Machine learning applications in astrophysics

Agnieszka Pollo
National Centre for Nuclear Research &
Jagiellonian University, Poland

with the team
Ola Solarz, Kasia Malek, Gosia Siudek, Artem Poliszczuk, Szymon Nakoneczny, Luis Suelves, Maciek Bilicki, Aditya Narendra and others

also: see the next talk by Hareesh Thuruthipilly & Margherita Grespan

„Astronomy related” Nobel prizes -

- 2020 – black holes
- 2019 – physical cosmology and exoplanets
- 2017 – gravitational waves
- 2015 – neutrino oscillations
- 2011 – accelerating expansion of the Universe
- ....
- 2006 – microwave background radiation
- 2002 – X-ray astrophysics
- 1992 – pulsar-based test of general relativity
- 1936-1983: only 5 „astrophysical” Nobel prizes

Changes in the data domain: Milky Way

- ~5,000 stars visible to the naked eye
- ~1004 stars in the Tycho Brahe/Johannes Kepler catalog (1627)
- 1993: Hipparcos Catalogue: 118 218 stars
- 2020: Gaia EDR3: 1,811,709,771 = 10^9 (mostly) stars
- (total in Milky Way: 100 thousand million = 10^11 stars)

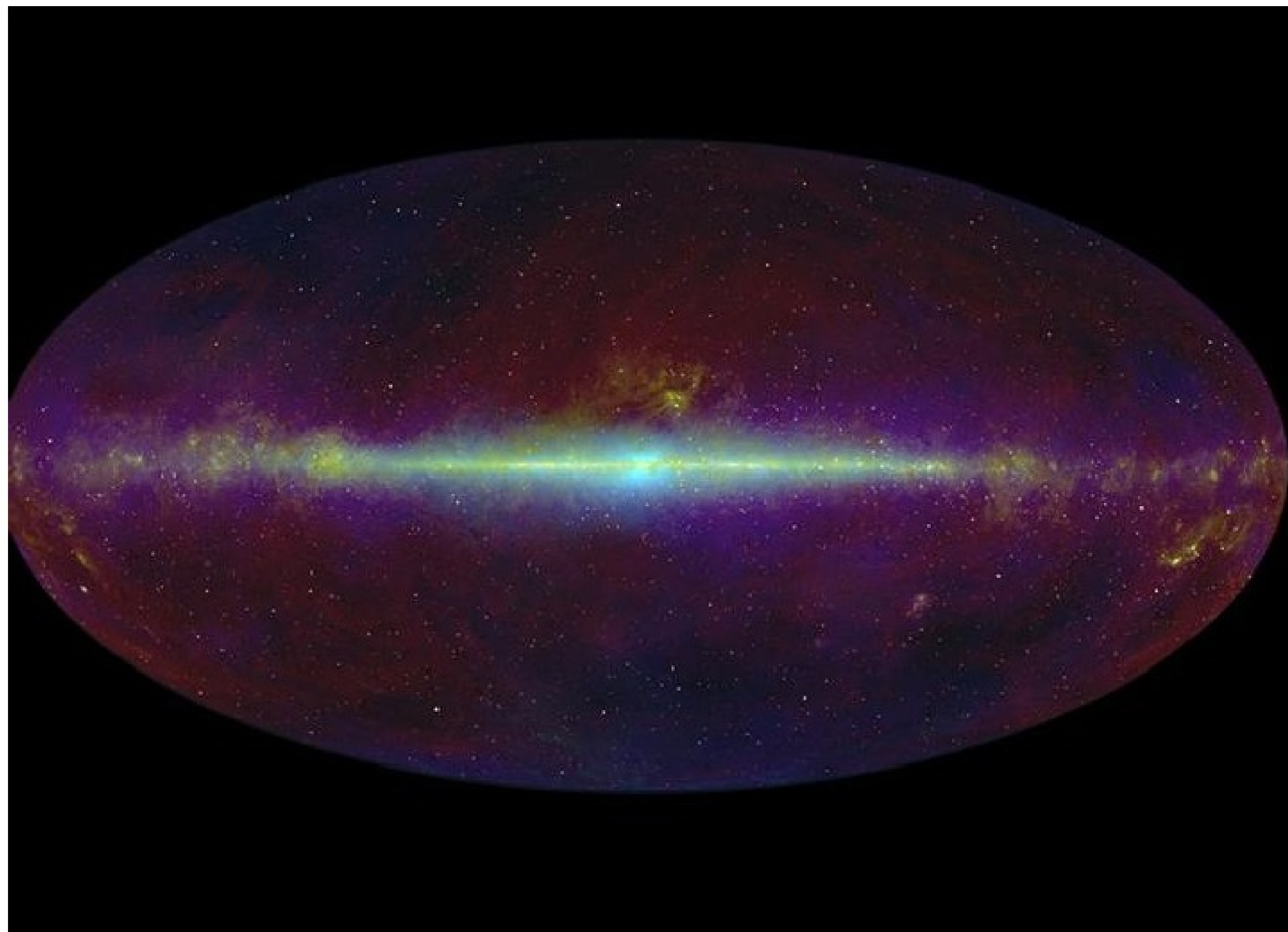
Changes in the data domain: extragalactic world

- ~3 galaxies visible to the naked eye
- ~110 „nebulae” (out of which 40 galaxies): Messier catalog (1774)
- 1888-1908: New General Catalogue of Nebulae and Clusters of Stars (NGC): 7,840 (+5,386)
- ~1990: the APM galaxy catalog: 14,681 (nearby) galaxies
- ~1990: CfA2 Redshift Survey: 18,000 (nearby) galaxies
- 1995: CFRS – deep surver of 700 galaxies
- ~2000: SDSS - ~150,000 (nearby) galaxies and quasars;
- mid-2000: deep surveys -> ~a few 10,000 galaxies
- mid-2010: deep surveys -> ~100,000 galaxies; local surveys (SDSS and cont.): milion(s) of galaxies
- near future: DESI with 8mln+ galaxies, LSST with one SDSS per 3 nights...
- estimate: 125 billion (1.25×10^11) galaxies in the observable universe

(Astronomically) Big Data
Wide-field Infrared Survey Explorer (WISE)

- All sky in the infrared
- over 747 mln sources

(15 PB of data: tables and images)



(http://wise2.ipac.caltech.edu/docs/release/allsky/)

(Astronomically) Big Data of near future: Vera Rubin Observatory

- Large Survey of Space and Time (LSST)
  - Deep and wide survey in time domain
  - mirror 8.4-m; 3200 megapixel camera
  - **37 bln** stars and galaxies

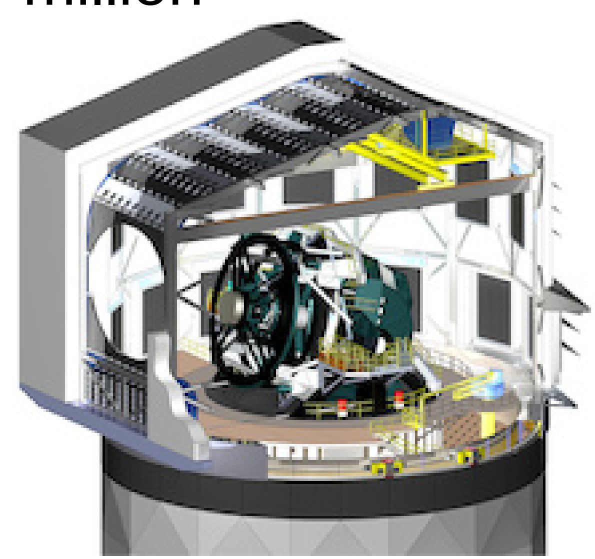
20B galaxies

17B resolved stars

6M orbits of solar system bodies

Average number of alerts per night: about 10 million

- 10-years long sky survey
- **15-30 TB of data (all SDSS) per night**
- After 10 years: ~**200 PB of data**



Machine learning for (mostly) extragalactic science

- Huge and soon much larger „big data” in the era of „precision cosmology”
- Goal(s):
  - source classification
  - source identification
  - reconstruction of properties
  - novelty search
- 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/source properties)
- Unsupervised (+semi-supervised) → clustering of sources into previously unknown and unexpected classes

Machine learning for astronomy – challenges

- Problems and challenges
  - Extrapolation (small and biased training samples)
  - Physical interpretability (do trends we see really mean something? No to black box approach – we would like to learn new physics)
  - Reproducibility
  - Resources

Machine learning for astronomy – challenges

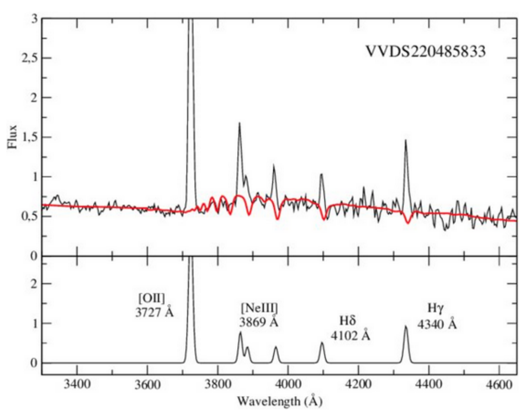
- Problems and challenges

  - observation vs experiment – we can see only as much as there is to see in the Universe

Machine learning for astronomy – challenges

- Problems and challenges
  - (relatively) small parameter space

- photometry + imaging (in different spectral ranges)



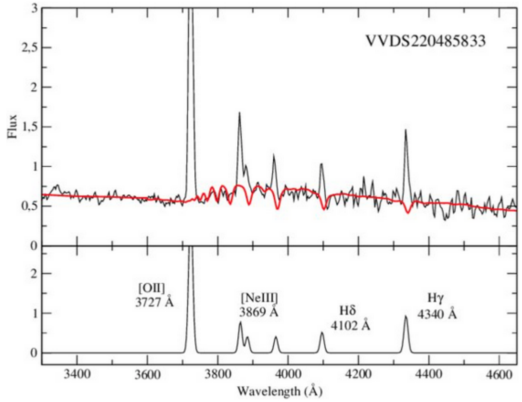
- spectra

„Multimessenger time domain astronomy”

- time variability
- polarization

Machine learning for astronomy – challenges

- Problems and challenges
  - (relatively) small parameter space



  - alternatively: a larger space of derived parameters (stellar mass, age, metalicity, star formation rate...) but at a risk of model dependence and resultant biases

Machine learning for astronomy – challenges

- Problems and challenges
    - transfer learning

JWST (space)

CFHT (ground-based)

