Challenges for Quasar Science with LSST

Gordon Richards

Drexel University

LSST AGN Science Collaboration Roadmap Development Meeting January 3, 2017

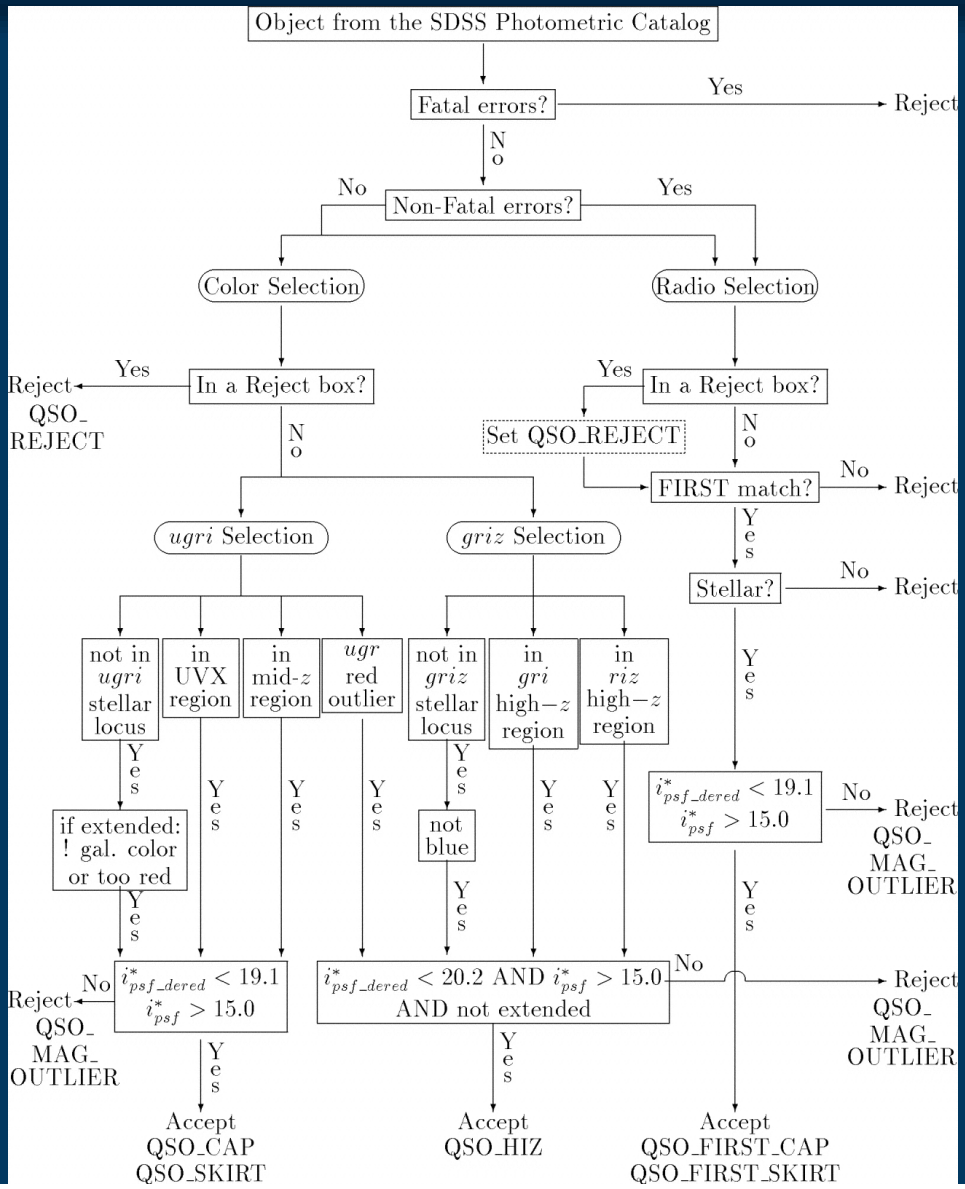
Finding AGNs: The Old Way

Complex, Glorified Color Cuts

Perfectly fine when coupled with spectroscopy to find specific needles in haystack.

