NEW DATA – NEW UNKNOWNS

(or how our notion of astronomical discovery is changing right now)

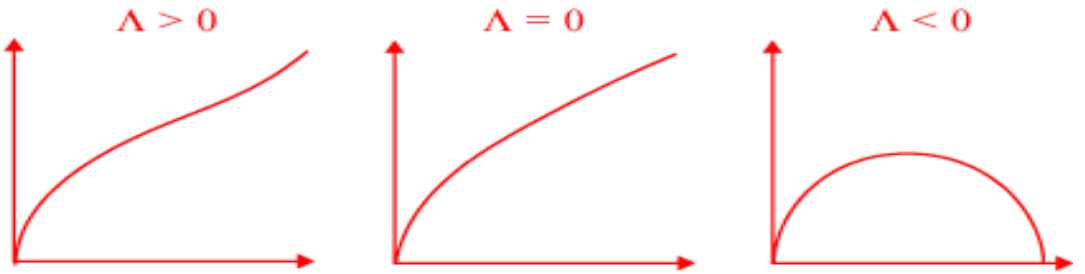
Agnieszka Pollo

NCBJ + OA UJ, Poland

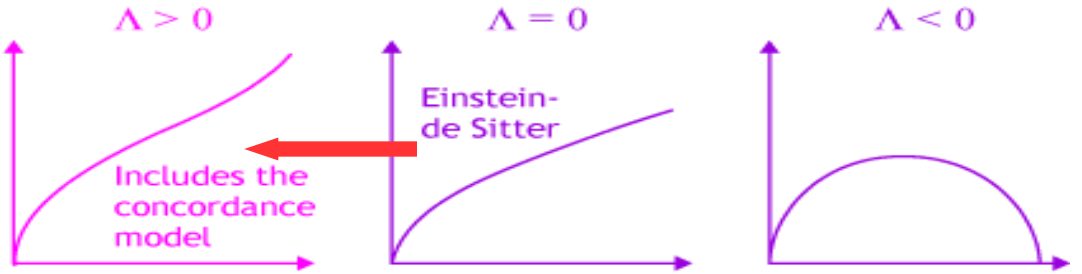
Physics and astronomy and areas of great discovery potential also today.

During our lifetimes we already moved to live in another Universe...

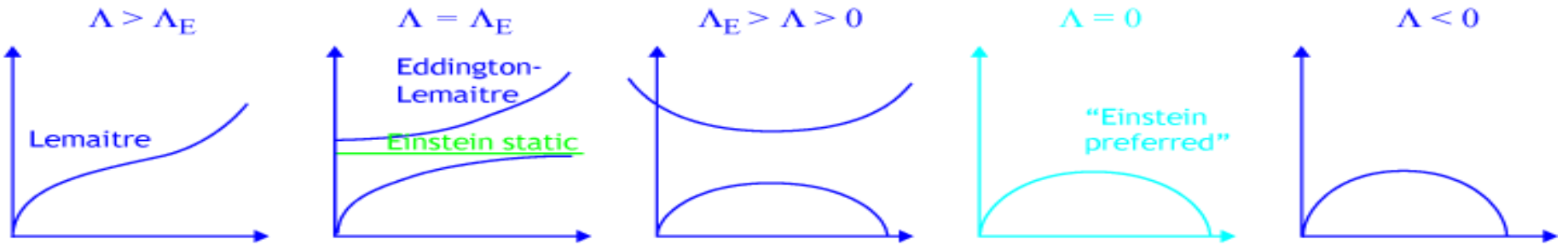
Negative Curvature Models: $k = -1,\ \Omega_k > 0$ (infinite space)



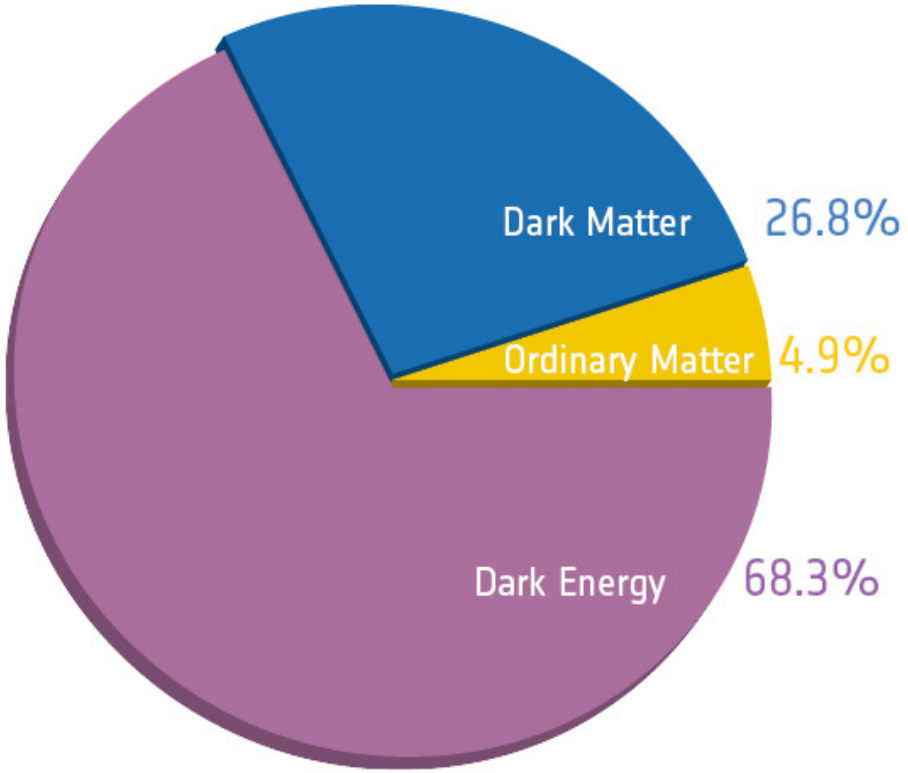
Flat Models: $k = 0,\ \Omega_k = 0$ (infinite space)



Positive Curvature Models: $k = 1,\ \Omega_k < 0,$ (finite space)



