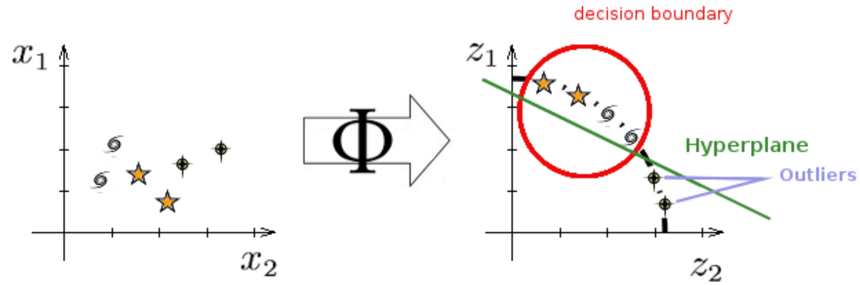
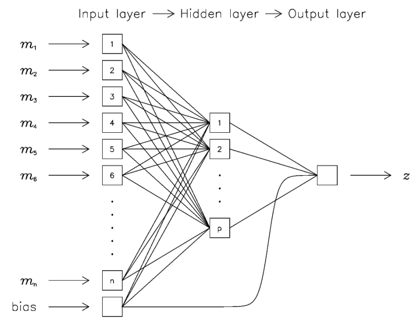
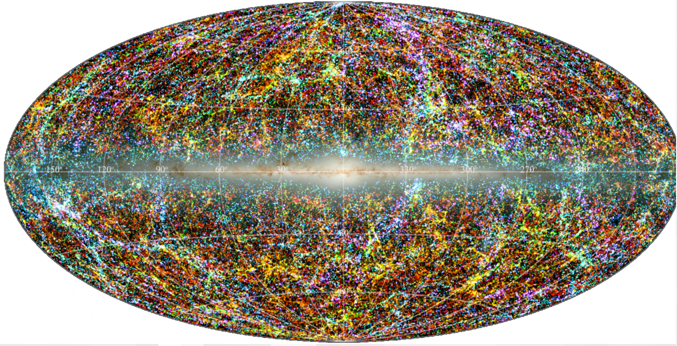






Large-scale structure and machine learning

A subjective review

Maciej Bilicki

Leiden University (the Netherlands),
National Centre for Nuclear Research (Poland),
& University of Zielona Góra (Poland)

Universiteit Leiden

UNIWERSYTET ZIELONOGÓRSKI

ASTRONOMER

What my friends think I do

What my parents think I do

What society thinks I do

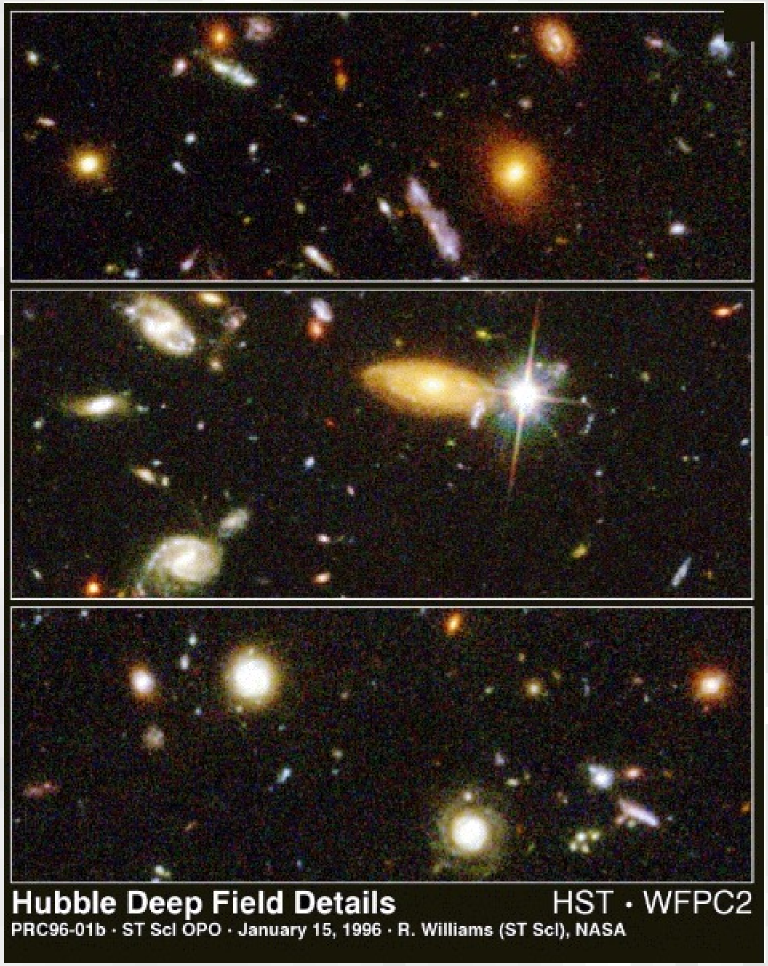
What the media thinks I do

What I think I do

What I actually do

What astronomers* really do

- We **search for**, **detect** and **study** astronomical **objects** (planets, stars, galaxies…), as well as various “backgrounds” (radiation, neutrino, ...)



- We **map the sky**: **surveys** at different electromagnetic wavelengths
- This is now done in **(semi-)automatized way**, including with instruments in space
- **Data “reduction”** (processing) is also getting automatized - “pipelines”
- End users of a survey will often obtain a product such as a **database of images, of spectra** and/or a **source catalog**

This applies mostly to* ***observational astronomy* *(unlike theoretical & computational)*

Surveying the sky

- **Two main approaches in sky surveys:**

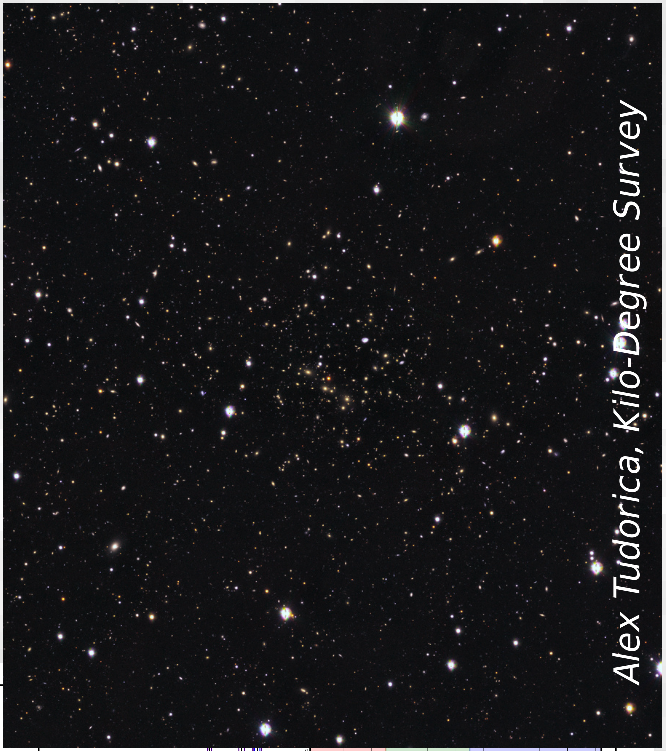
  1) **Photometric**: **imaging the sky** at various wavelengths (visual, infrared, ultraviolet, radio, ...) Often “blindly” on previously uncharted areas

  2) **Spectroscopic**: measuring **electromagnetic spectra** (i.e. energy distribution) of objects Needs input from 1) to know where the sources are

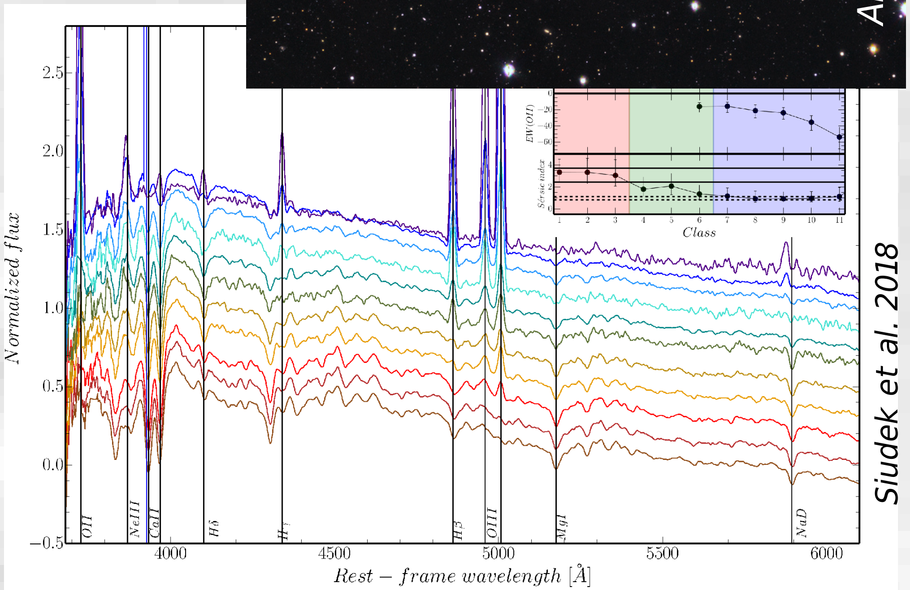
  Both 1) and 2) can be done repeatedly to look for **time variations**

- The data are obtained using (often sophisticated) **charge-coupled devices** (CCDs) and stored in a **digital form**

- **Current datasets range from $O(10^3)$ to $O(10^9)$ sources**; this keeps growing

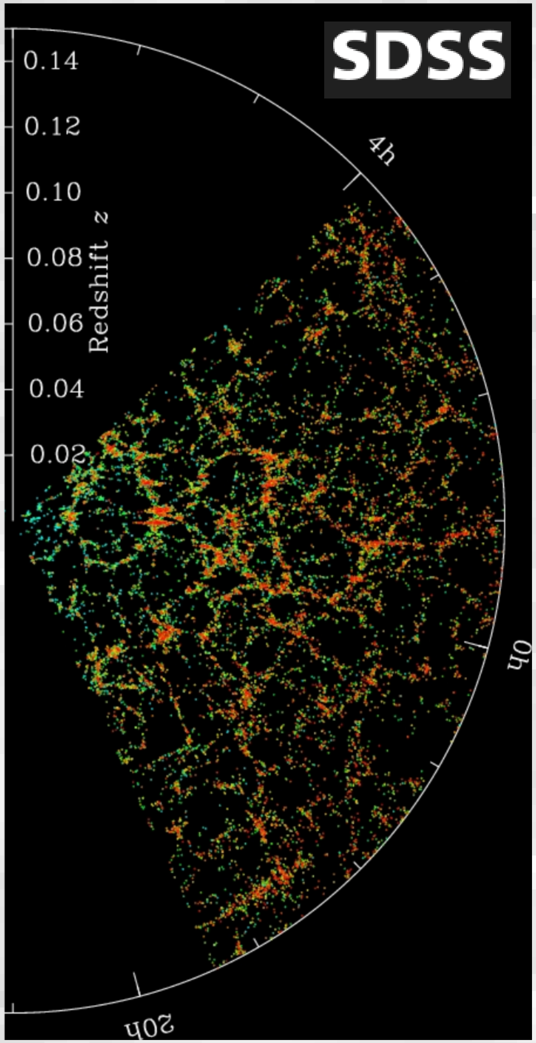


Alex Tudorica, Kilo-Degree Survey



Siudek et al. 2018

Data avalanche in modern astronomy



Some examples:

