Automatic classification of sources in large astronomical catalogues - how LSST will gain from experience gathered on the largest datasets available today

or

Preparing the Polish scientific community for LSST

Agnieszka Pollo

Poland

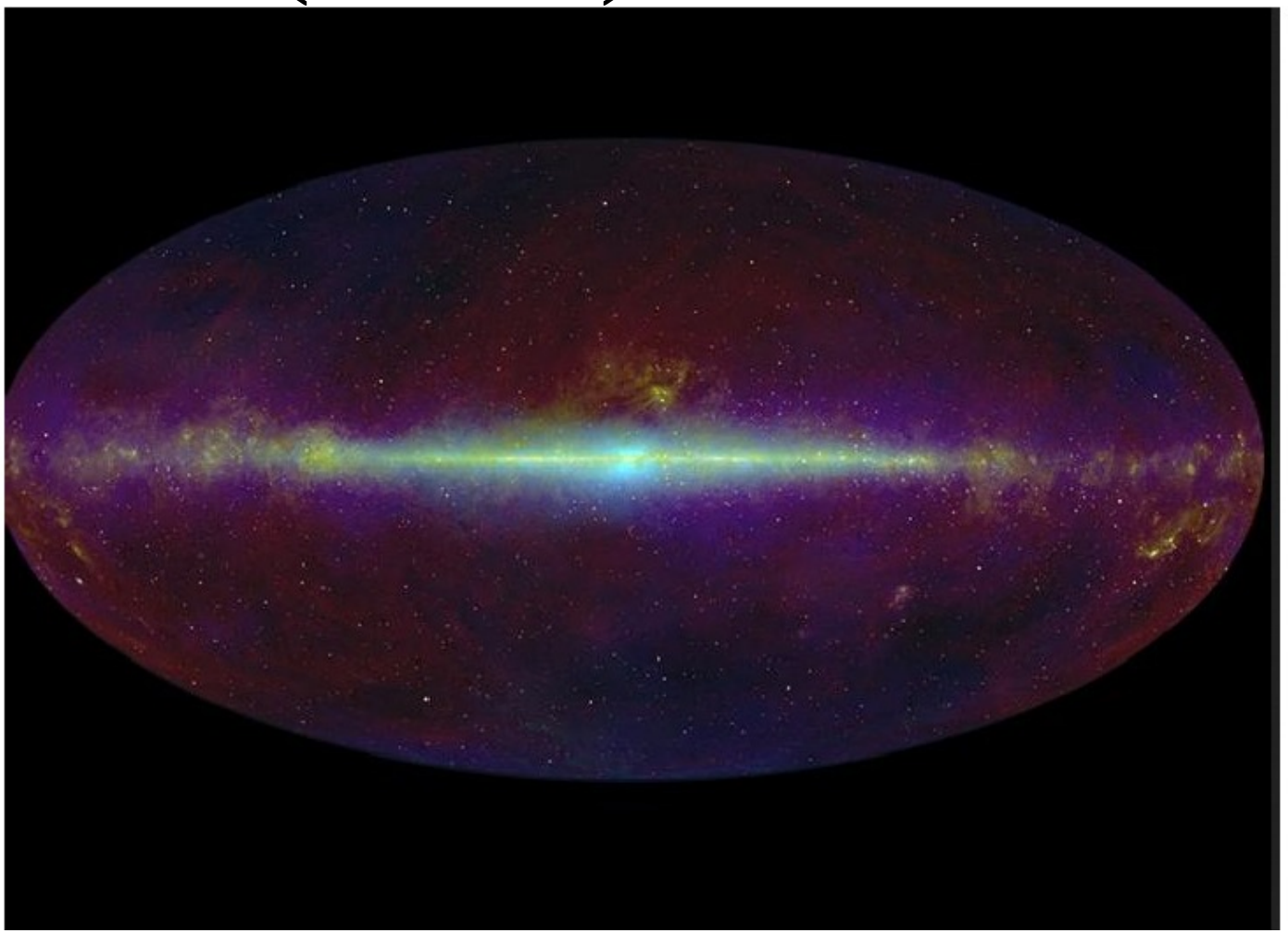
special thanks to

Ola Solarz, Kasia Malek, Maciek Bilicki, Gosia Siudek, Tomasz Krakowski, Agnieszka Kurcz, Magda Krupa, Artem Poliszczuk, Szymon Nakoneczny, Bozena Czerny, Lukasz Wyrzykowski

Machine learning for source classification:
Supervised → when we know a priori what sources we expect to find and we can use some datasets for training
→ classification (for separate groups) or
→ regression (for smooth transition)
Unsupervised → clustering of sources into previously unknown and unexpected classes

Source classification of very large data: Wide-field Infrared Survey Explorer (WISE)

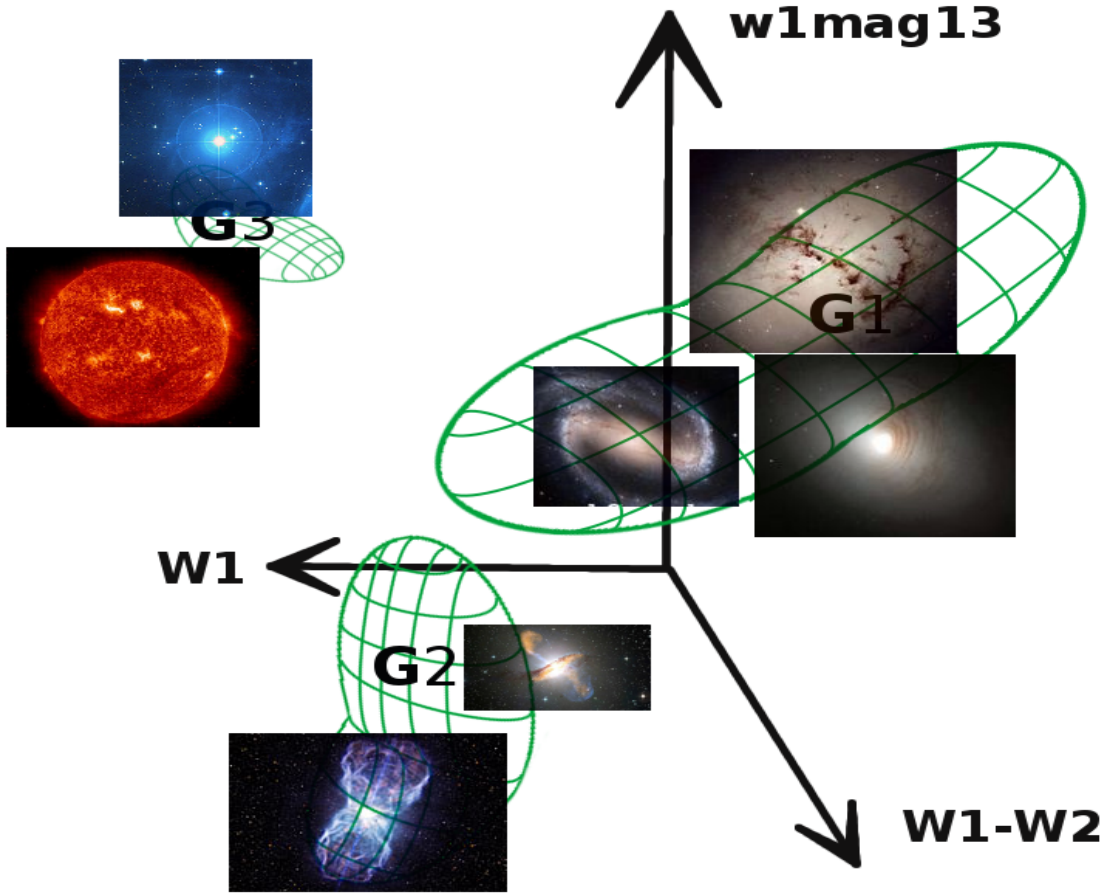
- All-Sky survey in IR
- Detected over 747 mln sources
- (15 PB of data; tables + images)
- Publicly available (position, photometry in 4 bands from 3.6 to 22 um)
- Low angular resolution (~6”)
- No redshift information so far



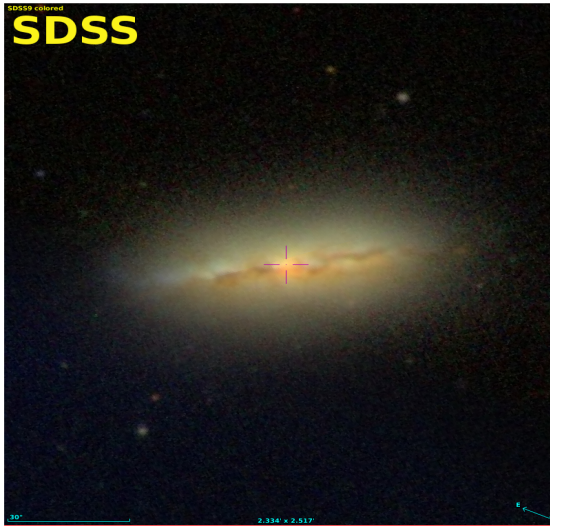
WISE: first step towards ML novel source detection

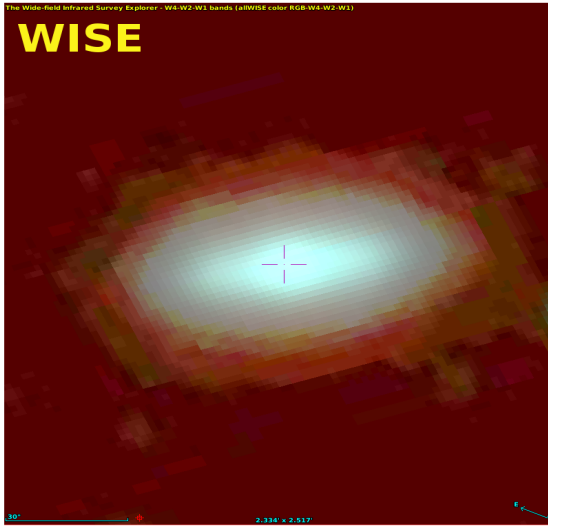
Training set (all what we expect):

AllWISE x SDSS (α,δ) with spectro-z (secure)