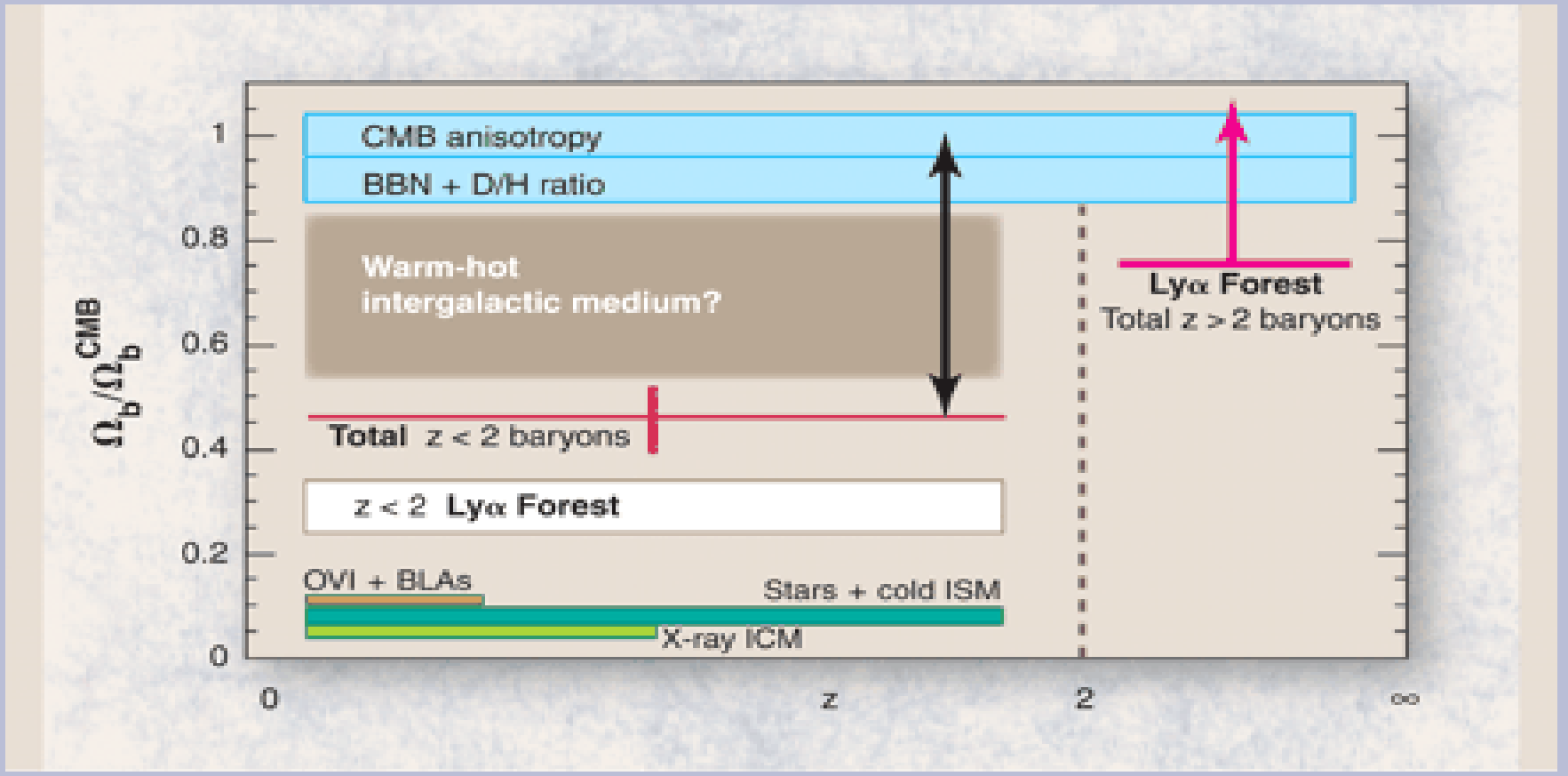
Classification of Friedmann Models



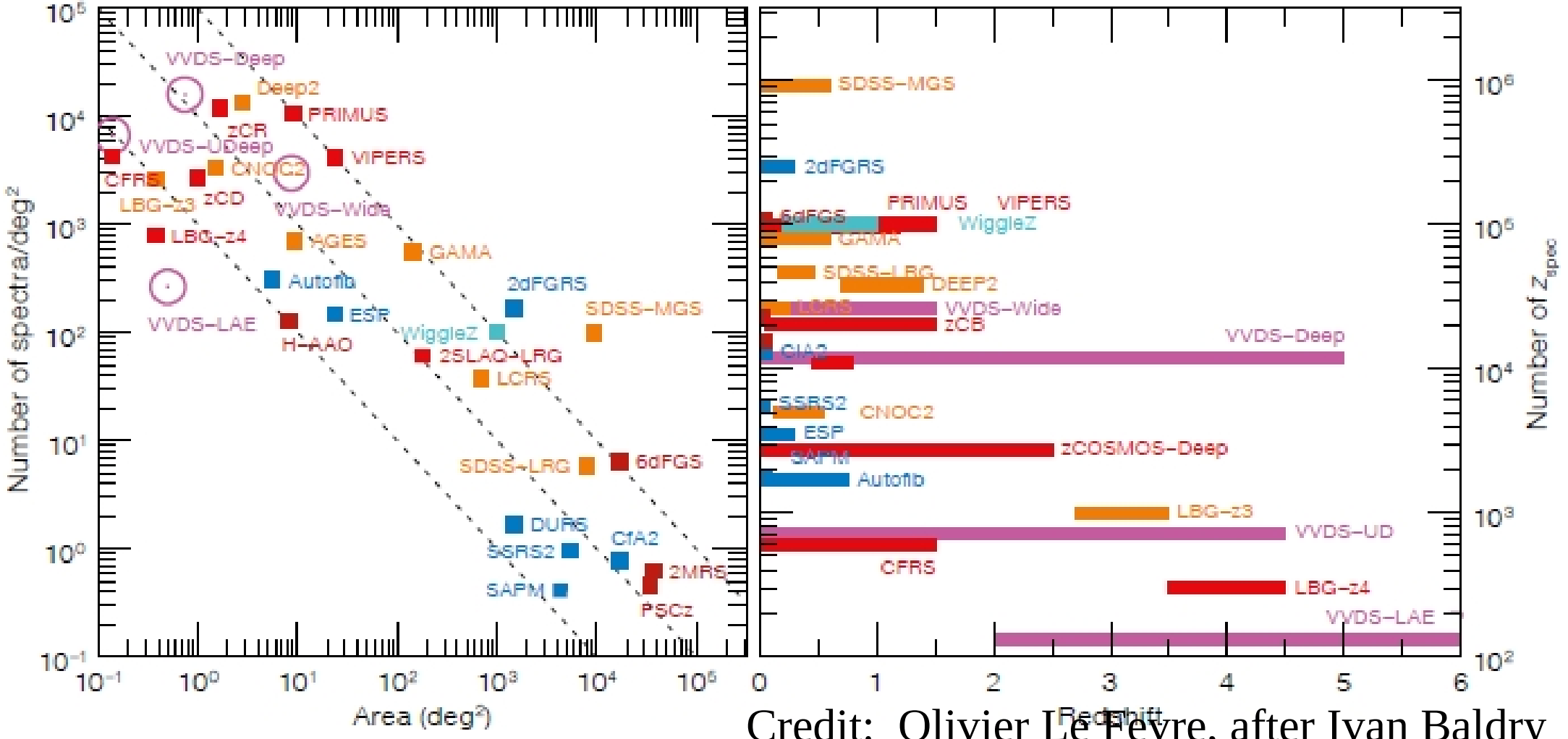
However…

- Almost 27% of the total matter-energy budget of this new Universe is Dark Matter and we do not know what it is
- More than 68% is Dark Energy whose nature we understand even less
- Less that 5% is supposed to be “baryonic matter” of the type we know and understand but...

…actually, in the present day Universe we have a problem
missing baryons;
stars and ISM contain only 10% of all baryons, and only 40% of all expected baryonic matter can be detected at all

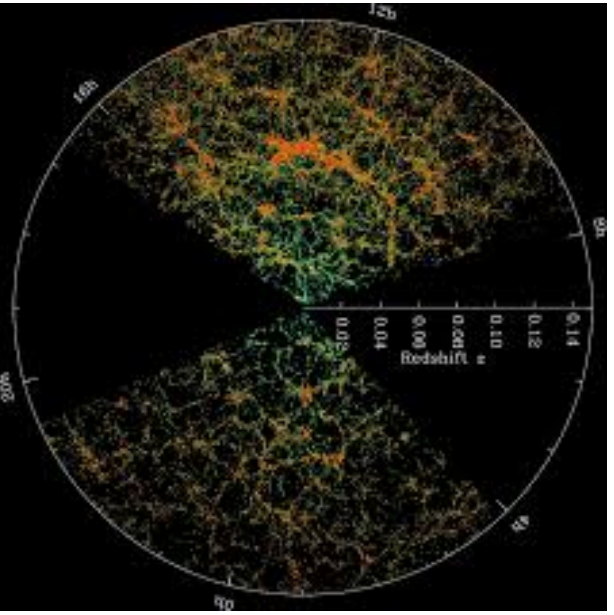


In the same time, our main source of information about the Universe are galaxy/QSO/star catalogs – showing at most 10% of less than 5% of the actual content of the Universe...

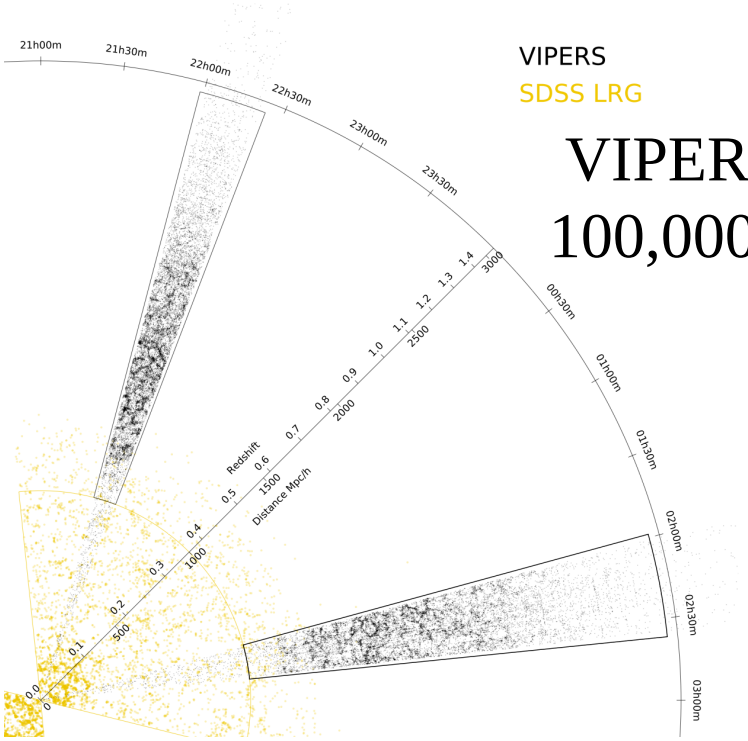


Credit: Olivier Le Fèvre, after Ivan Baldry (Messenger 2014)