- Sloan Digital Sky Survey (since ~2000): **115 TB** of data
- Zwicky Transient Facility (start 2018) → **1 PB** of imaging data, ~1 billion objects with time-domain information
- Large Synoptic Survey Telescope, to start in ~2020, ten years of planned operation → **30 TB PER NIGHT**
- The Square Kilometer Array (the largest planned network of radio antennas, to start in 2020s) → **~4.6 Zettabytes**

It is becoming unfeasible not only to process the data on the user's side, but even to store them or (soon) to transfer all of them from the instrument!*

** This is already the case for e.g. Gaia space telescope, where data is significantly filtered out onboard before being sent to Earth.*

Numbers gathered by Dr. Aleksandra Solarz (NCBJ Warsaw)

Data avalanche in astronomy: some challenges faced

Measuring distances to galaxies

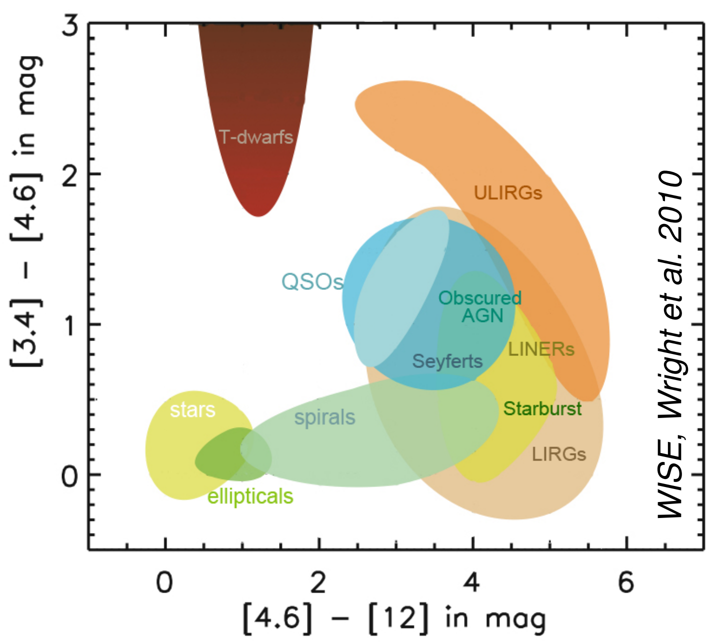
- **Galaxy distances** are essential for cosmology and extragalactic astronomy
- Known from the **redshift**: the farther galaxy is, the more its spectrum is shifted towards longer wavelengths due to the expansion of the Universe*
- Redshift is **measured using spectroscopy**: at present ~3 million such measurements ("*spectro-zs*")
- This may **grow by ~1 order of magnitude** in foreseeable future...
- ...but in imaging surveys we have already detected **O(10^8) galaxies** and this is likely to increase to even **O(10^{10})** in the coming decade(s)
- It is **highly unlikely to ever measure** spectroscopic **redshifts of most galaxies** detected by the humanity (technological limitations)

*redshift $z = \lambda_{observed} / \lambda_{emitted} - 1$; distance *d:* **Hubble law:** $\boldsymbol{c\,z = H_0\,r}$

Data avalanche in astronomy: some challenges faced

Classifying sources

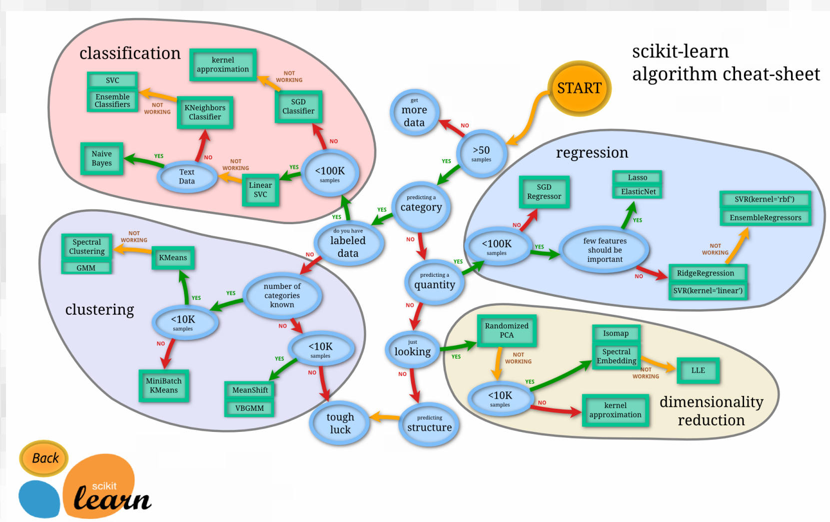
- We want to separate astronomical sources into **stars**, **galaxies**, their subtypes, ... but also detect **novel** or **unexpected** objects
- Most efficient by **combining imaging with spectroscopy** (point-like vs. extended, characteristic spectral features,...)
- The same **challenge to get spectra** for most of the already imaged objects as when measuring redshifts
- Traditional approach without spectroscopy: use "**colors**" – ratios of fluxes at different wavelengths Different source types will occupy different regions in **color-color spaces**



- Today's surveys image the sky at **many wavelengths** at a time – human brain not very good at operating in >3D spaces (visualization and projection issues, ...)

Enter machine learning

- Considered useful for astronomy since **~mid 1990s** (e.g. Fayyad et al. 1993), gained on popularity in **early 2000s** (e.g. Wolf et al. 2001, Collister & Lahav 2004)
- So far, mostly **supervised learning**: **training set** used to learn relations between **feature space** and searched **pattern(s)**, then the algorithm applied to **target set**
- Now also **unsupervised learning**, for instance to look for **clusters in multi-dimensional feature space**
- Most used in astronomy: **artificial neural networks**, **random forests**, **boosted decision trees**, **support vector machines**... – usually applied on **post-processed data** (source catalogs, spectra...)
- Recently also **deep learning** (convolutional neural networks), for instance applied directly to **digital images** (i.e. pixels)
- **Possible applications:**
  * <u>redshift estimation</u>
  * <u>source classification</u>
  * <u>rare object search</u>
  * data cleaning
  * ....



Cosmological distances and the redshift

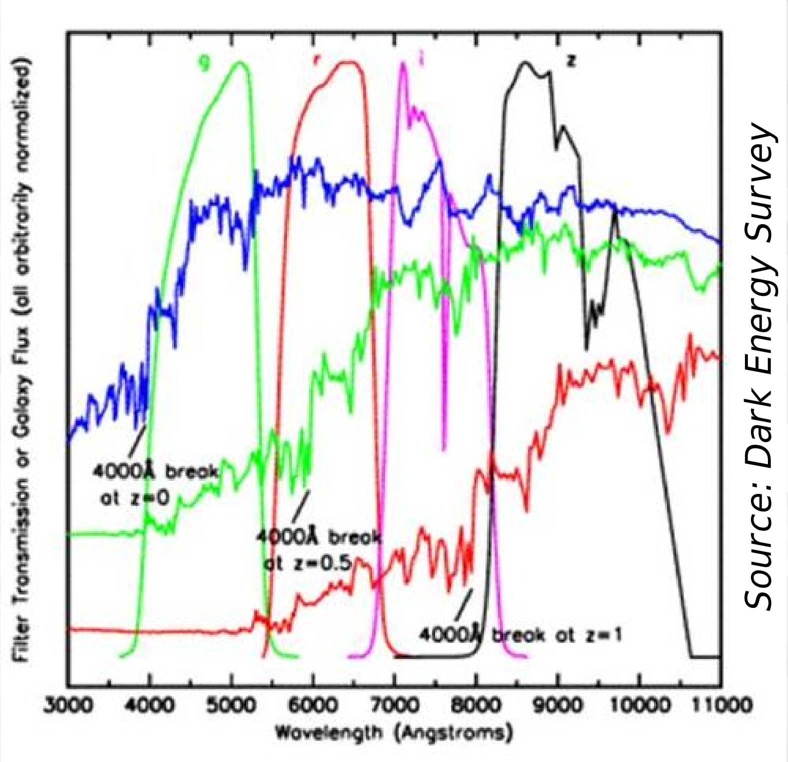
- Impossible to measure distances to galaxies with **direct methods** (e.g. parallax)
- Some galaxies have distance estimates via standard(ised) candles: **distance ladder**
- Generally, 3 coordinates of galaxies in catalogs: two angular ones and the **redshift**
- Redshift as a **proxy** for distance, via the **Hubble law**:

$$\mathbf{z} \approx H_0 \times \mathbf{d} / c \;\Rightarrow\; \mathbf{d \approx 4300\, z} \text{ [Mpc] for } H_0\text{=70 km/s/Mpc}$$

- Redshift can be precisely measured **only with spectroscopy**
- Vast majority of already detected galaxies **do not have spectroscopic redshifts**

Galaxy distances from *photometric* redshifts

- Galaxy **spectrum is shifted** towards longer wavelengths due to the **cosmological expansion**
- **Farther** galaxies are also on average seen as **fainter** (observed flux $\propto$ distance^{-2})
- Galaxies **evolve** with time, which is reflected in their spectra (gas is converted to stars, etc.)
- Therefore: **fluxes** of galaxies observed at different wavelengths **change** depending on galaxy redshift