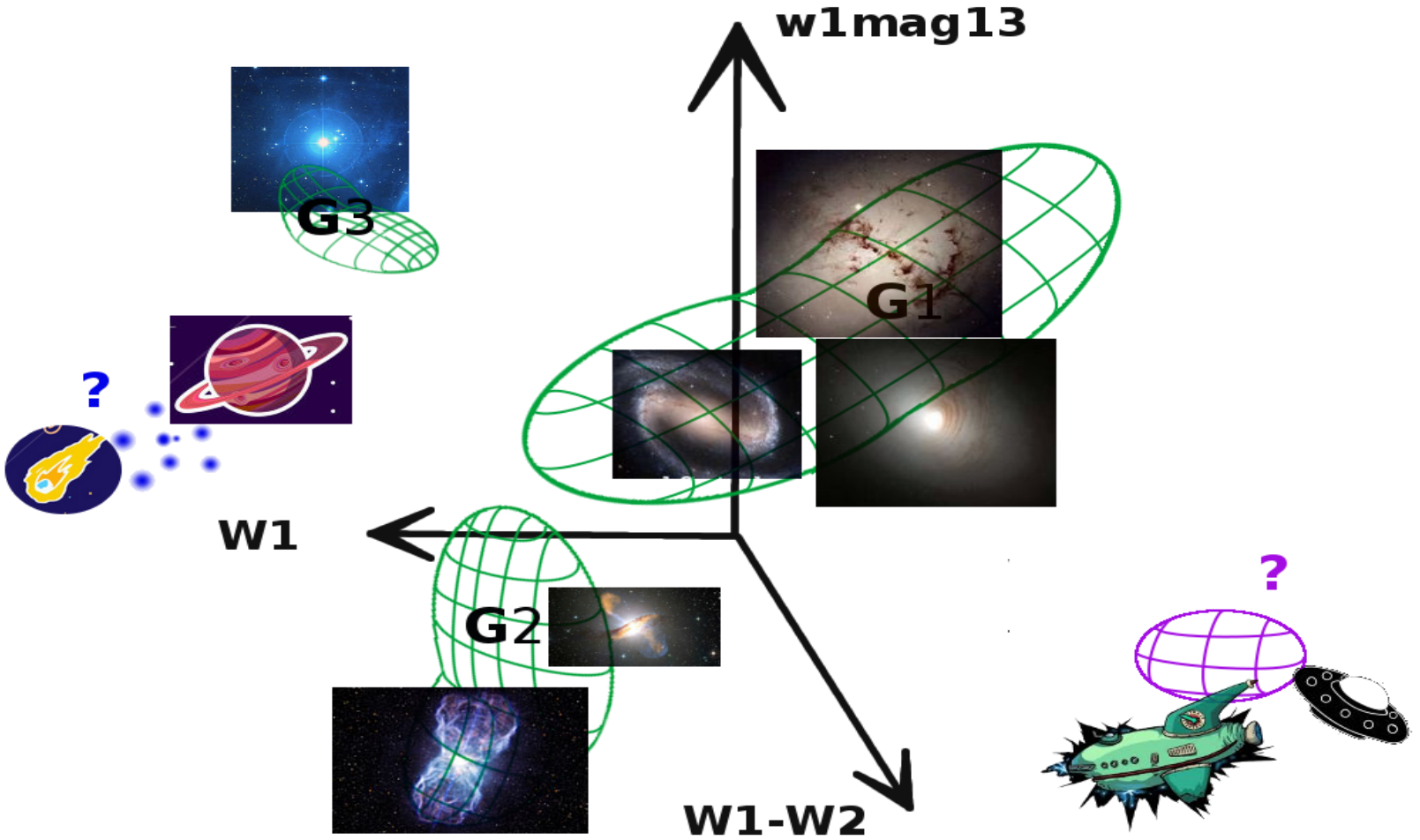


A. Solarz et al. 2017



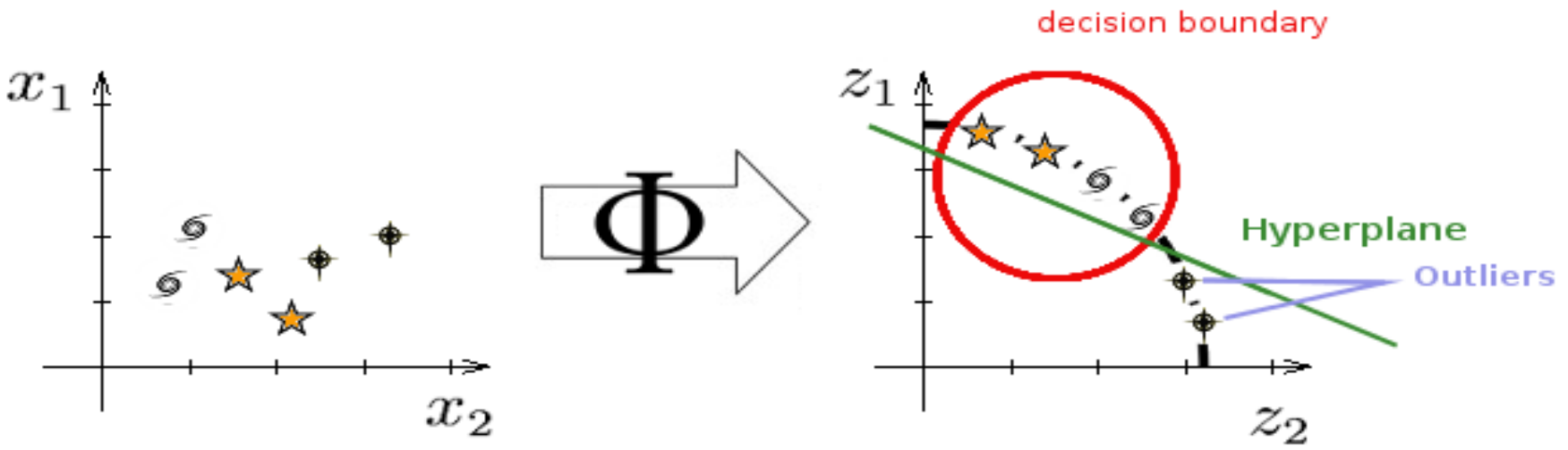


WISE: accounting for unknown unknowns



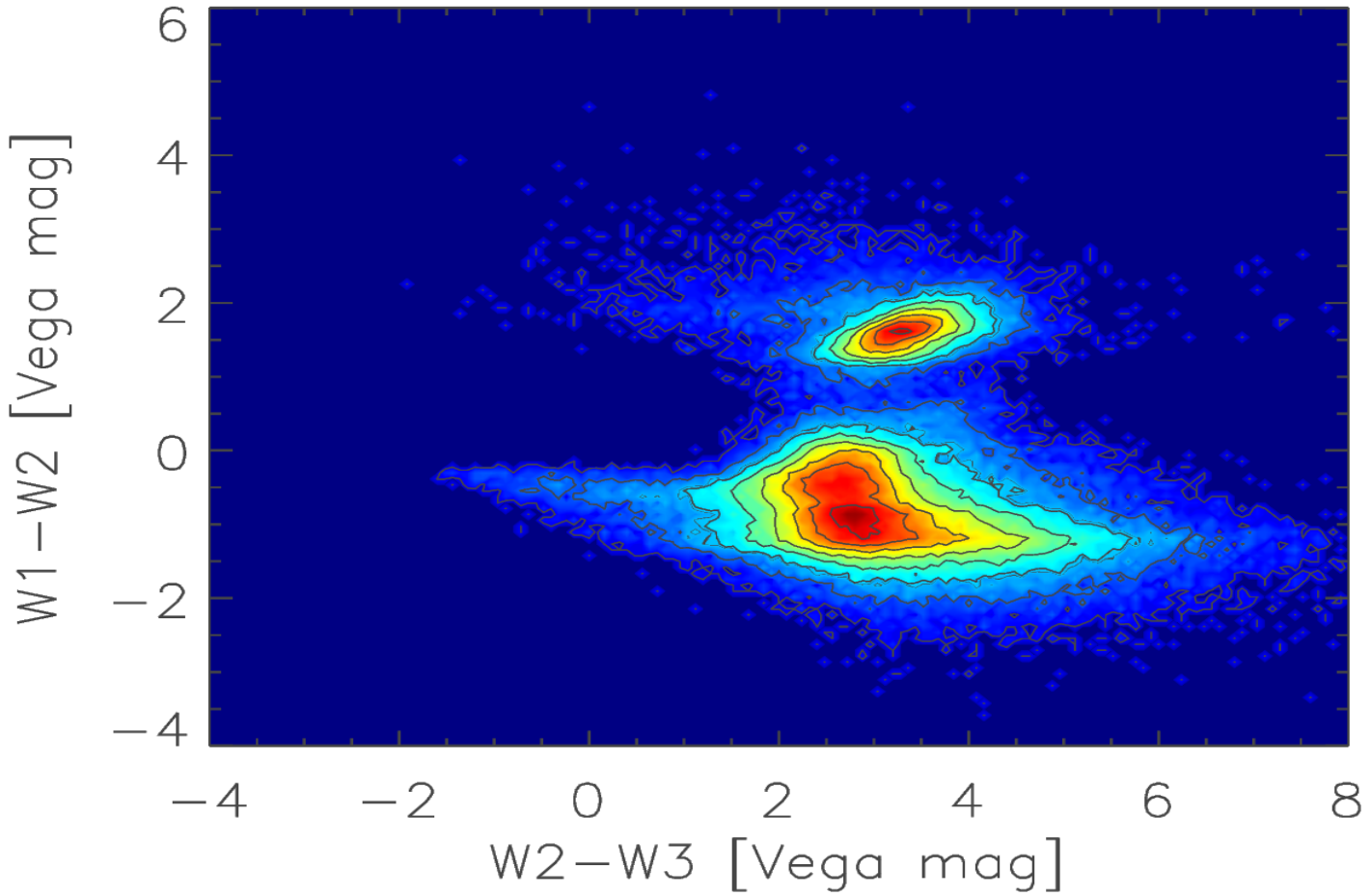
A. Solarz et al. 2017

Novelty detection with One-Class Support Vector Machines



Create one ‘known’ class (mix of AllWISE x SDSS galaxies, stars, QSOs)
Maps input data to a higher D parameter space (based on Kernel methods)
Hypersurface hugging the expected sources
Anything with ‘unknown’ patterns falls outside the hypersurface => novelties

A. Solarz et al. 2017



Results:

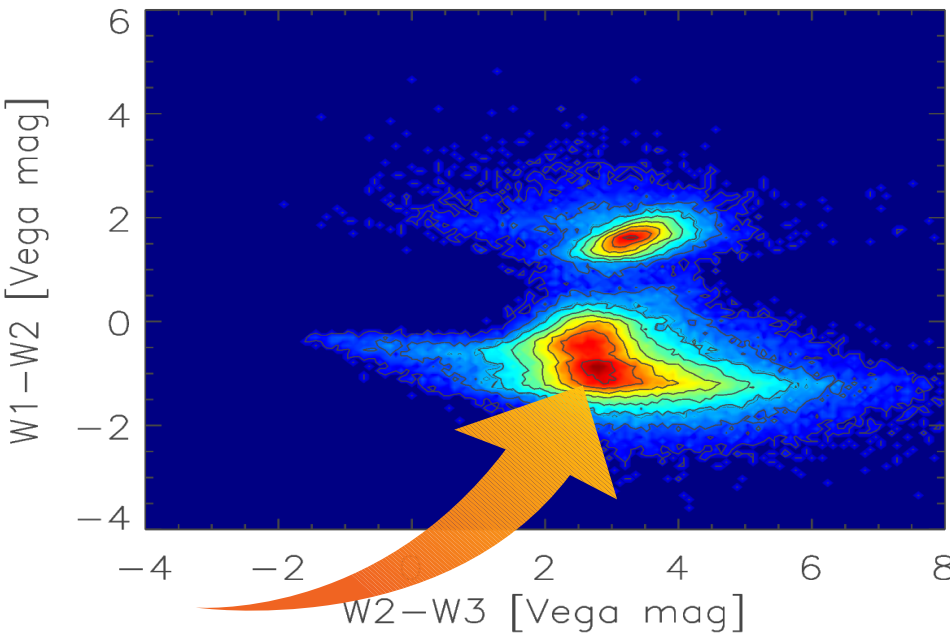
~650,000 anomalous sources

What are they?

