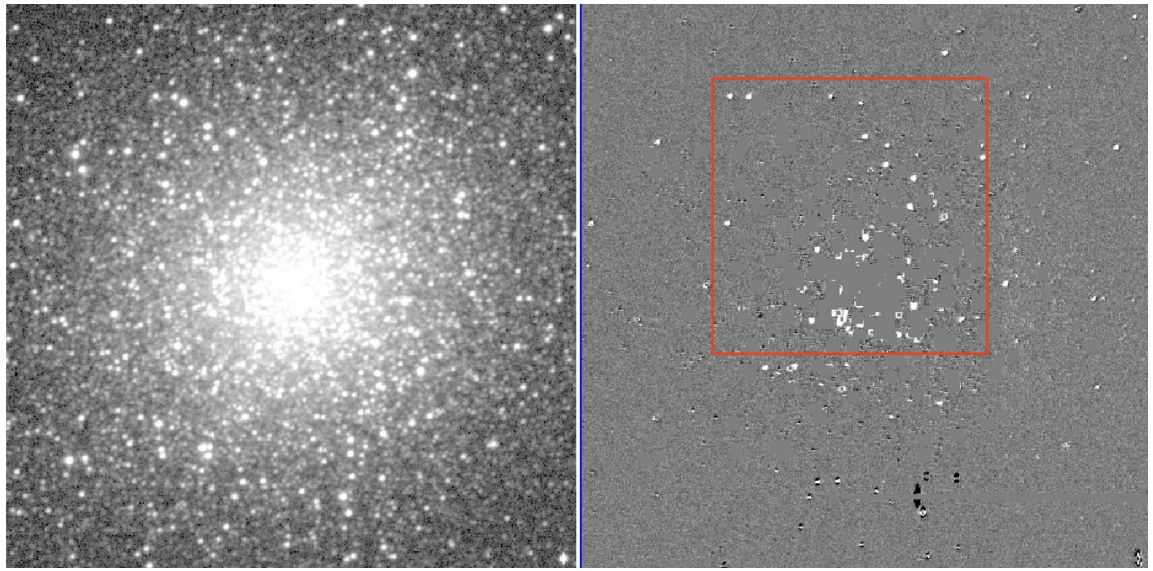


Image Differencing

Image Differencing equation:

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

Difference image

Image kernel:
Captures differences in image quality between the two input images

Solving for the Image Kernel

The Alard-Lupton Algorithm: minimize

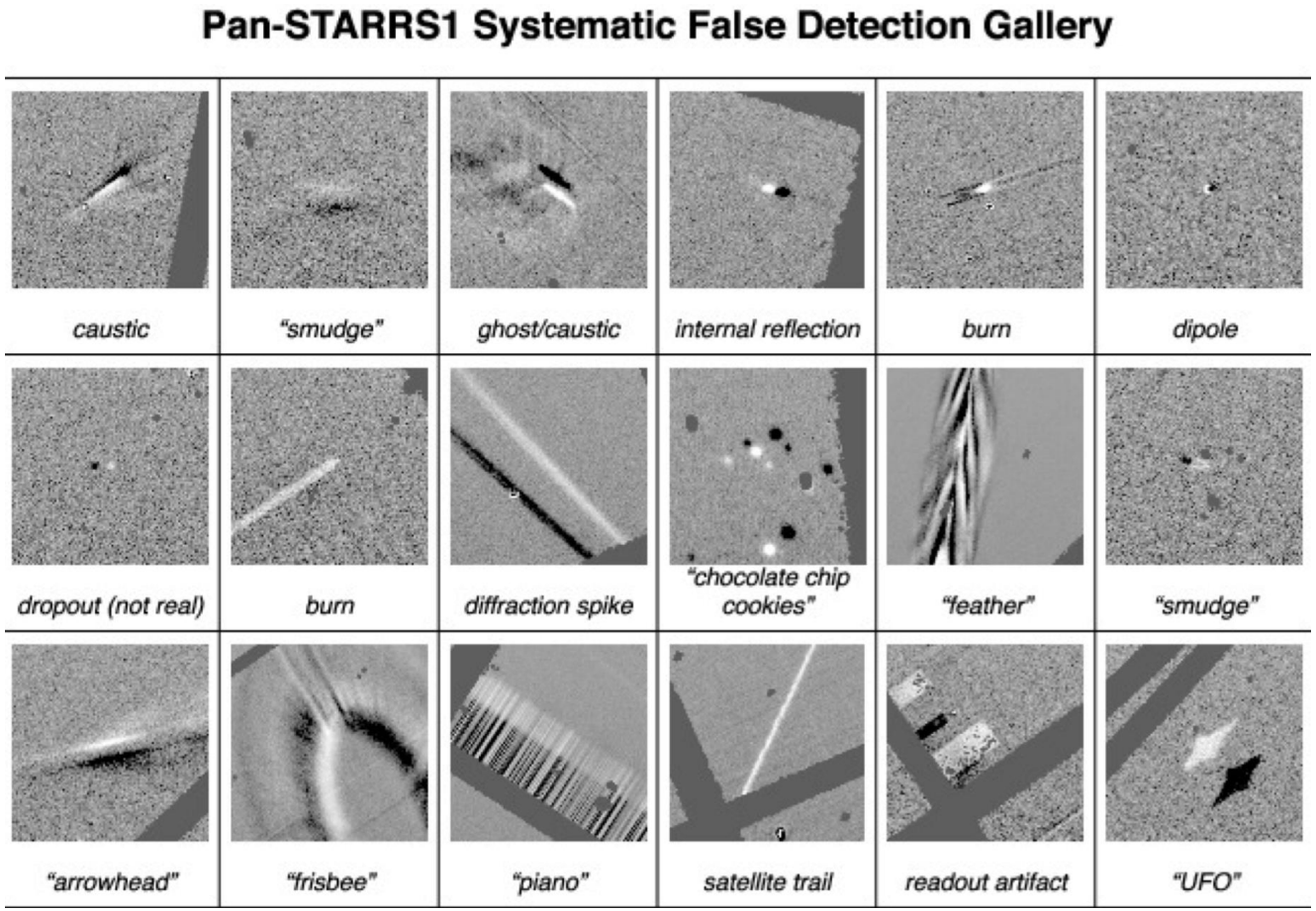
$$\sum_i \left([R \circledast K](x_i\ , y_i) \ - \ I(x_i, y_i) \right)^2$$

With an image convolution kernel K defined in terms of basis functions

$$K(u, v) \ = \ \sum_{ijk} a_n \, e^{-\left(u^2 + v^2\right)/2\sigma_k^2} u^i v^j$$

The Alard Lupton algorithm is used for ZTF – open question about scaling to the volume, velocity, and variety of LSST. See Zackay, Orek, and Gal-Yam, 2016 for an alternative approach based on likelihood ratio tests ==> statistically well posed!