Present day state of art spectroscopic galaxy surveys



SDSS: z ~ 0;
1.3 mln galaxies



VIPERS: z ~ 1;
100,000 galaxies

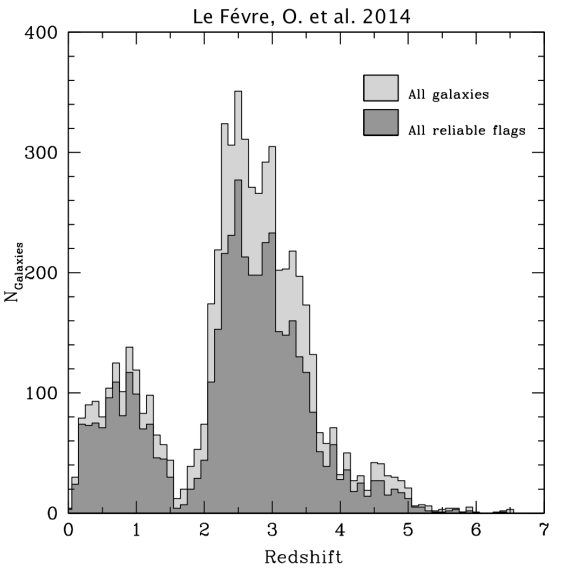
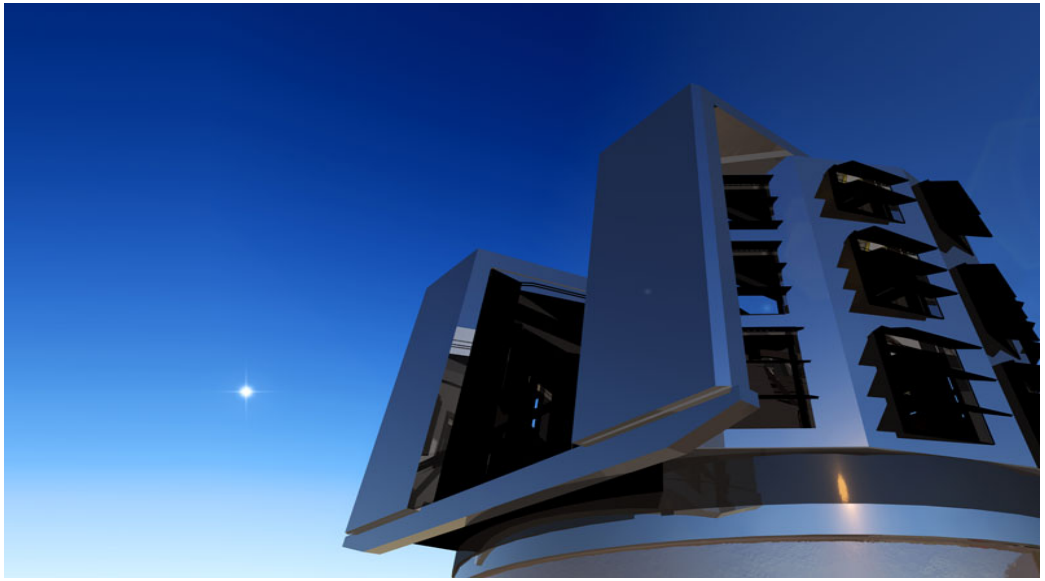


Fig. 16. The current redshift distribution from 6057 galaxies already processed in the VUDS survey, for all objects with a redshift measurement (light grey) and all objects with a > 80% reliable redshift measurement (flags 2, 3, 4, and 9; dark grey).

More and more difficult
to handle but discoveries
can be still done
in a more or less traditional
(which does not say manual!)
Way...

VUDS: z ~ 3;
10,000 galaxies

But now, we are expecting an avalanche of new data



https://www.lsst.org/lsst

- SDSS: ~115 TB in total
- Zwicky Transient Facility (ZTF; start 2017)
- 1 PB of image data ~1 billion objects
- Large Synoptic Survey Telescope (LSST; first light ~2020); 30 TB PER NIGHT
- The Square Kilometer Array (SKA) ~4.6 Zettabytes

Old methods do not suffice - need of automated tools to detect, characterize and classify gathered information

SKA; South Africa

Credit: A. Solarz 2018

Wide-field Infrared Survey Explorer (WISE)

Presently the largest and the deepest → perfect for testing efficient methods of fast and effective search for discoveries

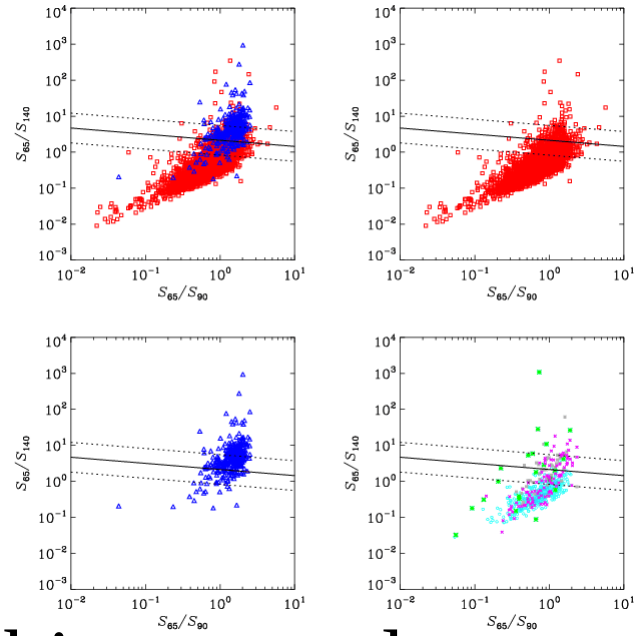
- All-Sky survey in IR
- Detected over 747 mln sources
- (15 PB of data; tables + images)
- Publicly available (positions, photometry in 4 bands 3.6-22 um)
- Low angular resolution (~6”)
- No redshift information so far

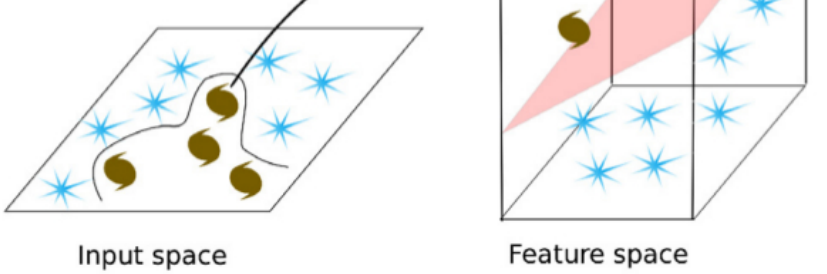
Credit: A. Solarz 2018

Popular machine learning tool for supervised classification: Suport Vector Machines

Basic idea: to move from classifications based on very limited number of parameters (like color-color plots or line-to-line ratio or sth. similar) to the feature space built from a larger number of parameters