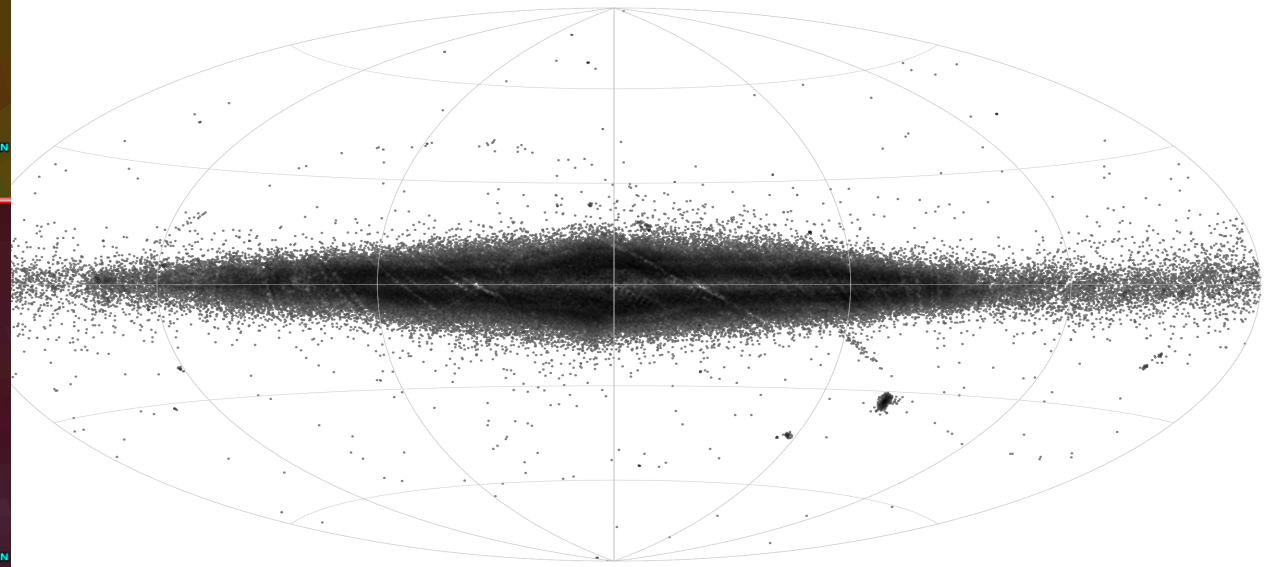
A. Solarz et al. 2017

Spurious sources

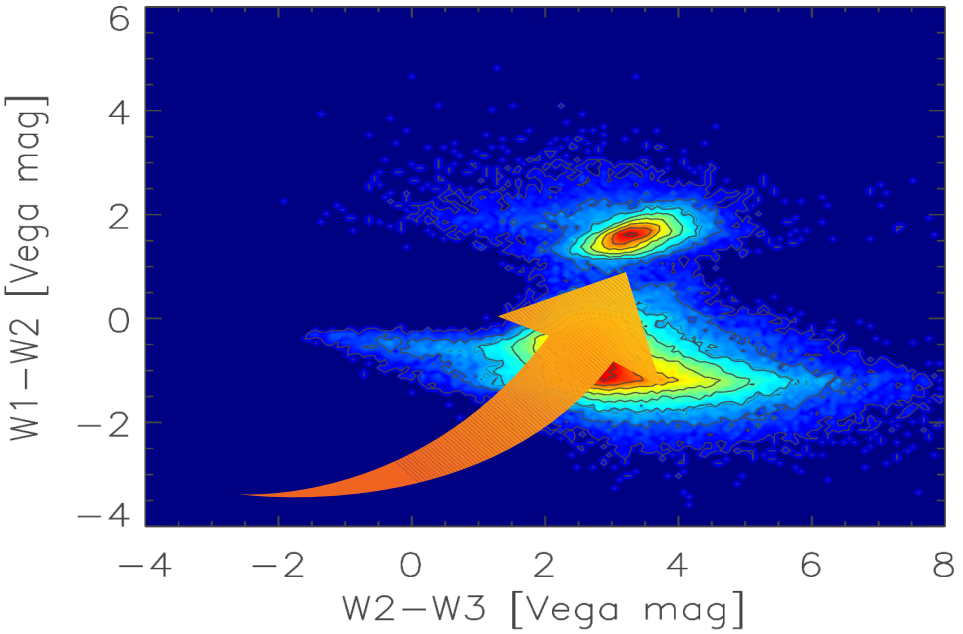
W1-W2 ~ -1 ; 80%
Spitzer GLIMPSE:
IRAC I1 [3.6 um], IRAC I2 [4.5 um]
Low WISE resolution (6”)
in crowded fields => blends



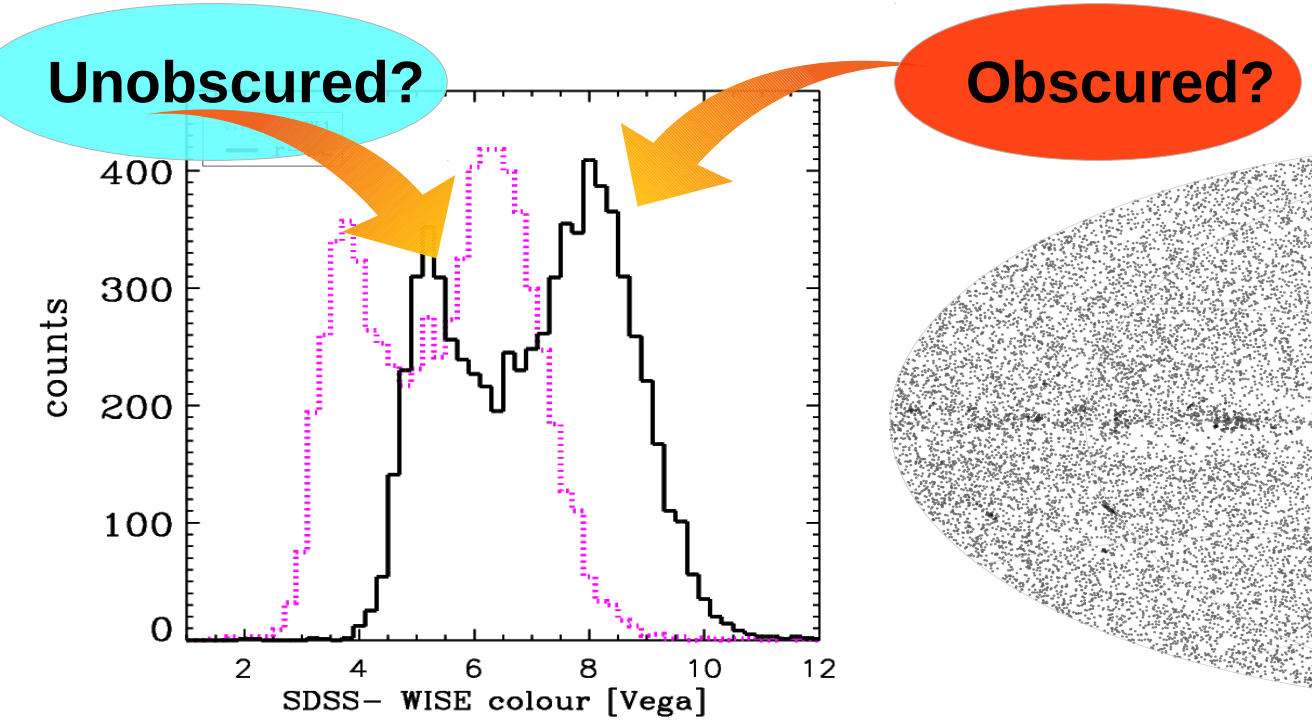
A. Solarz et al. 2017



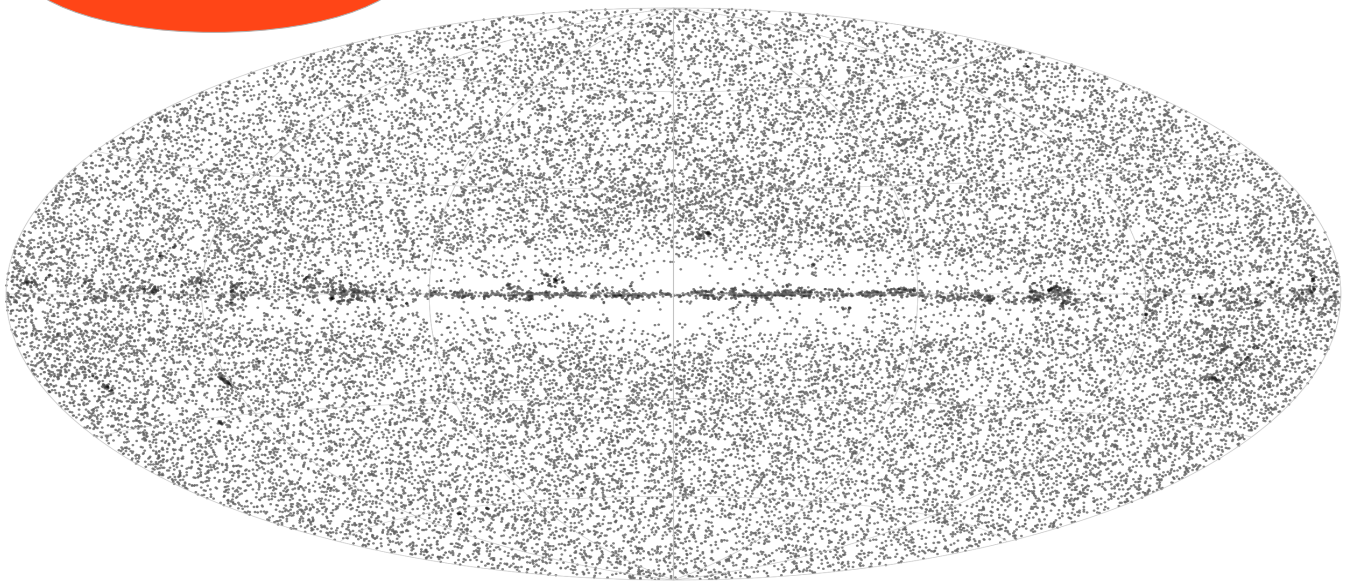
AGN candidates?



30,000 sources (Galactic Plane: mostly blends)
W1 [3.6 um]~ 16 [Vega mag], W3 [12 um] ~ 10 [Vega mag]
Warm dust emission/PAH emission lines
76% undetected at other wavelengths!
~7 000 objects with SDSS photometry (no spectro-z)



Solarz et al. 2017

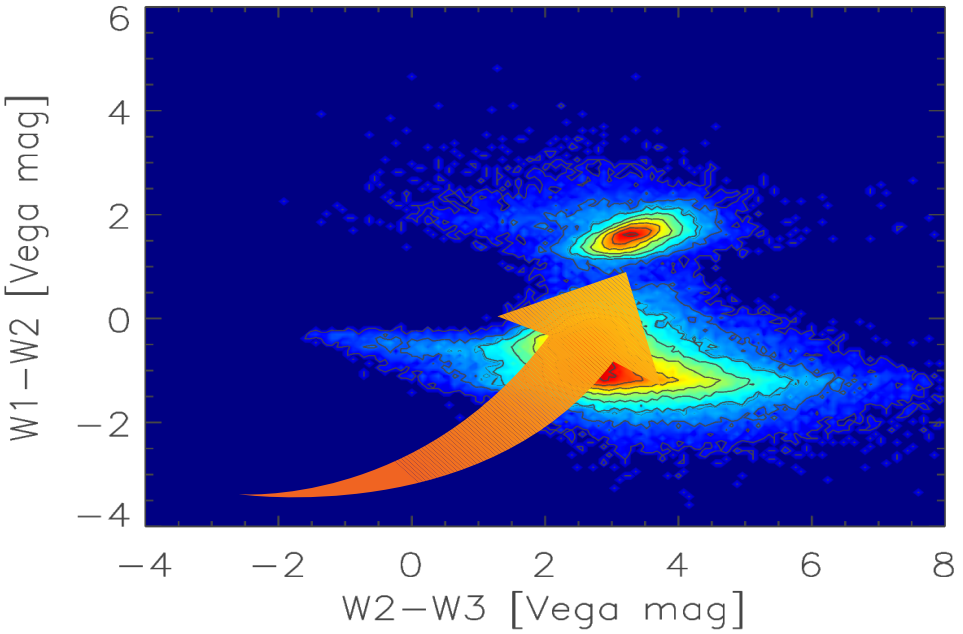


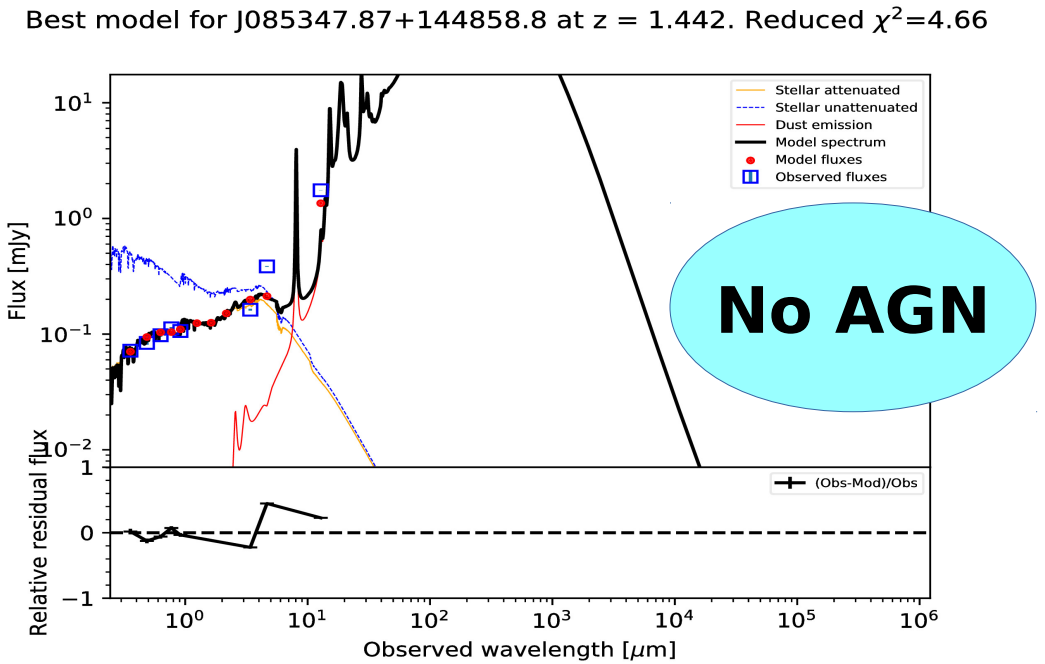
AGN candidates?