False Positives



LSST Image Processing Pipeline

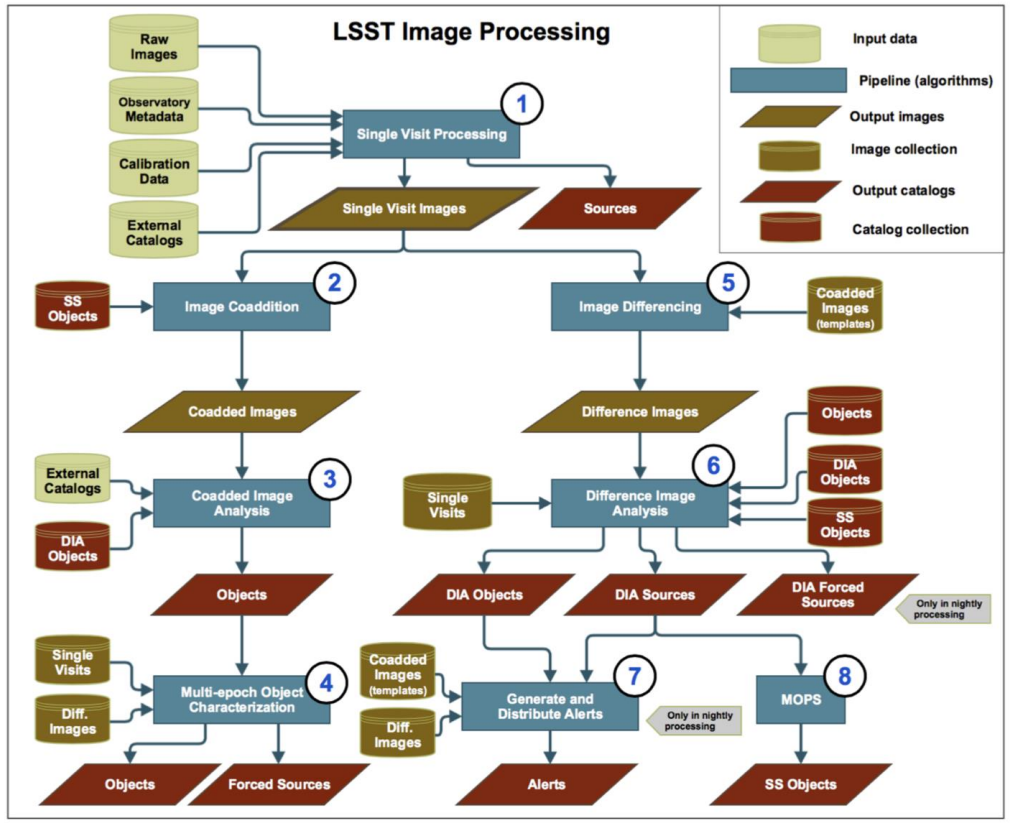
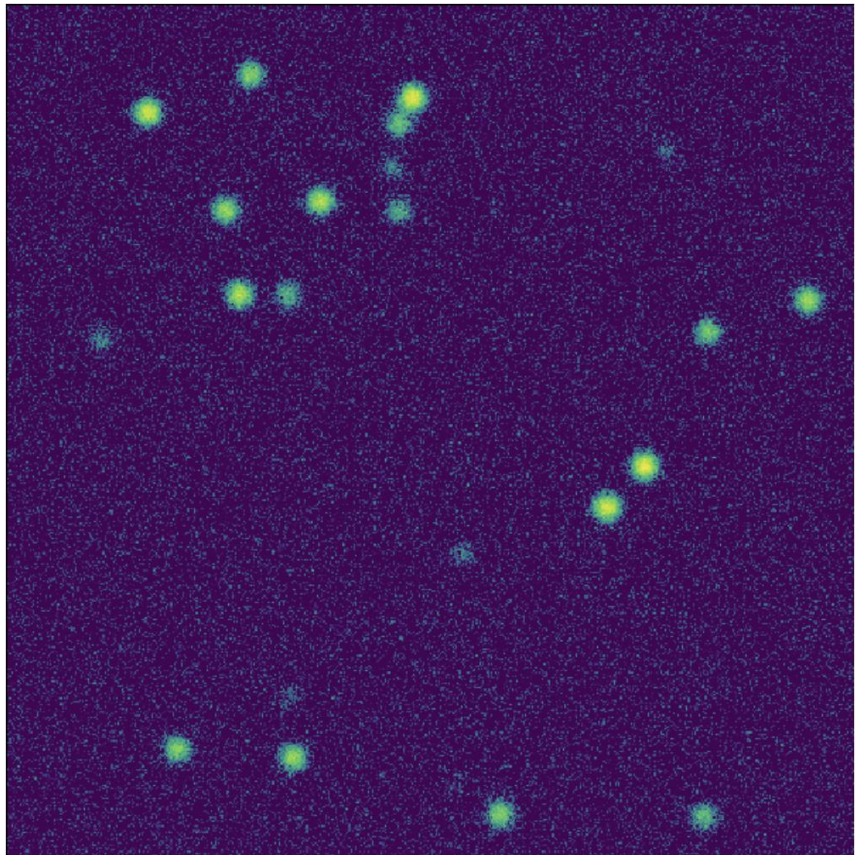