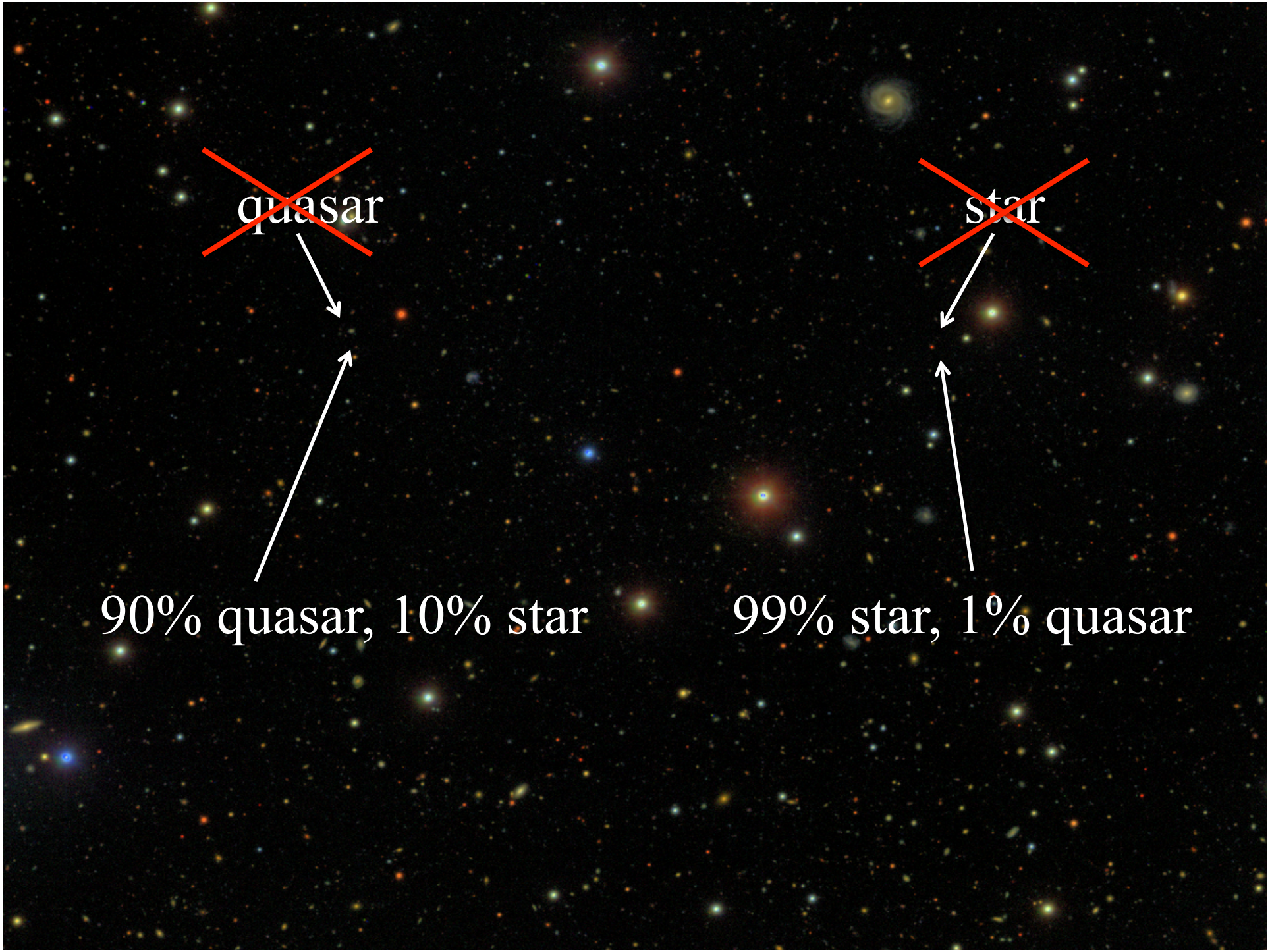
Completely inadequate for LSST

Richards et al. 2002



In a Nutshell

- Find AGN probability for every object on sky


quasar
star
90% quasar, 10% star
99% star, 1% quasar

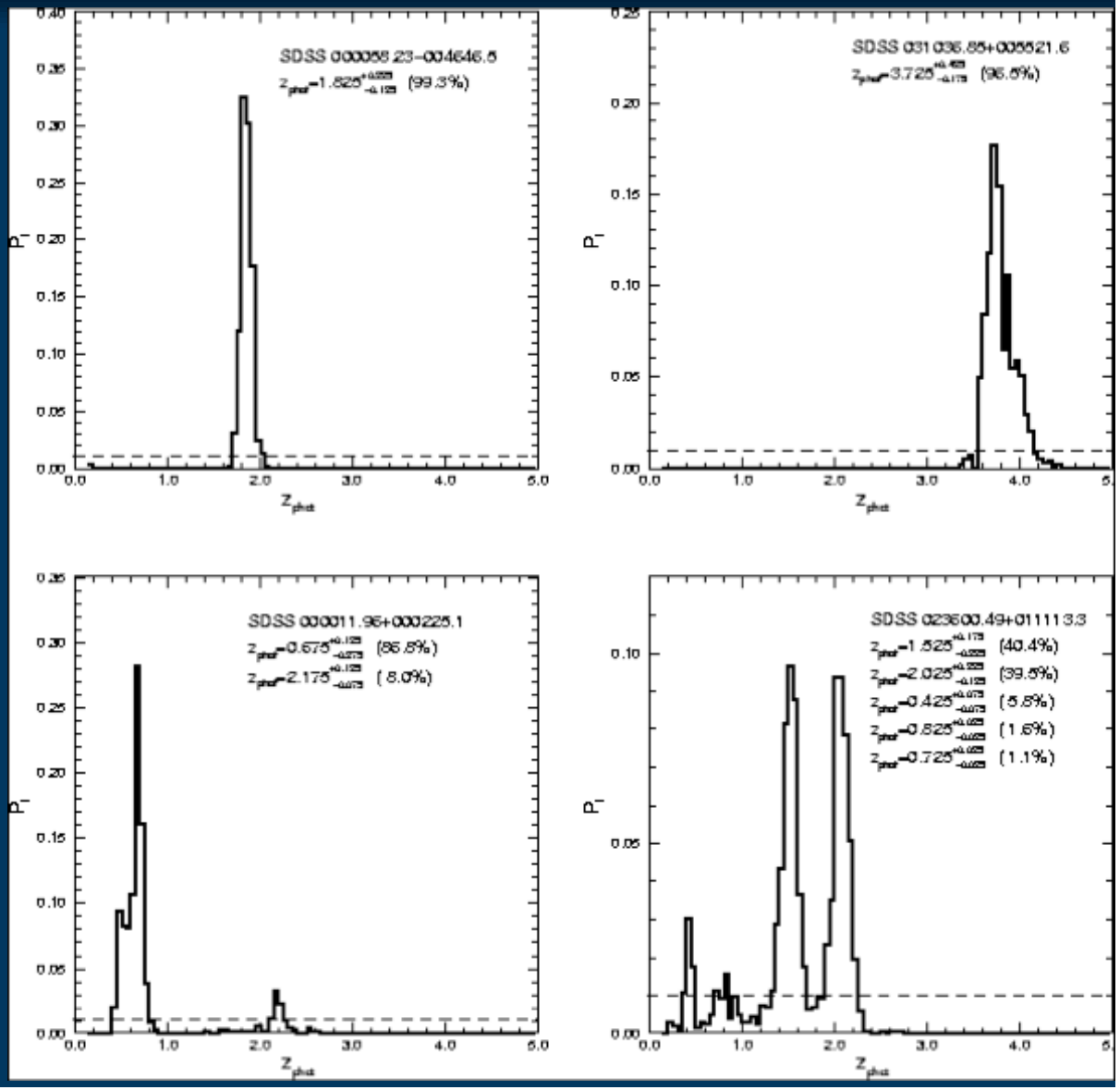
In a Nutshell

- Find AGN probability for every object on sky