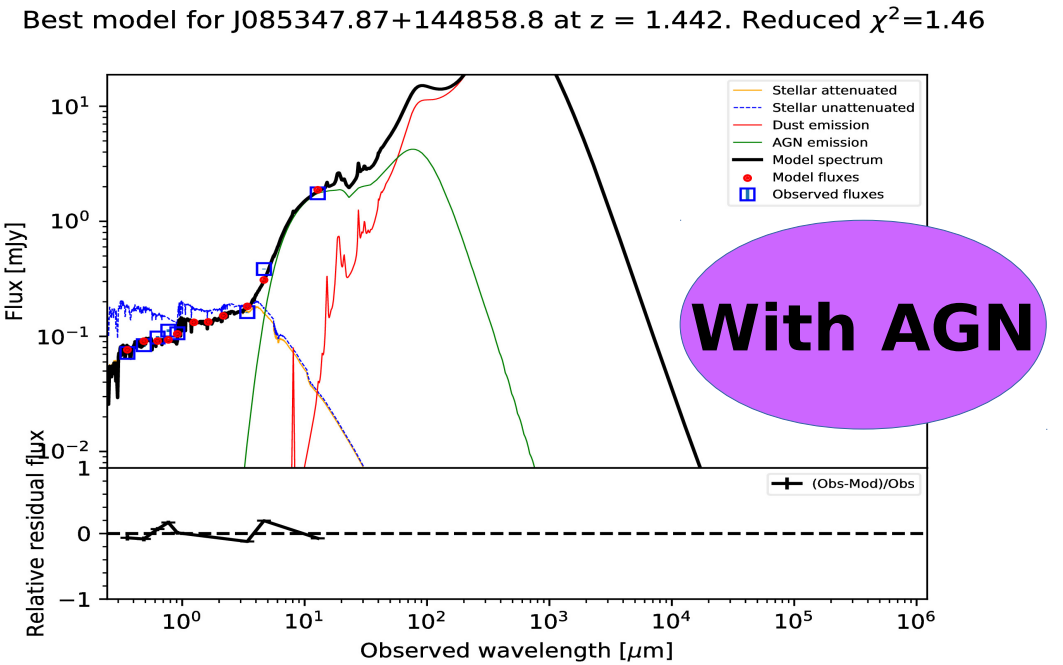
Photo-z for ~2 700 obj (Beck+16).
SDSS + WISE photometry
Spectral Energy Distribution with CIGALE

RESULTS:

AGN component necessary to explain IR fluxes
85% (Ultra)Luminous Infrared Galaxies





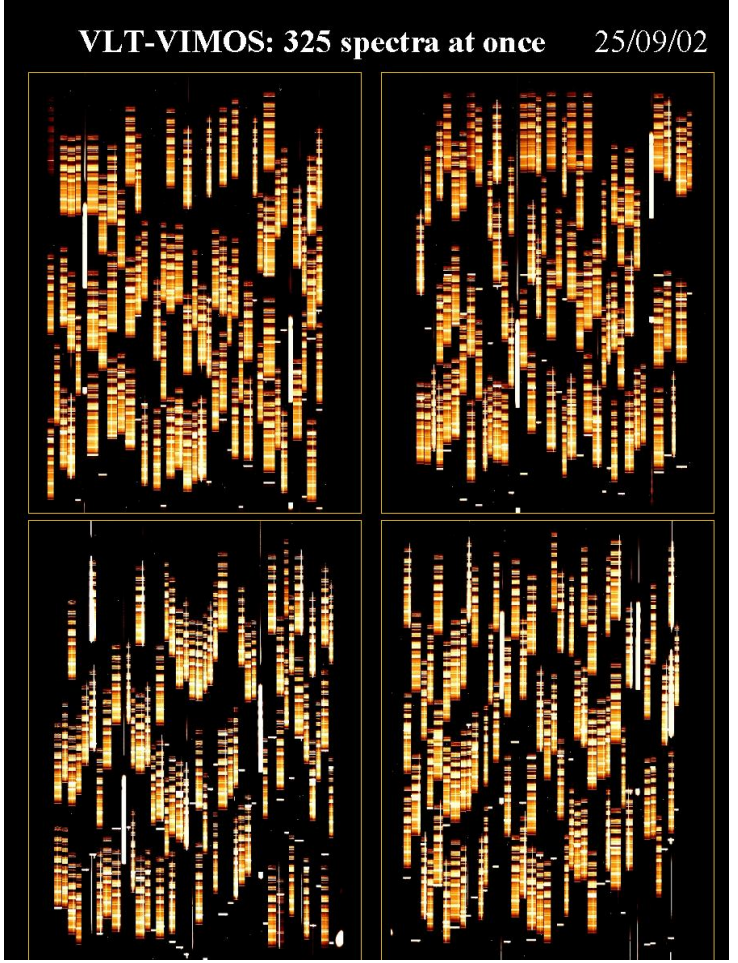


A. Solarz et al. 2017;

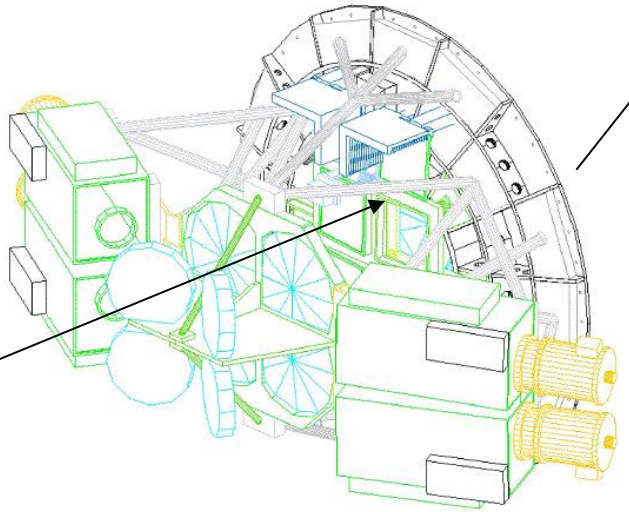
https://cigale.lam.fr/

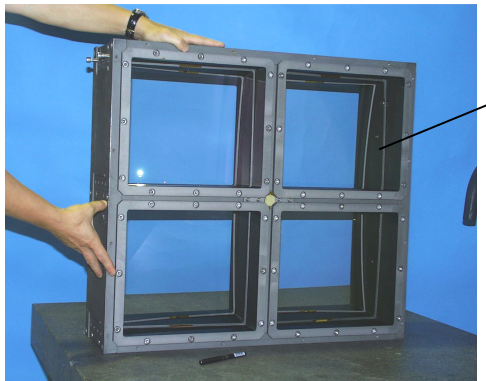
VIPERS

VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY



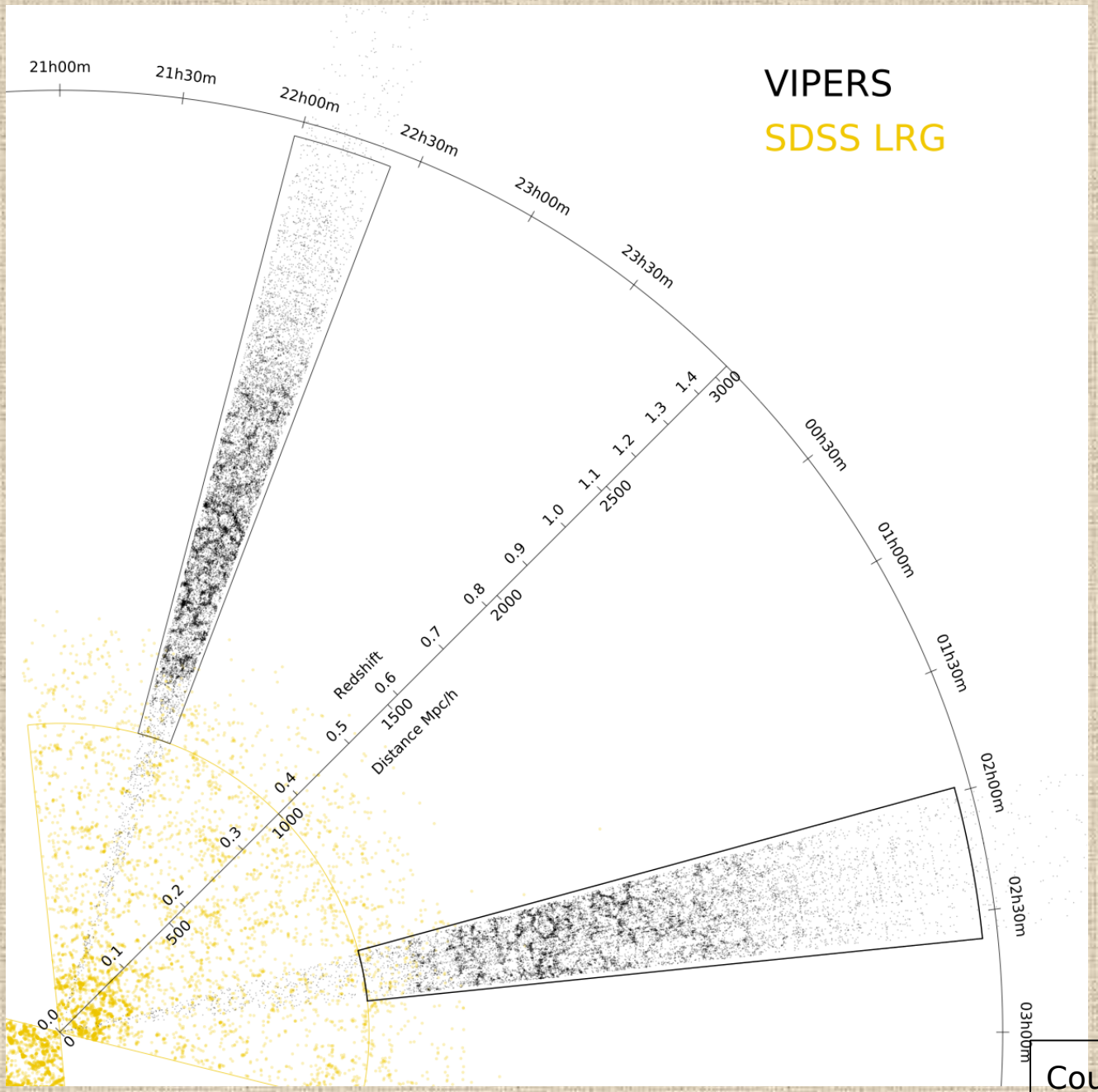
Large ESO Programme, started in 2008,
Data publicly released in the fall of 2016.
http://vipers.inaf.it/rel-pdr1.html





Goal: **100,000** spectra
of galaxies
at $0.5<z<1.2$

Guzzo et al. 2014, 2017, Scodeggio et al. 2017



Courtesy Ben Granett

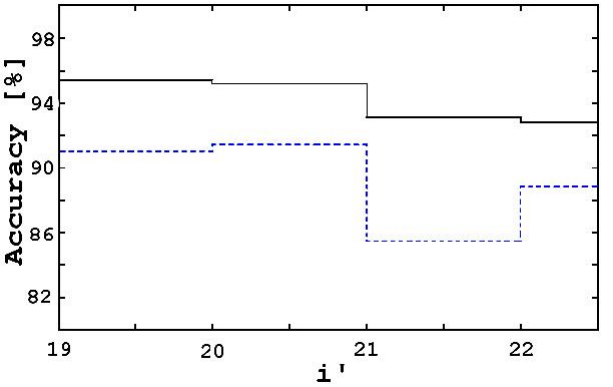
VIPERS: the case of rich data

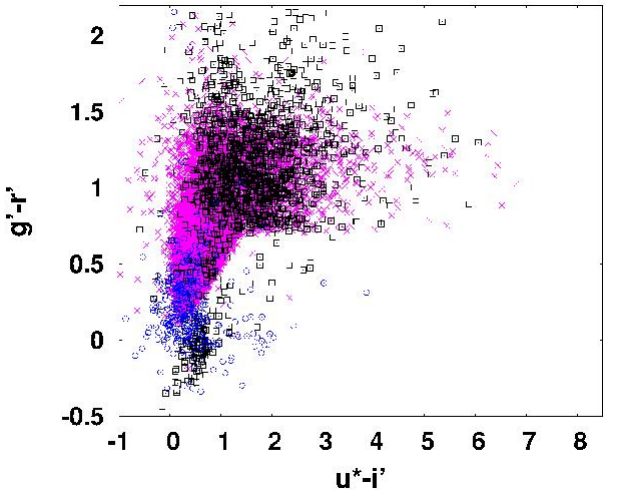
Question: based on the observed colors, but having a well defined training sample based on spectroscopic (VIPERS) data, how well can we pre-classify a sample into galaxies, stars and AGNs at $0.5<z<1.2$?