Figure 2: Illustration of the conceptual design of LSST science pipelines for imaging processing.

Fluxes and Backgrounds

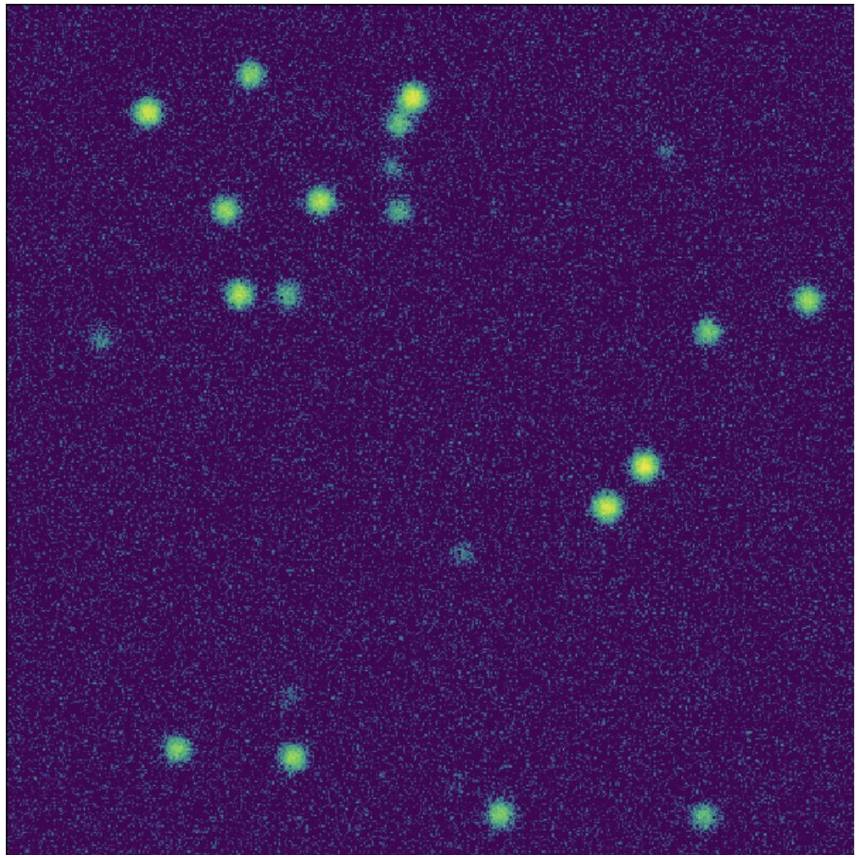
How do we measure a flux? We've seen this is a complicated question, but ultimately we do this by counting.

To do this – we need to identify sources.