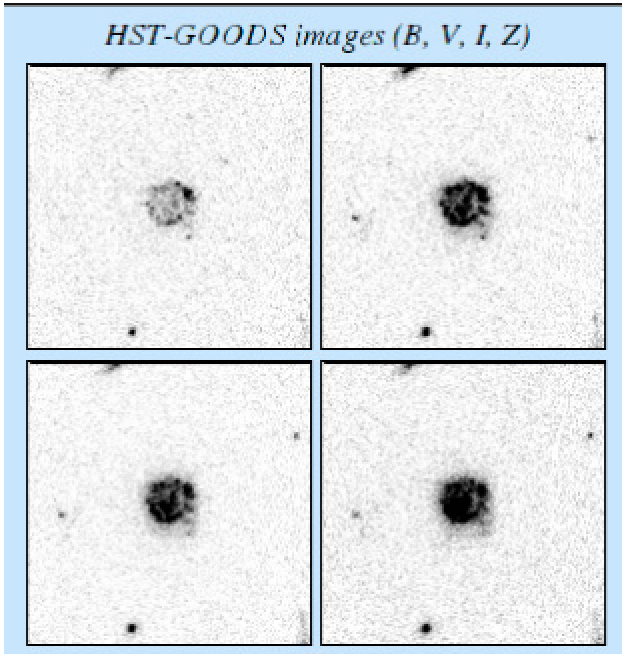
llustris (simulation)

Machine learning for astronomy – challenges

- Problems and challenges: data representability

HST (M31)

what we would like to see

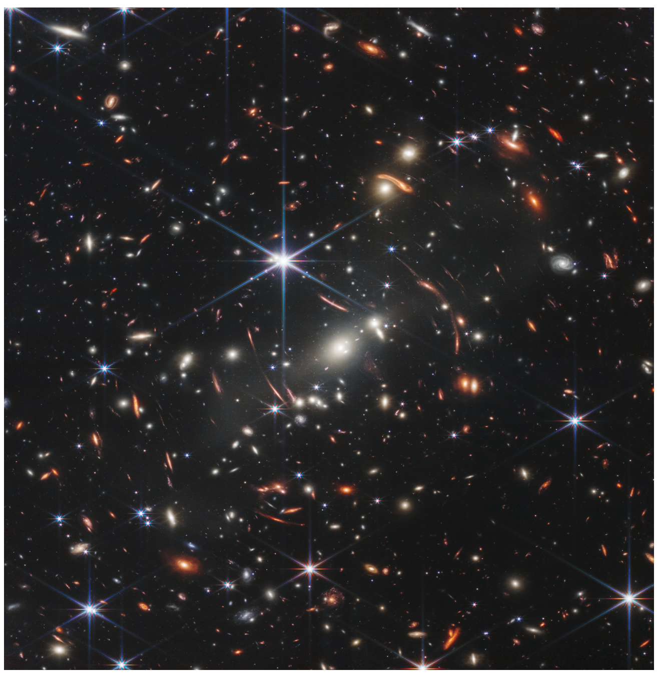


what we usually do see

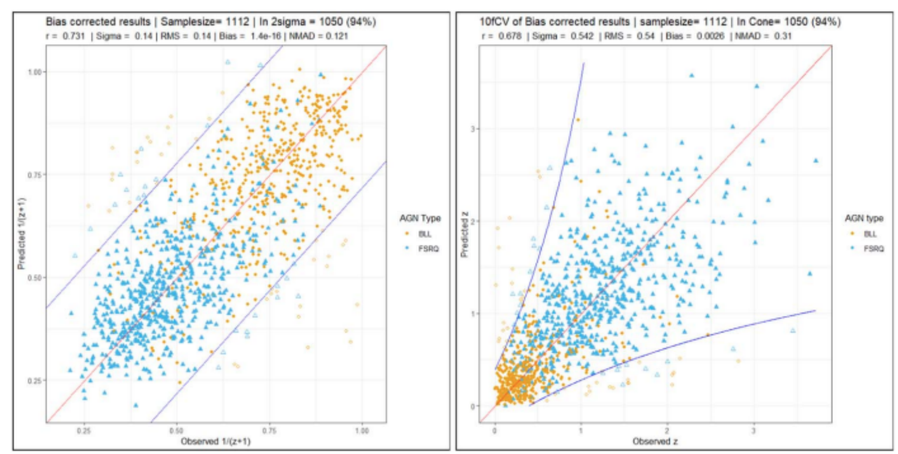
Machine learning for astronomy – challenges

- Problems and challenges: data representability

JWST



- training based on brighter objects to generalize over faint ones
- different distributions of properties of training and generalisation samples
- fainter objects are
  - intrinsically fainter – having different physical properties
  - more distant –> if in space, also in time – representing different evolutionary stages
  - more distant -> different rest frame



Narendra et al. 2022

Machine learning for astronomy – challenges

- Problems and challenges: model interpretability
  - I get a model but does it have any physical meaning?
  - But also: maybe I can find new physical information in the ML-based model?

Some examples of what ML can actually be used for (and how challenges can be met)

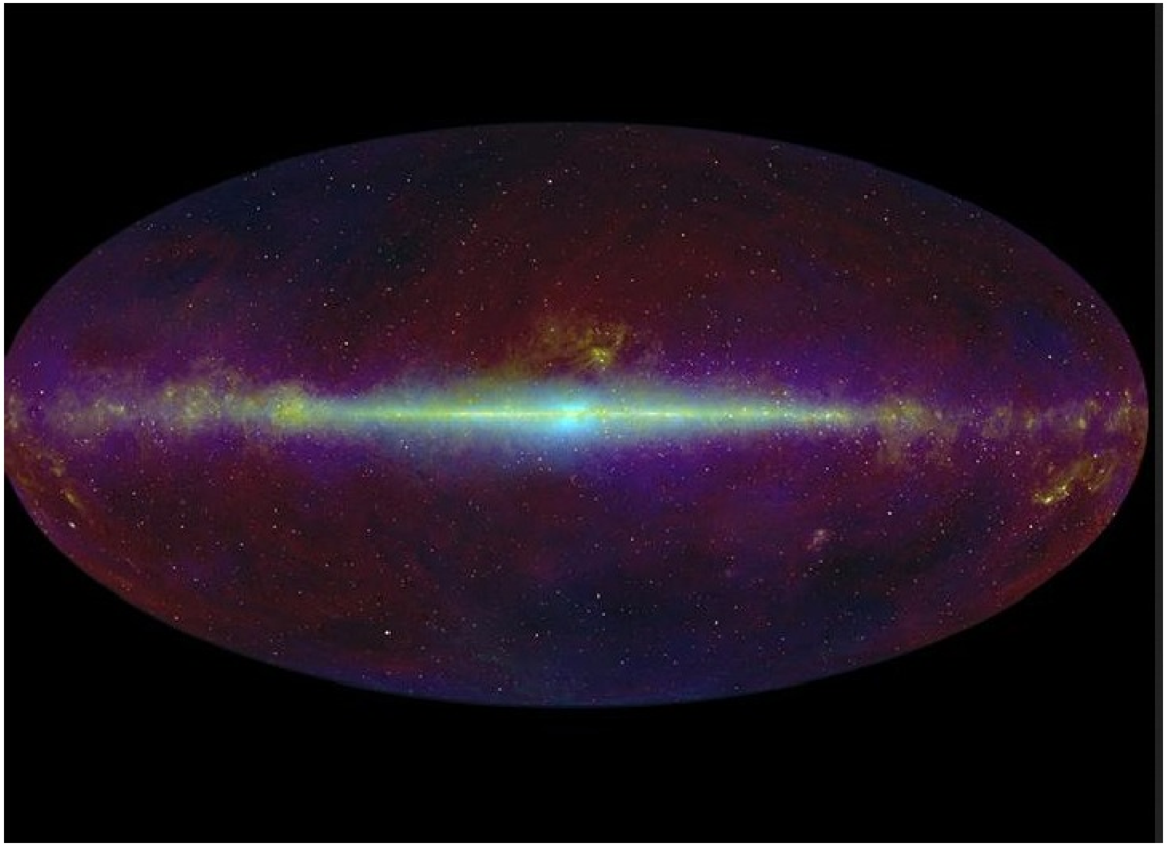
Looking for unknowns (novelty search)

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

Solarz et al. 2017

- All-Sky survey in IR
- Detected over 747 mln sources

  (15 PB of data; tables + images)
- Publicly available (position, photometry in 4 bands (3.6-22 um)
- Low angular resolution (~6”)
- No redshift information so far (i.e. - no clear identification for all!)
- The largest single astronomical catalog so far – training ground for search for unknowns



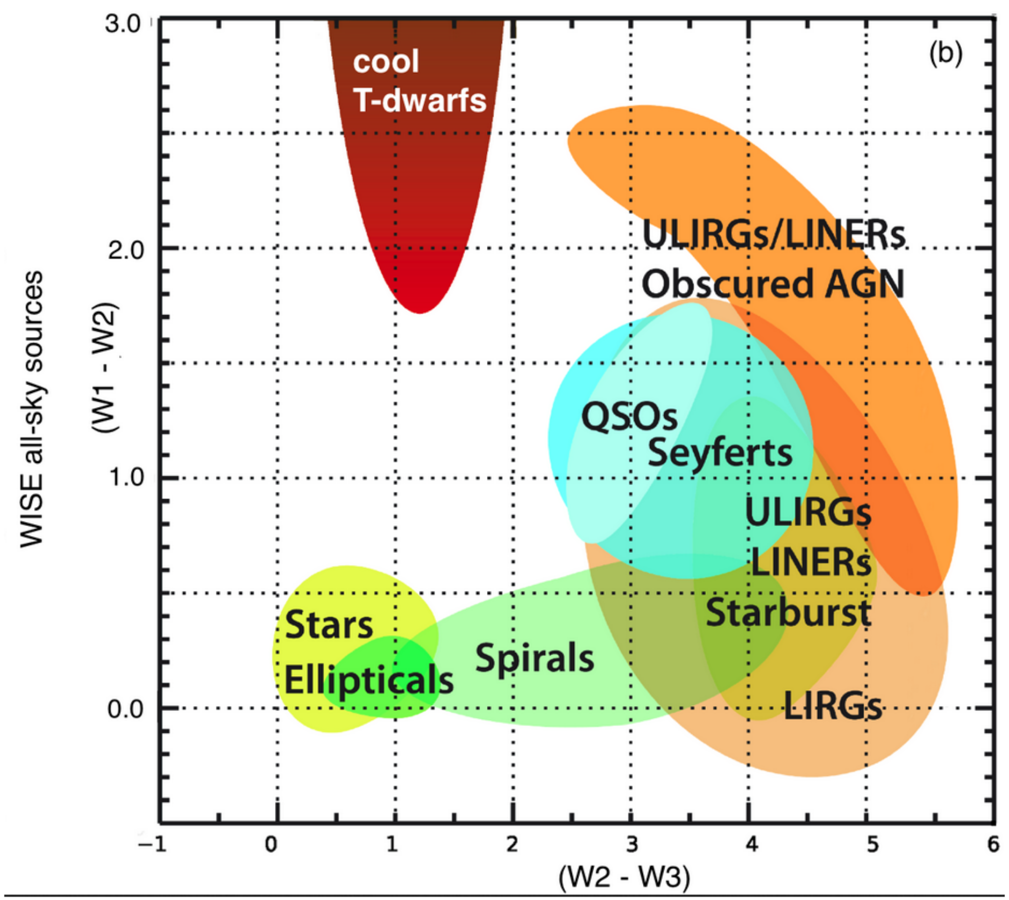
(http://wise2.ipac.caltech.edu/docs/release/allsky/)