- Determine Photo-z Probability Distribution Function (PDF)
- (Possibly recompute using initial information from other science collaborations?)

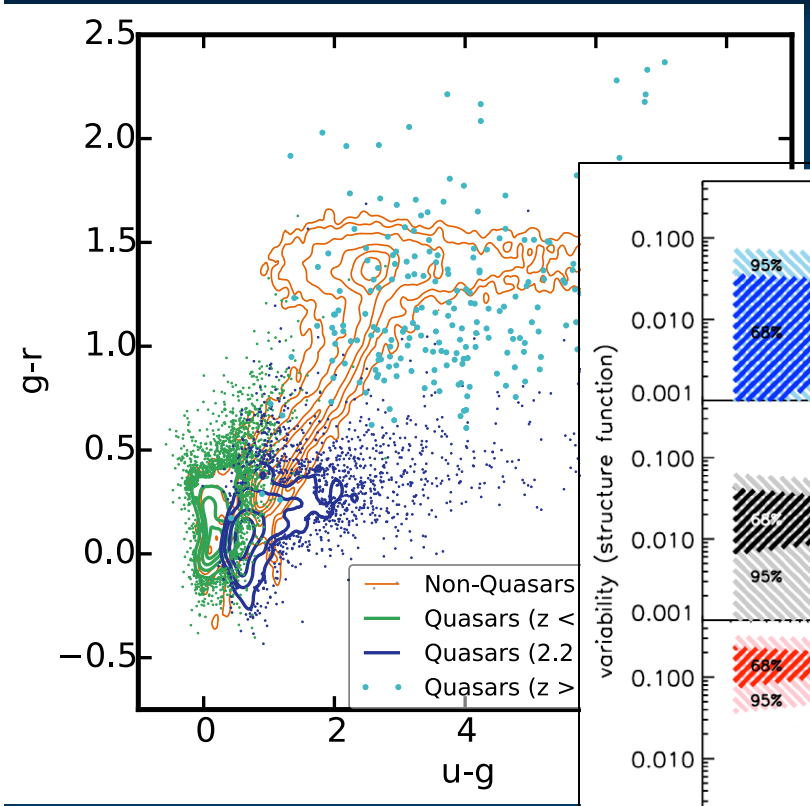
In the absence of spectroscopy these processes need to be maximally complete, efficient, accurate, and precise.