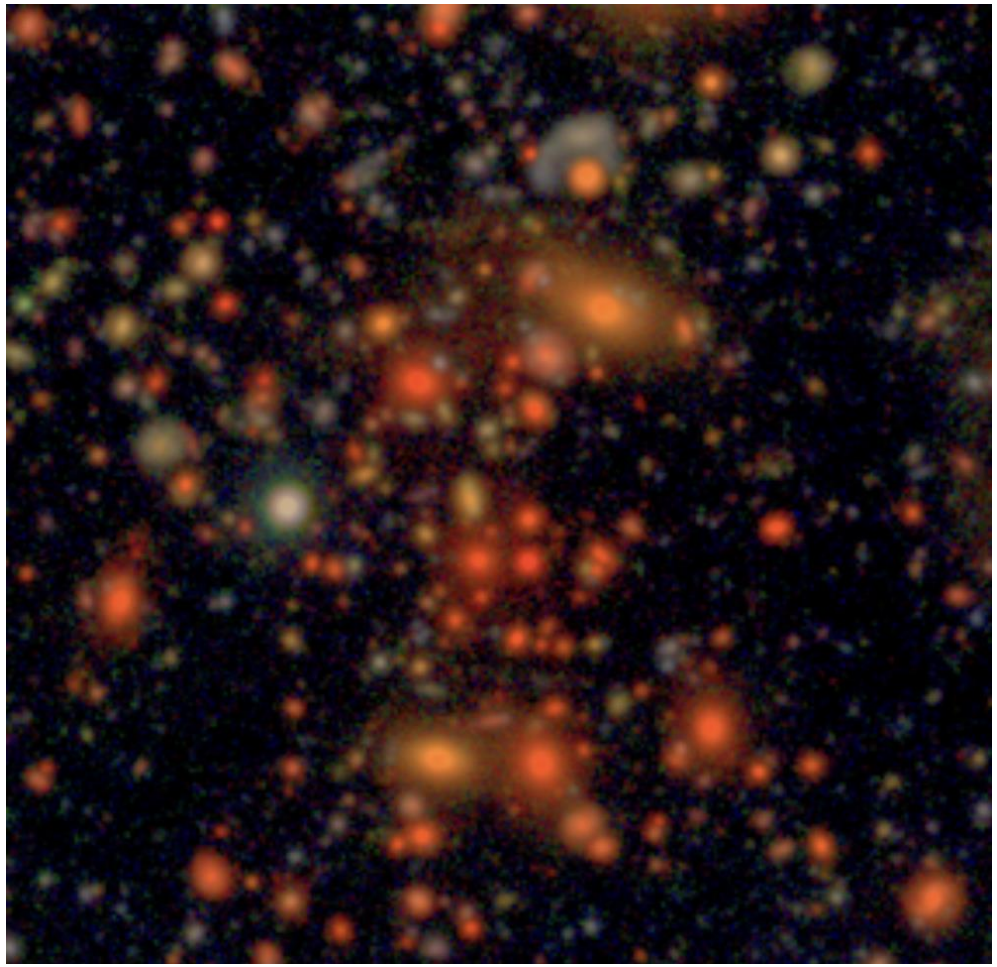
Fluxes and Backgrounds

How many sources are in this image?