Search for unknown among the knowns

- „Traditional" approach to source classification: color-color diagrams or similar
- Truely „novel" sources should deviate in properties but they may mimic the behaviour of known sources, especially when only few properties are taken into account

  → Search in multidimentional (as much as data permit, with feature selection on the way...) parameter space

Solarz et al. 2017



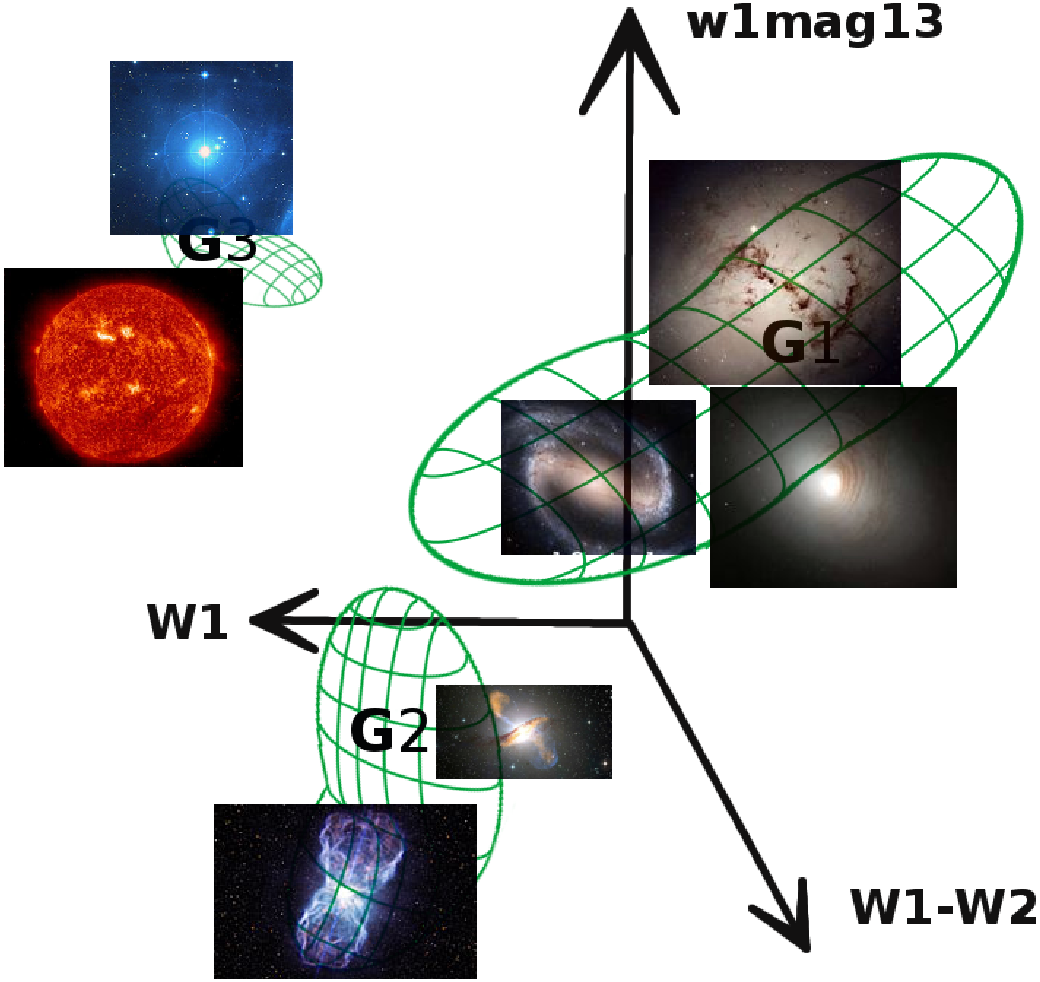
Credit: Wright+10

WISE: novel source detection

Solarz et al. 2017

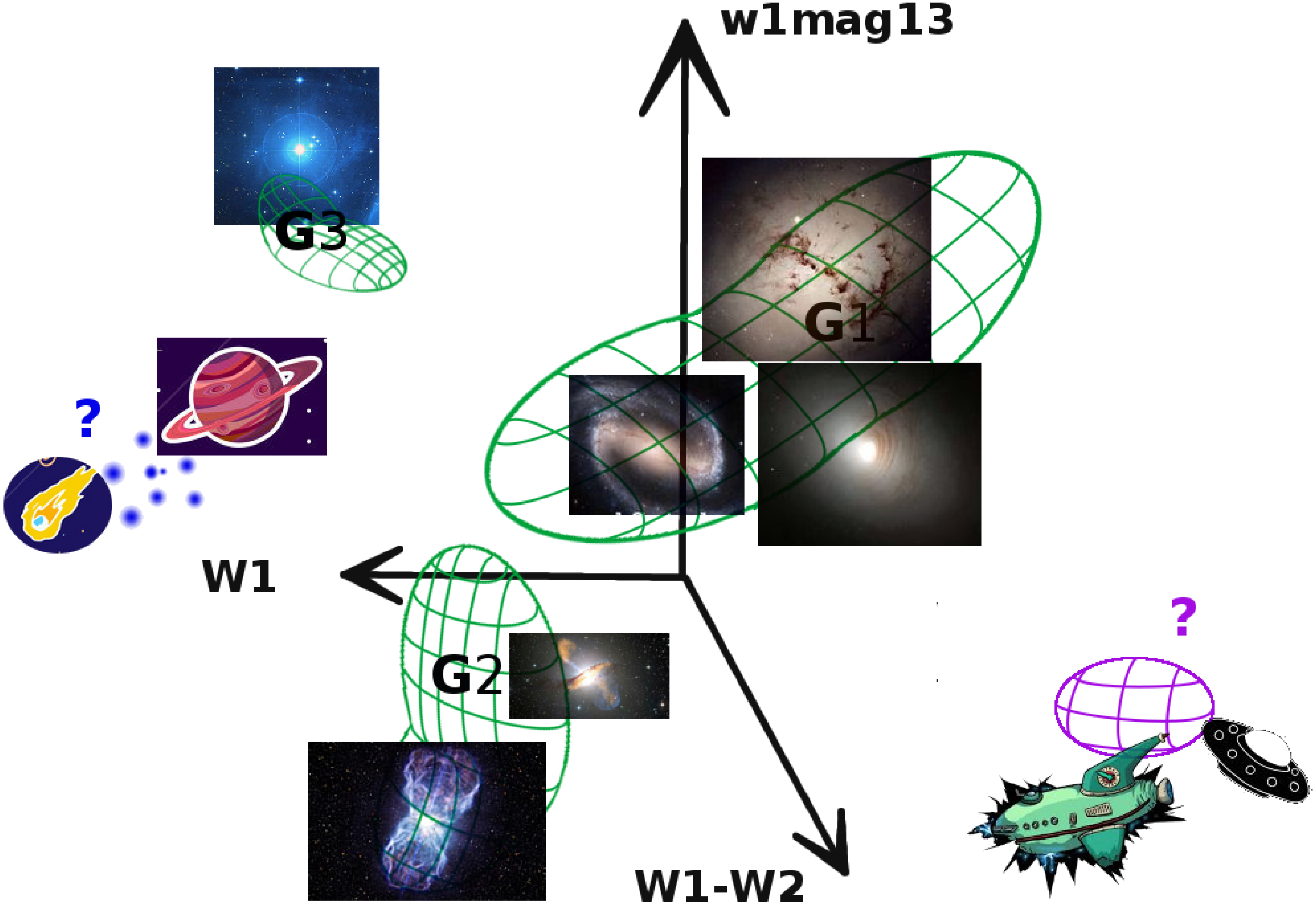
Training set (what we expect):