Photo-z PDFs

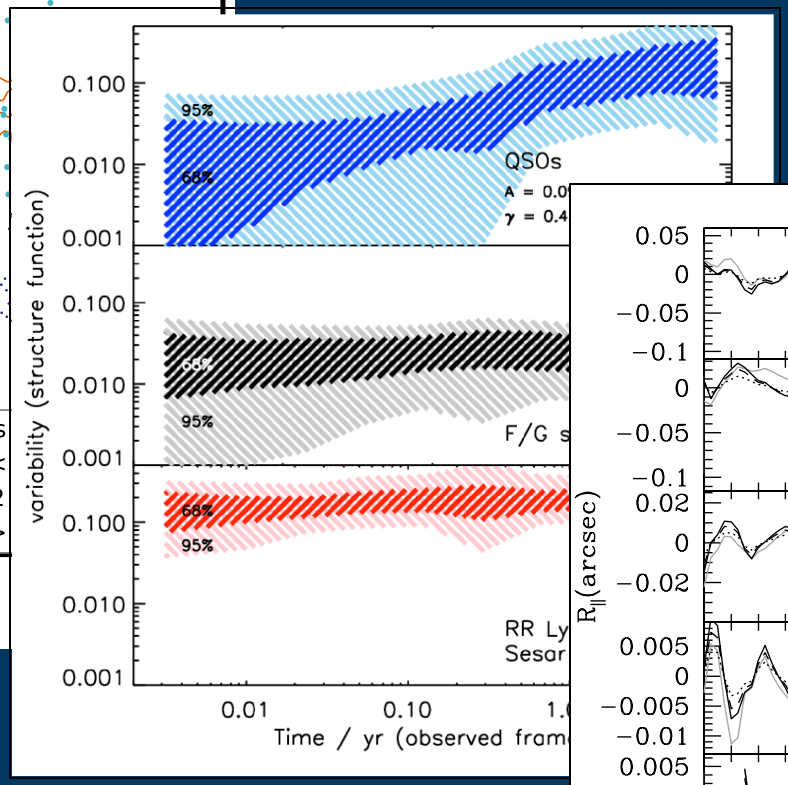
Photo-z's as full PDFs instead of best-fit single values.