Pollo et al. 2010

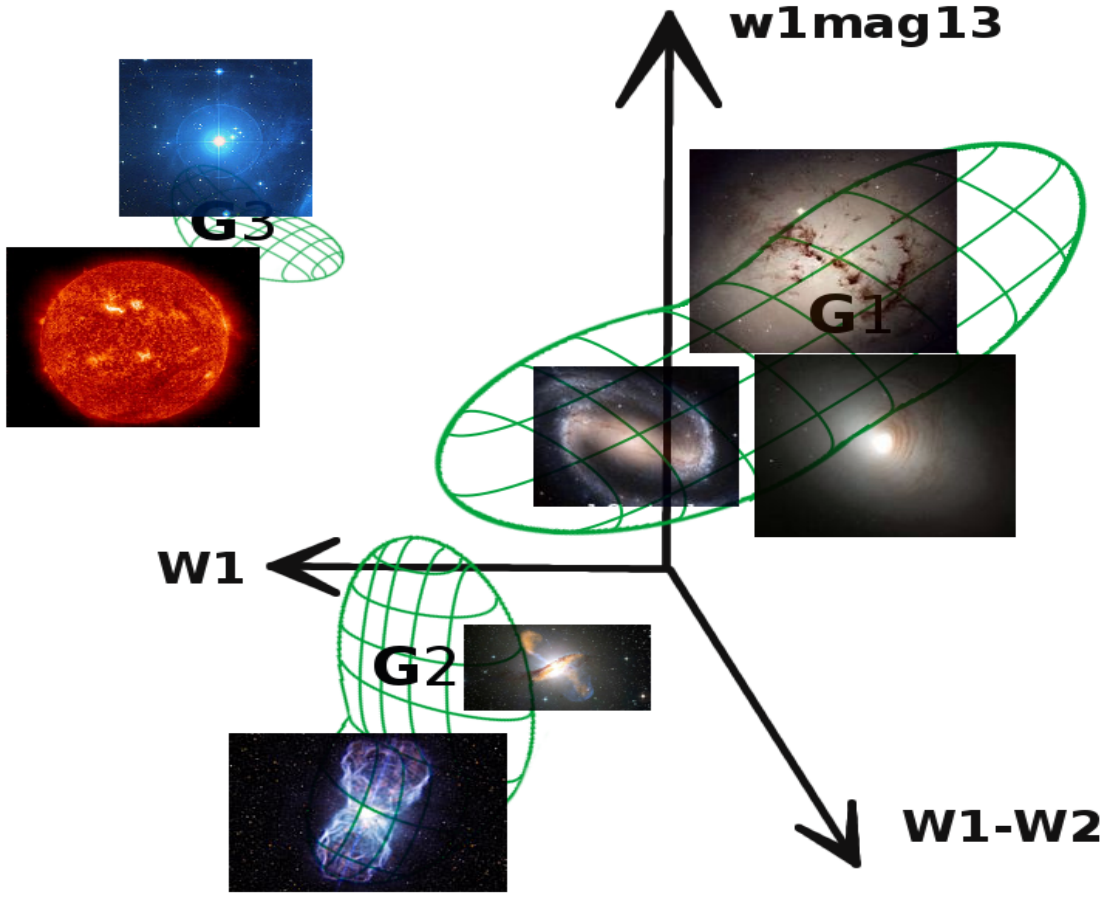




Malek et al. 2013

Objects poorly separated in the two parameter space can get well separated in a multi parameter space, and the problem is easier to linearize.

Machine Learning based **novel source detection**

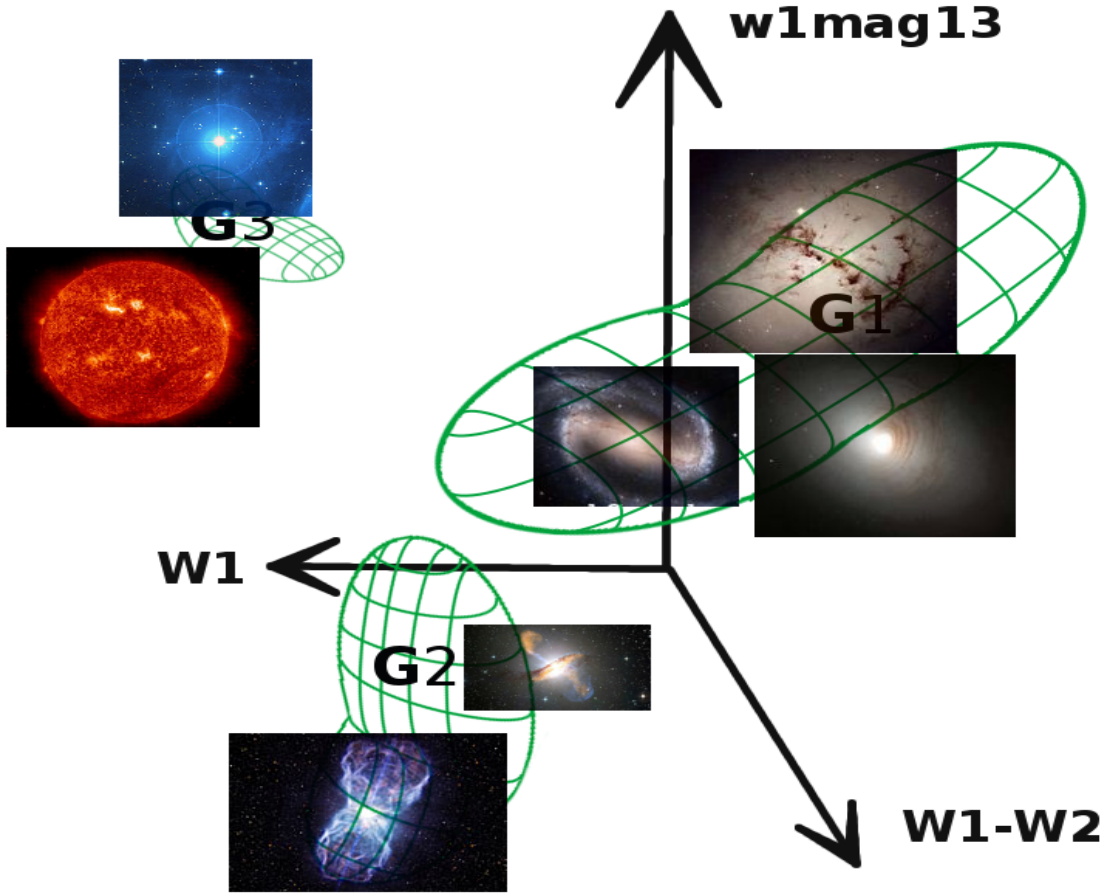


A. Solarz et al. 2017

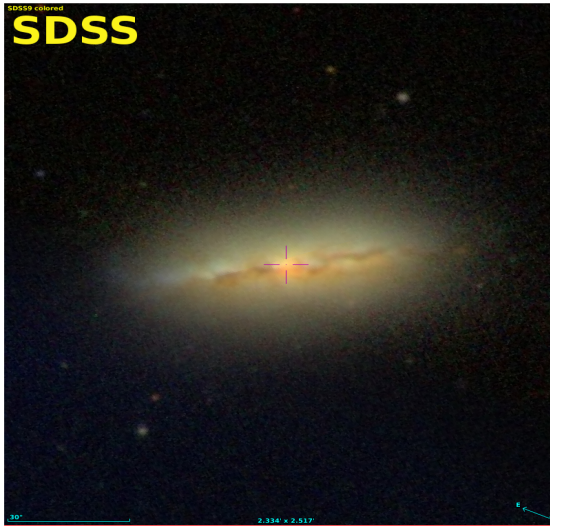
In each large dataset (especially of a type never done before), we can expect some sources of previously unknown properties – possibly a completely new class of sources.
An idea to search for them: classify a dataset into two main classes: “known” and “out of parameter space covered by known sources”

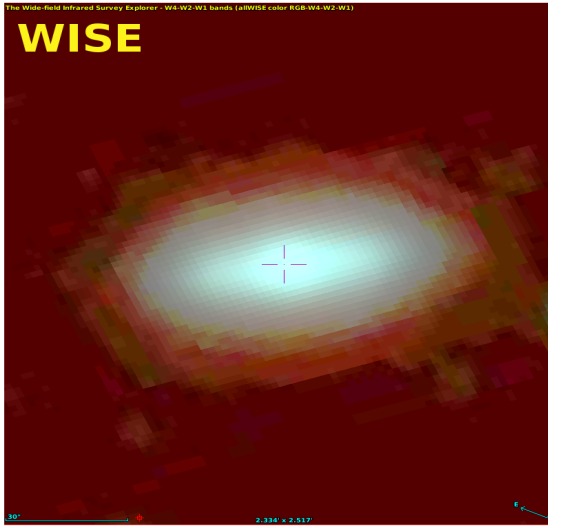
WISE: first step towards ML **novel source detection**

Training set (all what we expect):
AllWISE x SDSS (α,δ) with spectro-z (secure)