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



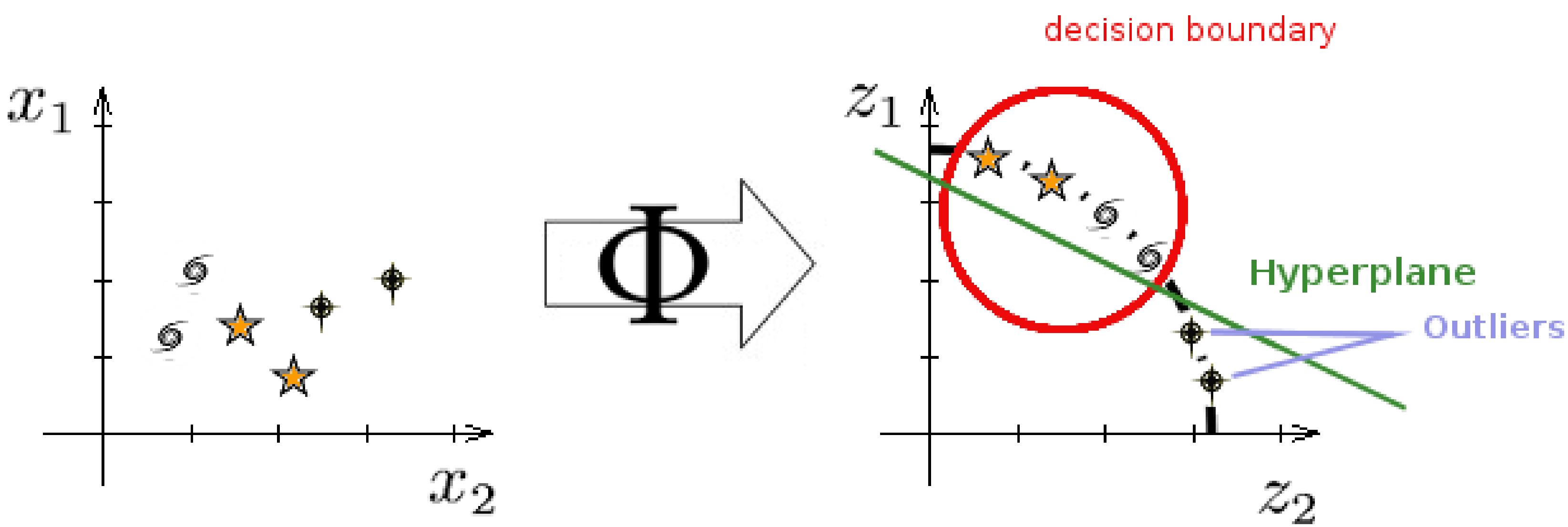
WISE: search for for unknown unknowns

Solarz et al. 2017

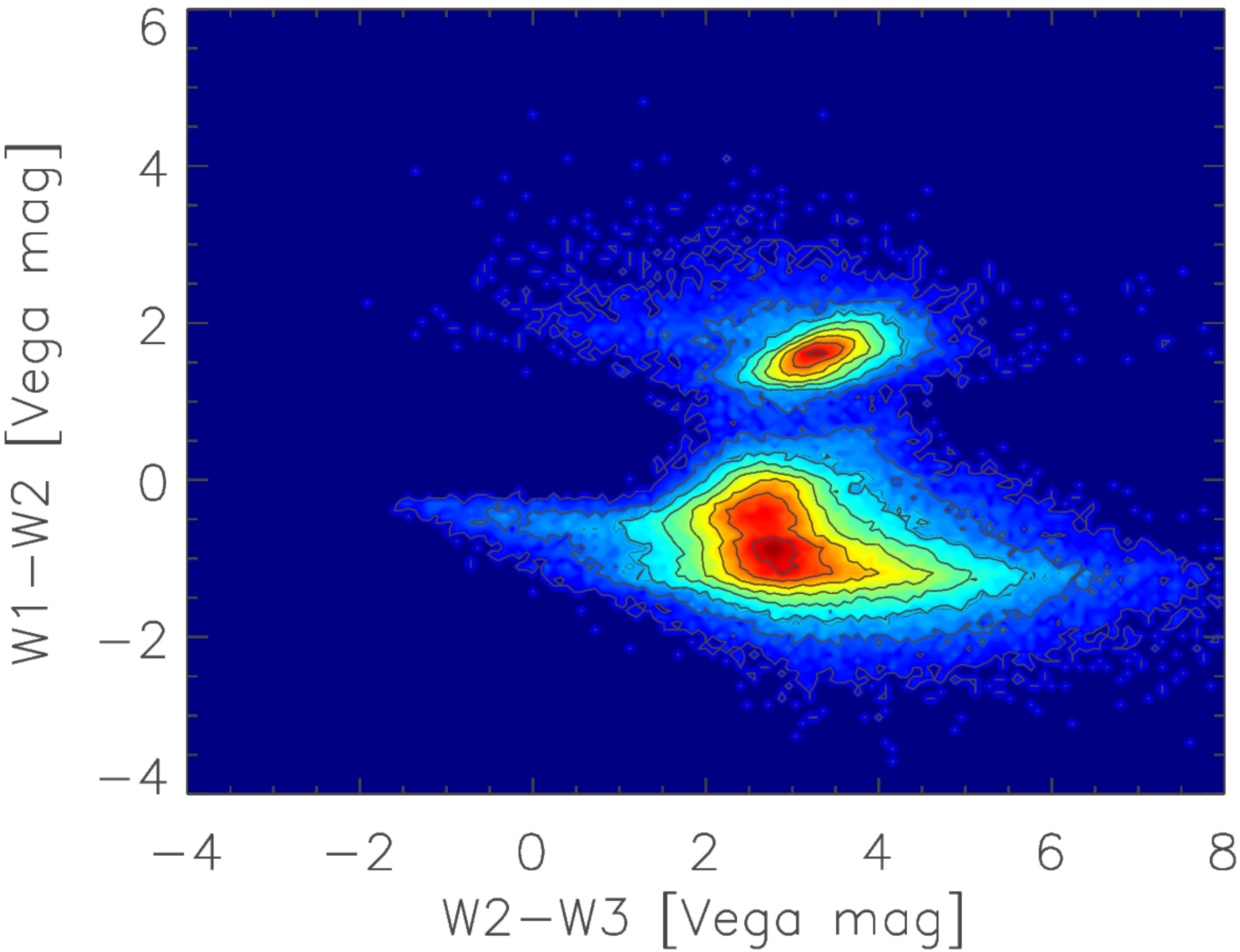


Novelty detection with One-Class Support Vector Machines

Solarz et al. 2017



- 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



<u>Results</u>:

~650,000 anomalous sources

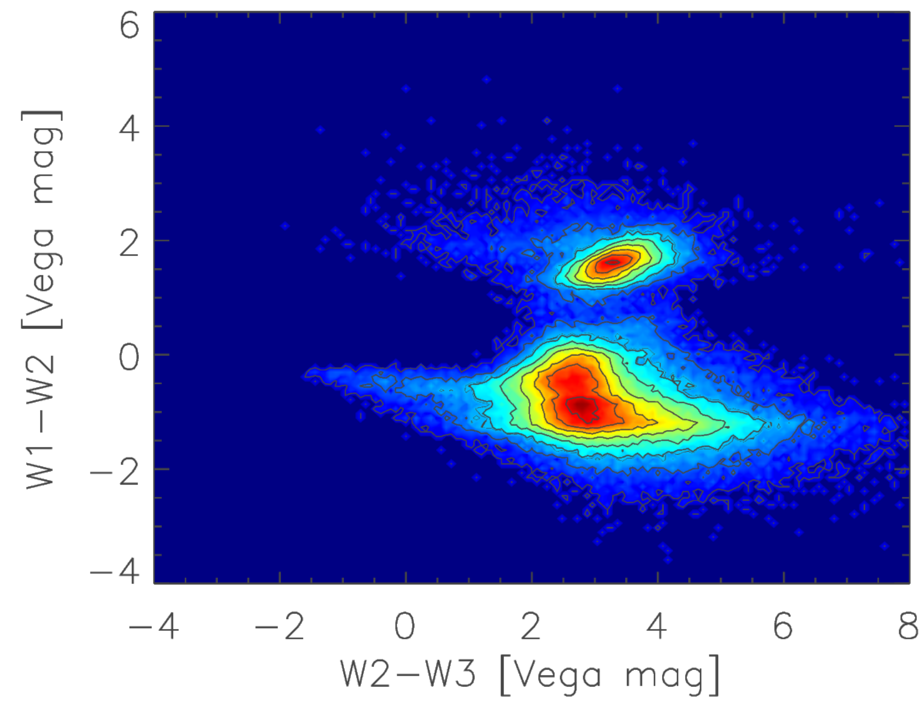
What are they?

Solarz et al. 2017

AGN candidates?

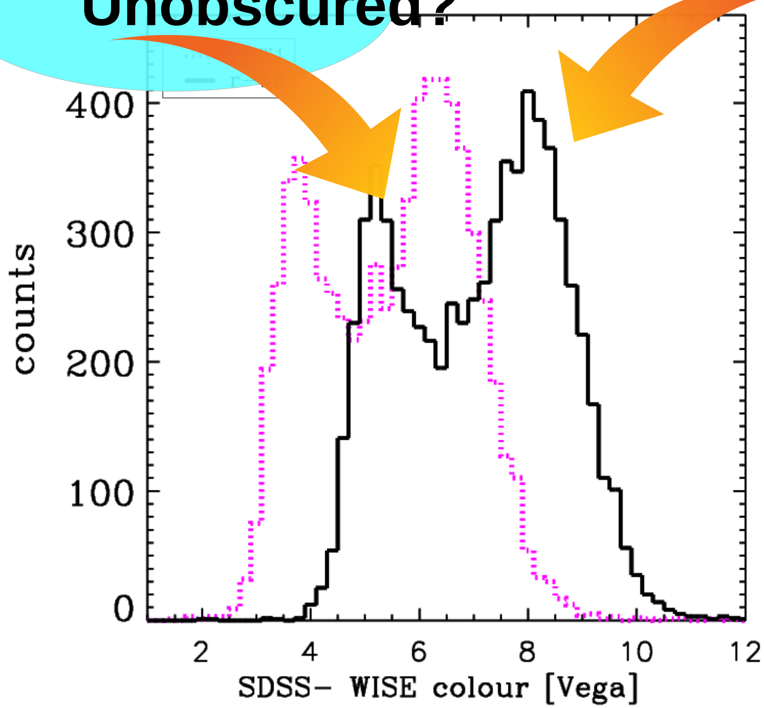
- 30,000 sources (Galactic Plane: mostly blends)
- 76% undetected at other wavelengths!
- ~7 000 objects with SDSS photometry (no spectro-z)
  - Peculiar (dusty)QSOs
  - Low-z very dusty galaxies
  - Very dusty Galactic objects

Solarz et al. 2020

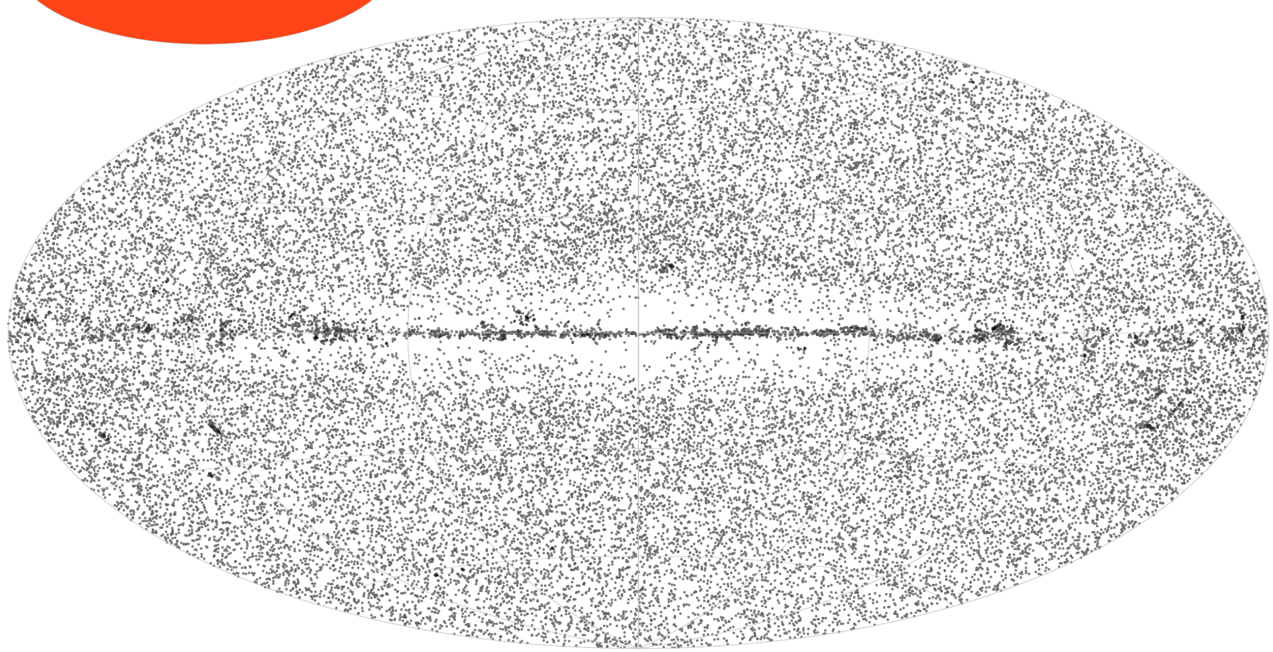


Unobscured?

Obscured?