Maximal Use of Attributes

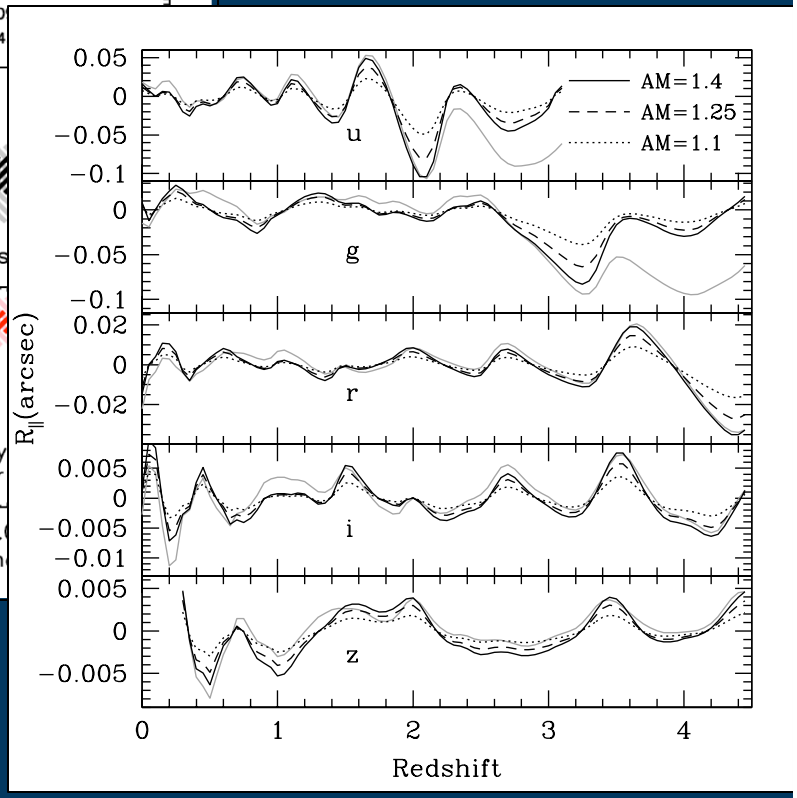


colors



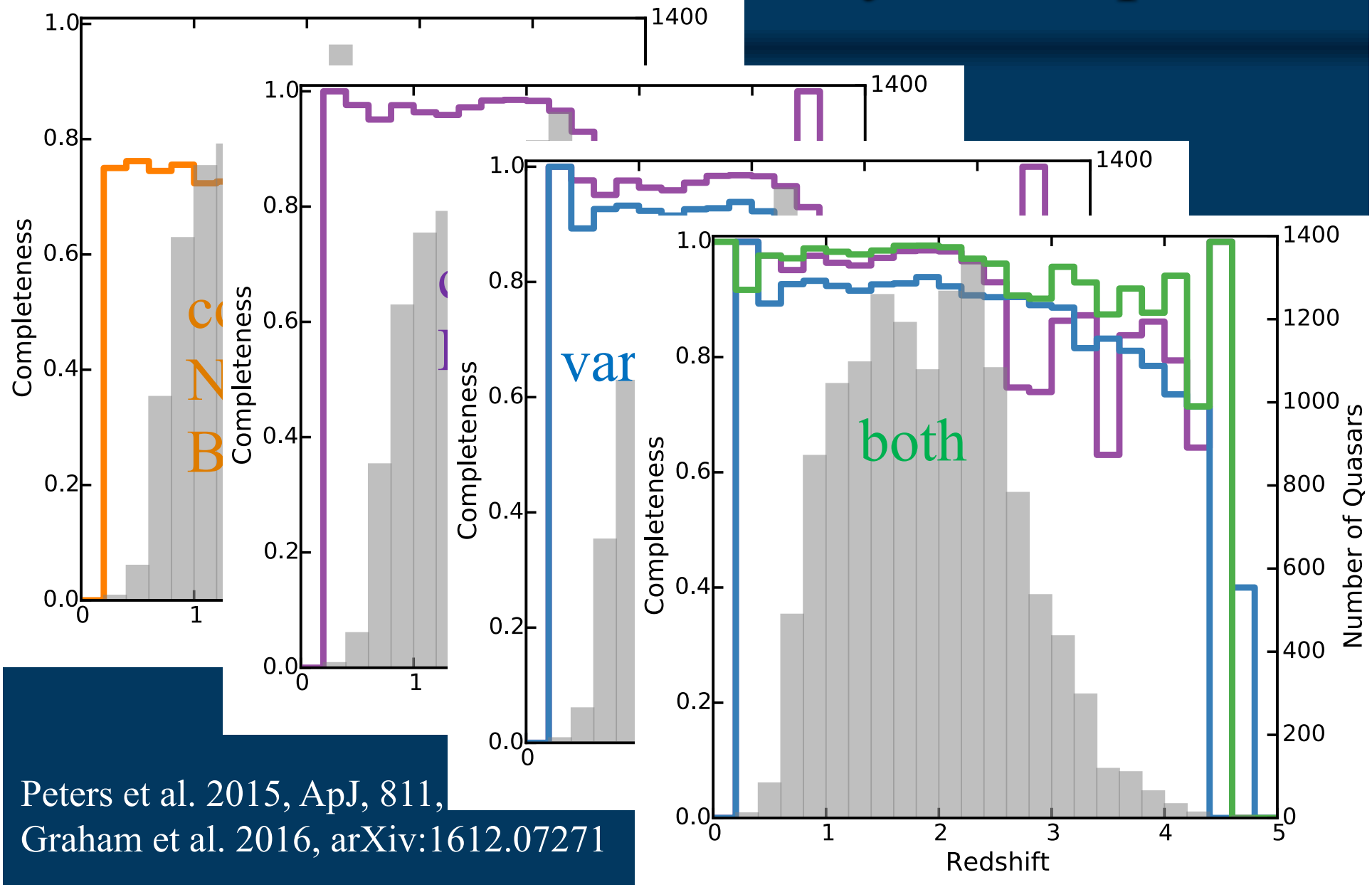
variability

astrometry



Also proper motion and multi-wavelength data.

Color AND Variability Examples



Peters et al. 2015, ApJ, 811,
Graham et al. 2016, arXiv:1612.07271

Machine Learning Algorithms

Need to start using Machine Learning algorithms (e.g., Scikit-Learn) They can handle mixed-origin attributes. Just need to be able to "whiten" them (transform to zero mean and unit variance).

In [1]:
```
from astropy.table import Table
import numpy as np
```

In [ ]:
```
# Read in training file
data = Table.read('GTR-ADM-QSO-ir-testhighz_findbw_lup_2016_starclean.fits')
Xtrain = np.vstack([ data['ug'], data['gr'], data['ri'], data['iz'], data['zs1'], data['s1s2']]).T
ytrain = np.array(data['labels'])
```

In [ ]:
```
# "Whiten" the data
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
Xtrain_scaled = scaler.fit_transform(Xtrain)
```

In [ ]:
```
# Instantiate the Random Forest classifier
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=10, max_depth=15, min_samples_split=2, n_jobs=-1, random_state=42)
rfc.fit(Xtrain_scaled, ytrain)
```

https://github.com/gtrichards/QuasarSelection
https://github.com/gtrichards/PHYS_T480

Bayes' Rule: Attributes vs. Priors

Might think to use mags/fluxes instead of colors, but not normal distributions, so can't whiten and training data won't match test data. So use as priors instead.

$$P(Star \mid x) = \frac{P(x \mid Star)P(Star)}{P(x \mid Star)P(Star) + P(x \mid QSO)P(QSO)}$$

Where

- x = N-D "attributes" (colors, variability params, etc.)
- P(Star|x) = probability of being a star, given x
- P(x|Star) = probability of x, drawing from stars training set
- P(Star) = stellar prior
- Star if P(Star|x)>0.5, QSO if P(Star|x)<0.5