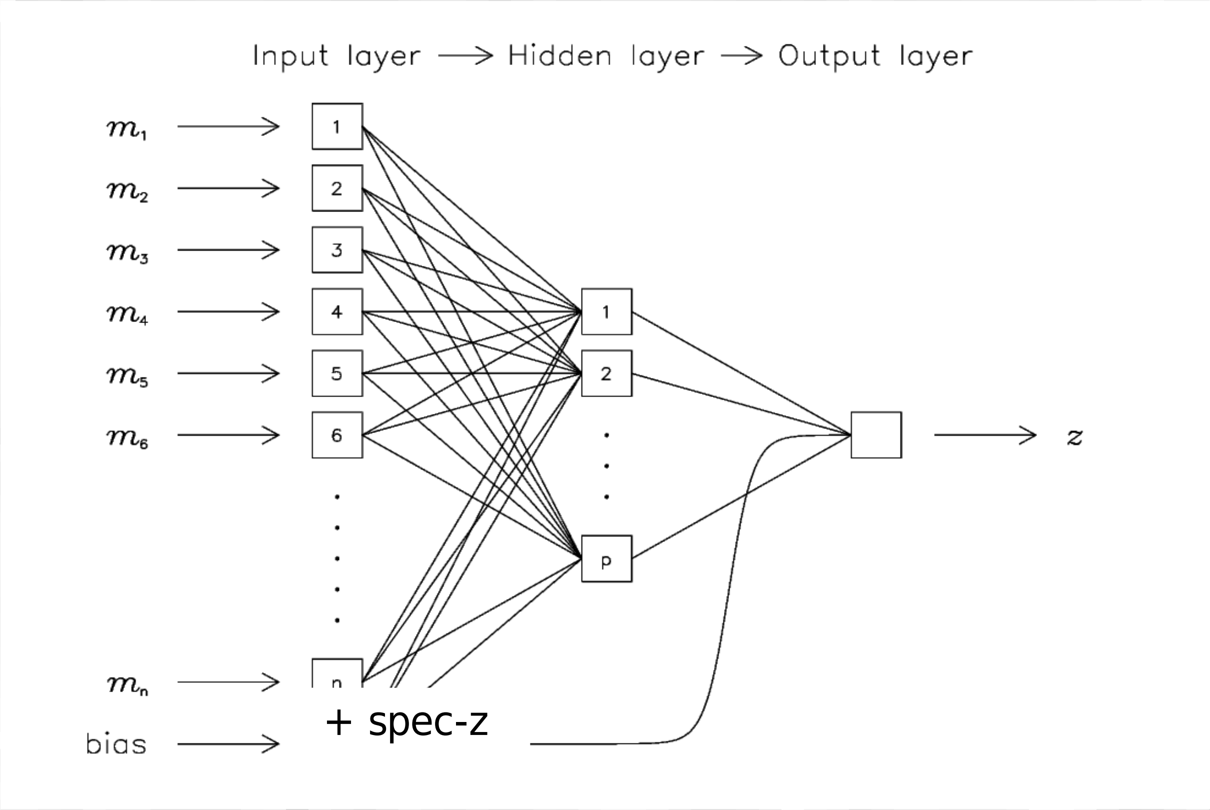
Source: Dark Energy Survey

- Redshifts can thus be **estimated** from multi-wavelength photometry: **photometric redshifts** (“photo-zs”) – for instance **using machine-learning***
- Photo-zs much **less *precise*** (scatter of ~10% or more) than spectro-zs but usually statistically ***accurate*** (overall bias in |zphot-zspec|~0)

**Photo-zs can also be estimated via spectral energy distribution (SED) fitting – workshop by Katarzyna Małek & Samuel Boissier next week*

Photometric redshifts with machine learning

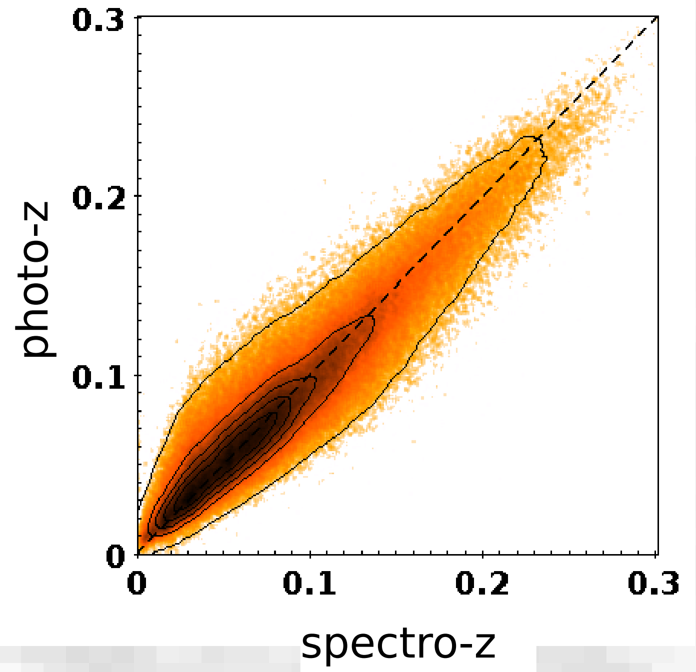
- **Machine learning** (ML) algorithms can be **trained** on **spectro+photo** data to derive best-fit **photo-zs** for a given set of passbands (regression problem)
- **Feature space** can include any quantities correlated with redshift: fluxes, sizes, colors...
- **Plethora of algorithms** applied: neural networks, random forests, support vector machines, Gaussian processes, ...
- ML photo-zs require **representative spectroscopic calibration** datasets (subsamples of the target photometric data) – usually the main limitation



A simple example of an
artificial neural network
scheme for photo-zs
(ANNz, Collister & Lahav 2004)

Photo-zs in practice in the "local" Universe: 2MASS Photometric Redshift catalog (2MPZ)

- We cross-matched three **all-sky photometric catalogs**:
  2MASS XSC (ground-based near-IR, J H K_s); **WISE** (space-based mid-IR, 3.4μm and 4.6μm) and **SuperCOSMOS** (digitised scans of photographic plates, B R I)
- We calculated **photometric redshifts** with an *artificial neural network* algorithm (Collister & Lahav 2004), trained on a representative spectroscopic subsample
- **2MPZ catalog** with **1 million galaxies**, **⟨z⟩=0.08**, covering **most of the sky**
- Some statistics of the photo-z estimates:
  → 1-sigma scatter $\boldsymbol{\sigma_{\Delta z} = 0.015}$
  → median error $\boldsymbol{|\Delta z|/z = 13\%}$
  → only **3% of outliers** $\boldsymbol{>3\sigma_{\Delta z}}$
- 2MPZ is **available for download** from http://surveys.roe.ac.uk/ssa/TWOMPZ