Solarz et al. 2017

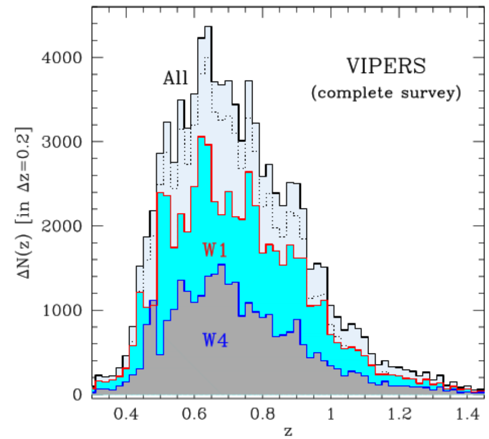


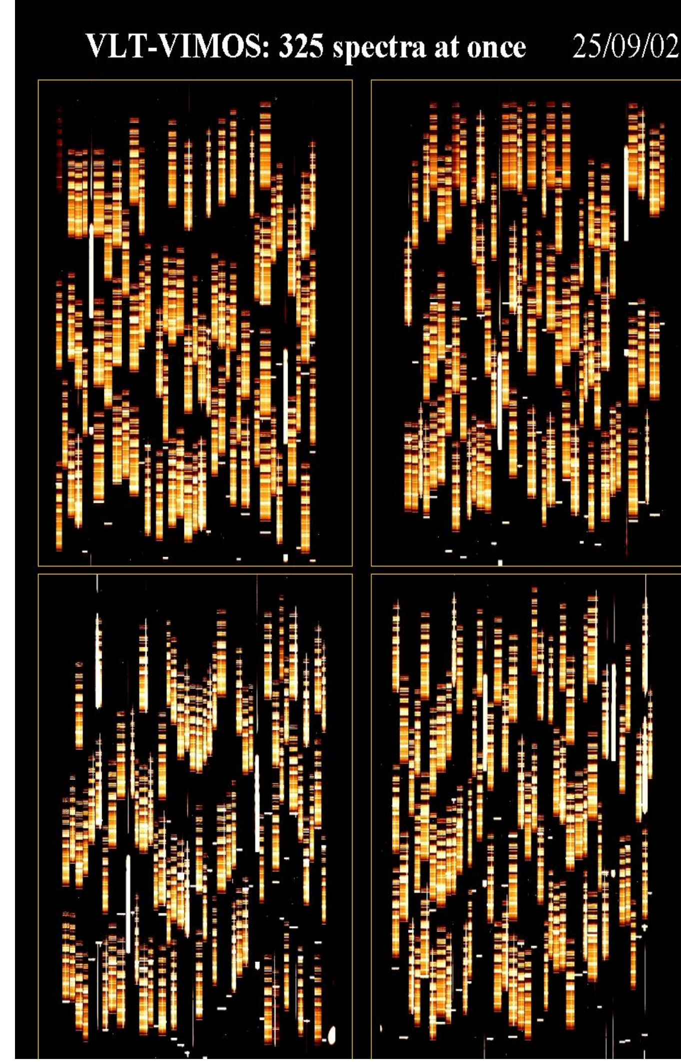
(Previously) unknown classes inside known data and long history of interpretability

SURVEY STATUS AS OF 06/11/2016

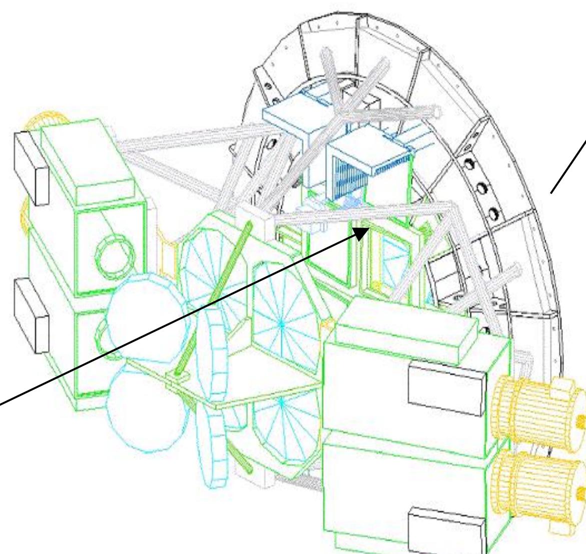
EFFECTIVE TARGETS	MEASURED REDSHIFTS	STELLAR CONTAMINATION	COVERED AREA
93252	88901	2265 (2.5 %)	100.0 %

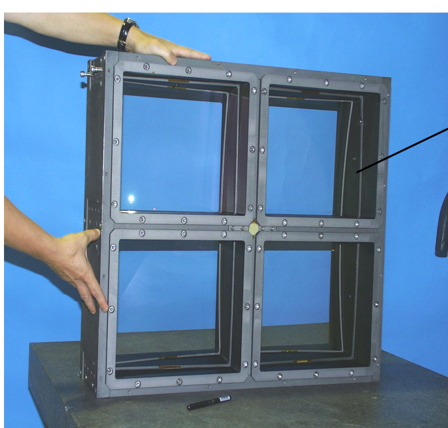
EFFECTIVE TARGETS (ET) are all the primary targeted objects with the exclusion of the ones flagged as -10 (undetected). MEASURED REDSHIFTS (MR) are the fraction of ET for which a redshift has been measured. STELLAR CONTAMINATION are the MR objects which have been identified as stars.





Large ESO Programme, 2008-2016



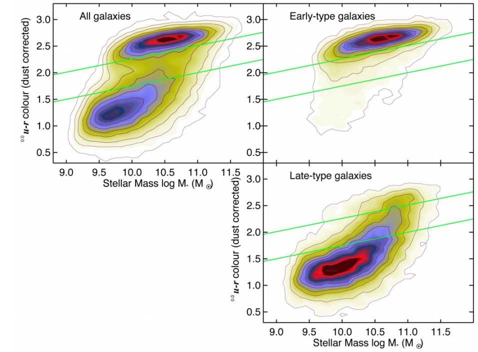


Goal: **100 000** spectra
of galaxies
at 0.5<z<1.2
2 fields on the sky, 24 deg^2

Guzzo et al. 2014, 2017, Scodeggio et al. 2018

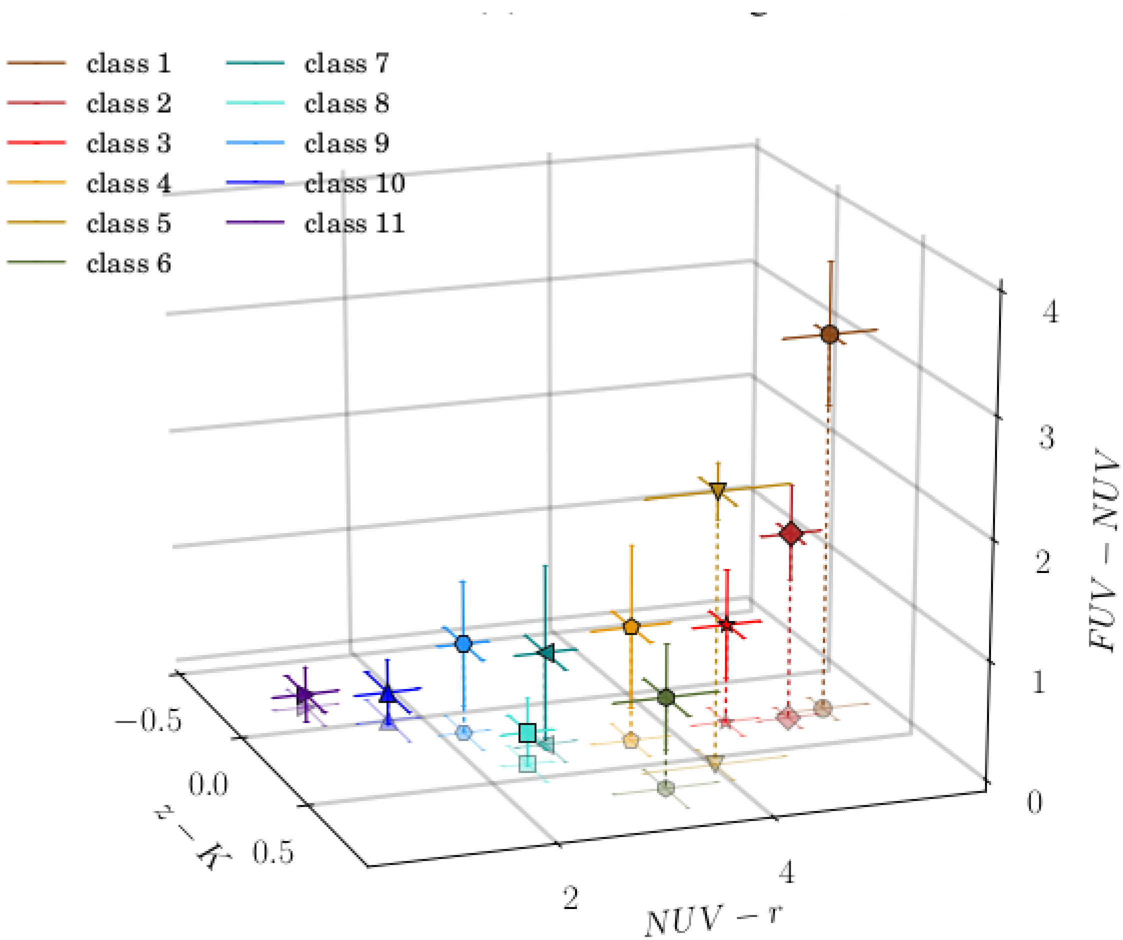
VIMOS PUBLIC EXTRAGALACTIC REDSHIFT SURVEY

Schavinski et al. 2016



- Method: unsupervised - FEM - Fisher Expectation-Maximization (Bouveyron & Brunet 2011);
- Parameter space: of 12 rest-frame optical magnitudes and a spectroscopic redshift