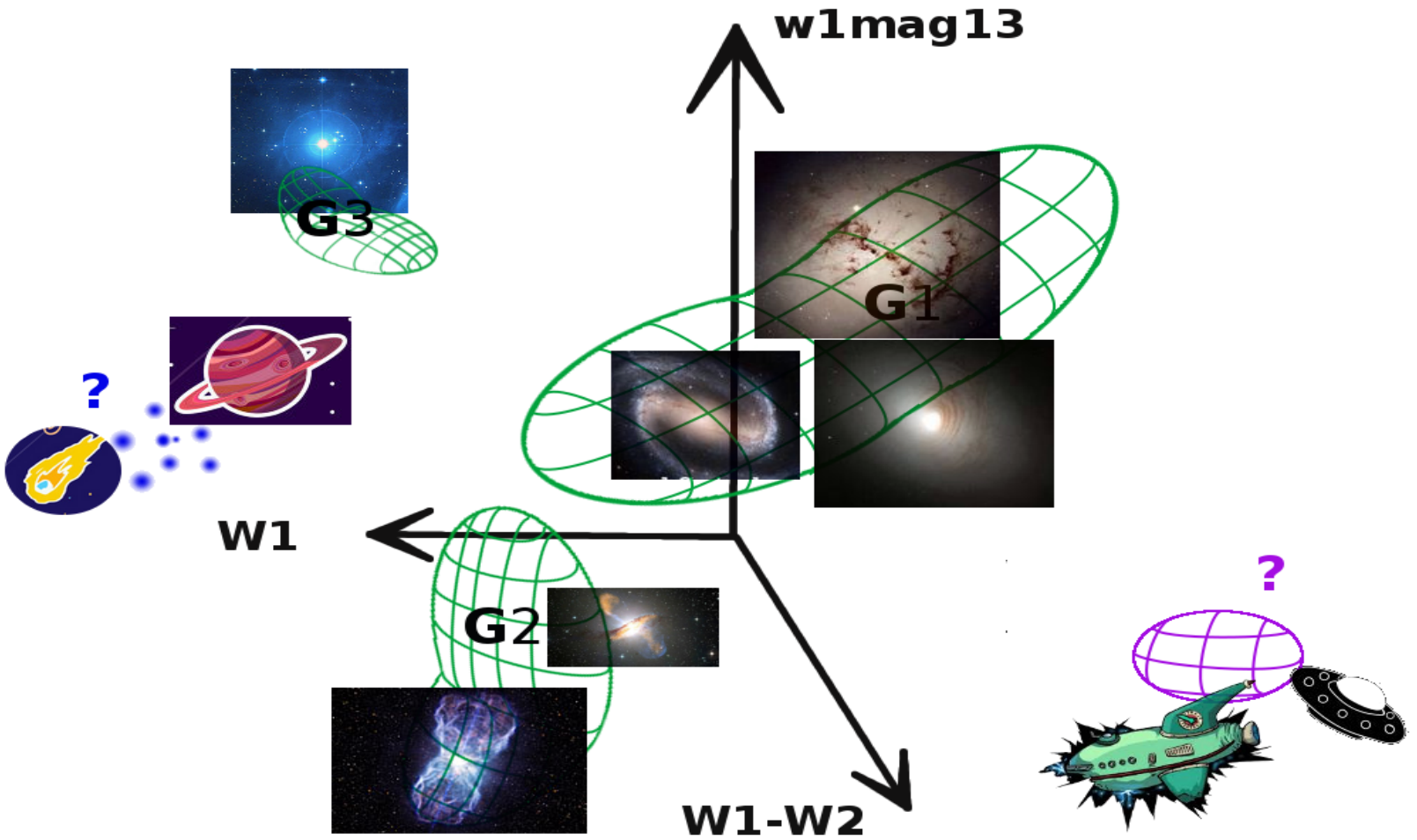


A. Solarz et al. 2017



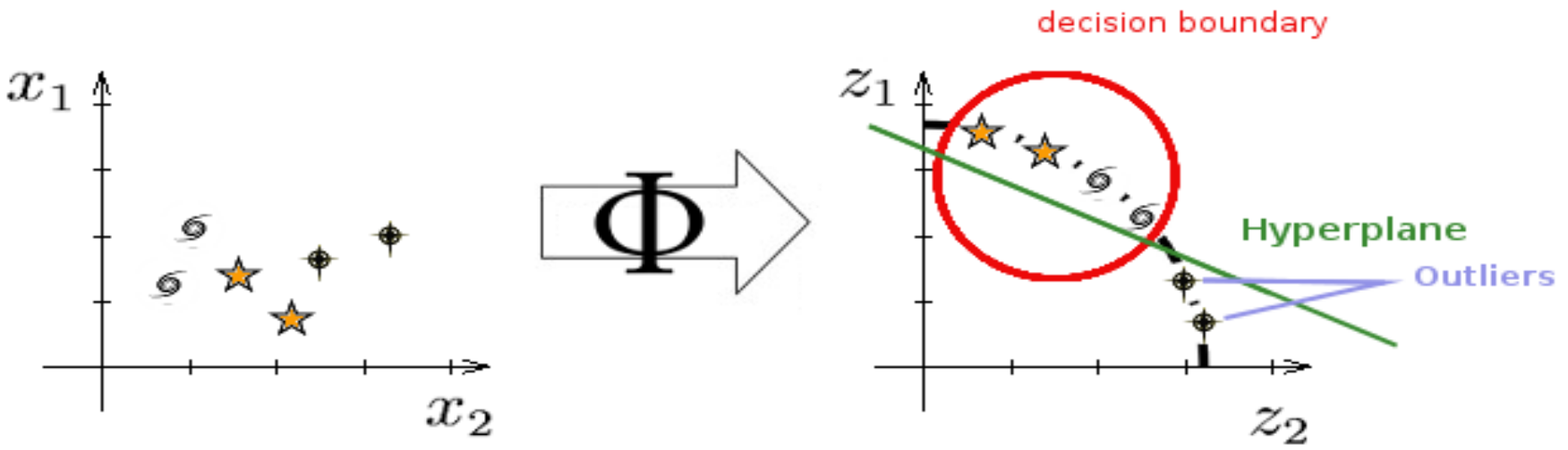


WISE: accounting for unknown unknowns



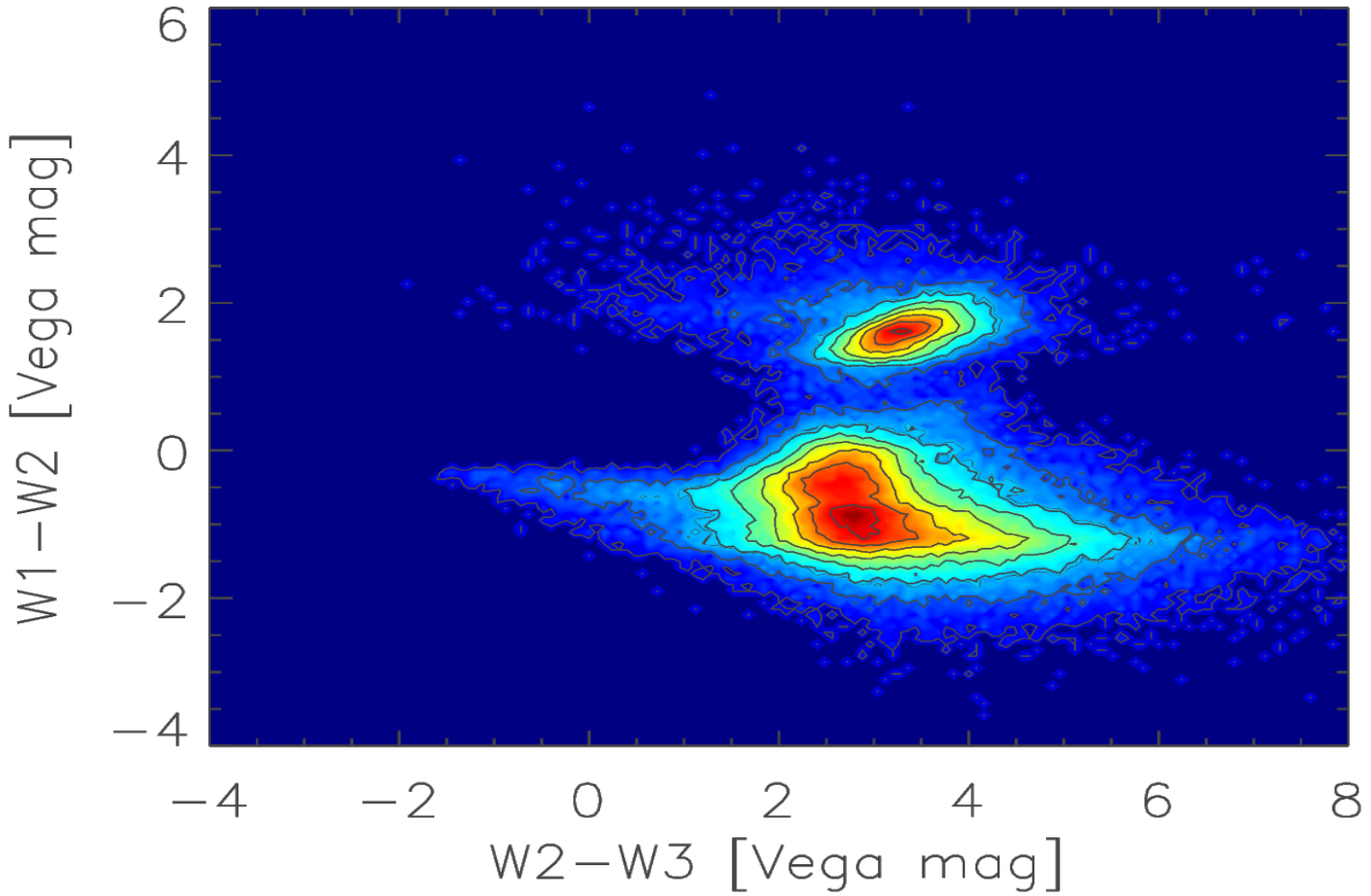
A. Solarz et al. 2017

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:

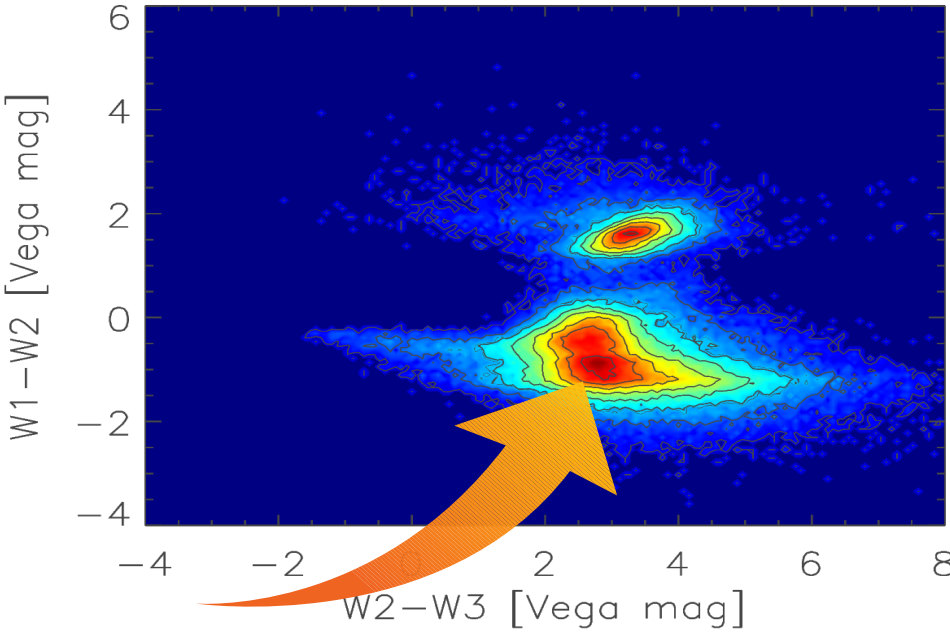
~650,000 anomalous sources

What are they?