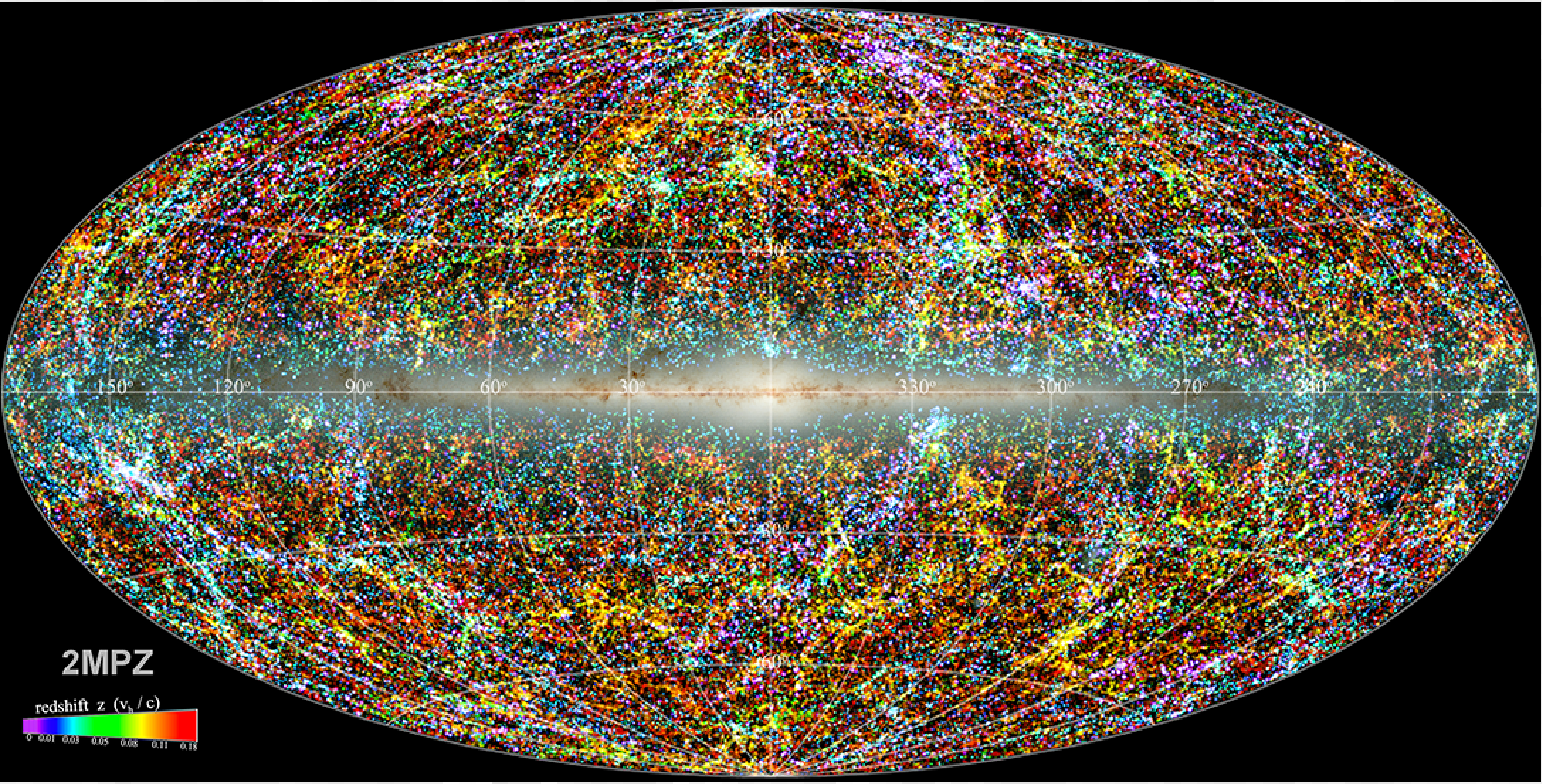


MB, Jarrett, Peacock, Cluver & Steward, ApJS, 2014

2MASS Photometric Redshift catalog

1 million galaxies in 3D

Color-coded by photometric redshifts

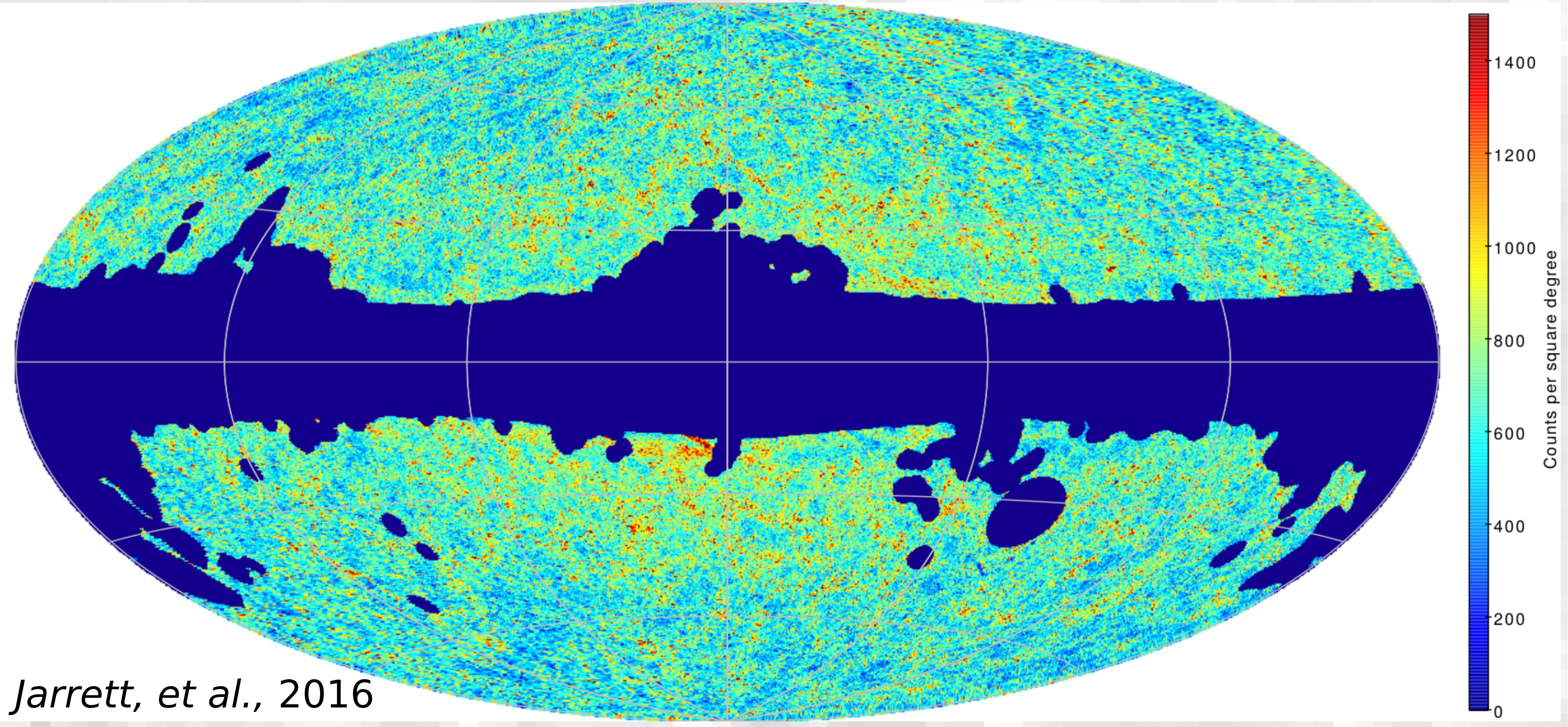


Plot by Tom Jarrett

Going deeper over 75% of sky:

20 million galaxies from WISE x SuperCOSMOS

- **All-sky galaxy sample much deeper than 2MASS:**
  Mid-IR **WISE** paired up with optical **SuperCOSMOS**, R_{AB}<19.5, $[3.4\mu]_{Vega}$<17 mag
- Cross-match at |b|>10° gives **170 million sources**, but mostly stars / blends
- A color-based **clean-up** of star blends leaves almost **20 million galaxies**
- Separate work on **automated selection of galaxies** (Krakowski et al. 2016)

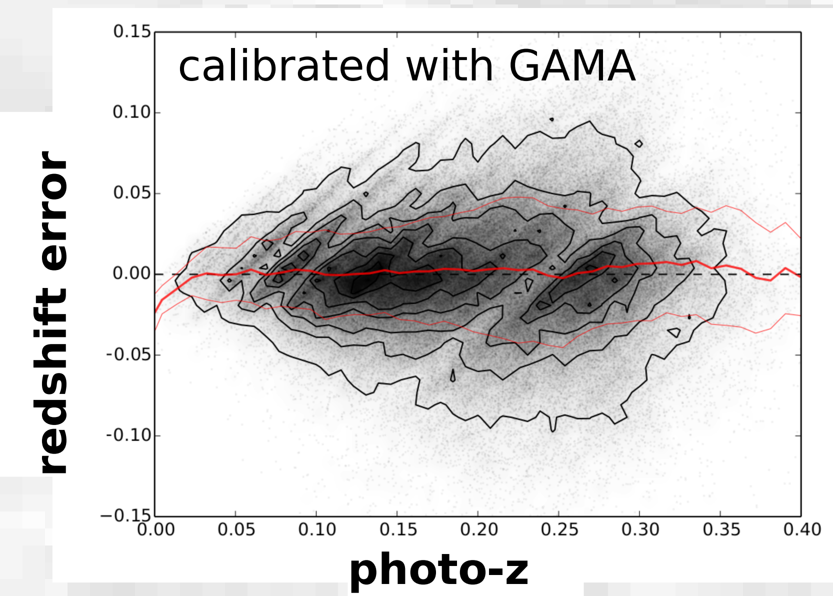


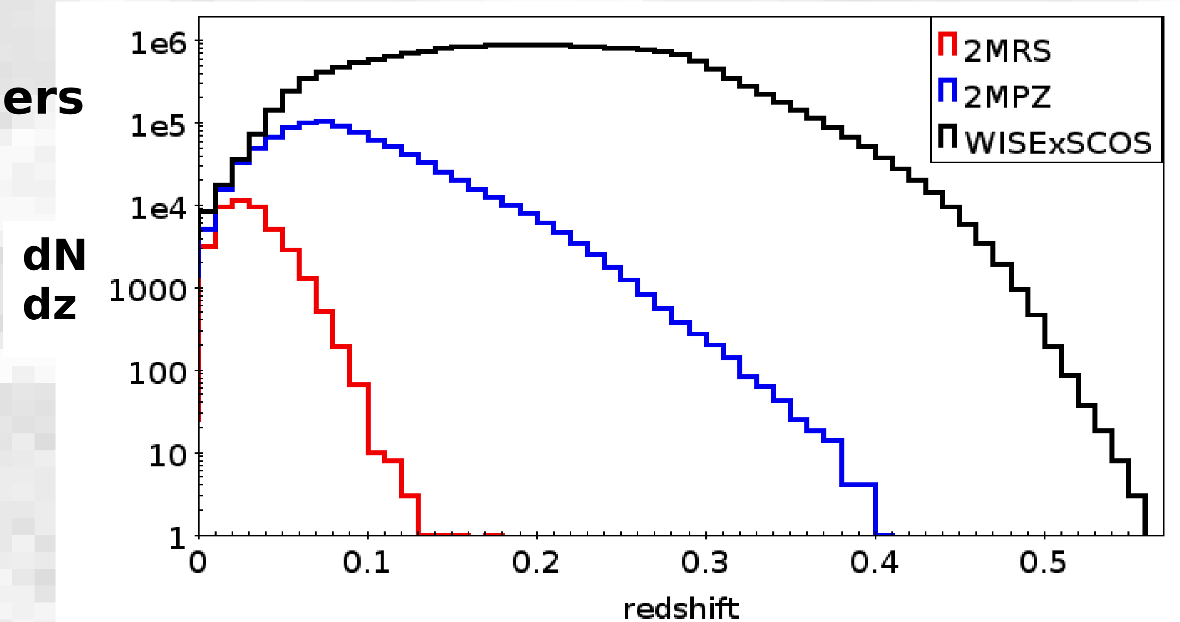
MB, Peacock, Jarrett, et al., 2016

The largest "all-sky" ~3D sample

20 million galaxies from WISE x SuperCOSMOS

- **WISE x SuperCOSMOS photo-z catalog:** <u>much</u> deeper than 2MPZ
- **Four photometric bands** for photo-z's: optical **B,R,** infrared **3.4 & 4.6 μm**
- Training set: **GAMA-II** spectroscopic (*r*<19.8 in 3 equatorial fields; Liske et al. 2015)
- **WIxSC** has median **z~0.2**, but probes the LSS **to z~0.4** on ~70% of sky
- Photo-z performance: $\boldsymbol{\sigma_{\Delta z} = 0.03}$**,** median **error 14%** and **3% outliers**





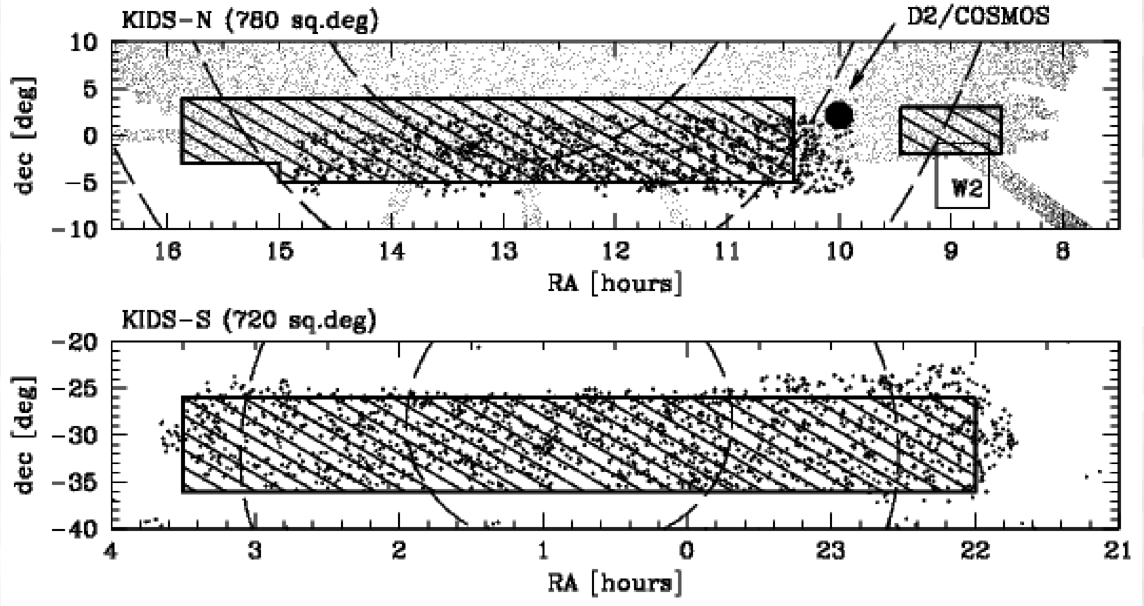
MB, Peacock, Jarrett, et al., ApJS, 2016