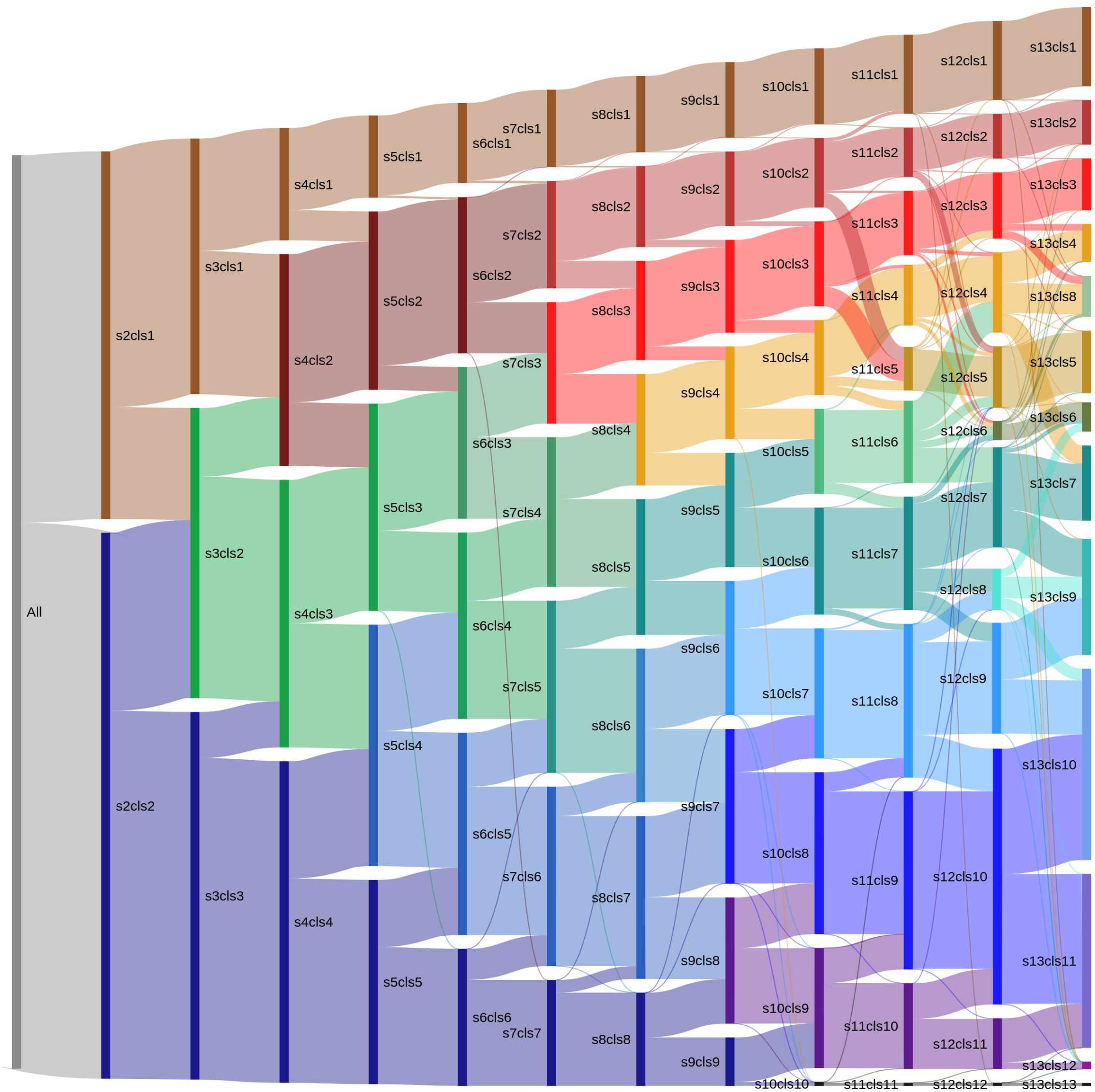
Beyond bimodality: how many galaxy populations can be blindly selected at z~1?

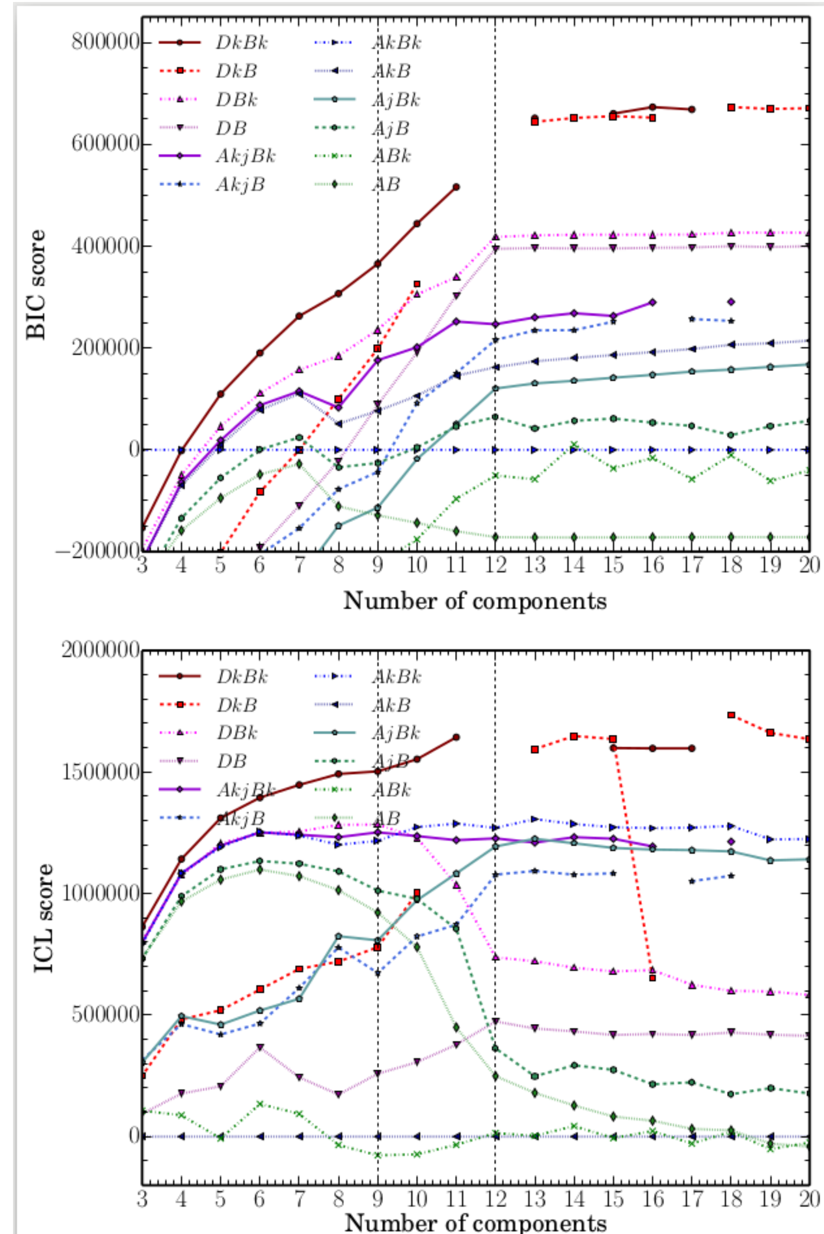


11 *well separated* classes of galaxies at 0.5<z<1 (+ a 12th class of outliers), forming the sequence of: 3, 3, and 5 subclasses of early, intermediate and late types, respectively. well reproduced in SDSS (local Universe)

Siudek et al. 2018
Turner et al. 2021

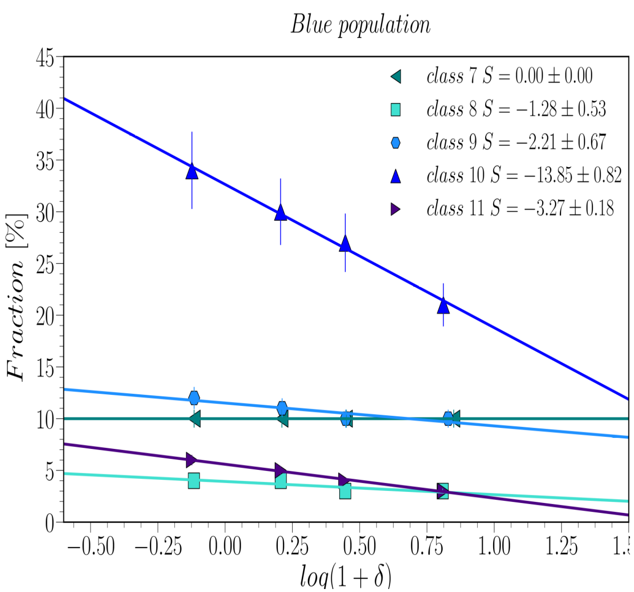
How many galaxy populations can be blindly selected at z~1?



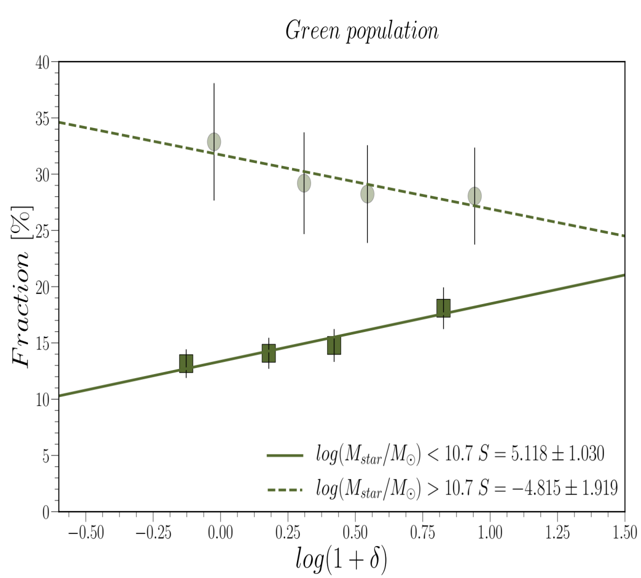


Siudek et al. 2018

Does this 11 class division reflect actual physical information?
→ Traces of different galaxy evolutionary paths seen in multi-color space?
→ See what happens when quantities not related to classification are introduced (environment!)...



For blue galaxy populations: the downsizing trend is mostly driven by only one (admittedly, the largest) subpopulation (consistent with mass-driven passive evolution)
while the fractions of other blue SF galaxies are much less mass/environment-dependent

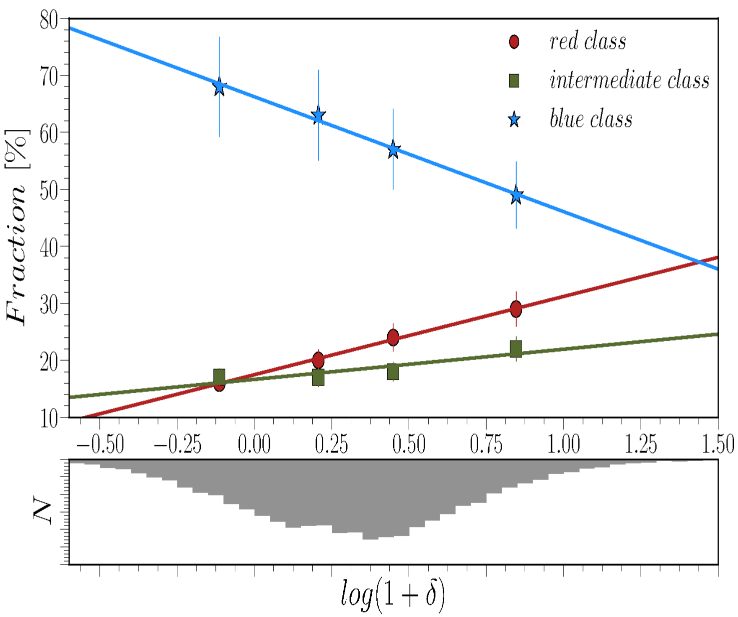


For intermediate and dusty populations the environmental trends are reversed depending on stellar mass: low mass ones behave like passive galaxies; high mass ones - like active galaxies

- a variety of galaxy populations is physical, and indicates a variety of their evolutionary paths

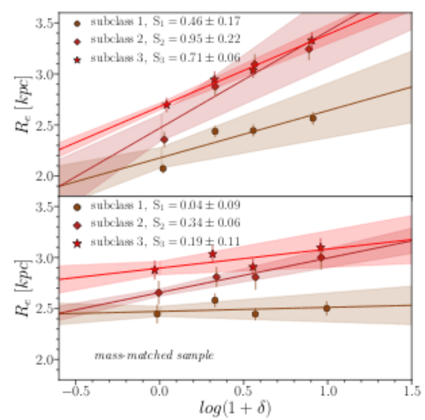
(**Siudek et al. 2022**)

Does this 11 class division reflect actual physical information?
→ Traces of different galaxy evolutionary paths seen in multi-color space?
→ See what happens when quantities not related to classification are introduced (environment!)...