Malek, Solarz, Pollo et al. 2013

VIPERS-trained SVM classifier for AGNs, stars and galaxies at z>0.5

- Trained on almost 20,000 VIPERS sources with the best spectroscopic measurement
- Optical (based on 4 apparent magnitudes in u', g', r' i' bands) and NIR+optical classifiers trained
  - NIR measurement dramatically increases the classifier's accuracy
- Classification pattern which is not obvious from color-color plots
- → Similar approaches will work for LSST and should work well; NIR follow-up can improve the results significantly





Malek, Solarz, Pollo et al. 2013

Question: having such an unprecedented wealth of spectroscopic data, can we classify galaxies better than just traditional blue-red-green valley galaxies?

Method: unsupervised classification based on a feature space of absolute magnitudes + redshifts.

Siudek et al. 2018 a, b

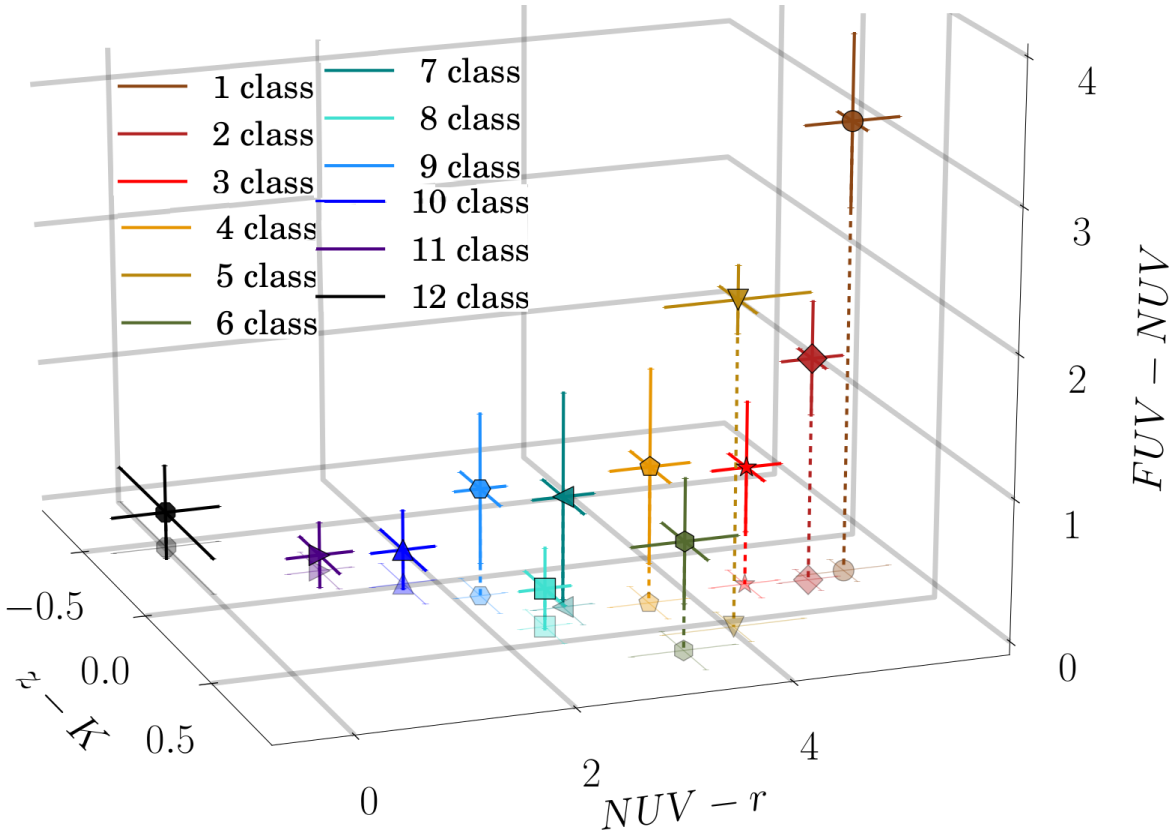
Unsupervised classification of galaxies at z>0.5

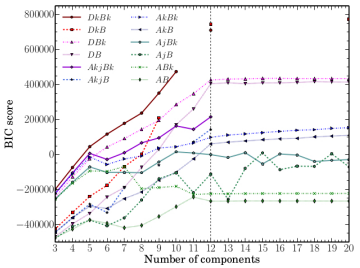
Unsupervised classification of VIPERS galaxies based on their distribution in a multidimensional absolute magnitude space

12 dimensions: absolute magnitudes + zspec

→ **blind separation** (no training sample or hints) into **12 classes**, which are well separated in the multidimentional feature space.

Method: Fisher expectation maximization algorithm (FEM; Bouveyron& Brunet 2011).

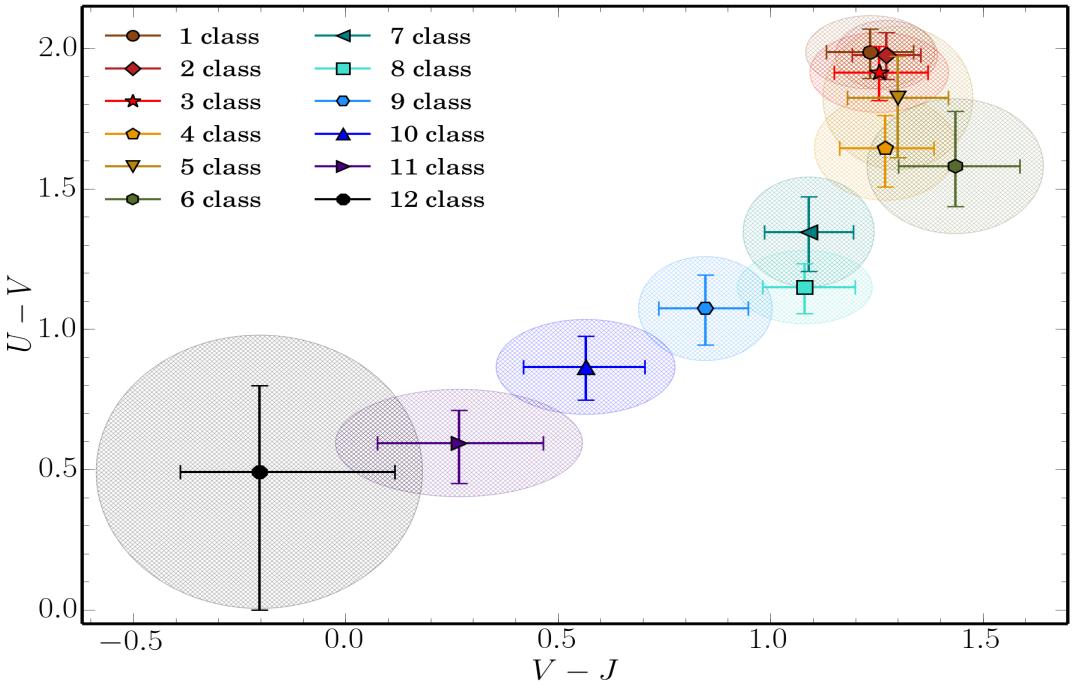


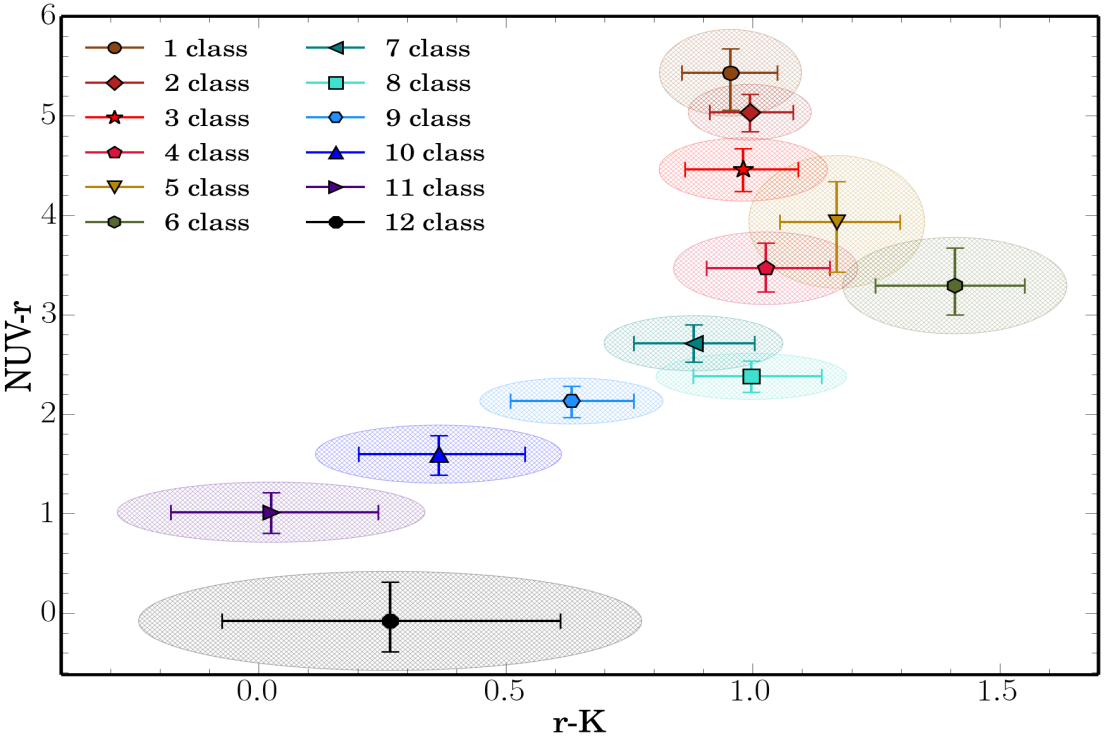


An optimal clustering model and number of groups choise based on a combination of statistical (BIC, AIC, ICL) and physical criteria.

Siudek et al. 2018,
arxiv: 1805.09904

Unsupervised classification





Multidimensional approach allows to achieve better separation, while on standard 2D color-color diagrams classes overlaps, e.g. red passive galaxies (classes 1, 2, and 3) are not distinguishable on UVJ diagram.