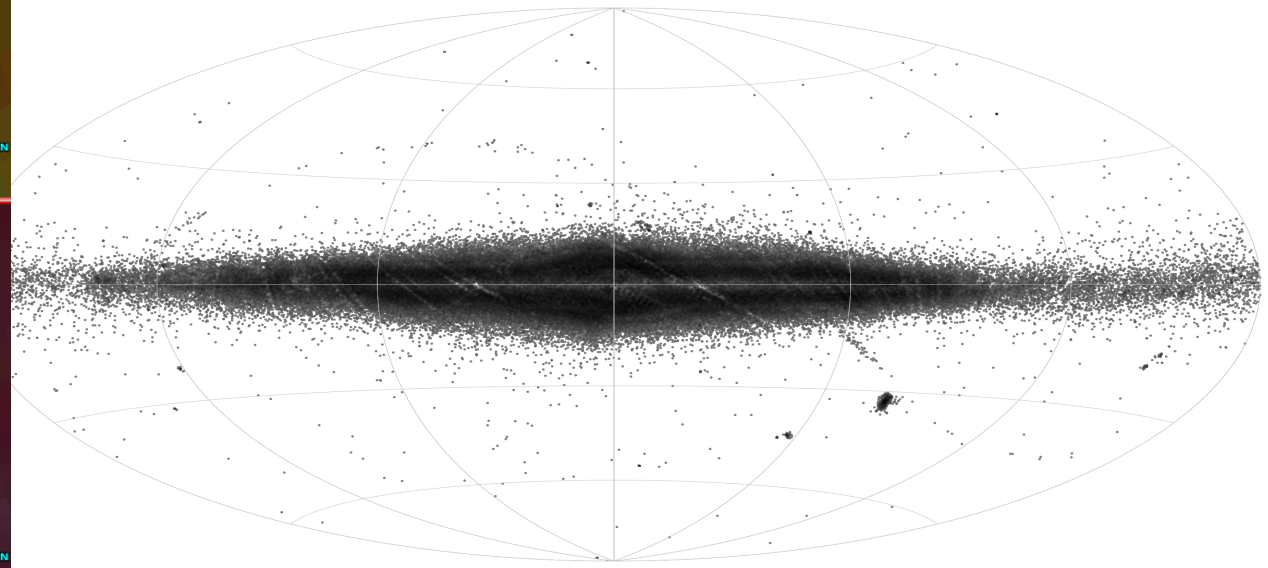
A. Solarz et al. 2017

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

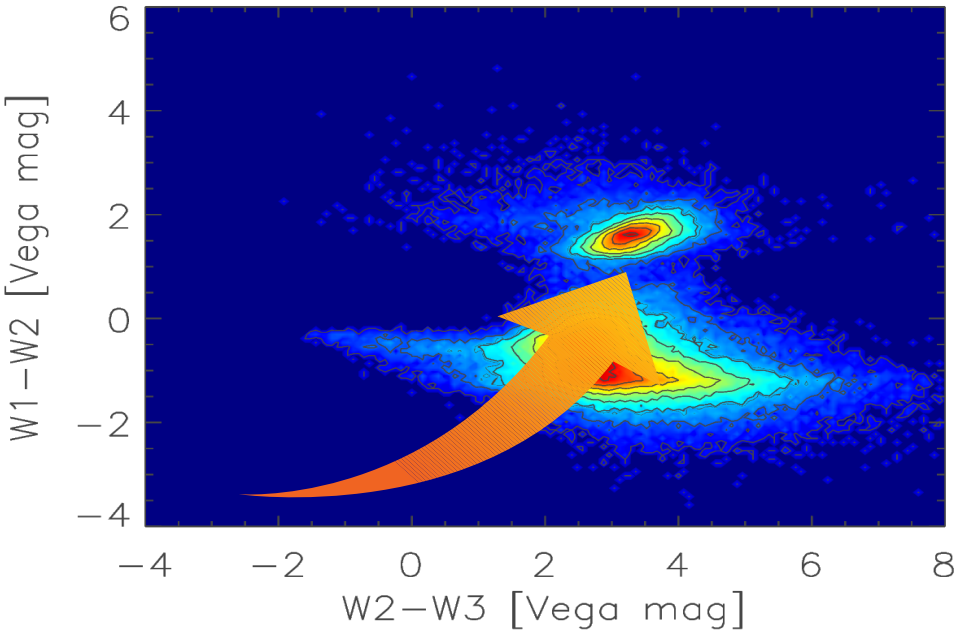


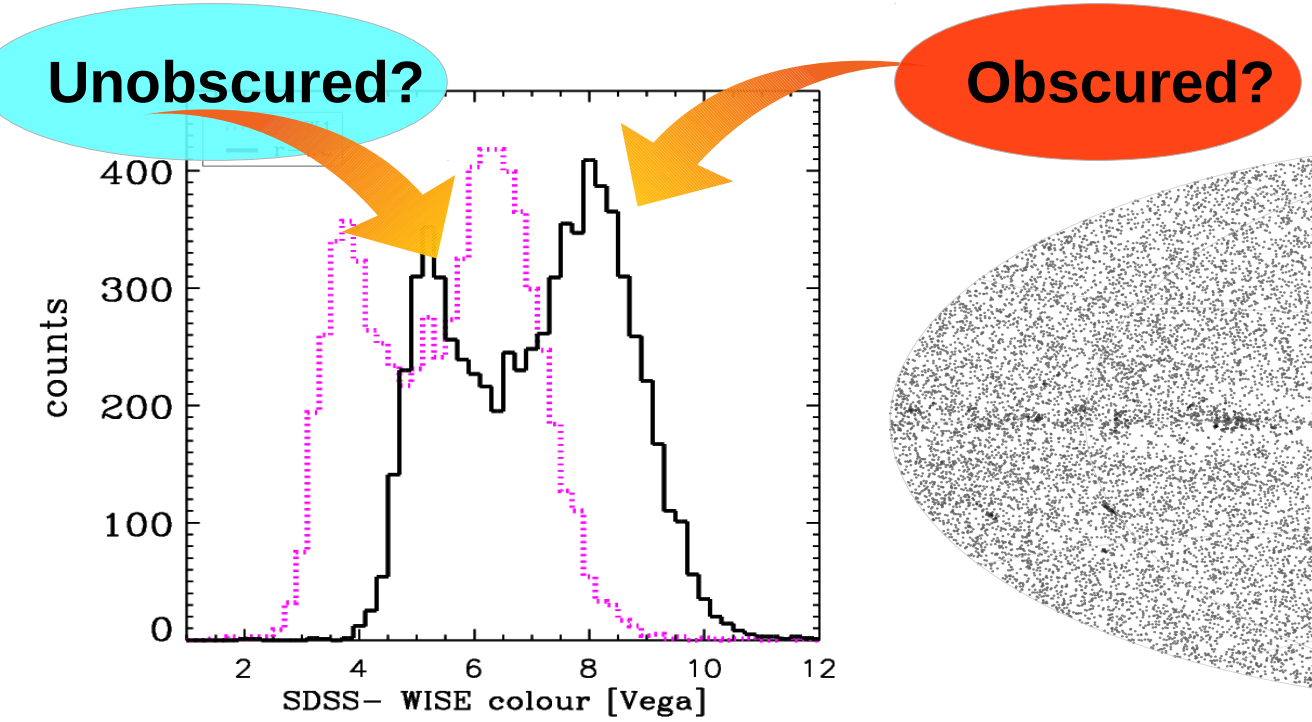
A. Solarz et al. 2017



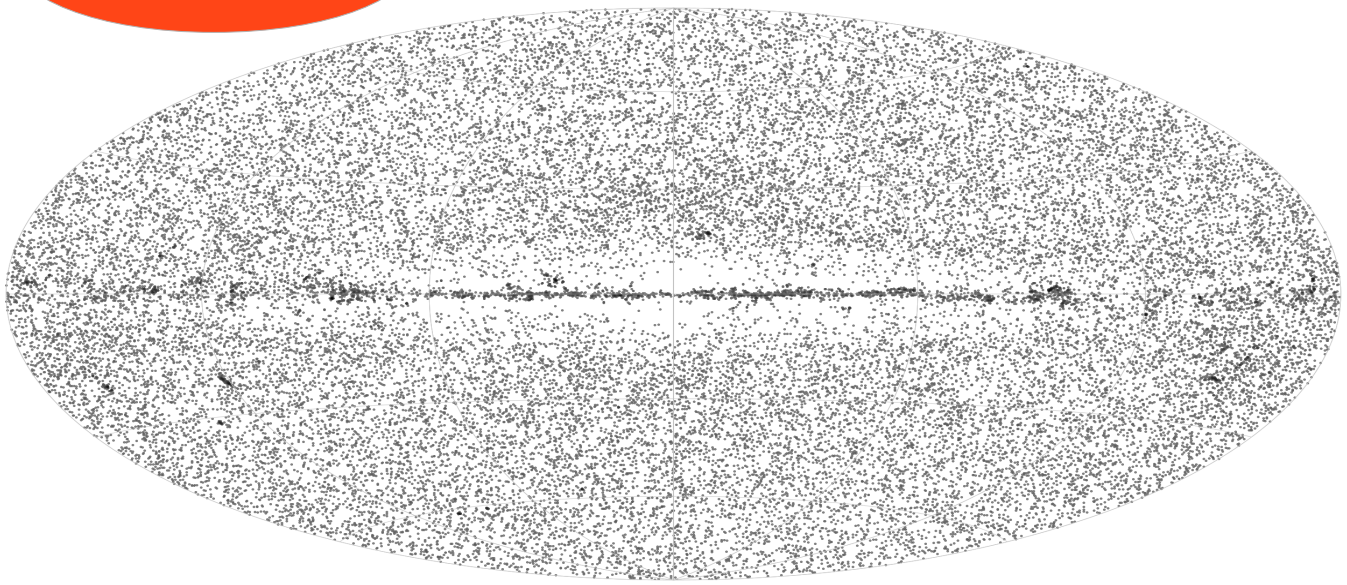
AGN candidates?

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

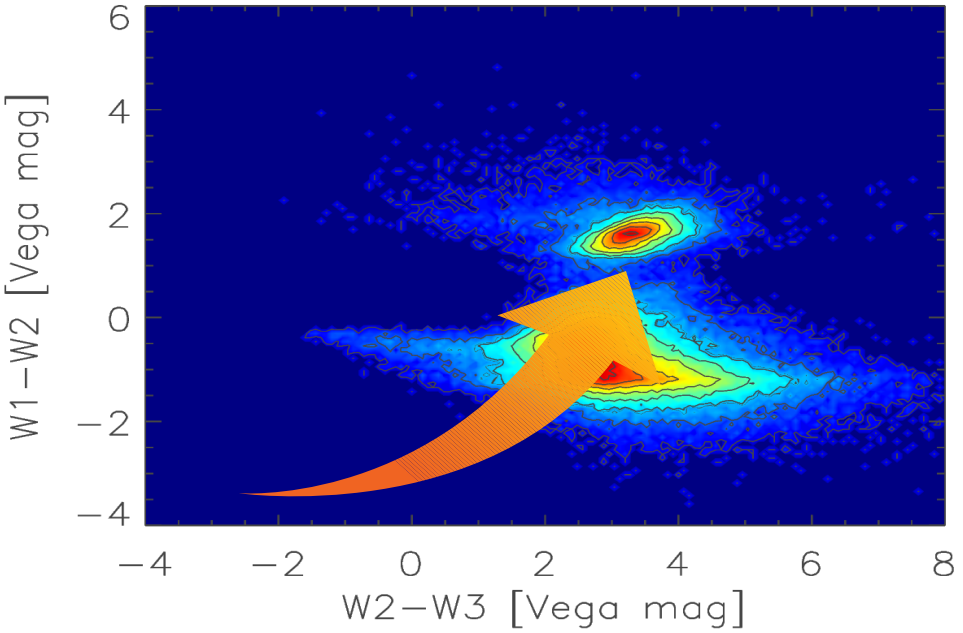
Spectral Energy Distribution with CIGALE

RESULTS:

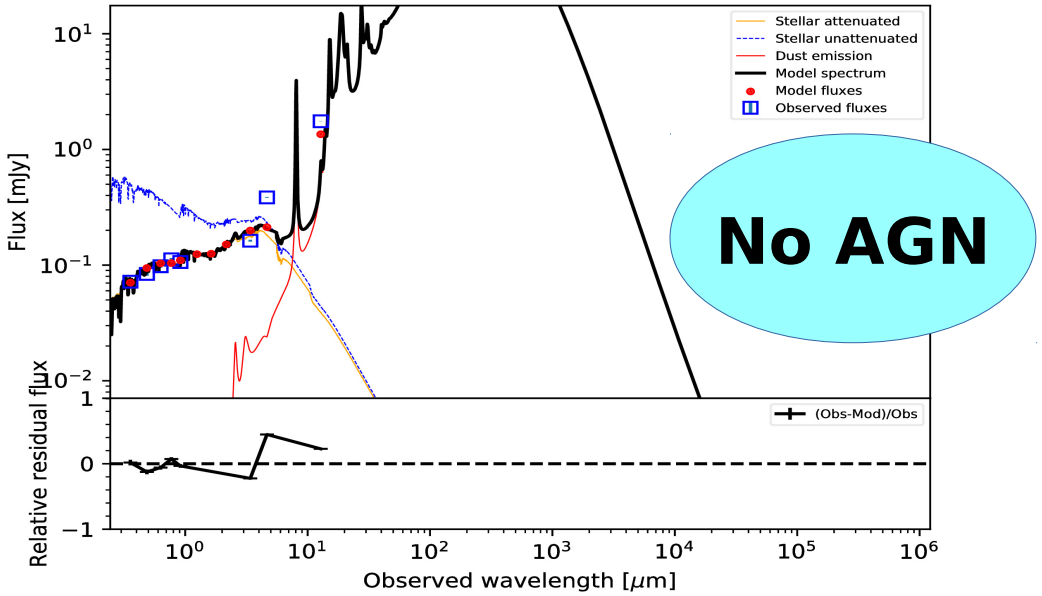
AGN component necessary to explain IR fluxes