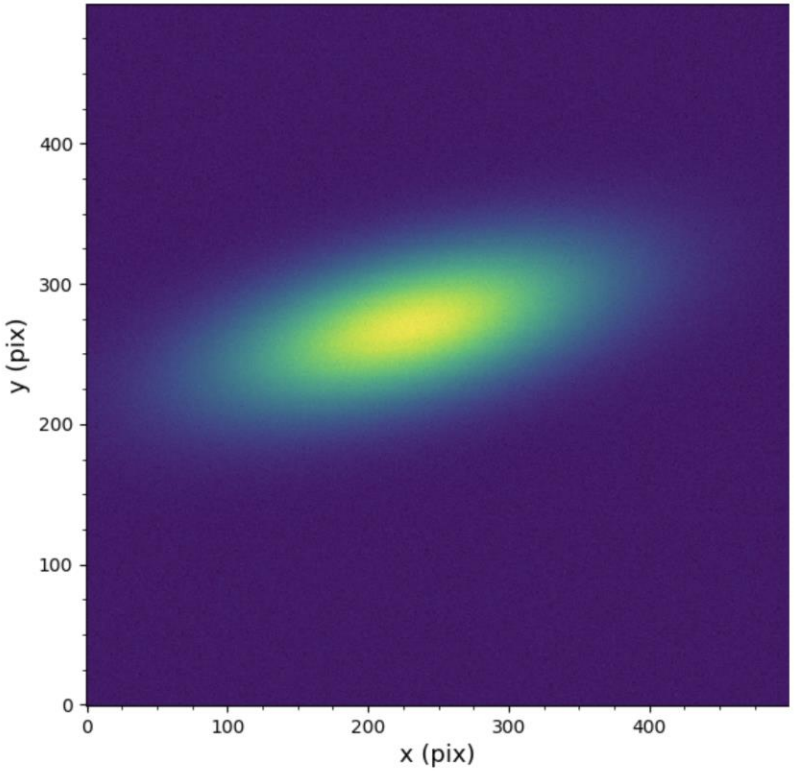


How about this one?

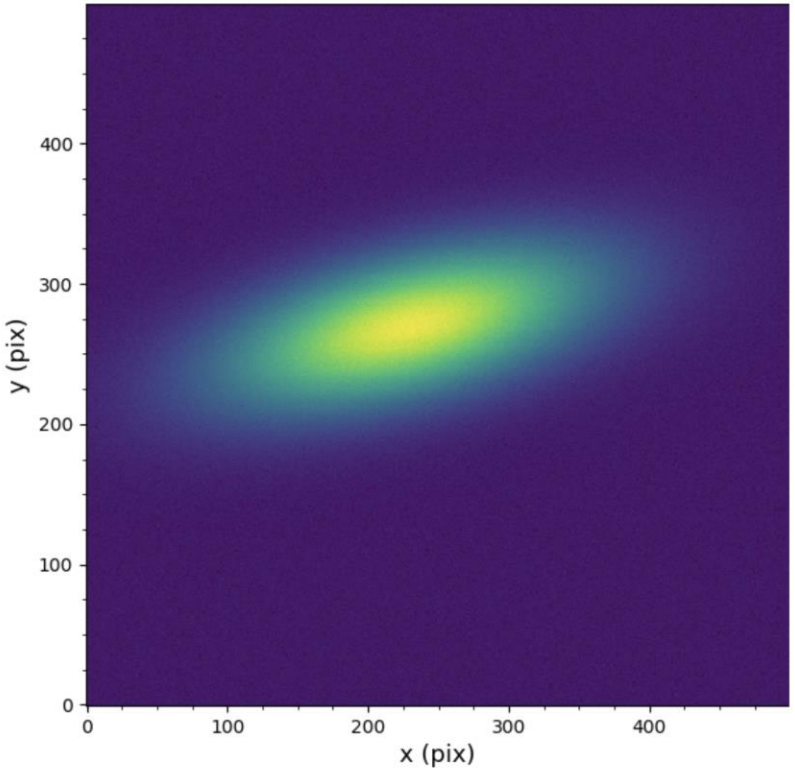
Or this one?