Other Issues/Complications

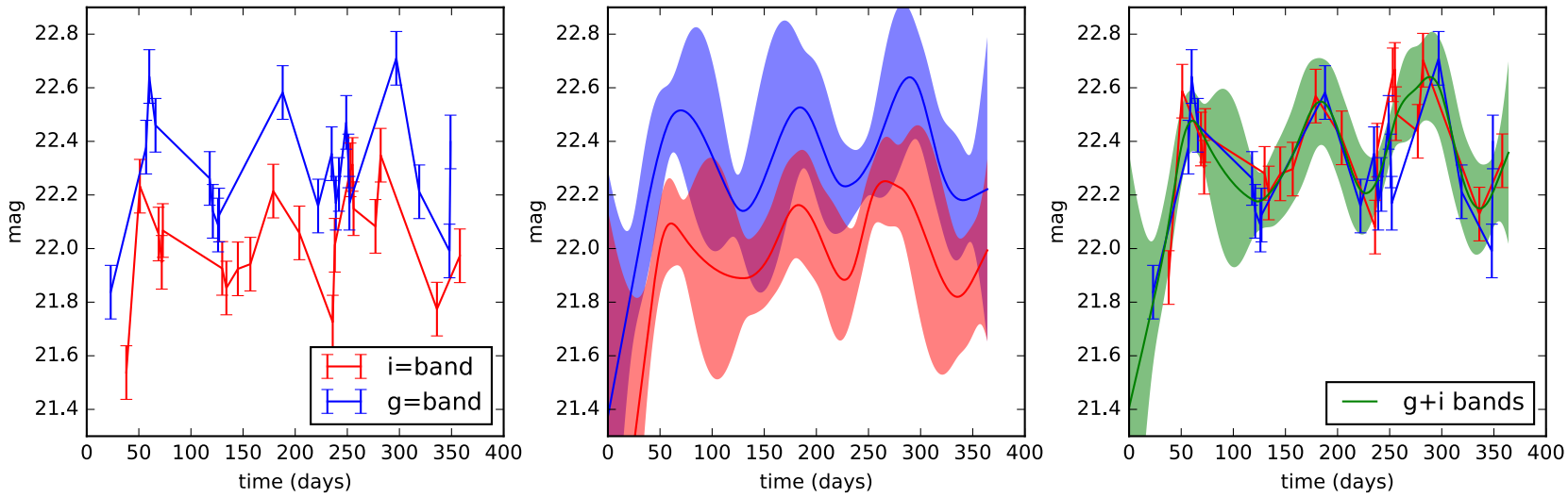
How to determine colors from non-simultaneous data in multiple epochs?
How to incorporate variability?
How to incorporate astrometry?
How to incorporate multi-wavelength data?
Star-galaxy separation?
Using existing data to simulate LSST data.
Using OpSim.

Variability

- Does object vary like an AGN?
- If we had fixed time sampling, we could just treat like magnitudes (but we don't).
- So, probably have to do a parametric fit (despite the resulting loss of information).
- But now have a normally distributed "attribute".
- Most important of all is how to merge different bandpasses to create a more well-sampled light curve.

Merged Light Curves

Would like to treat g-band data as modified i-band data and double the cadence:



Above is an example of "kriging" together two light curves using Gaussian Random Processes.

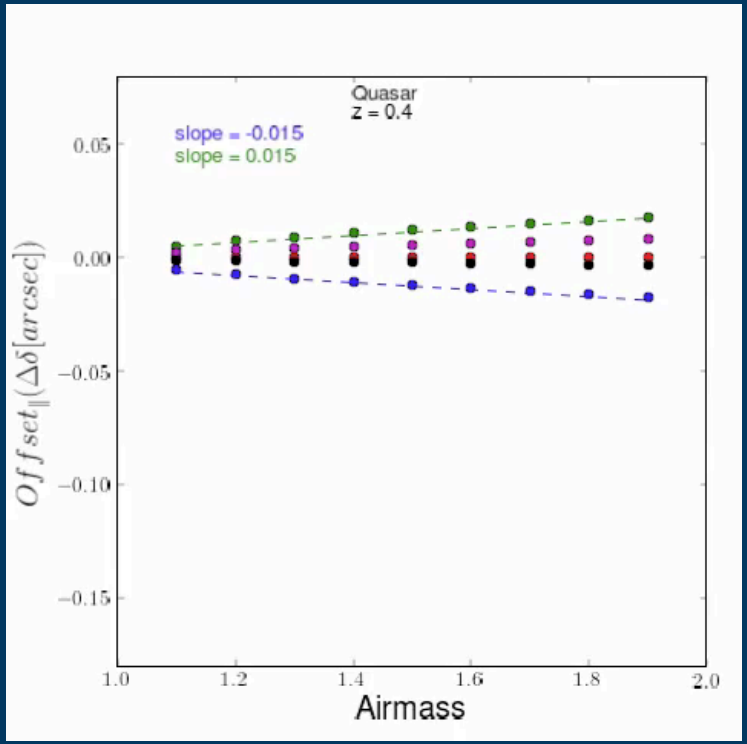
Question: How to handle "brighter=bluer" and time delays.