Siudek et al., arxiv: 1805.09904

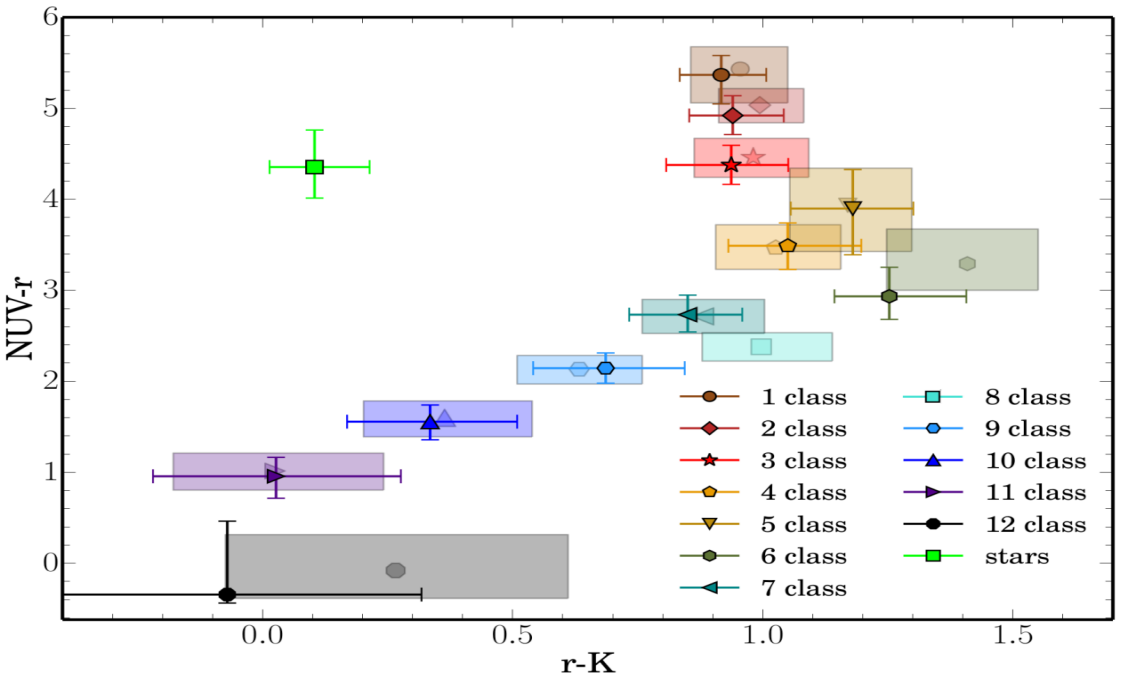
Unsupervised classification

Unsupervised classification of VIPERS galaxies based on their distribution in a multidimensional absolute magnitude space

12 dimensions: absolute magnitudes + zphot

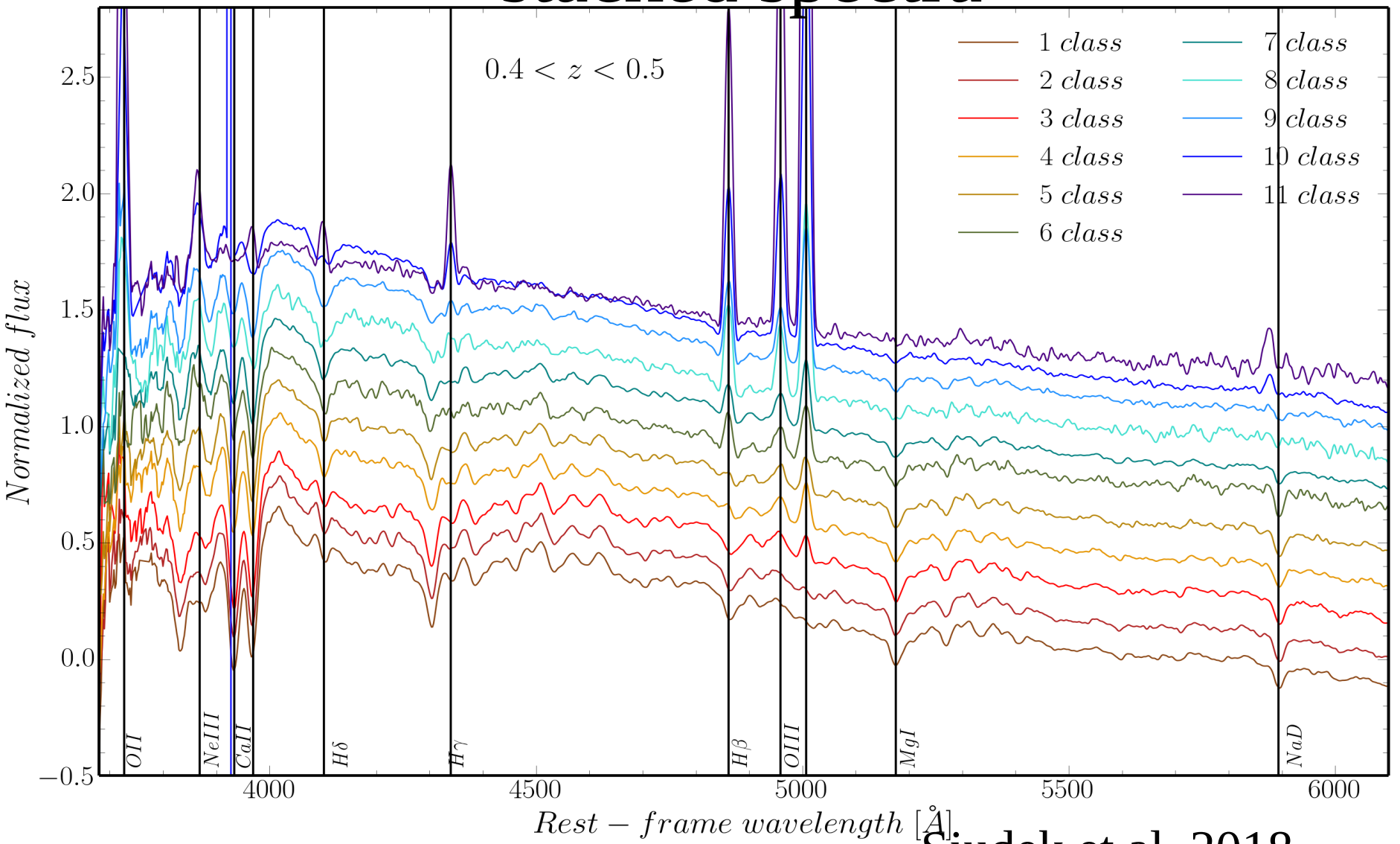
→ large photometric samples can be used to distinguish different galaxy classes at z > 0.5 with an accuracy provided so far only by spectroscopic data

→ we should be able to make such a fine classification on the LSST data as well



Siudek et al., arxiv: 1805.09905

Unsupervised classification stacked spectra

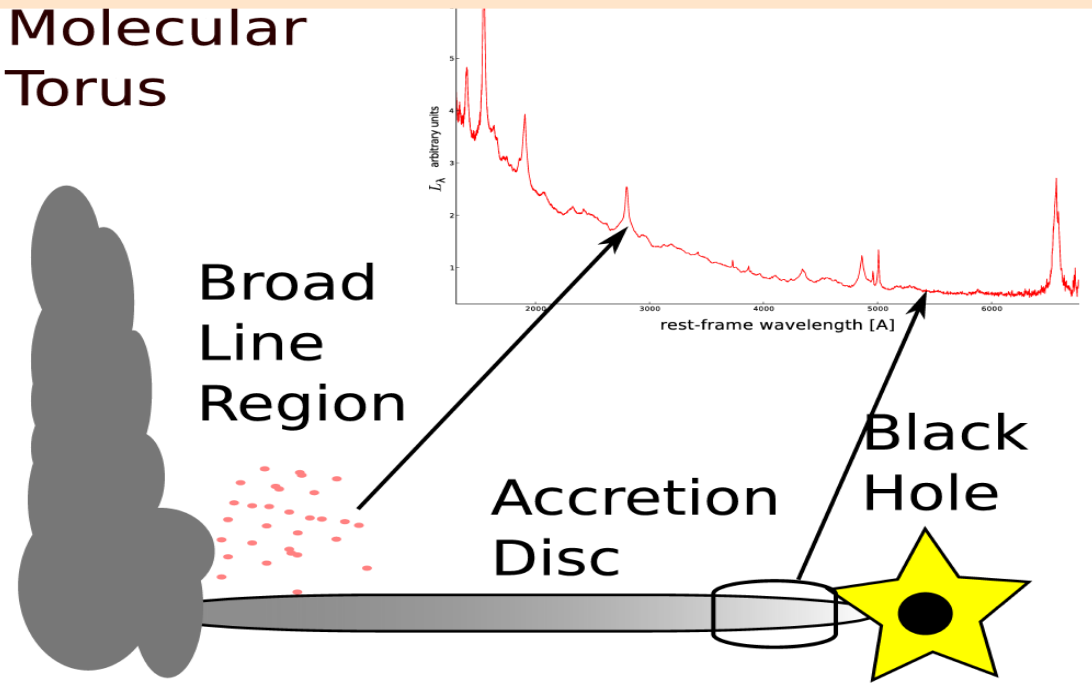


Siudek et al. 2018

Since we are at the AGN session – a few glimpses from to other Polish groups involved (or planning to) in the LSST

Quasar monitoring as a method of testing cosmology (Bozena Czerny)

Quasar variable continuum emission comes from the central parts of accretion disks. Distant Broad Line Region clouds reprocess part of the radiation and respond with a delay. The measurement of this delay opens a way to use quasars for cosmology.



Monitoring of the nearby AGN showed clearly that the measured delay is proportional to the square root of the **absolute monochromatic luminosity** of a quasar. The result is supported by theoretical BLR model of Czerny & Hryniewicz (2011).
Thus, measurement of the

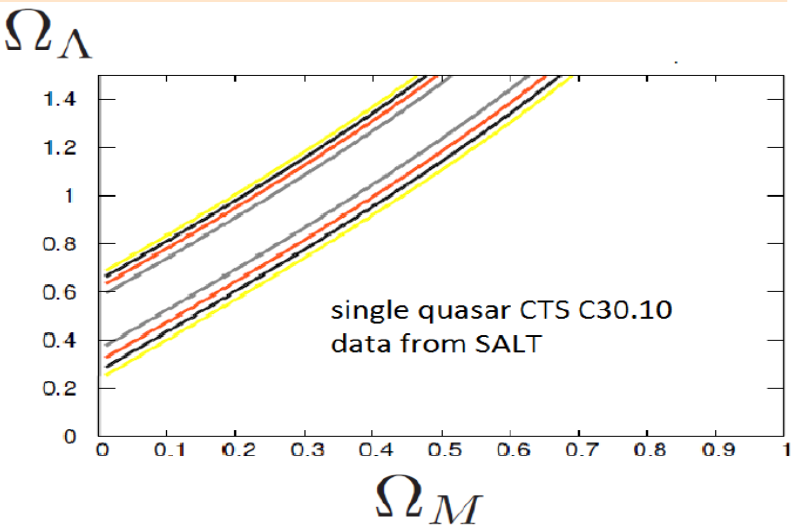
- redshift
- observed monochromatic flux
- time delay of a line with respect to the continuum

allows to obtain the redshift and the corresponding luminosity distance.

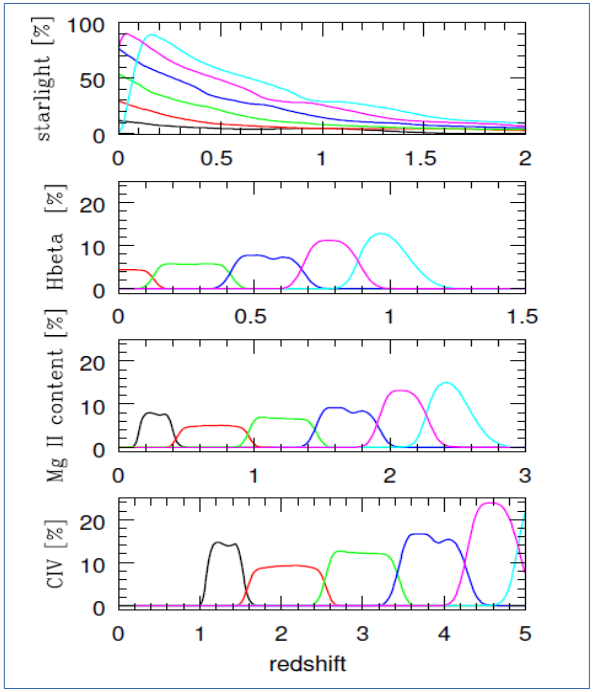
The relation between the redshift and luminosity distance depends on cosmology, thus measuring this relation we can obtain cosmological constraints.

Quasar monitoring as a method of testing cosmology

Spectroscopic monitoring even of a single quasar gives relatively good constraints. Here we show an example of a single quasar monitored in Mg II line with 11-m SALT telescope.



However, spectroscopic monitoring requires a lot of observing time. LSST willl monitor many quasars but in a photometric mode. Still, multi-channel photometry can replace photometry (e.g. Chelouche et al. 2014).