Data at http://ssa.roe.ac.uk/WISExSCOS

Kilo-Degree Survey (KiDS)

- **New era in imaging surveys**: excellent photometry at large depths and wide angles
  Kilo-Degree Survey, Dark Energy Survey, Hyper-SuprimeCam SSP

- **KiDS:** imaging of ~**1500 deg^2** in ***ugri*** bands at **depth *r*~24.9** (5σ) with **seeing<0.8"** in the *r* band



- **Data Release 3** includes ~50 million sources on **~450 deg^2** (full depth) (*de Jong et al.* 2017)

- Main science goal: cosmology with **weak gravitational lensing** but used for many other applications – **unprecedented depth/coverage/seeing** combination

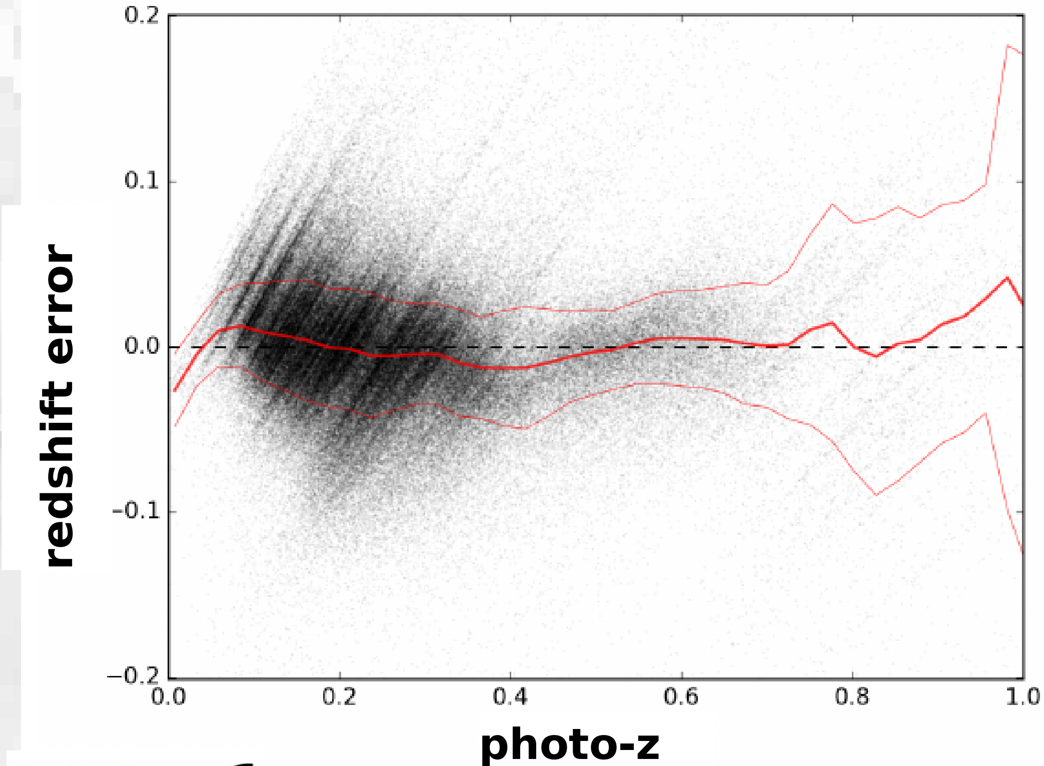
- KiDS area already covered with **VIKING** near-IR ***zyJHK$_s$*** to a similar depth as *ugri*

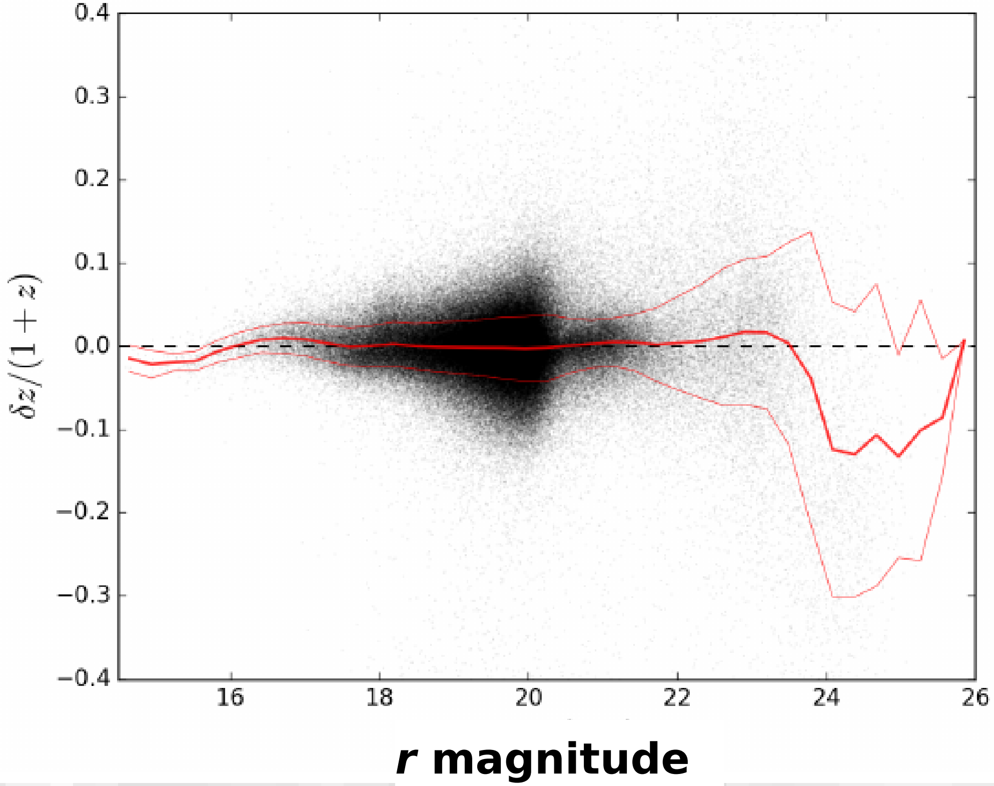
- Next KiDS releases will include 9-band photometry (from DR4: over ~1000 deg^2)

- **Photometric redshifts** crucial: most of KiDS galaxies do not have spectroscopy

KiDS machine-learning photo-zs DR3 full-depth catalog

- Machine-learning **photo-zs derived with ANNz2** (Sadeh et al. 2016)
- Magnitude-space **weighting of the training set** implemented (Lima et al. 2008)
- KiDS DR3 **public photo-z catalog** for all the sources with 4-band *ugri*
- Photo-*z*s judged **reliable to $z_{phot} < 0.9$ and $r < 23.5$**



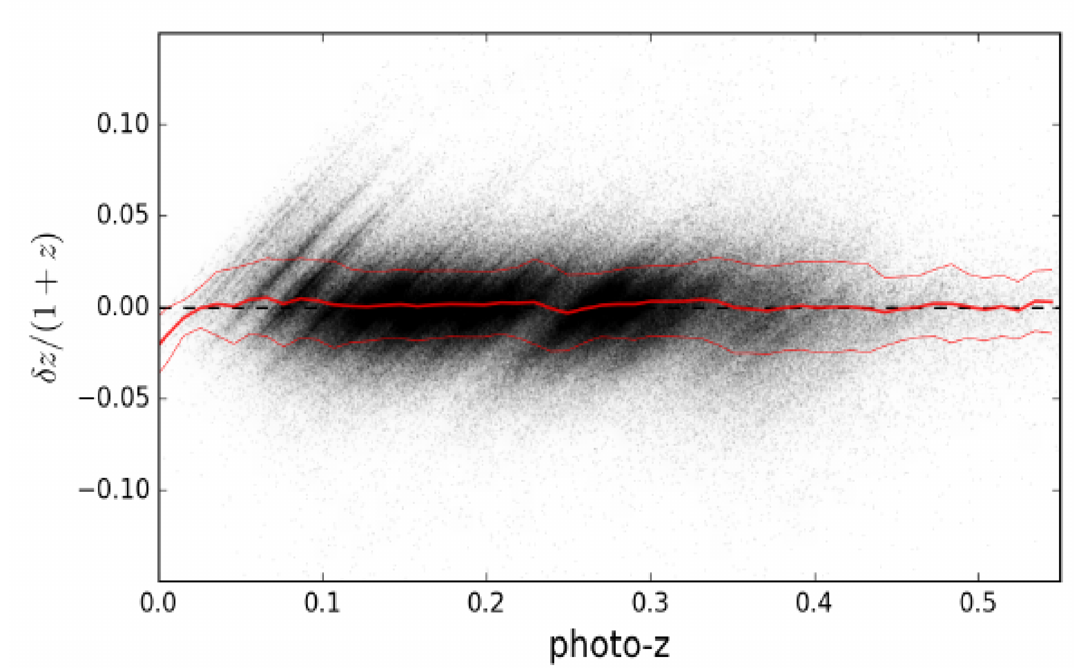


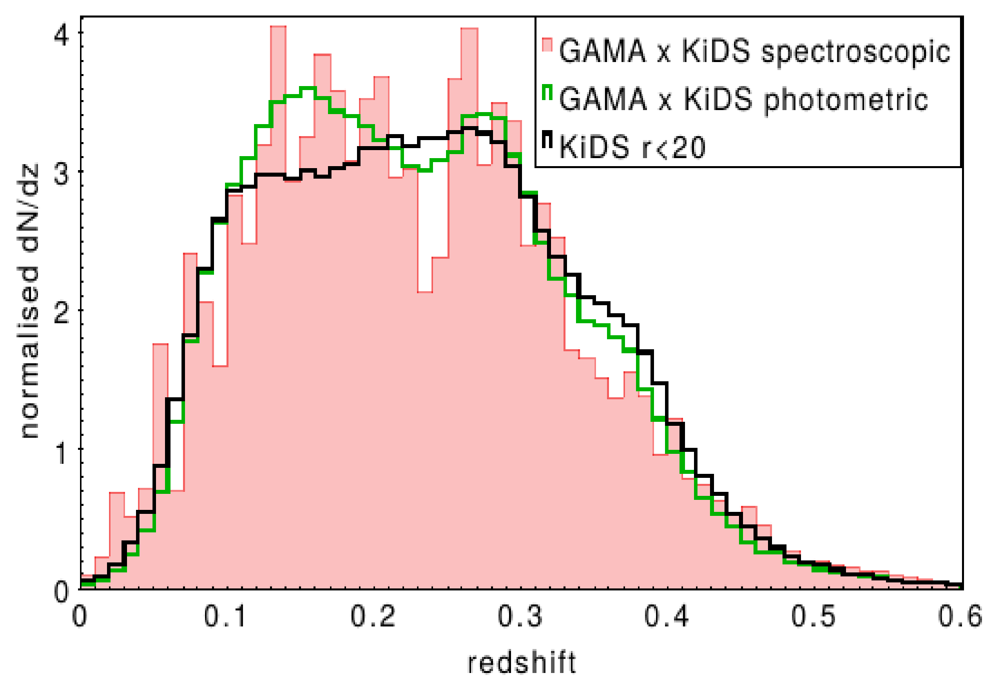
KiDS

MB, KiDS, GAMA & 2dFLenS teams, 2017

KiDS machine-learning photo-zs
Public GAMA-depth DR3 catalog

- ANNz2 **trained on GAMA** equatorial+G23
- Used KiDS *ugri* **magnitudes, colors, and semi-axes** as parameters
- Limited to ***r* < 20**: ~800,000 galaxies in DR3
- Very **precise and accurate** photometric redshifts ($\sigma_{dz/(1+z)}$=0.02)