Astrometric Data

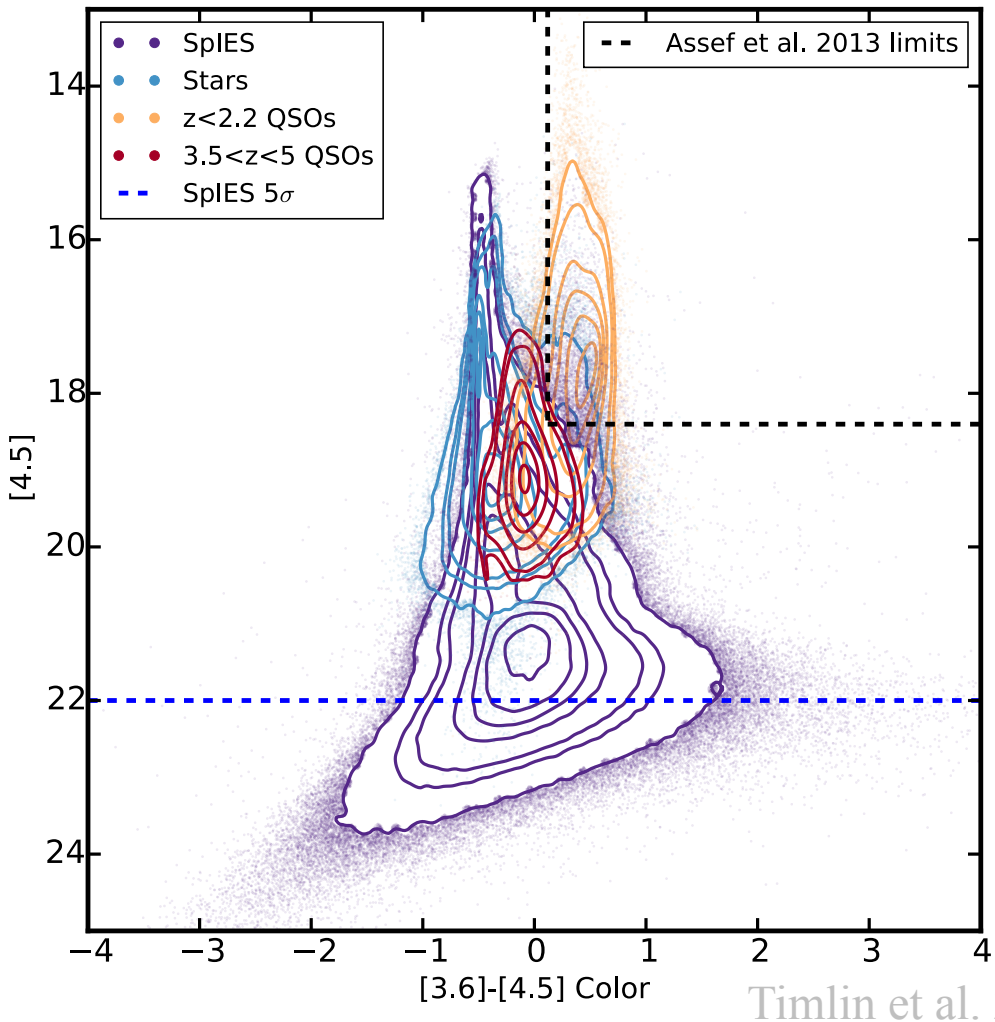
Is the object moving?
If not can we learn something from Differential Chromatic Refraction?

Astrometric data has similar issues as variability.
Specifically, need to parameterize:

- Proper Motion (mas/yr)
- Slope of offset with airmass (Kaczmarczik+09, Peters+15)