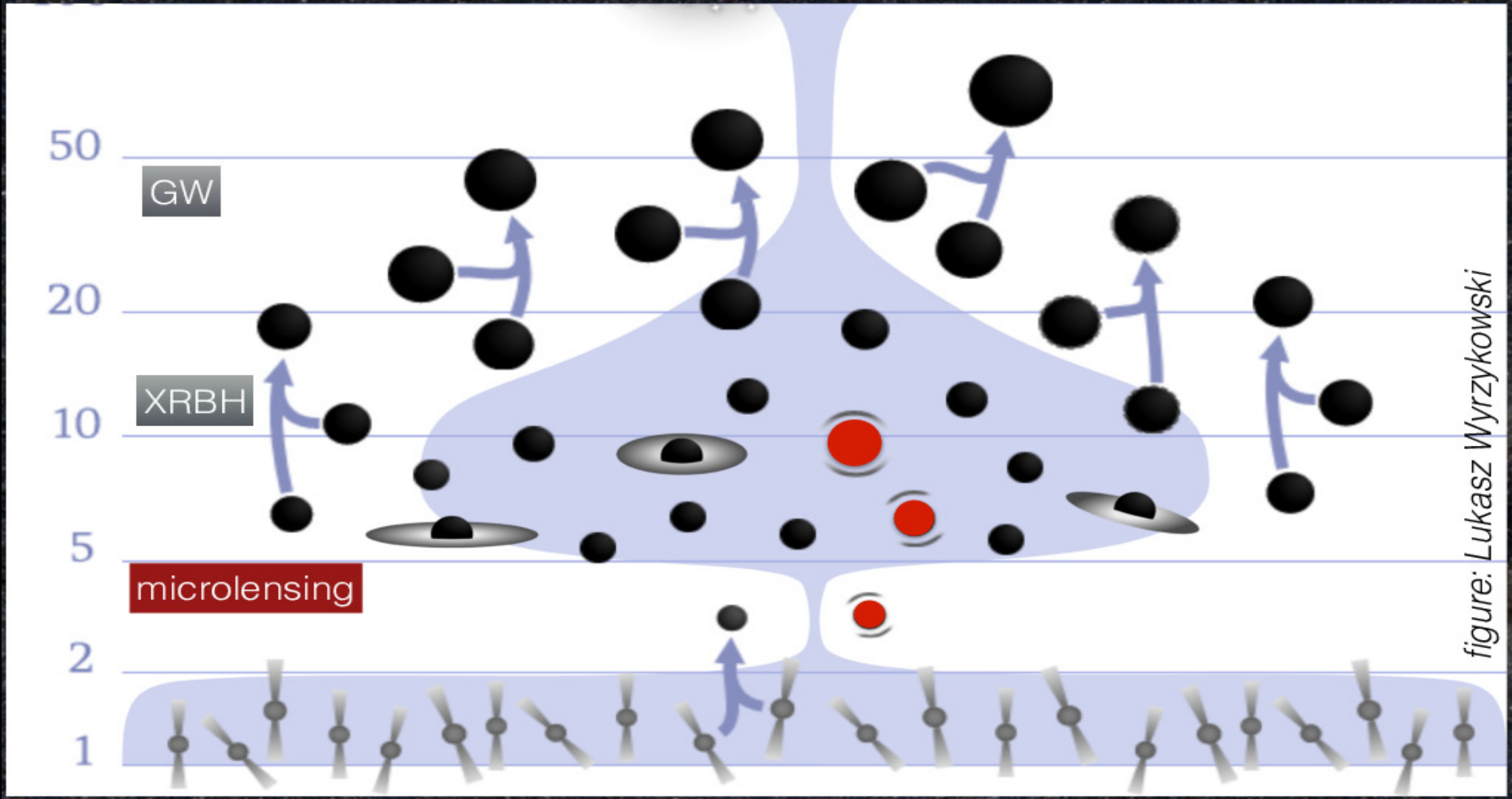
The contribution of the starlight and important lines (Hbeta, Mg II, CIV) to LSST channels is significant, so the time delay measurements can be performed for sources with photometry better than 0.01 mag. Long monitoring is essential – time delays are hundreds of days for distant objects.

GALACTIC BLACK HOLES WITH THE LSST

Lukasz Wyrzykowski
Warsaw University Astronomical Observatory
Poland

figure: Łukasz Wyrzykowski

POPULATIONS OF KNOWN BLACK HOLES AND NEUTRON STARS



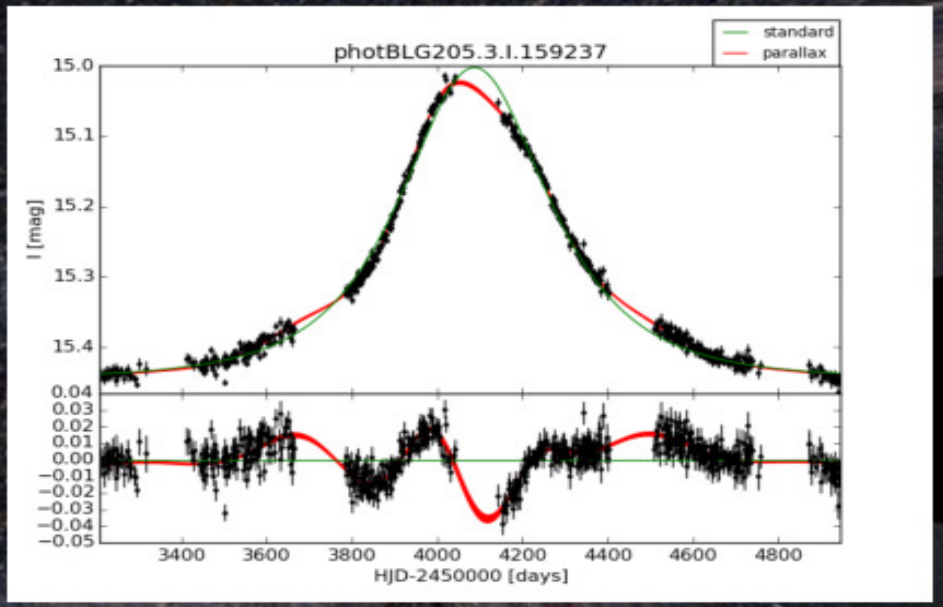
Microlensing probes all range of masses!

Łukasz Wyrzykowski

HOW TO RECOGNISE A BLACK HOLE LENS?

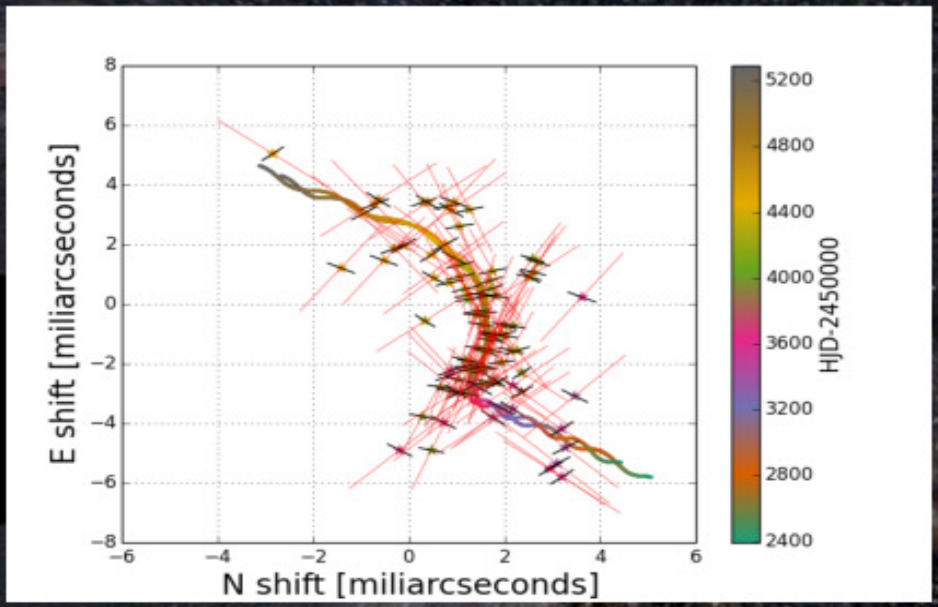
example: OGLE3-ULENS-PAR-02 ~9Mo BH candidate

OGLE photometry and parallax model



Wyrzykowski+2016

simulated Gaia astrometry



Rybicki+2018

Photometry + astrometry = mass, distance, luminosity

Events last from months to years.
~30d sampling of the LSST in the Galactic Plane enough.
Astrometry at ~1mas required.

Łukasz Wyrzykowski

MICROLENSING WITH GAIA SINCE 2016

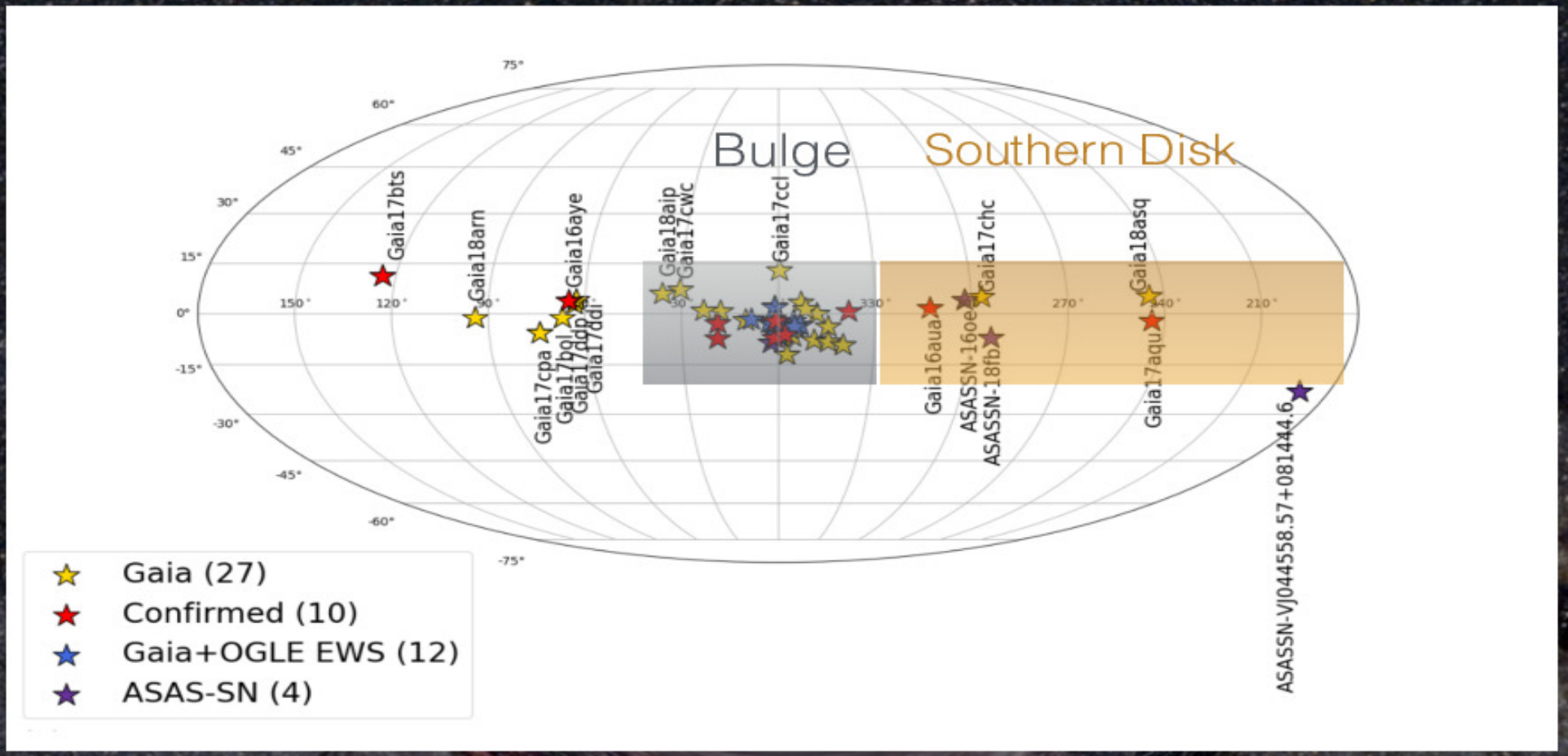


figure: K.Kruszynska

Galactic Plane is the best place to look for black hole lenses!

BH lensing 30 times more likely than in the Galactic Bulge due small star-star lensing probability.

Łukasz Wyrzykowski

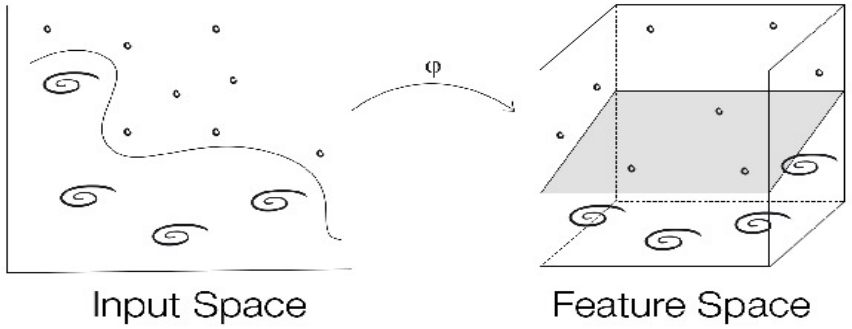
Summary

- The existing datasets provide a very good training ground for future much larger sky surveys, in particular LSST
- Looking for unknown with OCSVM
- Unsupervised galaxy classification should be possible based on photo-zs
- But: importance of the NIR follow-up

Searching for QSO in AllWISE data with SVM

SVM Principle

- In most cases we can't find a suitable separation in an input space
- Solution: mapping the input space into a higher dimension feature space and searching the optimal hyperplane



Fuzzy SVM

Classification improvement: applying fuzzy membership of th data based on measurement uncertainty or distance from the class center.