MB, KiDS, GAMA & 2dFLenS teams, 2017

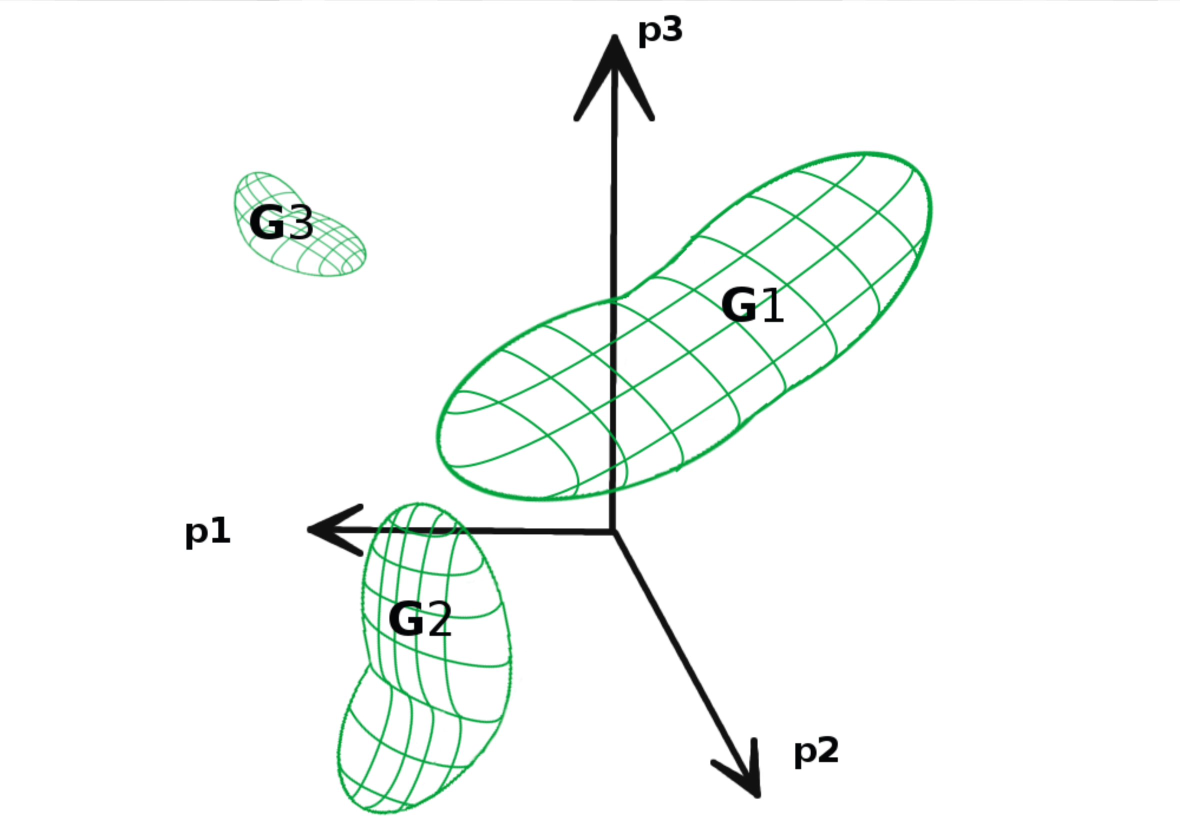
Machine learning for astronomical source classification

- ML algorithm learns to **recognize different types** of astronomical data (G); in the **supervised case** this is based on **training examples**

- ML works in a **parameter/feature space** (p) based on **discriminating properties** of the data

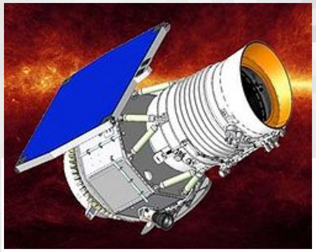
- In astronomy, the parameter space is usually source **fluxes** at various wavelengths and related **colors** – but could also be **redshifts**, **spectra** or **time-domain** information

- Popular algorithms: **support vector machines** (SVM), **random forests**, **neural networks**...

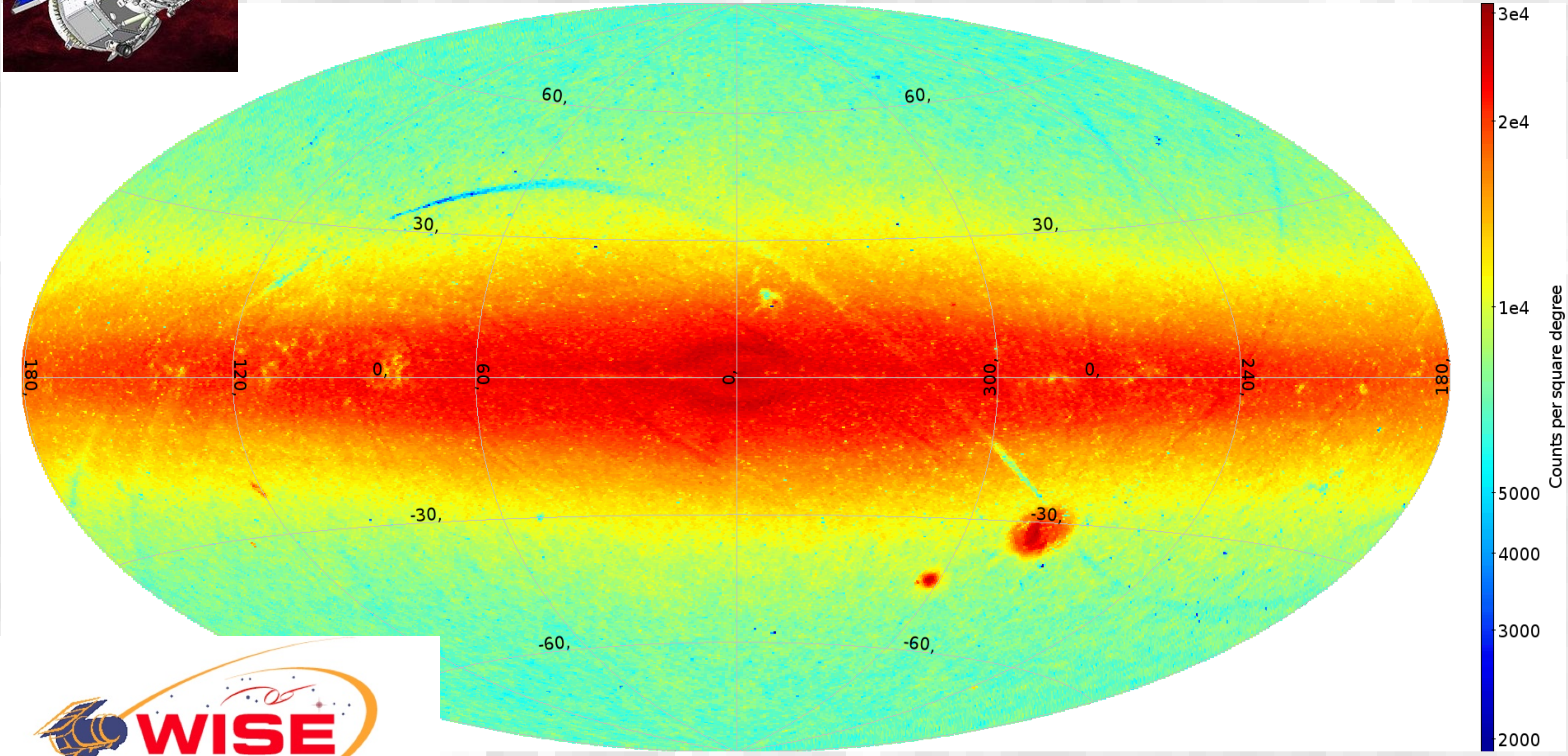


Slide courtesy of Dr. Aleksandra Solarz

An example of astronomical big data: Wide-field Infrared Survey Explorer (WISE)

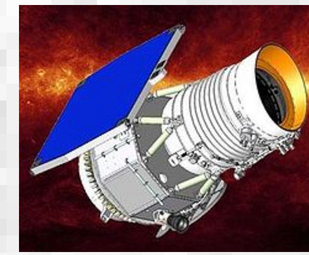


600 million sources within ~uniformity flux limits



The potential of

- **Wide-field Infrared Survey Explorer** (WISE) satellite data: **all-sky photometric catalogue** in 3.4, 4.6, 12 and 23 μm



- **One of the largest all-sky samples**: 750 million sources ...of which **~100 million** are **galaxies and QSOs**

- **WISE** itself is **much deeper** than 2MASS (by ~3 mag): another “layer” for all-sky cosmology (**galaxies even at z>1**; e.g. Jarrett et al. 2017)