Multi-wavelength Selection



Timlin et al. 2016

Adding X-ray, IR, radio, other optical/ UV, and even temporal data

Issues:
Balkanization
Dropouts
Resolution
Bandmerging

Need ambassadors

Photometric Redshifts

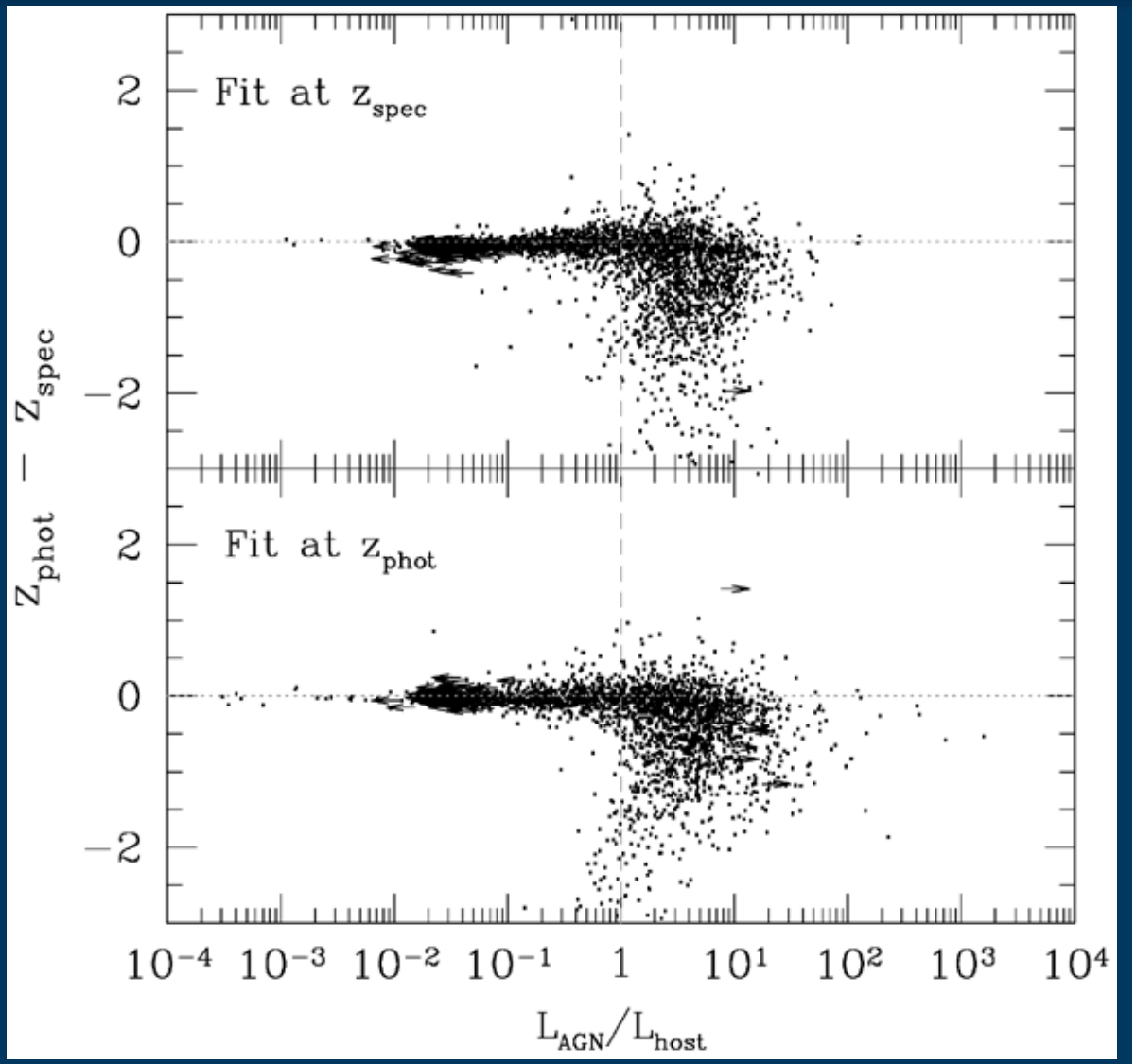
In many ways, photo-z estimation parallels classification. All of the attributes used for classification can/should be used for photo-z estimatation.

Indeed, might even want to handle both problems simultaneously (i.e., don't just given AGN probability, but AGN @ z=0.1, 0.2…7.0).

Lots of good Machine Learning algorithms for regression.

Photo-z: Parametric vs. Non

Template (parametric) methods good at low-L, empirical (non-parametric) methods good at high-L



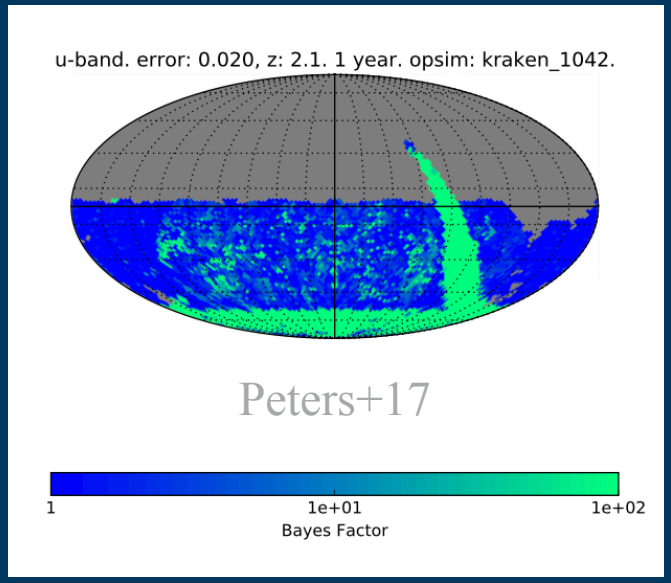
Assef+10

Simulations

Key to all of these things is working in the so-called "Metric Analysis Framework (MAF) and using OpSim outputs.

This includes using existing data sets to simulate LSST. E.g., Catalina data

A problem is that many people aren't familiar with the MAF (including me). So, need to get everyone on board.



Truth from Deep Drilling Fields

The Deep Drilling Fields will be our **Truth Tables**.

We should be honing our algorithms in these fields NOW using existing data.

Then blanketing them with spectroscopy. Again NOW.

Pre-Science

A goal should (must) be to use simulated data to produce keys science results in **advance** of the data.

Conclusions

- All objects need to be assigned an AGN probability (and redshift PDF)
- Need to maximize use of machine learning algorithms and all available attributes.
- Will need ambassadors to guide the integration of data from different facilities, computation of parametric attributes, algorithm development, etc.
- Need to work in MAF context and develop DDFs.

– https://github.com/LSSTScienceCollaborations/ExtragalacticScienceRoadmap/bh.tex