Difficulties of AllWISExSDSS selection effects

Previous attempts

- (Kurcz et al. 16) SVM-based classification of the AllWISE data gave different distribution than that of SDSS QSO.

Solution

Iterative training: adding estimated probabilities as an input features and repeat training on the new sample.

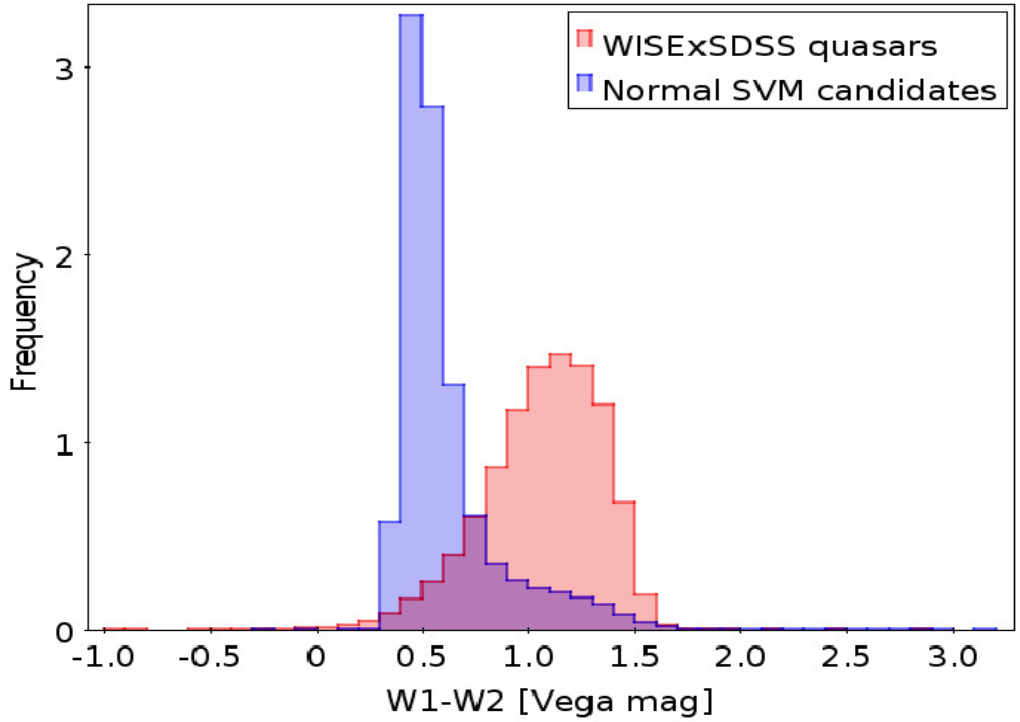


Figure: Classification from Kurcz et al. 2016

Result of the iterative training

Completness: 94% $\rightarrow$ 80%, purity: 83% $\rightarrow$ 97%

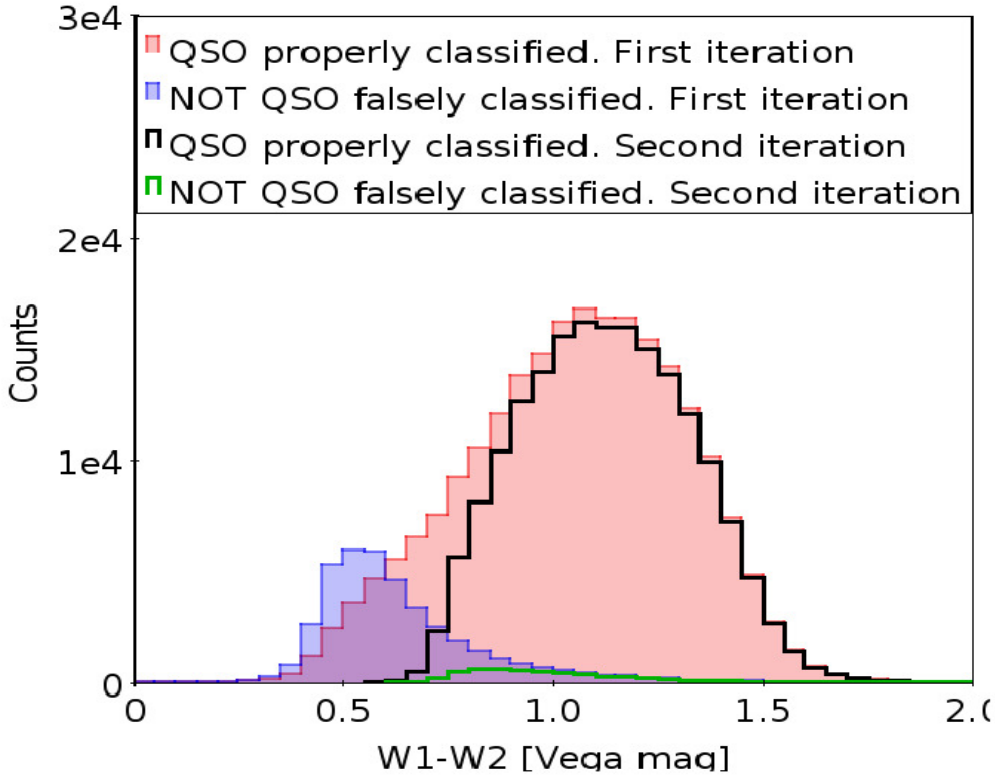


Figure: Training

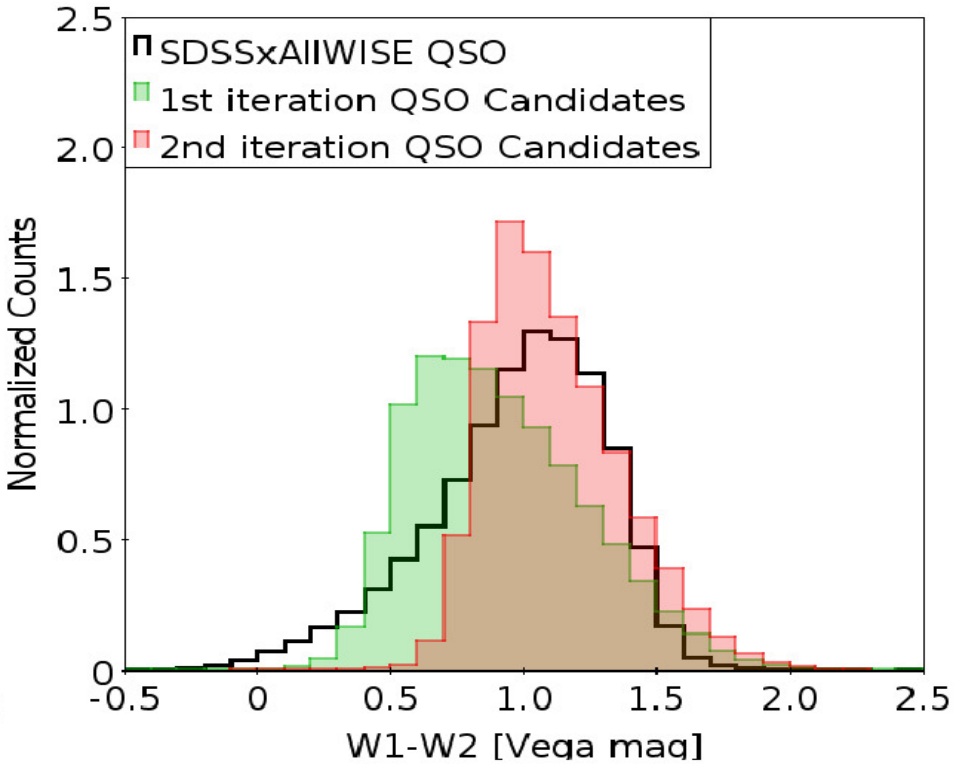


Figure: Generalization

Initial results

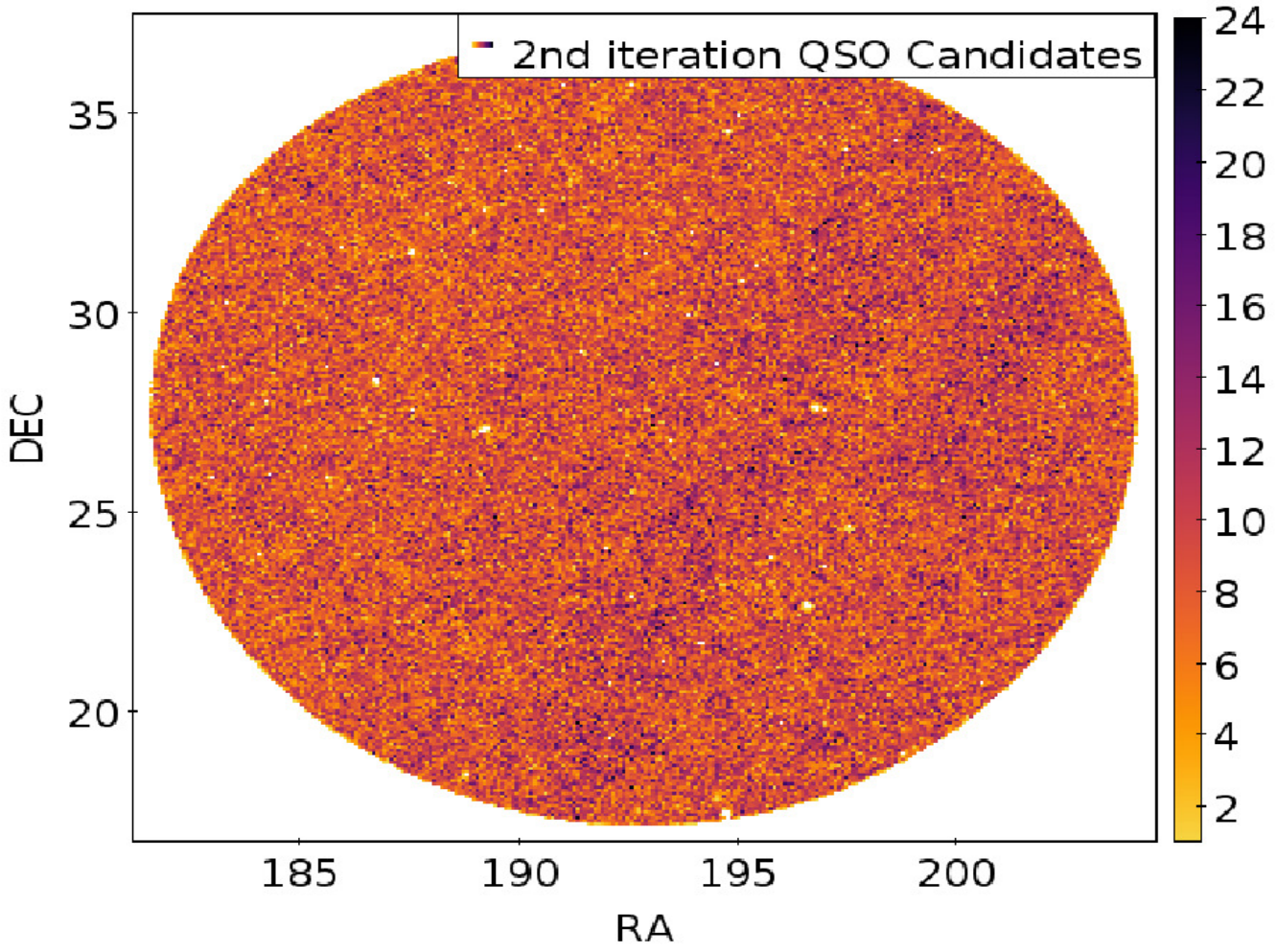


Figure: 10^6 QSO candidates in the north galactic pole area