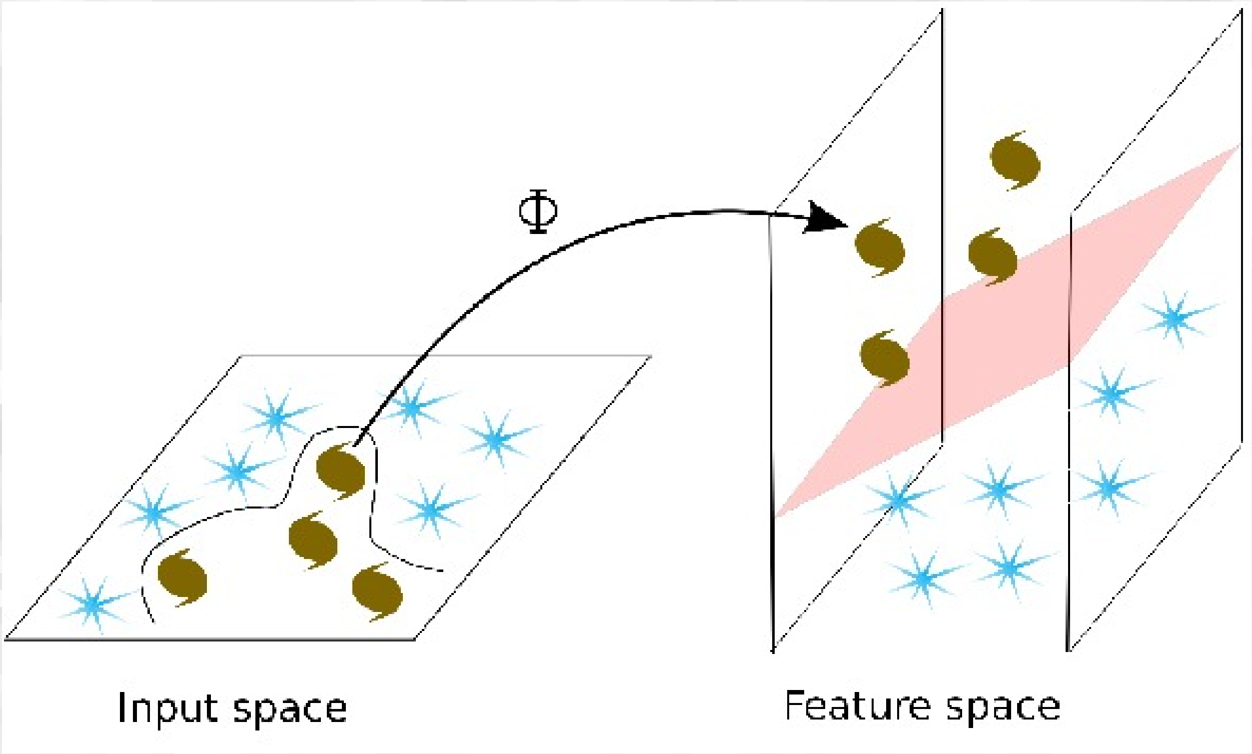
- Full **cosmological potential of WISE** still to be explored: galaxies very difficult to extract; stars dominate even at high latitudes

- **Difficulties in star/galaxy separation** due to blending (>6” resolution) and limited feature space (only 3.4 and 4.6 μm measurements at full depth)

Automated source classification with support vector machines

SVM: segregate data into categories based on training examples

- Use kernel functions to map input data onto a higher-dimensional feature space
- Find a hyperplane separating two classes in the feature space
- Output source classes assigned based on their position relative to the boundary



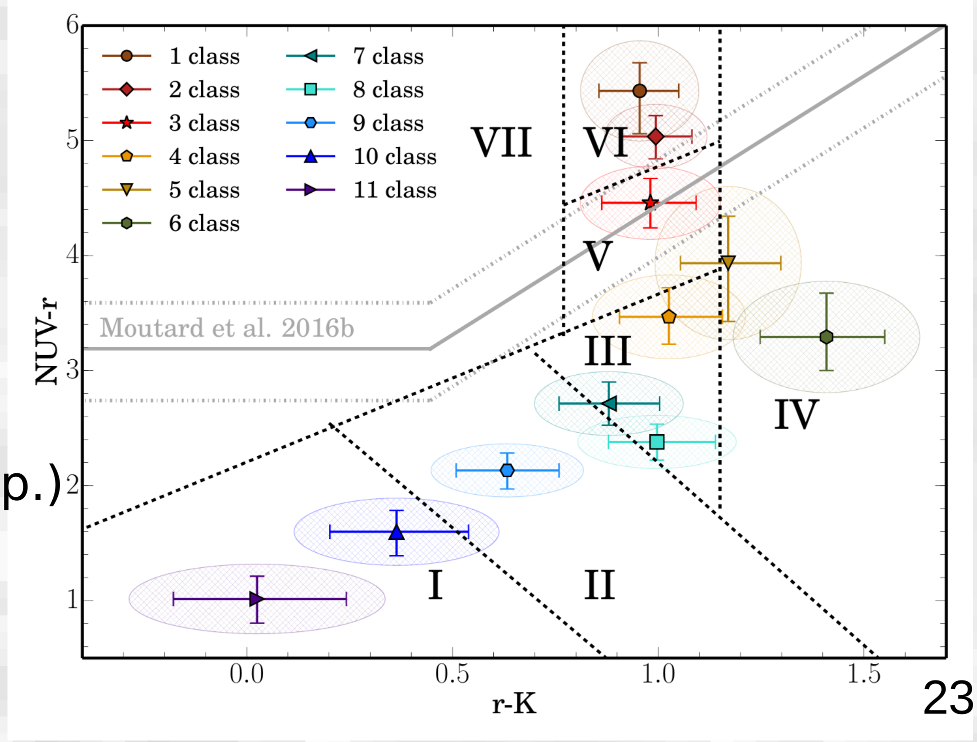
Małek, Solarz and VIPERS team, 2013

Slide courtesy of Dr. Aleksandra Solarz

Machine learning for source classification: applications to (big) astronomical data

Recent examples (subjective selection):

- First attempt at **3-class selection** (star/galaxies/quasars) in the all-sky WISE dataset of **over 300 million sources**, using **SVM** (Kurcz et al. 2016)
- **SVM-based galaxy selection** in WISE x SuperCOSMOS photometric data of **~50 million objects** (Krakowski et al. 2016)
- **Unsupervised classification** of galaxies in the VIPERS dataset, using **Fisher Expectation-Maximization** algorithm (Siudek et al. 2018)
- **Quasar search** in KiDS data using **random forests** (Nakoneczny et al. in prep.)
- Many more various applications by different teams to numerous datasets

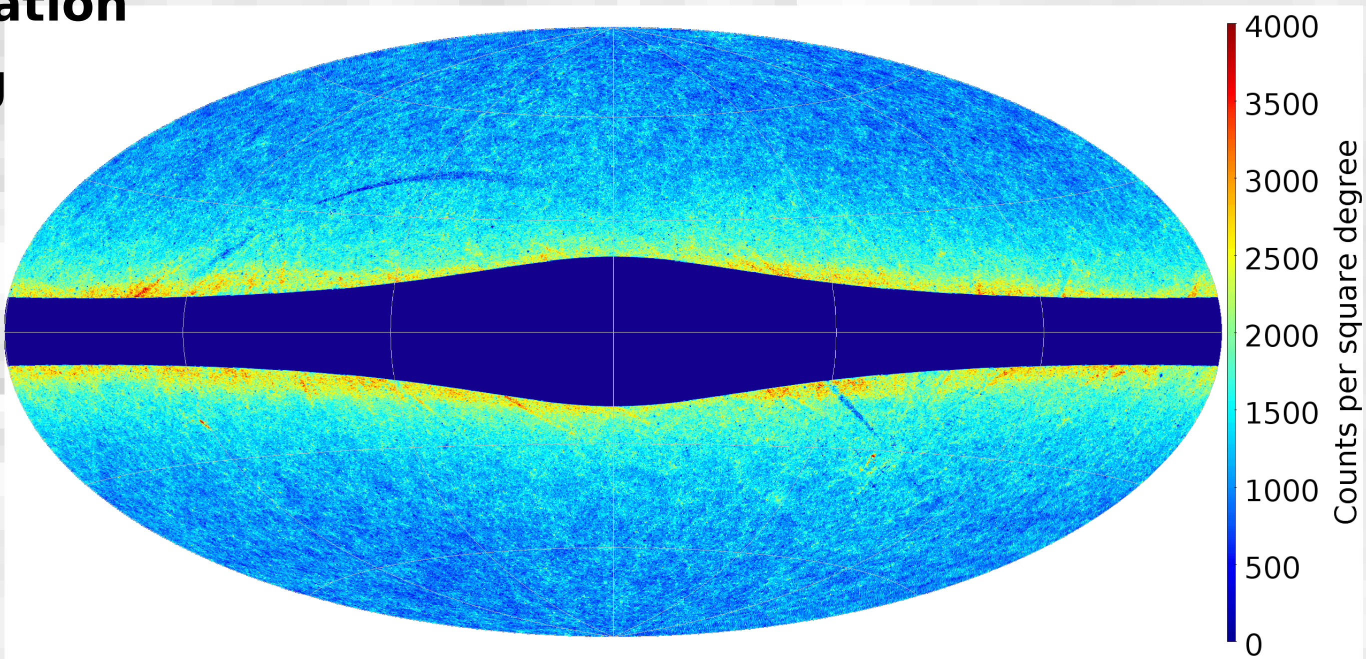


WISE source separation with support vector machines

- We used the SVM algorithm **trained on SDSS x WISE** spectroscopic sources (**stars / galaxies / quasars**)
- Current **results for W1<16** Vega (1 mag brighter than WISE flux limit) due to limitations of the training set (practically no SDSS galaxies at W1>16)
- **45 million galaxy candidates on ~80% of sky**

Kurcz et al. 2016

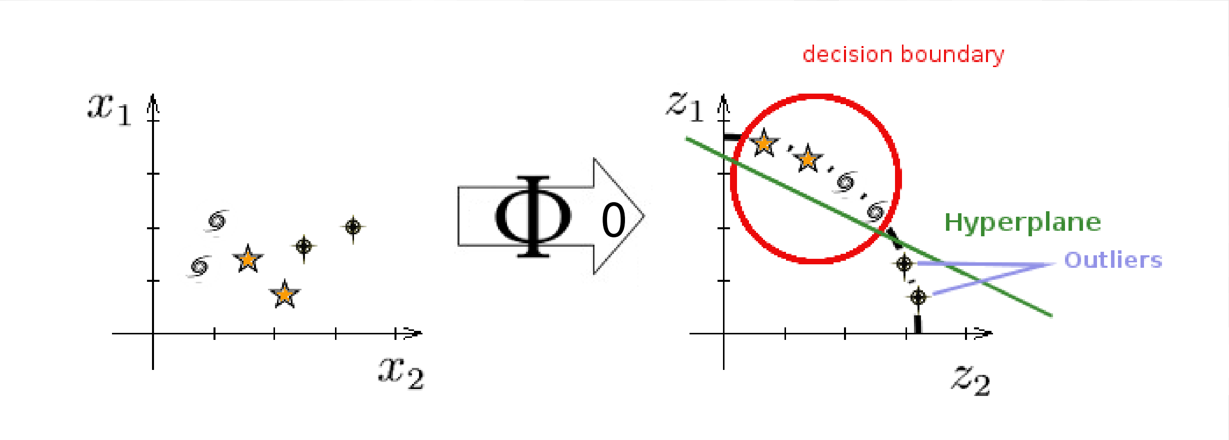
- Inevitable stellar **contamination at low latitudes** – blending due to 6” WISE beam
- Work in progress using **refined methods and extended samples** (Poliszczuk et al. in prep.)