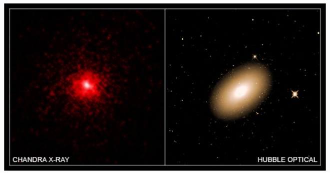
- a variety of galaxy populations is physical, and indicates a variety of their evolutionary paths

(**Siudek et al. 2022**)



...the reddest red class: small and size does not depend on environment (independently on stellar mass): a product of early fast quenching (while the other two might have grown also through mergers)

a catalog of 77 „red nuggets” (relic galaxies which never merged in their lives) at z~0.7 (**Lisiecki et al. 2022**)



Mrk12 16 – not one of ours but as ours would look „today”

Merger in the background or a history of unexpected interpretability

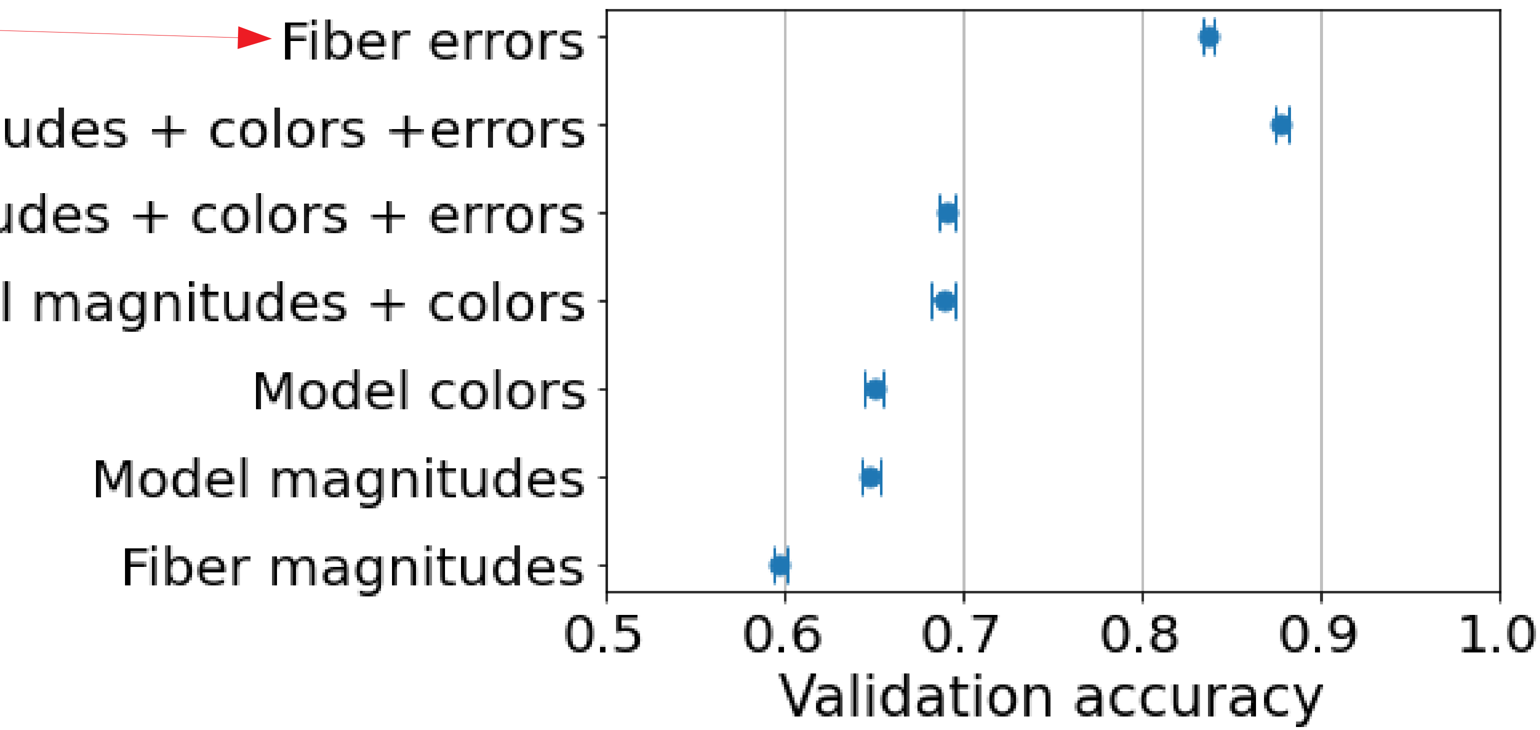
(talk to Luis Suelves)

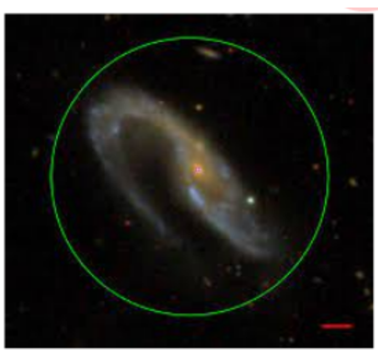
Credit: HST, NASA/ESA

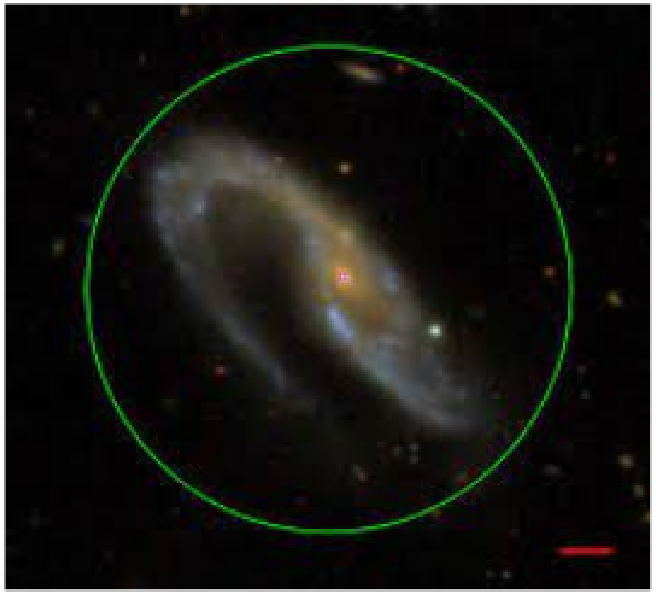
How to automatically find merging galaxies?

→ People very often use Deep Learning (with moderate success)

→ Concept: see if we can do any good (but faster/easier/more interpretable) with photometry only (fluxes, colours, errors)

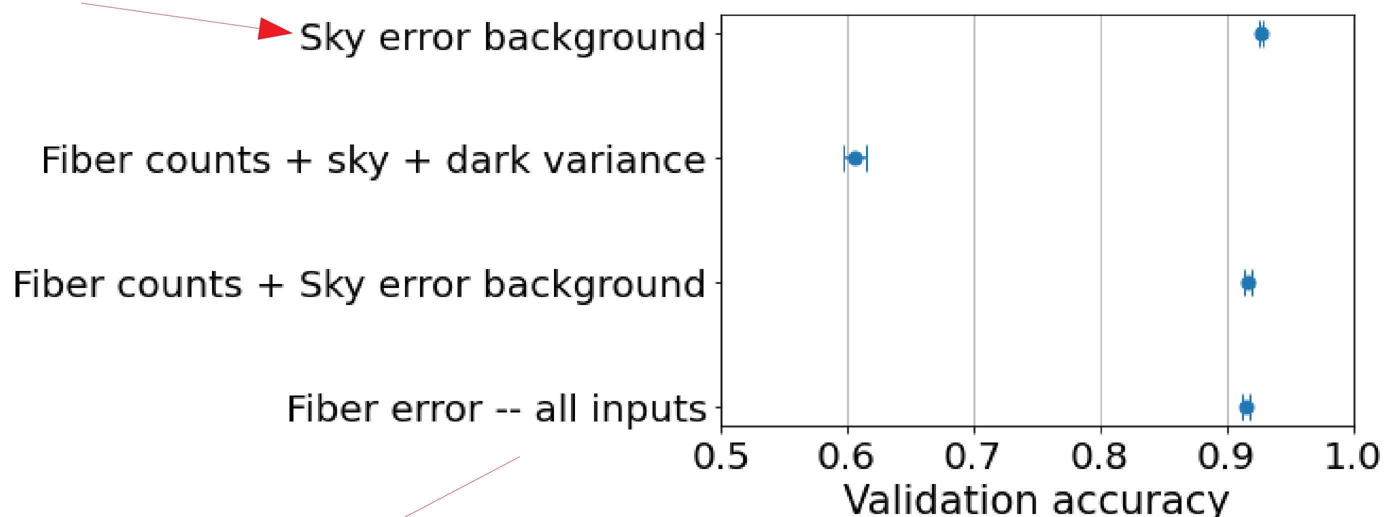


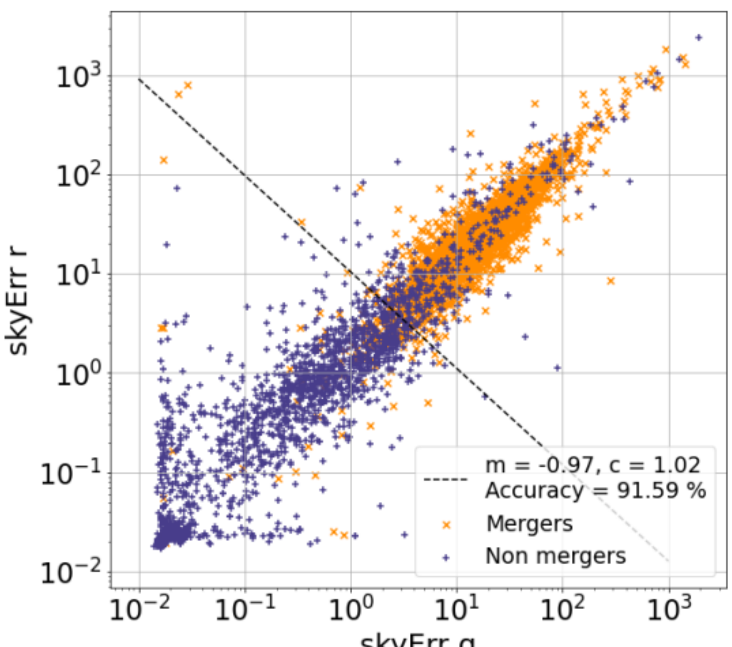




How to automatically find merging galaxies?