85% (Ultra)Luminous Infrared Galaxies

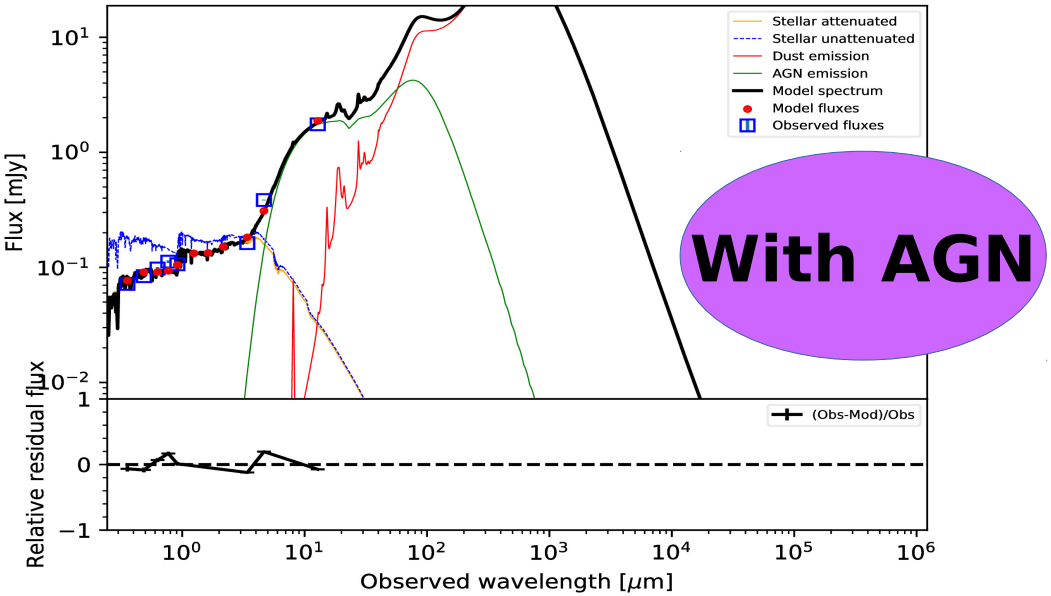
Spectroscopic surveys ongoing (proposals: GEMINI-FLAMINGOS2; EFOSC2/SOFI @ La Silla observatory: time allocated!!!, ALMA next in line)



Best model for J085347.87+144858.8 at z = 1.442. Reduced χ^2=4.66



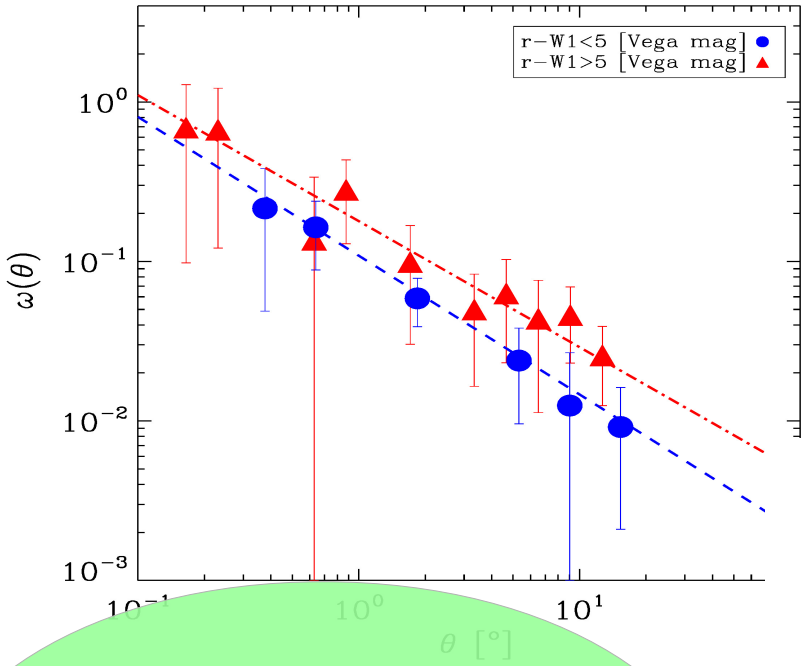
Best model for J085347.87+144858.8 at z = 1.442. Reduced χ^2=1.46



Solarz et al. 2017;

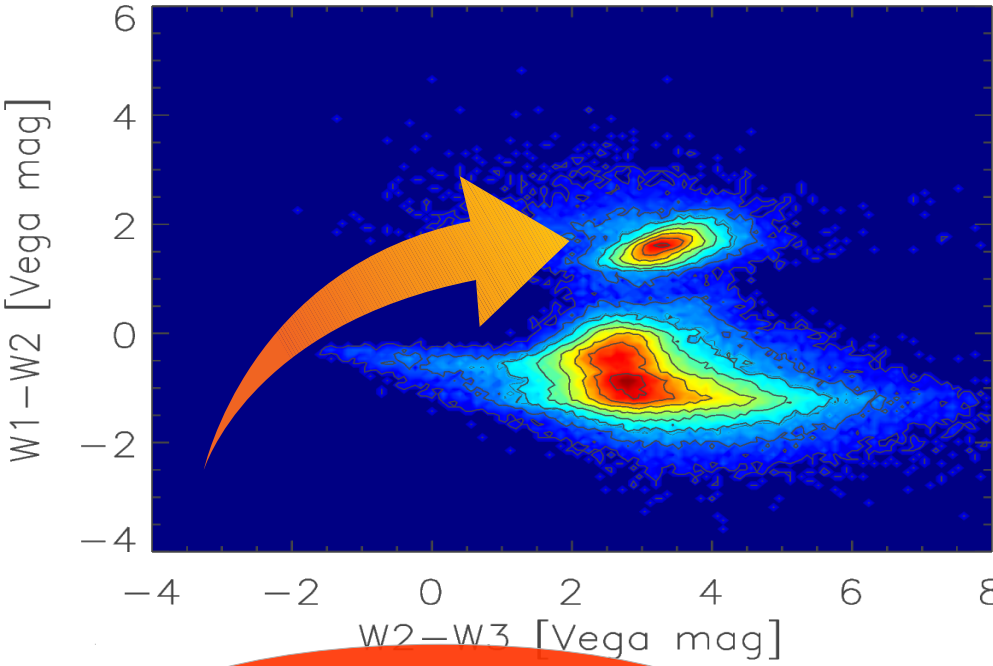
https://cigale.lam.fr/

AGN candidates for cosmology

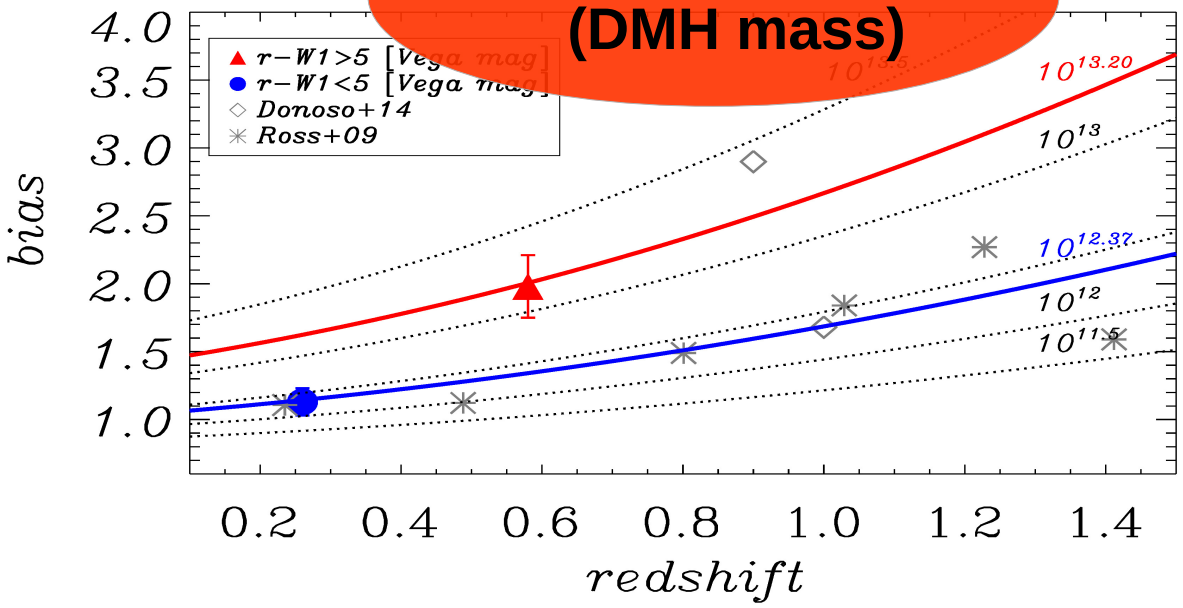


Correlation function: "excess probability" of finding another galaxy at a distance r from some galaxy

Solarz et al., 2018



Typical environment (DMH mass)



Summary

- New large astronomical datasets with huge numbers and time domain (synergy with CREDO!) will open completely new discovery area
- For these datasets automated machine learning-based methods will be more and more a necessary standard
- We need to use existing datasets to learn and be prepared…
- But we also need to learn a new approach to discoveries (as our colleagues particle physicists already did): discoveries without hand-on data work, and sometimes with triggers only, and no real data even available