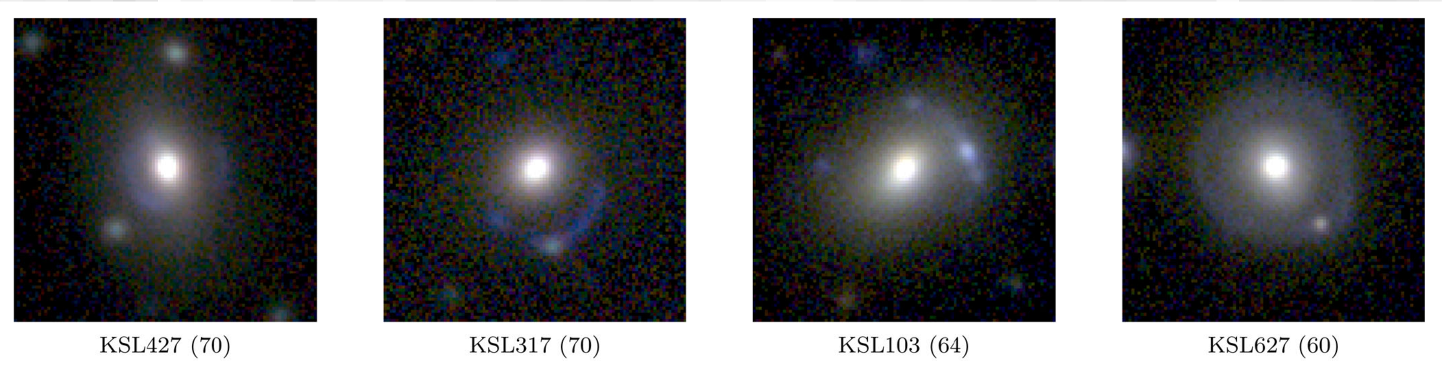
Machine learning for rare object search: applications to (big) astronomical data

Two recent examples (subjective selection):

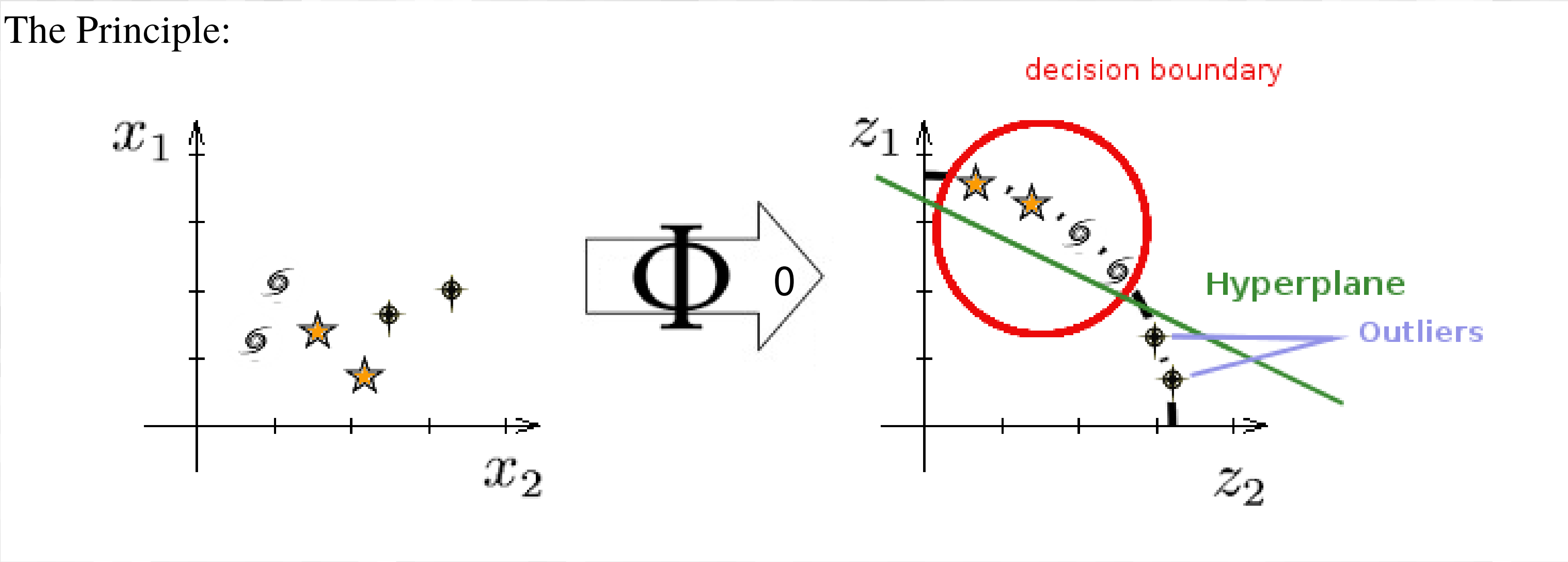
- A **One-Class-SVM** algorithm to search for **data anomalies** different from the training; first application to all-sky WISE (Solarz et al. 2017)



- A **convolutional neural network** application to imaging in Kilo-Degree Survey to search for **strong gravitational lenses** (Petrillo et al. 2017)



Novelty detection with One-Class Support Vector Machines

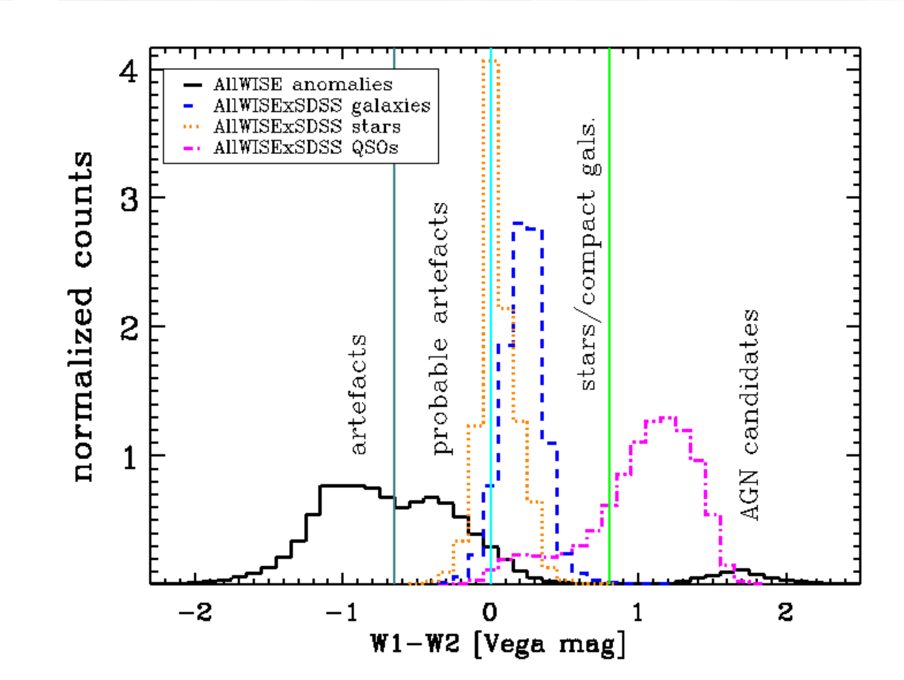


- Create one 'known' class (sources with e.g. spectroscopic labels)
- Map input data to a higher-dimensional parameter space
- Define a hypersurface encapsulating the expected sources
- Anything with 'unknown' patterns falls outside the hypersurface → Novelties

Slide courtesy of Dr. Aleksandra Solarz

Rare object detection in WISE with machine learning

- **Support vector machines** were used in "one-class" mode: training set as "known" sources, the rest as "unknown" (***anomalies***)
- Training data derived from optical SDSS → detected *anomalies* have **specific WISE mid-IR colors**



Solarz *et al.* 2017

- An all-sky population of **very "red" objects** [3.4μ]-[4.6μ] > 0.8 mag Vega
- Properties consistent with highly obscured dusty **quasars** at (maybe) large redshifts
- **Spectroscopic follow-up** needed to confirm their nature – observations in Chile starting soon!

The present and near future of wide-angle galaxy surveys

Some surveys happening now:

* **SDSS** (currently stage IV): galaxies, quasars (spectroscopy)

* Dark Energy Survey (**DES**): optical photometry on 5000 deg^2

* Kilo-Degree Survey (**KiDS**): precise optical and near-IR (**VIKING**) photometry on 1500 deg^2 (ESO)

* **H**yper **S**uprime-**C**am SSP Survey: excellent optical and NIR photometry on 1400 deg^2 (Japan+Taiwan+Princeton)

* and many, many others

Terabytes of data

Near and more remote future of wide-angle galaxy surveys

Planned surveys (examples):

- **TAIPAN** – spectroscopy of ~2 mln. galaxies at z<0.4 (from 2018/19)
- Dark Energy Spectroscopic Experiment (**DESI**) – spectroscopy of ~30 million galaxies (from 2018?)
- Square Kilometer Array (**SKA**) – array of radiotelescopes in South Africa and Australia; millions of galaxies at (emitted) 21 cm wavelength (from ~2020s?; precursors already operating/built)
- **Euclid** – European space-bourne near-IR telescope; slitless spectroscopy and deep photometry on ~1/4 of the sky; 2020s(?)
- Large Synoptic Survey Telescope (**LSST**) – photometric survey on an 8.4-m telescope in Chile; ~40 billion(?) sources (~2020?)

Petabytes of data

Astronomers as big data specialists

- The sizes and complexity of future astronomical datasets will require more automatised approaches towards data analysis
- This is happening already now in some cases
- Machine learning tools will be essential
- “Standard” supervised learning now, but unsupervised l. as well as deep l. may (will?) take over
- ML for photo-zs well settled and new ideas being developed (e.g. derivation of probability density functions)
- ML for astronomical classification still in its infancy – the best time to contribute significantly!