→ What is a magical ingredient of fiber errors?





→ We do not need any ML do get ~92% accuracy – it was just about finding the key data

→Physics: merging galaxies (today) do not differ that much from other galaxies – what makes them different are their surroundings (tidal tails etc.)

→new generation of DL for background only (never forget the power of differential analysis...)

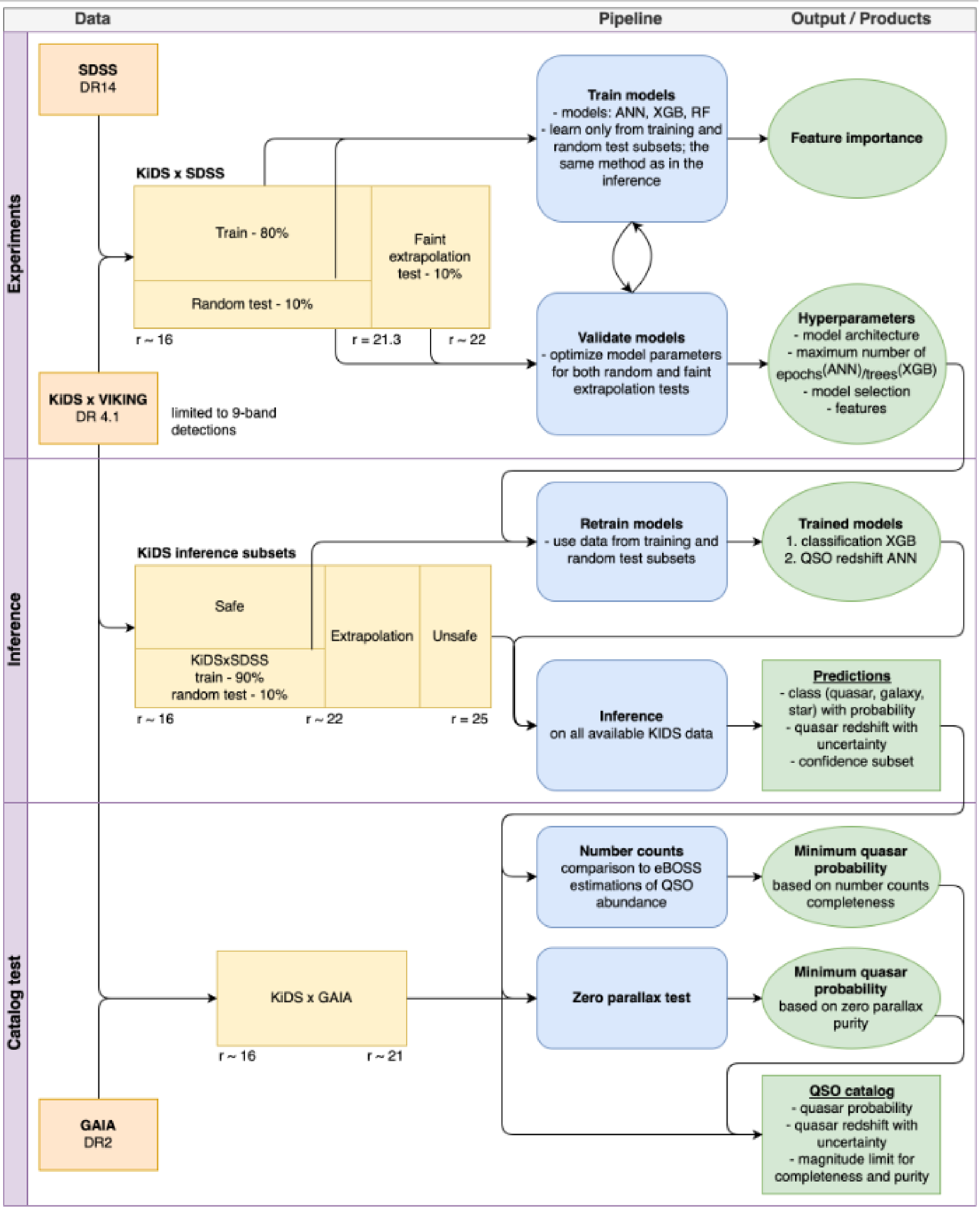
What is an AGN and how to find (and measure) them

- Challenges: different types and different diagnostics
- Big Data, big search: need for large and varied training samples!
- AGN properties, including photo-zs are tricky to recover even using „traditional” techniques
- Bright but training data available for low z/the bright end of LF: extrapolation problem.

- However, if we have a big training sample and are smart to use ML methods, we can get a reliable AGN sample and its properties (KiDS: Nakoneczny et al. 2019, Nakoneczny et al. 2021)

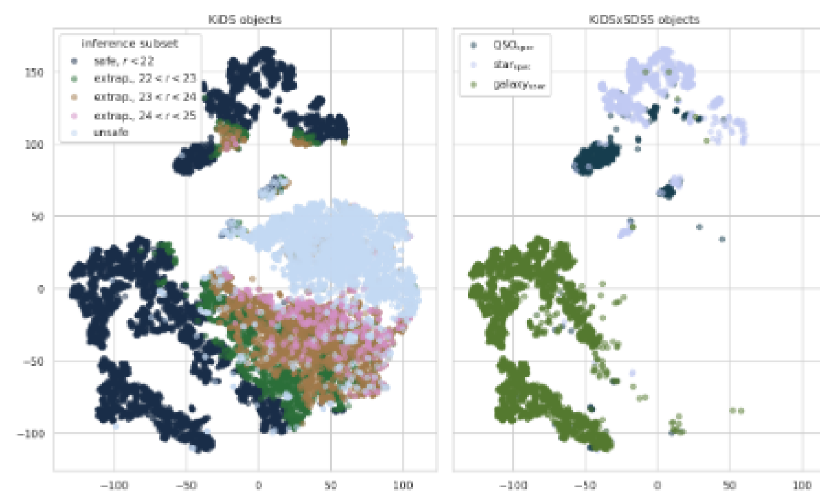
Nakoneczny et al. 2019

Nakoneczny et al. 2021



45 million objects of the KiDS photometric data limited to 9-band detections

-> 158,000 quasar candidates in the safe inference subset (r < 22) and an additional 185,000 candidates in the reliable extrapolation regime (22 < r < 23.5)



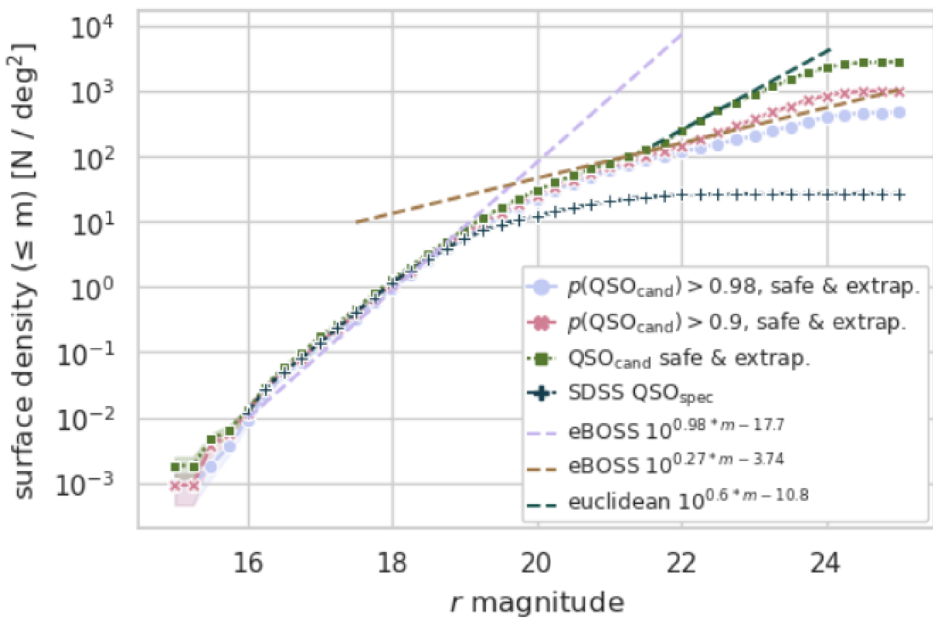
KiDS quasar candidates: how to make sure they are what we think they are

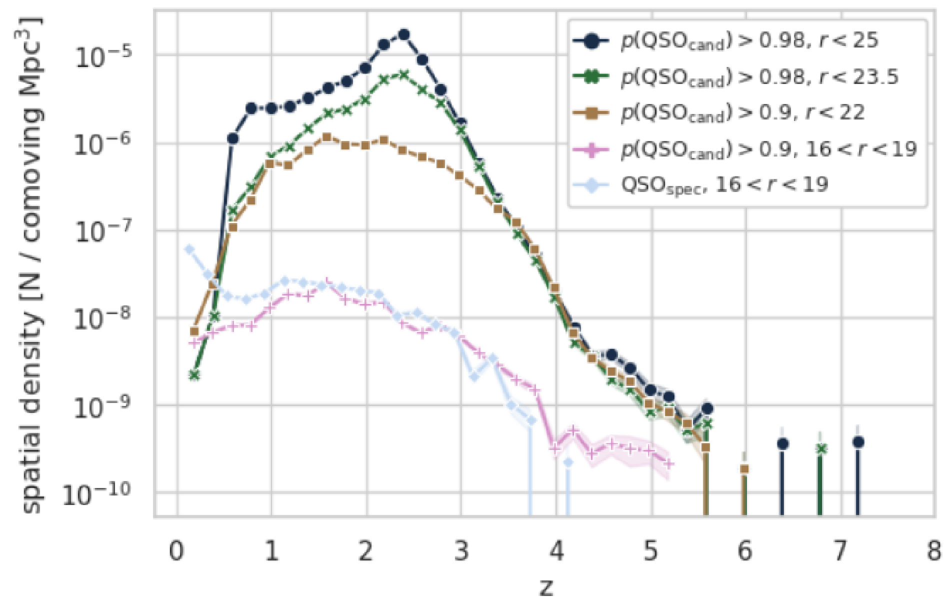
Nakoneczny et al. 2019

Nakoneczny et al. 2021

ID: Xboost
photo-zs: ANN

	safe $r < 22$	safe & extrap. $r < 23.5$	safe & extrap. $r < 25$
QSO_{cand}	266k (100%)	1.6M (100%)	3M (100%)
$p(QSO_{cand}) > 0.90$	**158k (59%)**	637k (39%)	1.1M (36%)
$p(QSO_{cand}) > 0.98$	127k (48%)	**311k (19%)**	507k (17%)





Summary

- Extragalactic Big Data
  - now more and more necessary to introduce new automated methods to study new large data, especially those coming soon (e.g. LSST)

- Problems and challenges
  - Extrapolation (small and biased training samples; limited parameter spaces)
  - **Physical interpretability** (do trends we see really mean something?)
  - Reproducibility
  - Resources