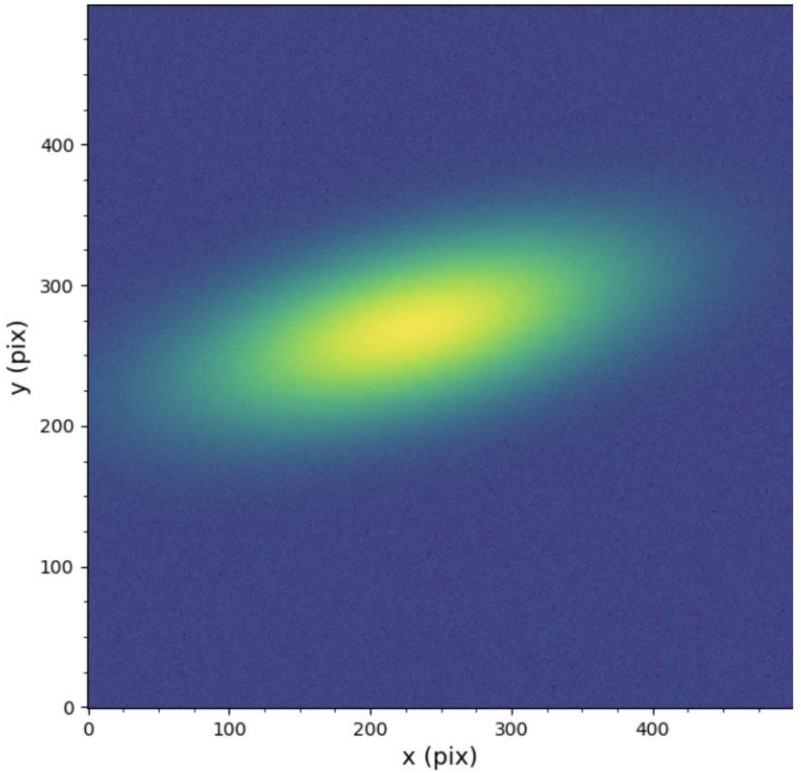


How big is this galaxy?

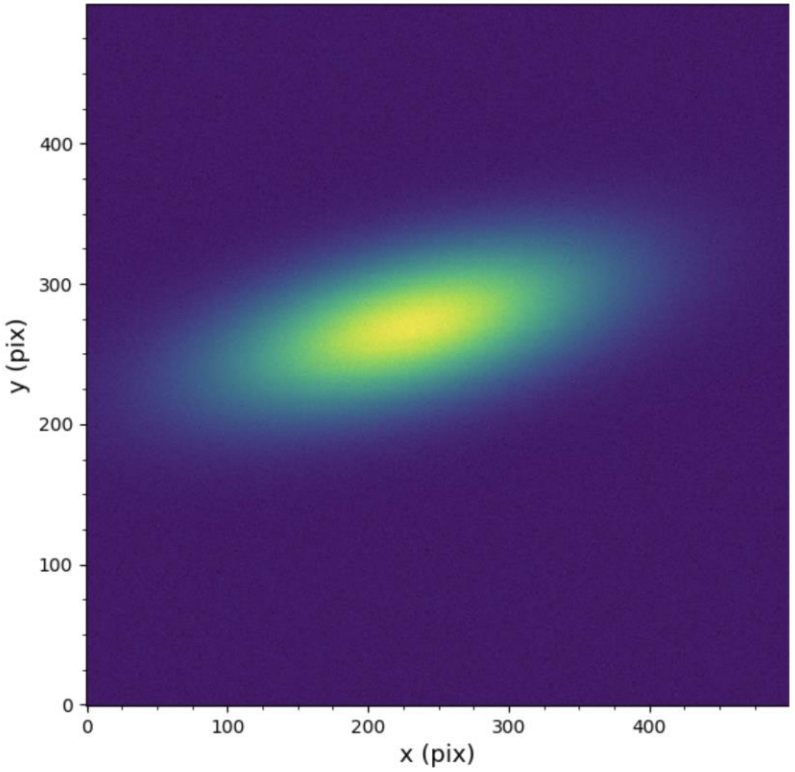


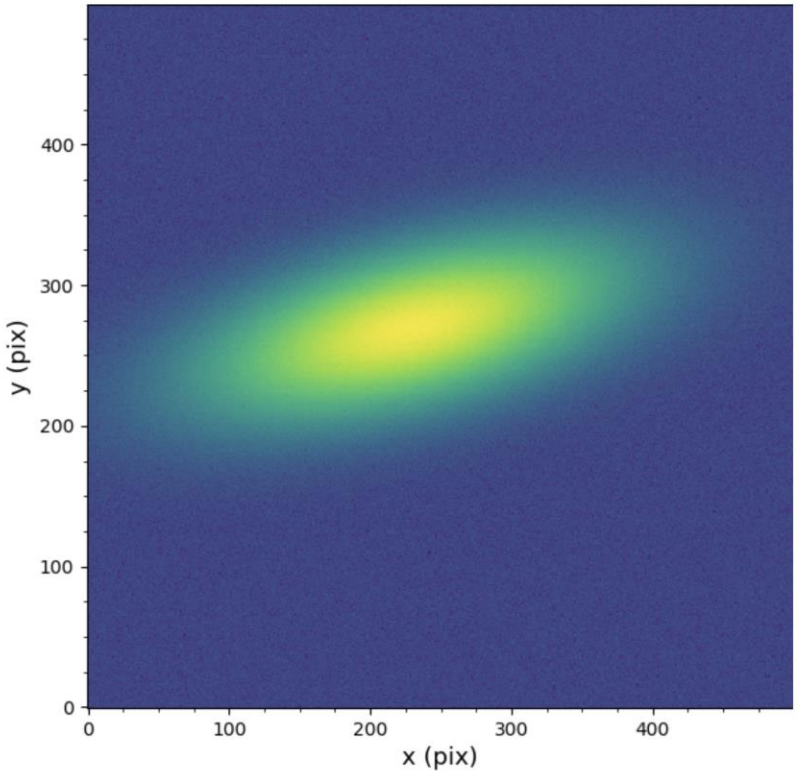
How big is this galaxy?

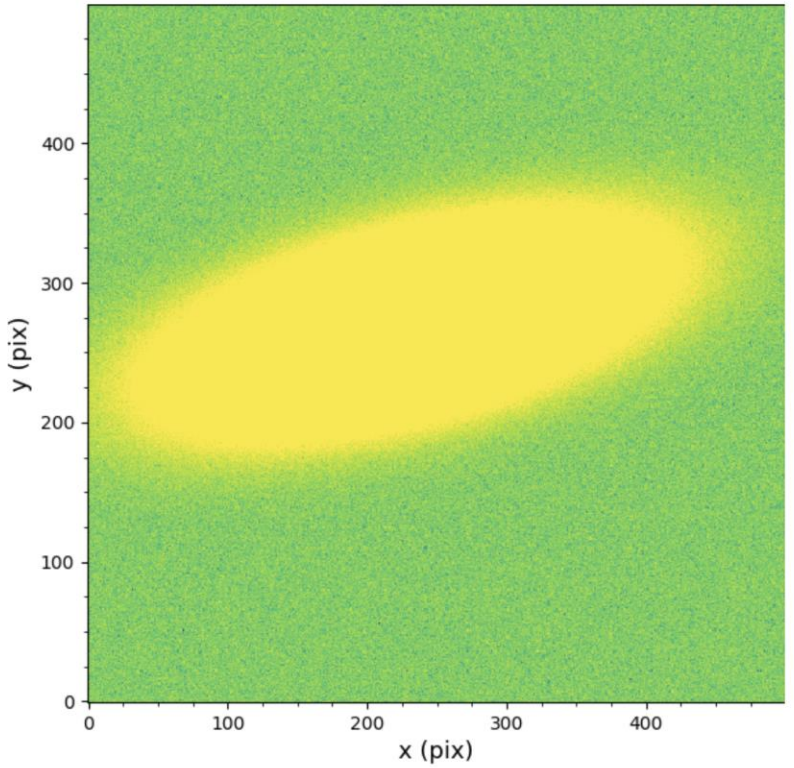




How big is this galaxy?







Summary: Understanding --> Information

Key takeaways:

1. Surveys require domain specific knowledge from design to analysis and interpretation. The theme of this week is the connection between *domain* and *data* science.

2. Information loss isn't due to the physical instrument, but due to the image processing methods (or the "choice of data compression/representation")

3. Improving our processing approach, can do up to a factor of several (or an order of magnitude) better just by reprocessing the same dataset. *Will future surveys just be software?*

An important point: Discovery oriented science at the frontiers of an instruments capability often obscures 'normal' but excellent science that can be done with less overhead. Spending more time with archival data can lead to insights with less overhead than working with/proposing for new datasets.

Notebook Overview: