This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 721463

SUNDIAL
H2020 Innovative Training Network

DATA SCIENCE UNINA

ARTIFICIAL INTELLIGENCE IN ASTROPHYSICS

I.e. Some Machine Learning successes and problems

Giuseppe Longo
DATA SCIENCE LAB
University of Napoli Federico II - Italy
2019 M. Kingsley fellow - CALTECH
CINI Consortium

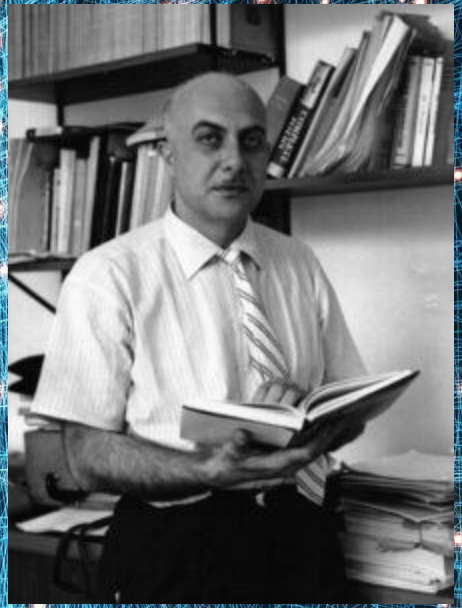
longo@na.infn.it

Special Thanks to the group:

Massimo Brescia
Stefano Cavuoti
Michele delli Veneri
Giuseppe Longo
Oleksandra Razim
Giuseppe Riccio
Olena Torbaniuk

\+ & Many great students

Eduardo Caianiello (1921-1993)



Prime Video: La | data science tes | Outline of a the | Eduardo R. Caia | Eduardo Caianie | Presentazioni m | 2019_Ukraine_c | DSF_INFN - Go | infn_redshift - C

sciencedirect.com/science/article/abs/pii/0022519361900467

App | Home | Data Science M.D. | Giuseppe | YouTube | Oggi: Todoist | Arxiv Sanity Preser... | My Account | WIRED | CuriosityStream -... | overleaf | Riviste & Blog | Altri Preferiti

Get Access | Share | Export | Search ScienceDirect | Advanced

Outline

Abstract

References

ELSEVIER

Journal of Theoretical Biology

Volume 1, Issue 2, April 1961, Pages 204-235

Outline of a theory of thought-processes and thinking machines

E.R. Caianiello

Show more

https://doi.org/10.1016/0022-5193(61)90046-7

Get rights and content

Abstract

Thought-processes and certain typical mental phenomena are schematized into exact mathematical definitions, in terms of a theory which, with the assumption that learning is a relatively slow process, reduces to two sets of equations: “neuronic equations”, with fixed coefficients, which determine the instantaneous behavior, “mnemonic equations”, which determine the long-term behavior of a “model of the brain” or “thinking machine”. A qualitative but rigorous discussion shows that this machine exhibits, as a necessary consequence of the theory, many properties that are typical of the living brain: including need to “sleep”, ability spontaneously to form new ideas (patterns) which associate old ones, self-organization towards more reliable operation, and many others. Future works will deal with the quantitative solution of these equations and with concrete problems of construction—things that appear reasonably feasible. With a transposition of names, this theory could be

Recommended articles

Nanosheet-constructed transparent conducting...
Journal of Alloys and Compounds, Volume 561, 2013, ...
Purchase PDF | View details

The LIML estimator has finite moments!
Journal of Econometrics, Volume 157, Issue 2, 2010, p...
Purchase PDF | View details

Abstract phase space networks describing reacti...
Physica A: Statistical Mechanics and its Applications, V...
Purchase PDF | View details

1 2 Next

Citing articles (274)

Article Metrics

Citations

Citation Indexes: 307

Captures

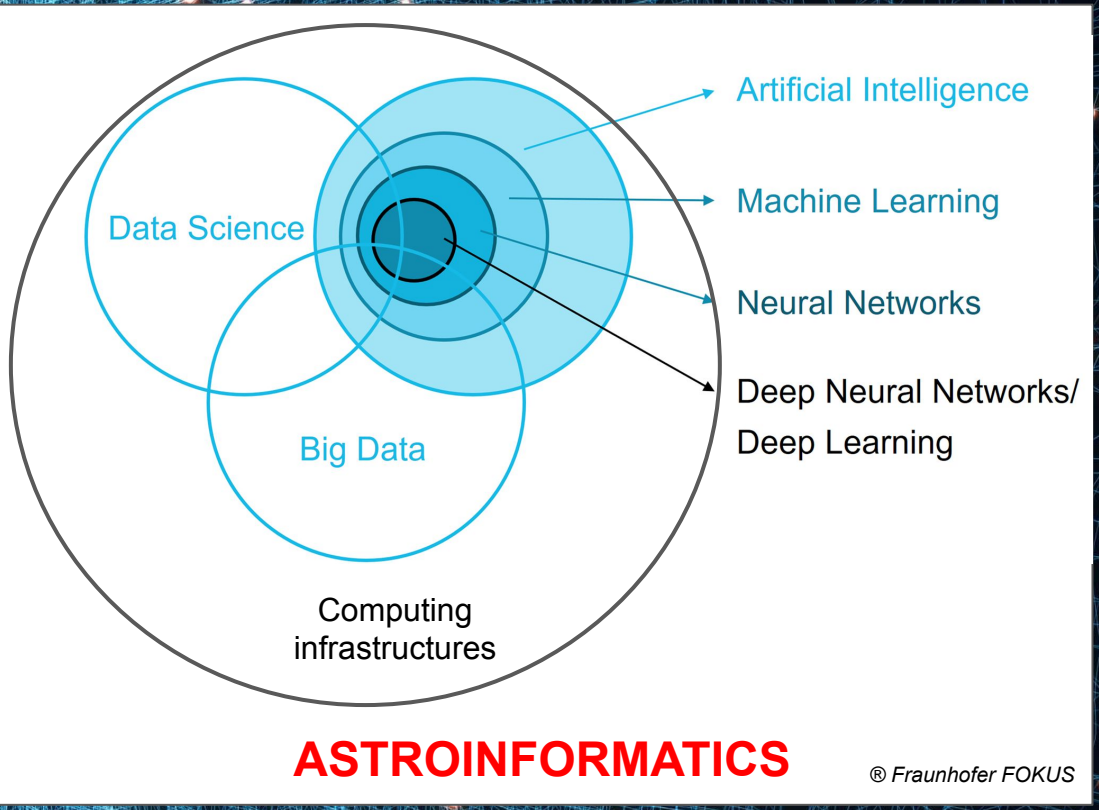
Exports-Saves: 2

Readers: 31

Mentions

Feedback

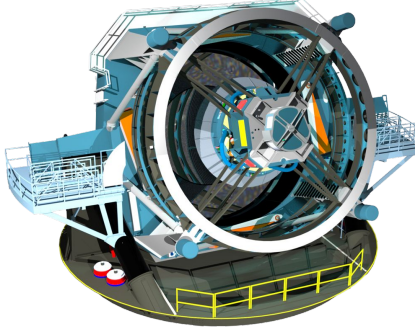
Some considerations



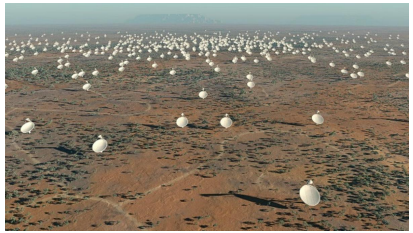
- **In the Physics domain Artificial Intelligence is just a buzzword** (recently resurrected for marketing purposes. Bureaucrats like it.)
- **Deep learning** is a subset (rather old stuff) of machine learning (became popular because of google).
- Machine learning, data mining, KDD, and statistical pattern recognition are different "nuances" of the same stuff
- More than 100 new methods of ML every year

Some Examples

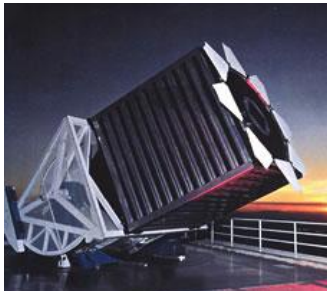
Survey	Volume	Velocity	Variety
SDSS *Sloan Digital Sky Survey*	50 TB	200 GB/day	Images multiband catalogues, spectra
GAIA	100 TB	40 GB/day	Images, catalogues, spectra
PANSTARRS *Panoramic Survey Telescope and Rapid Response System*	5 PB (5 years)	5 TB/day	Images multiband, catalogues
LSST *Large Synoptic Survey Telescope*	130 PB (10 years)	10 TB/day	Images multiband , catalogues, multi-epoch
SKA *Square Kilometer Array*	3 ZB	10 PB/day (raw) 150 TB/day (processed)	Radio images, multi-l multi-epoch

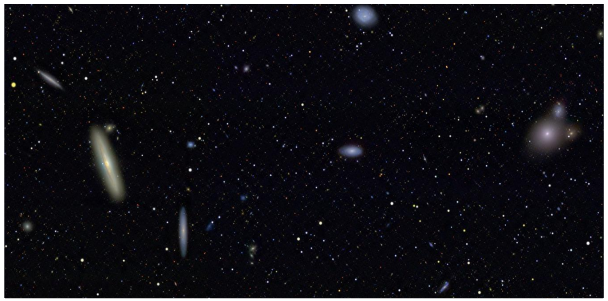


LSST
3.2 Gpixels camera



In all sciences 99.9% of the data will never be seen by humans



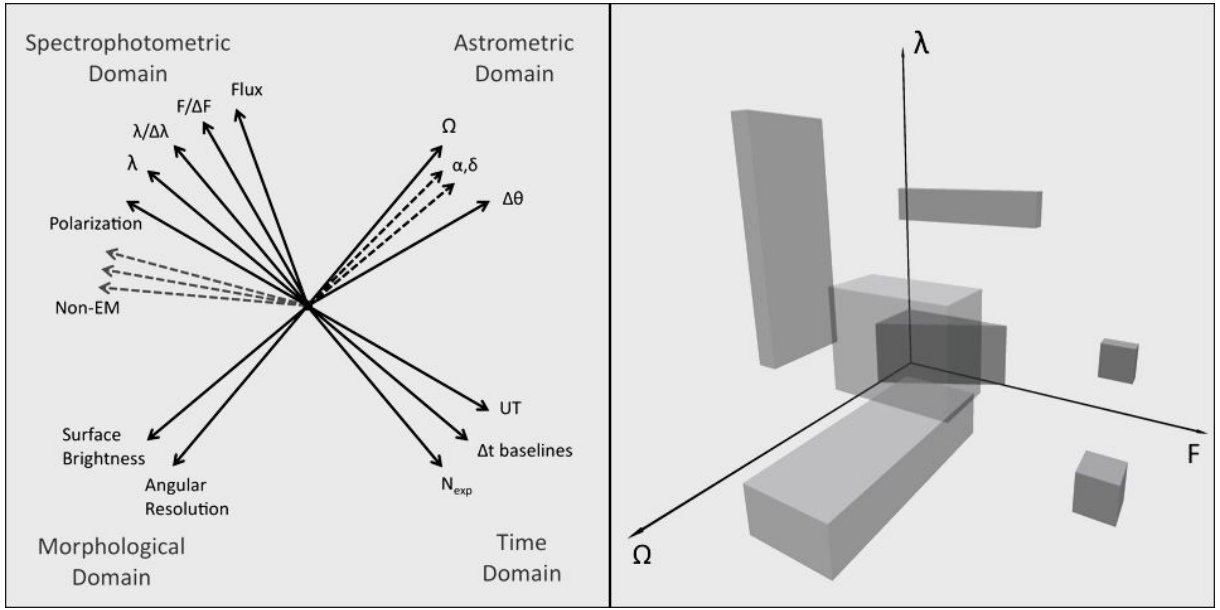


Modern sky surveys obtain ~ 10^{12} – 10^{15} bytes of images,
catalogs ~ 10^{8} – 10^{9} objects (stars, galaxies, etc.),
and measure ~ 10^{2} – 10^{3} numbers (features) for each
Time series 10^10 per year

- Both data volumes and data rates grow exponentially, with a ***doubling time ~ 1.5 years***
  - Even more important is the growth of ***data complexity***

The observed parameter space (OPS) has exploded not only in size but also in complexity

The OPS axes are defined by the observable quantities



Every observation, included, carves out a hypervolume in the OPS

The Growth of Astrophysical Understanding

A stroll through three millennia of astronomical speculation and discovery reminds us that inspired guesses are not enough. Progress comes primarily from the introduction of new observational and theoretical tools.

Martin Harwit

Perhaps the most remarkable aspect of the growth in our understanding of the universe is that we understand anything at all. Beyond the obvious regularities of the seasons, the Assyrians noted, as early as 700 BC, that the planets appeared to move in a complex semiregular pattern and that solar eclipses were possible only at the new moon, whereas lunar eclipses occurred only at the full moon. But what did all that tell the ancients about the structure of the universe?

Around 250 BC, the Greek natural philosopher Aristarchus of Samos worked out the distance of the Moon and its size. He proposed a method for determining the Sun's distance, but he was able to conclude only that the Sun was much farther away than the Moon and much larger than Earth. That led him to postulate, 18 centuries before Nicolaus Copernicus, that Earth revolves around the Sun.[1]

Aristarchus's theory was largely discredited, especially by Claudius Ptolemaeus of Alexandria. Ptolemy's *Almagest*, which appeared in about 150 AD, dominated Western astronomical thought for a millennium and a half. Ptolemy argued that Earth could not be rotating. Rotation, he thought, would throw anything not firmly attached off the surface, and "animals and other weights would be left hanging in the air." Moreover, Earth's rotation would be so fast that "never would a cloud be seen to move toward the east."[2]

That sounds quaint today, but it wasn't illogical. Ptolemy was a great scientist. The first lesson in astrophysics, however, is that every cosmic phenomenon is governed by competing effects—in this case, gravity, centrifugal forces, and friction. Unless we know the order of magnitude of each, we are likely to draw wrong conclusions.

The observers

When Copernicus revived the notion of a heliocentric system in 1543, he could offer no observational confirmation. The ground for a final resolution had to be prepared by Tycho Brahe (1546–1601), the greatest of the pre-telescope observers. Tycho constructed astronomical instruments more precise than any previously known. Over a 20-year period, he assembled the most accurate, systematic data that had ever been compiled on the positions of the planets.

The young Johannes Kepler, a theorist if ever there was one, dogged Tycho, intent on getting his hands on the data, which the great observer was jealously guarding so he could deduce the orbits of the planets himself. When Tycho was banished in 1597 from his island observatory in Denmark and sought political refuge in Prague, Kepler followed him. But it was not until after Tycho's death that Kepler inherited and began analyzing the data.[3]

One sees parallels to today's theorists impatiently seeking to get an early look at the data from the Wilkinson Microwave Anisotropy Probe's mapping of the cosmic microwave background. The WMAP data were, until just a few months ago, embargoed pending the publication of a full year's set of observations.[4] (See PHYSICS TODAY, April 2003, page 21.) As soon as the data were released, new theoretical analyses began to appear within days on the World Wide Web.

Kepler reduced Tycho's data and arrived at his three laws of planetary motion:

- The planets move in elliptical orbits—rather than in circles and epicycles.
- The rate at which a planet sweeps out area within its orbital ellipse is constant.
- The periods of the planetary orbits increase as the $^{3}/_{2}$ power of their semimajor axes.

The last of these findings was the first quantitative relationship between two observational parameters in astronomy. It constituted what one would call a well-posed question: Why does Kepler's third law hold?

With the advent of the astronomical spyglass in 1609 (the word *telescope* was not coined until the following year), Galileo Galilei quickly discovered an extraordinary new set of phenomena: mountains on the Moon, moons orbiting Jupiter, and the moonlike phases of Venus. To Galileo, those three observations meant that Earth is just one of the planets, all of them orbiting the Sun. For him, that clinched the Copernican theory. The Church, however, forbade Galileo to teach the theory and eventually confined him to house arrest until his death in 1642.

Why did it take until the 17th century for the great discoveries of Kepler and Galileo to come about? Today the answer is clear. Tycho's precision instruments and the spyglass, invented in Holland in 1608 and, a year later, improved by Galileo and pointed at the heavens, provided observational data that had simply been unavailable before.

Although Tycho's instruments gave the best positional data ever assembled, they were still limited by the abilities of the unaided eye, which cannot discern the moons of Jupiter. Galileo's telescopes provided a breakthrough in angular resolution and light-gathering power, a path astronomers are still treading as they build ever larger telescopes and interferometers.

A brief history of instrumentation and its successes illustrates the ability of new instruments to promote astronomical discovery.[5] (See also my article in PHYSICS TODAY November 1981, page 172.) There is a vast range of wavelengths, from the radio domain to the very highest gamma-ray energies, about which Tycho could know nothing. He had only his eyes to rely on, and they merely covered the minuscule visual portion of the electromagnetic spectrum. The naked eye provides a light-gathering aperture of only a few millimeters and a resolution of about an arcminute.

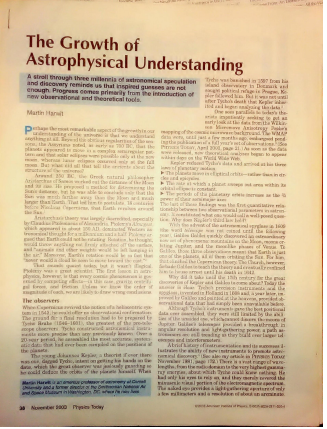
Martin Harwit is an emeritus professor of astronomy at Cornell University and a former director of the Smithsonian National Air and Space Museum in Washington, DC, where he now lives.

Why to understand the OPS is particularly important in science?

Because science is based on observations and experiments…. and

The history of discoveries can be reconstructed in terms of better coverage or better sampling of the observed parameter Space

M. Harwit, Physics Today, 2003

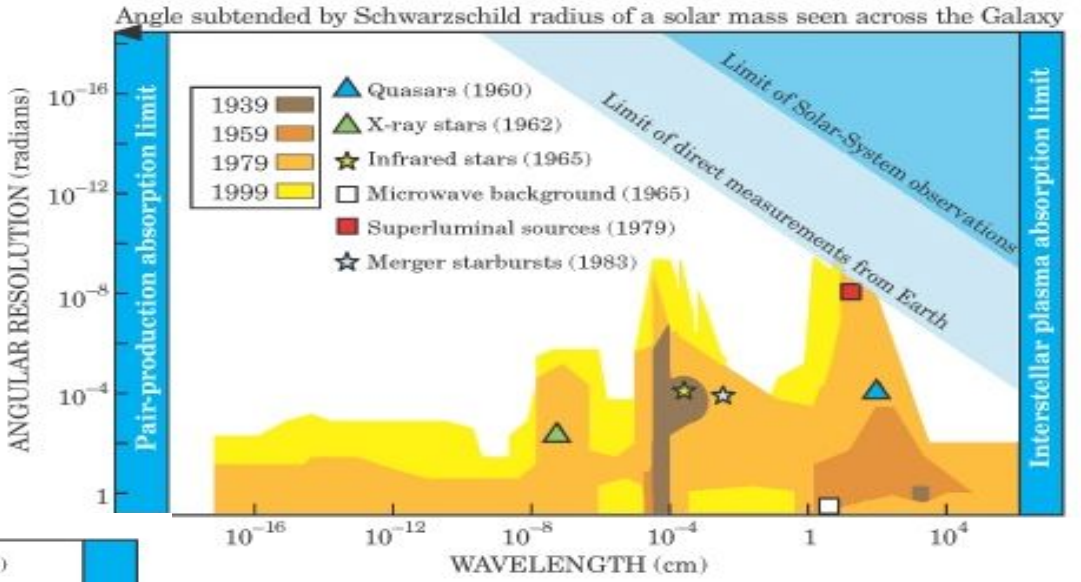


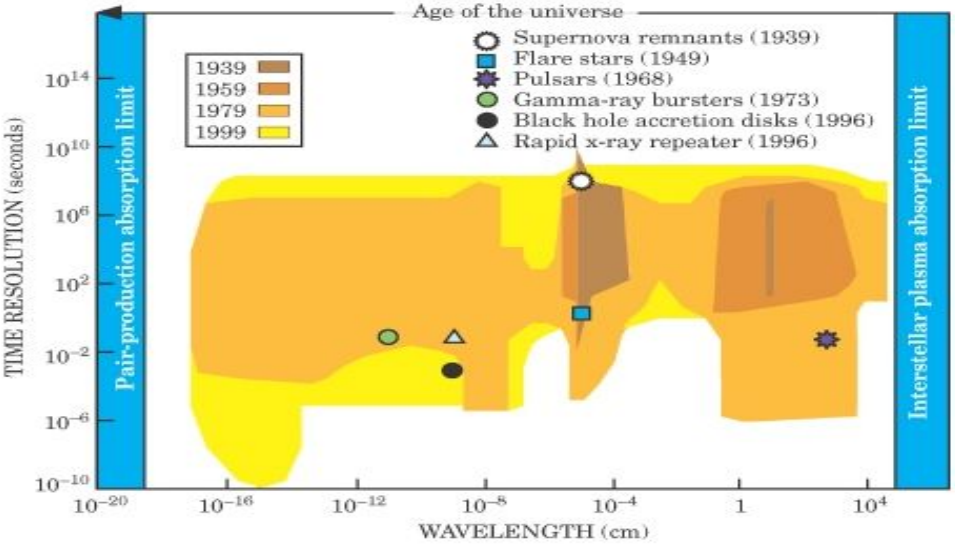
The Growth of Astrophysical Understanding

A stroll through three millennia of astronomical speculation and discovery reminds us that inspired guesses are not enough. Progress comes primarily from the introduction of new observational and theoretical tools.

Martin Harwit

M. Harwit, Physics Today, 2003

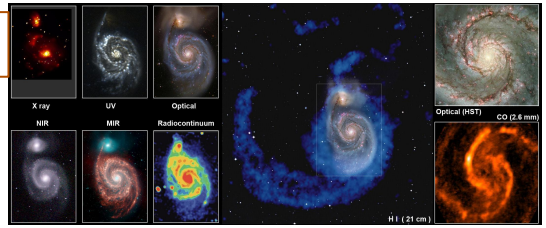


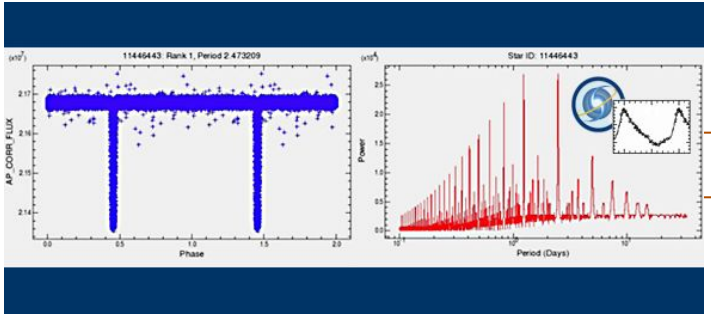


Where is the next methodological and technological advance?

In astronomy the MOST INTERSTING SCIENCE WILL COME FROM…

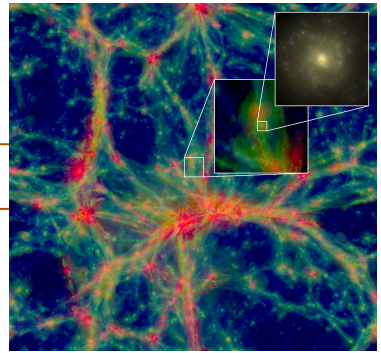
1. Cross-correlating data obtained at different wavelenghts

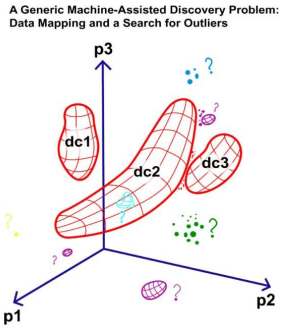




2. Studying the temporal behavior of sources

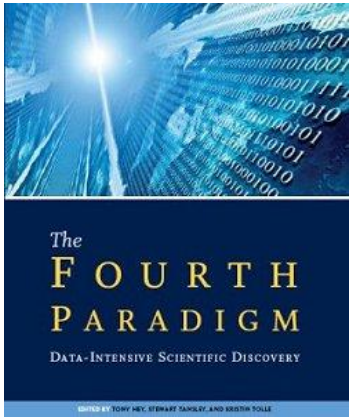
3. Possibility to compare the data with the outcome of large simulations





4. **Looking in the OPS for higher (than 3) dimensionality pattern and trends**

A METHODOLOGICAL SHIFT IS TAKING PLACE IN ASTRONOMY and in SCIENCES in general



The four legs of modern science

1. **Experiment** (ca. 3000 yrs)

2. **Theory** (few hundreds yrs) mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)

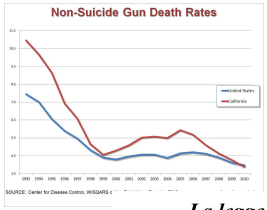
3. **Simulations** (few tens of yrs) Complex phenomena

4. **Data-Intensive science** (now!!!)

Physical, social, economic, biological laws are derived from data patterns (empirical laws)

$$f(x,y,z) = 0$$

No empirical law depends on more than 3 independent parameters !!!



La legge dei gas ideali

Unendo le leggi di Boyle, Charles e Avogadro si ha un'unica legge, approssimativamente valida per tutti i gas

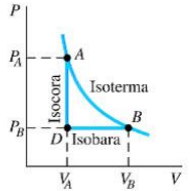
$$PV = nRT$$

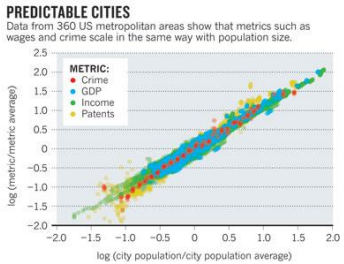
la costante R, costante dei gas, ha il medesimo valore per ogni gas (cioè è «universale»).

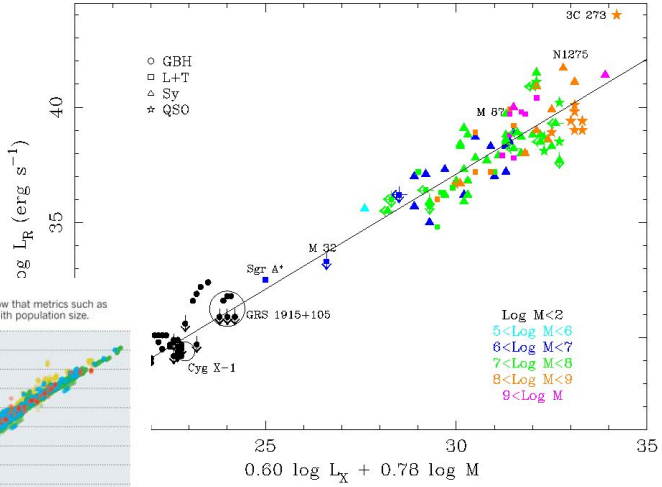
Alla temperatura di 0°C (273,15 K) e alla pressione di 1,00 atm, 1,00 mol di un gas occupano il volume di 22,414 L,

R = 0.08206	L . atm/(K . mol)
R = 62.37	L . Torr/(K . mol)
R = 8.314	L . Pascal/(K . mol)

$$\mathrm{SFR} \propto \begin{cases} (1+z)^{\beta} & \text{for } z < 1\,, \\ (1+z)^{\alpha} & \text{for } 1 \le z < 5\,, \\ 0 & \text{for } z \ge 5\,. \end{cases}$$



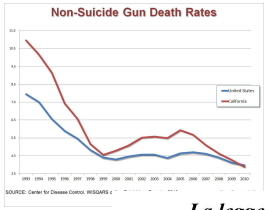




Physical, social, economic, biological laws are derived from data patterns

$$f(x,y,z) = 0$$

No empirical law depends on more than 3 independent parameters !!!



La legge dei gas ideali

Unendo le leggi di Boyle, Charles e Avogadro si ha un'unica legge, approssimativamente valida per tutti i gas

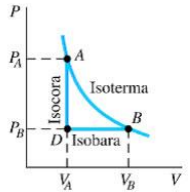
PV= nRT

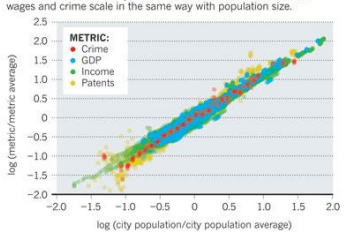
la costante R, costante dei gas, ha il medesimo valore per ogni gas (cioè è «universale»).

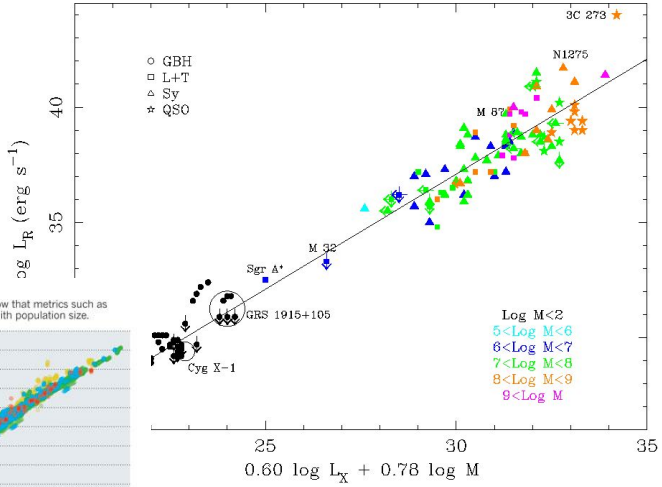
Alla temperatura di 0°C (273,15 K) e alla pressione di 1,00 atm, 1,00 mol di un gas occupano il volume di 22,414 L,

R = 0.08206	L . atm/(K . mol)
R = 62.37	L . Torr/(K . mol)
R = 8.314	L . Pascal/(K . mol)

$$\mathrm{SFR} \propto \begin{cases} (1+z)^{\beta} & \text{for } z < 1\,, \\ (1+z)^{\alpha} & \text{for } 1 \le z < 5\,, \\ 0 & \text{for } z \ge 5\,. \end{cases}$$







A simple universe… or rather **an intrinsic human bias …**

… affecting our knowledge and our understanding of the physical laws

The astroinformatics field is old(est in physics) **and exploding (2003 vs 2019)**

2003 - Special issue of the International Journal of Neural Networks on "Neural Network Analysis of Complex Scientific Data", Eds. Tagliaferri R., Longo G., D'Argenio B.
2010 - N.M. Ball and R.J. Brunner, 2010, arXiv:0804.3413

2019 - Focus Issue on Machine Learning in Astronomy, Publications of The Astronomical Society of the Pacific, Eds. Longo G., Merenyi E. & Tino P.
2019 - Papers presented at "Astroinformatics 2019", Pasadena July
2019 - C. Fluke & Jac obs, review (in press)

TASKS AND SCIENCE CASES - I

Task	2003	2003-2009	2009-2019	superv.	Unsuperv.	DL	SOME METHODS
S/G separation	yes	Yes	yes	Y	y	?	ANN, CNN
Galaxy properties Morphology Properties SFR Evolution	yes	yes	yes	Y	y	y	ANN, SVM, PPS; CNN, +many other
Spectral classification	yes	yes	yes	Y	y	y	ANN, SVM, RF
Image segmentation	yes		yes	y	y	y	ANN, GAN
Noise removal	yes		yes	Y	y	no	SVM, ANN, CNN, GAN, SLTN + many
Photometric redshifts (galaxies)	yes	Yes	yes	Y	y	y	SVM, ANN, RF, CNN, KNN, + many other
Variable objects	yes	Yes	yes	y	y	y	SVM, DT, ANN, RF, CNN + other
Stellar evolution models	yes		yes	y	n	n	ANN
Outlier detection		Yes	yes	Y	y	y	ANN, RF, CNN + all
Search for AGN		Yes	yes	Y		y	SVM, ANN, CNN + many

Task	2003	2003-2009	2009-2019	superv.	Unsuperv.	DL	Notes (most used methods)
Solar activity		yes	yes	Y	n	n	
Galactic studies Interstellar Medium Open clusters Stellar associations			yes	y	y	Y	GAME, ANN, GNG, DBSCAN,
Planetary studies Surface morph		yes	yes	Y	Y	n	SVM, ANN, ADABOOST, CNN
Asteroids			yes	Y		Y	CNN
Exoplanets			yes	y	y	y	DBSCAN, ANN
Gravitational lensing			yes	y		y	GAN, CNN
Dark matter			yes	Y		Y	GAN
Magnetic fields			yes	Y			ANN
Instrumentation Monitoring & control			yes	Y	Y	Y	SVM, ANN, expert systems
Data reduction and data logs			yes	Y	Y		ANN

ML in physics

- **How to evaluate performances:** statistical indicators are always ambiguous (RMS; NMAD; SIMA; PIT; ROC, etc...) & need to be selected and fine tuned on the datasets, methods, etc
- **How to evaluate effects of errors (i.e. PDFs?)? ML Methods don't do it for free...**
- Not all features are significant for the task, hence the **need to reduce dimensionality** (most relevant, all-relevant, Data Driven Approach?)
- **Proper coverage of OPS (defines biases, systematic errors, selection effects, etc....) hence:** how to control biases in the training set
- **Missing data are a problem (for us they are actually a double problem: lack of measurements or upper limits?)**
- How to minimise catastrophic outliers (byproduct: anomaly

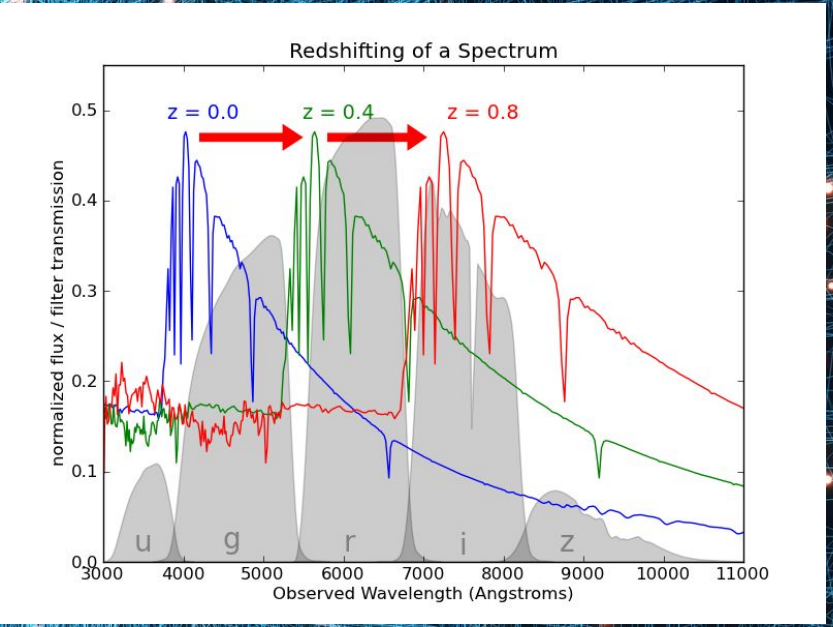


Photo-z as a template case of supervised regression

(i.e. the simplest possible case in ML)

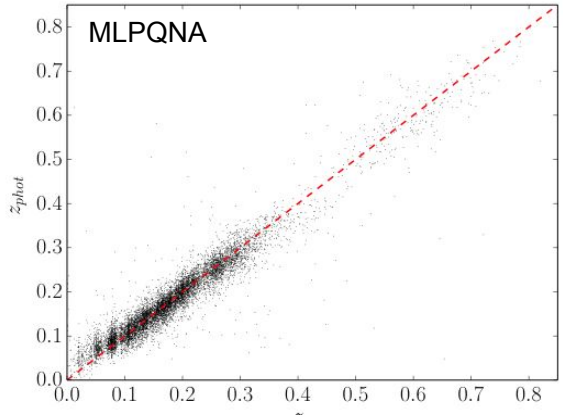
- More than 220 papers in the last 10 years
- Different surveys (almost all), many wavelengths
- Different coverages of OPS

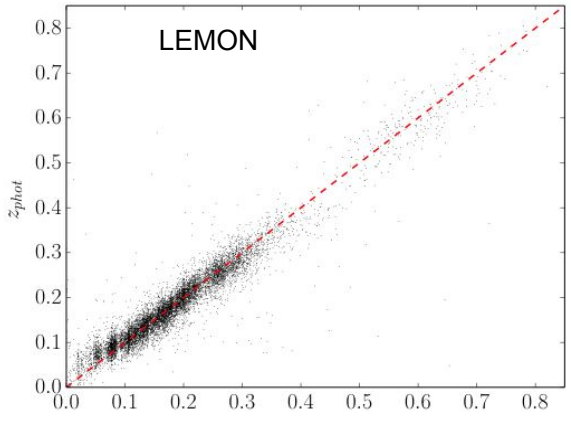
Summarising the (decadal) work by many:

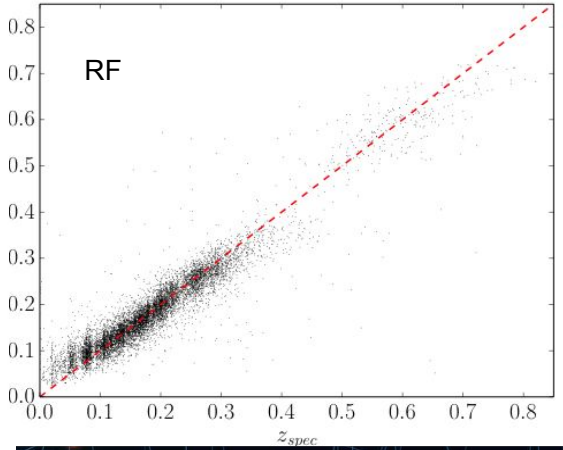
Massimo Brescia, Stefano Cavuoti
& Valeria Amaro, Alex Razim, Giuseppe Riccio, Michele delli Veneri and others.

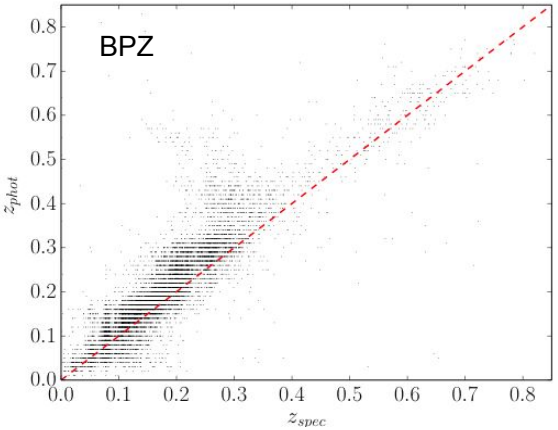
DATA RICH REGIME (large training set)

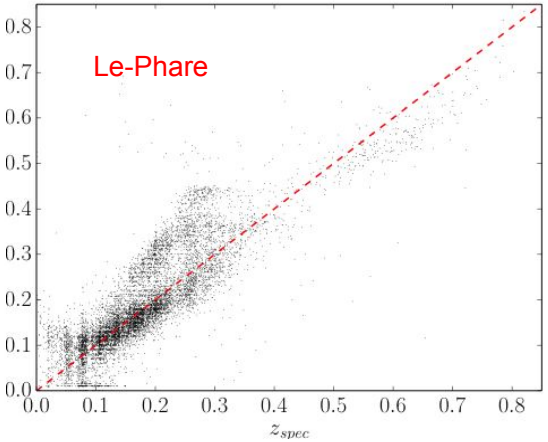
All methods have been applied: decision trees, random forest, SVM, SOM, MLP in different nuances, genetic algorithms, deep learning, autoencoders, ADABOOST, etc... (more than 50 different methods)











More or less, different ML methods are equivalent if properly used

and ... usually perform better than alternative approaches

E.g. Cavuoti, et al., MNRAS, 2016 on KiDS data

DATA RICH REGIME

ALL METHODS PERFORM WELL, BUT....

- **FEATURE SELECTION**

  Modern digital surveys produce huge amounts of measured parameters (e.g. SDSS ca. 550, KiDS more than 400, etc.)

  Merging more data sets makes the number of parameters explode.

  Number of examples is and will be forcefully limited

  different strategies to cope with it but no clear cut, unique solution....

FEATURE SELECTION

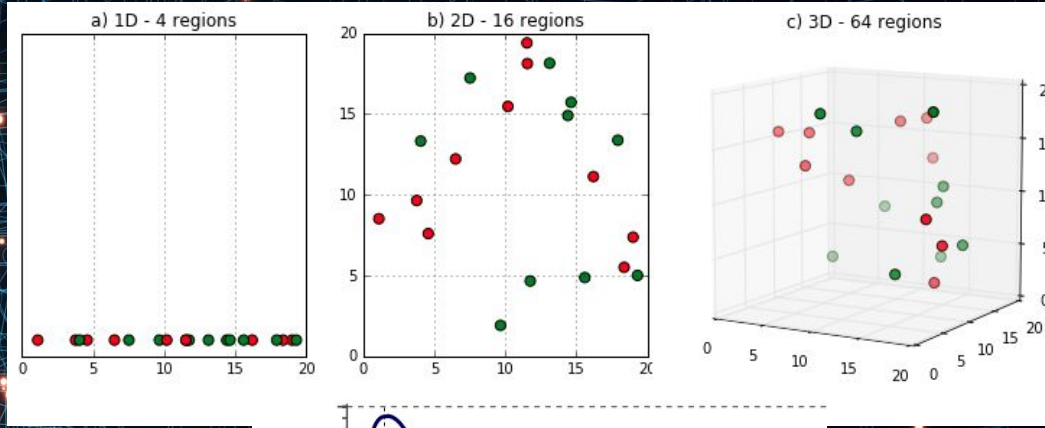
Finding optimal number and combination of parameters for a given task

- Increasing the number of parameters means that the density of training points (examples) decreases
- This leads to a loss in interpolation capabilities

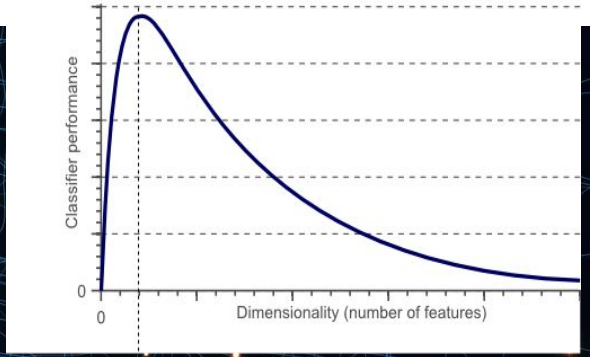
At the same time the volume of an inscribing hypersphere of dimension d and with radius 0.5 can be calculated as:

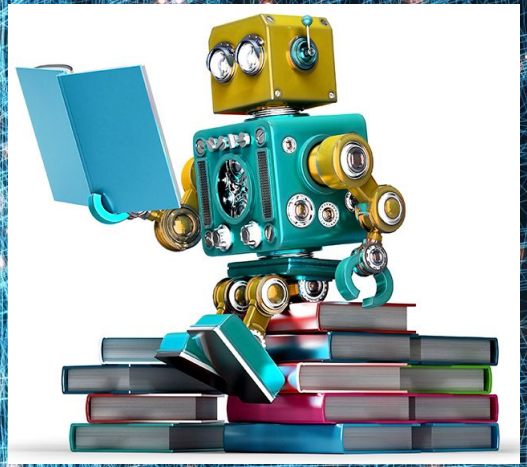
$$V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} 0.5^d.$$

Figure shows how the volume of this hypersphere changes when the dimensionality increases:



- The performance changes when the dimensionality increases, we have a peak and then a decrease, this leads to the importance of a "feature selection"





Feature selection

- Preselection based on common sense or on the **opinion of the experts**

- **Empirical** (try all) → **Most relevant**
  - **Forward selection**
  - **Data driven approach**
- **All relevant**

Brescia et al 2013, ApJ, 772, 140

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso mag_Aper_1 mag_Aper_2 mag_Aper_3 mag_auto and kron_radius	Near and Far UV total and isophotal mags phot. through 3, 4.5 and 7.5 arcsec apertures magnitudes and Kron radius in units of A or B
SDSS	u, g, r, i, z	psfMag	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y, J, H, K	PsfMag AperMag3, AperMag4, AperMag6 HallMag, PetroMag	PSF fitting magnitude in Y, J, H, K bands aperture photometry through 2, 2.8 & 5.7″ circular aperture in each band Calibrated magnitude within circular aperture r_hall and Petrosian magnitude in Y, J, H, K bands
WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro	W1: 3.4 μm and 6.1″ angular resolution; W2: 4.6 μm and 6.4″ angular resolution; W3: 12 μm and 6.5″ angular resolution; W4: 22 μm and 12″ angular resolution. Magnitudes measured with profile-fitting photometry at the 95% level. Brightness upper limit if the flux measurement has SNR< 2
SDSS	-	z_{spec}	Spectroscopic redshift

Traditional (empirical) approach:
First selection of features based on expertise
Trial and error on different combinations

Hundreds of experiments
Very demanding in terms of time

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{clean}(\Delta z_{norm})$ and $NMAD(\Delta z_{norm})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$. The standard deviation $\sigma_{clean}(\Delta z_{norm})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{norm}| \leq 2\sigma(\Delta z_{norm})$

Exp	n. obj.	$\sigma(\Delta z_{norm})$	% catas. outliers	$\sigma_{clean}(\Delta z_{norm})$	$NMAD(\Delta z_{norm})$
SDSS	41431	0.15	6.53	0.062	0.058
SDSS + GALEX	17876	0.11	4.57	0.045	0.043
SDSS+UKIDSS	12438	0.11	3.82	0.041	0.040
SDSS+GALEX+UKIDSS	5836	0.087	3.05	0.040	0.032
SDSS+GALEX+UKIDSS+WISE	5716	0.069	2.88	0.035	0.029

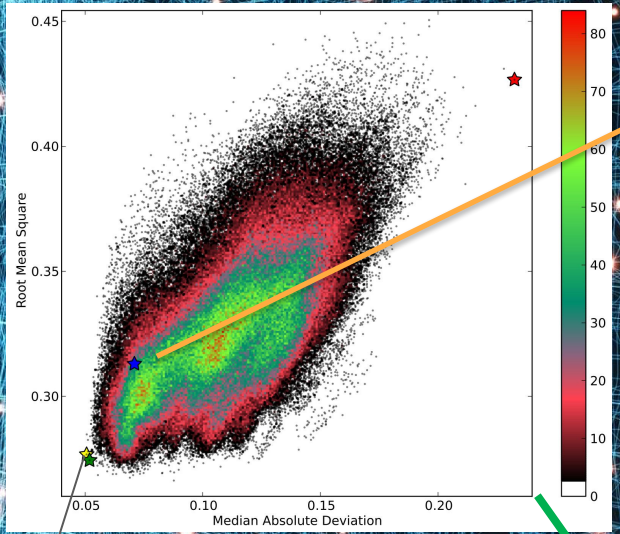
A brute force approach (K. Polsterer et al., 2015)

QSOs from SDSS

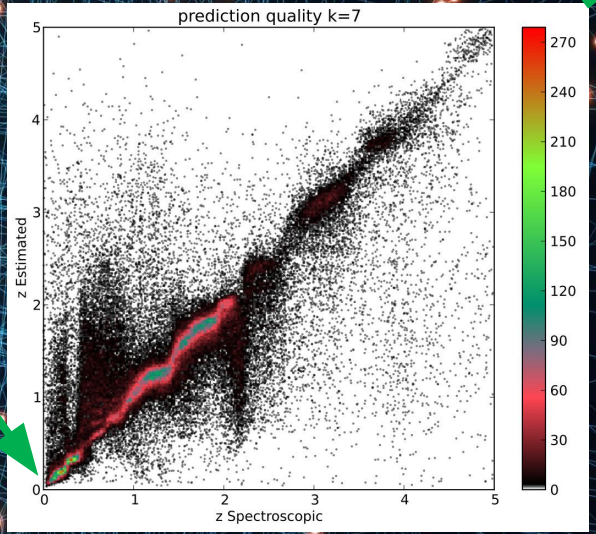
One does hot know a-priori which features are the most relevant

Use all 55 significant photometric features to select the most significant 4

$$\frac{n!}{(n-r)!r!} = 341,055 \quad \text{combinations}$$



Laurino et al 2011
Traditional feature selection



Best combination

$u_{model} - g_{model}$

$g_{psf} - r_{model}$

$z_{psf} - r_{model}$

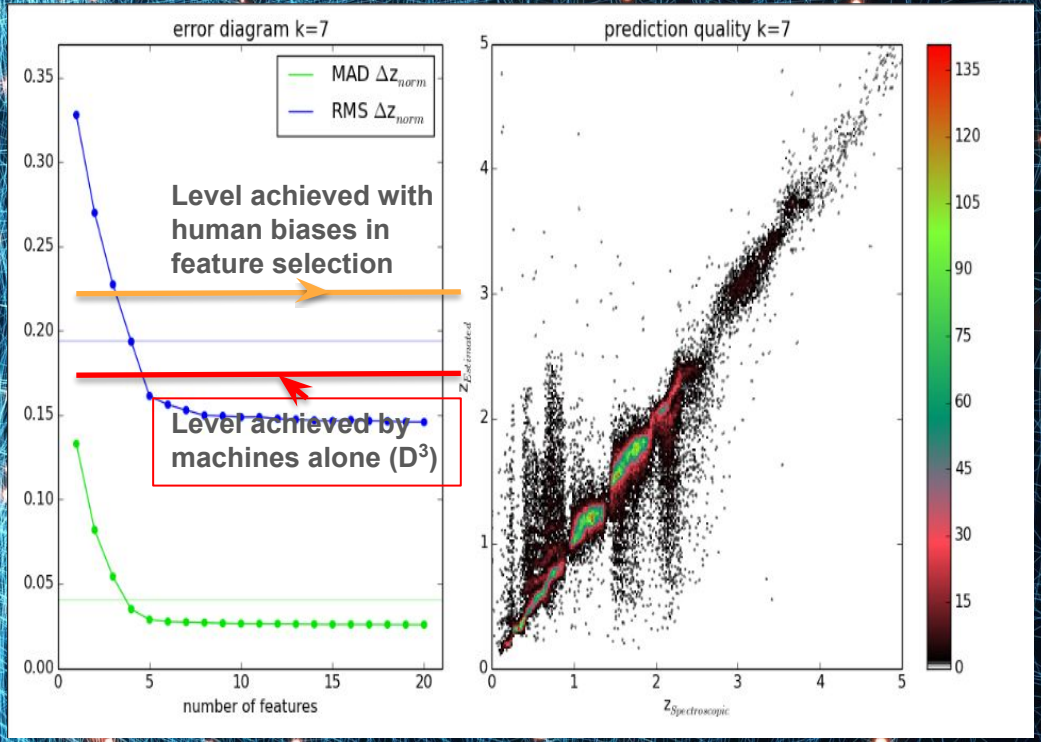
$i_{psf} - z_{model}$

Best results are comparable to Beck et al. 2014

Photometric redshifts for SDSS QSO (From K. Polsterer)

PSF, Petrosian, Total magnitudes + extinction + errors ….. 585 features…. Let us find the best combination of 10, 11, 12 etc… using FEATURE ADDITION

For just 10 features ….. 1,197,308,441,345,108,200,000 combinations (therefore just add the most significant feature strategy)



You hit a plateau at 10 features.

Accuracy twice better

These 10 features do not make sense to an astronomer

(**afterwards ... there might be some explanation**)

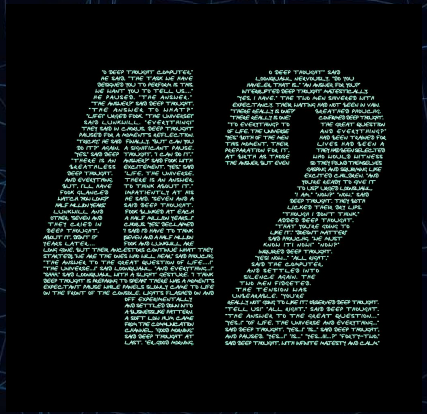
$$u_{psf} - g_{petr}$$
$$dered\left(z_{pdf}\right) - dered\left(i_{petr}\right)$$
$$dered\left(g_{psf}\right) - dered\left(r_{\mathrm{mod}}\right)$$
$$dered\left(r_{psf}\right) - dered\left(z_{\mathrm{mod}}\right)$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{\mathrm{mod}el}}}$$
$$dered\left(r_{\mathrm{mod}}\right) - dered\left(i_{\mathrm{mod}}\right)$$
$$i_{psf} - i_{petr}$$
$$dered\left(z_{psf}\right) - dered\left(r_{petr}\right)$$
$$g_{\mathrm{mod}} - g_{petr}$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{petr}}}$$

1. Most features are useless (actually dangerous) ... hence FS

2. Features which are measured on the basis of human judgement are not the most informative for a machine

3. **Optimal performances require to redefine approach to data reduction and analysis) and to extract features (measures) optimised for ML/AI**

4. ....



Return of the features, D'Isanto, Cavuoti et al. 2018

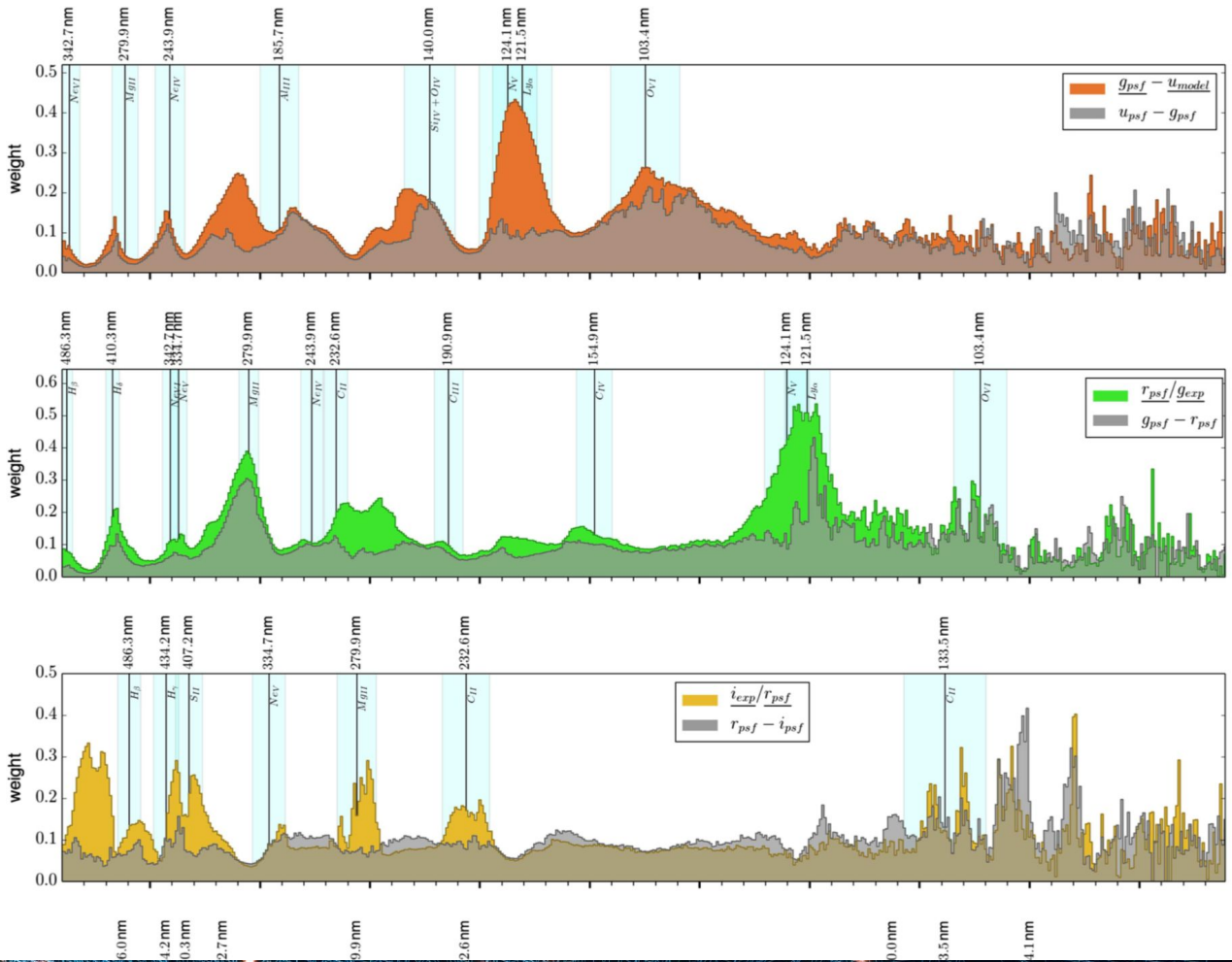
- Same data set... 4250 features
- Method: KNN in GPU Implementation
- Greedy forward selection strategy

- $r_{psf} - r_{petro}$
- $i_{psf} - i_{dev}$
- i_{psf} / z_{model}
- r_{psf} / i_{exp}.

Table 3. Summary of the scores obtained with the RF and DCMDN models in the three experiments.

Exp	Set	# Features	Mean	RMSE	NMAD
DR7a	$Classic_{10}$	10	−0.024	0.163	0.051
	$Best_4$	4	−0.023	0.163	0.080
	$Best_{10}$	10	−0.014	0.124	0.044
	DCMDN	65 536	−0.020	0.145	0.043
DR7b	$Classic_{10}$	10	−0.030	0.180	0.059
	$Best_4$	4	−0.027	0.183	0.087
	$Best_{10}$	10	−0.019	0.145	0.050
	DCMDN	65 536	−0.024	0.171	0.032
DR7+9	$Classic_{10}$	10	−0.033	0.207	0.073
	$Best_4$	4	−0.032	0.206	0.100
	$Best_{10}$	10	−0.023	0.174	0.060
	DCMDN	65 536	−0.027	0.184	0.037

Notes. The DCMDN automatically extracted 65 536 features for each experiment. The resulting scores are also given.



An example of why these features are relevant.

Feature importance of some features in the Best10 set composed by magnitudes from neighbouring bands.

The results are compared to the classic features using PSF magnitudes of the same bands.

Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2).

In optically selected samples and in presence of large knowledge base, the photo-z problem is saturated by ca. 10 features whose nature strongly depends on the data (no transfer from one data set to the other)

Computationally intensive (extremely), and difficult (if not plain impossible) for large data sets

The Features which carry most of the information are not those usually selected by the scientists but....

... physicists prefer to understand the selected features (and if possible to associate them to physical properties)...

Feature selection - All relevant

Brescia et al. 2018

PHiLAB (Parameter Handling investigation LABoratory)

Aims at finding tall the features with carry useful information for a given problem

Based on two concepts: «**shadow features**» and **Naïve-LASSO regularization** and exploiting Random Forest model as importance computing engine.

SHADOW FEATURES represent the noisy versions of the real ones and their calculated importance can be used to estimate the relevance of the real features.

LASSO penalizes regression coefficients with an L_1-norm penalty, shrinking many of them to zero. Features with non-zero regression coefficients are "selected".

A shadow feature for each real one is introduced by randomly shuffling its values among the N samples of the given dataset.

Kursa & Rudnicki 2010, Journal of Statistical Software, 36, 11

Regularization in Machine Learning is a process of introducing additional information to solve learning overfitting or to perform Feature Selection in a sparse Parameter Space. The regularization is a penalty term added to any loss function L.

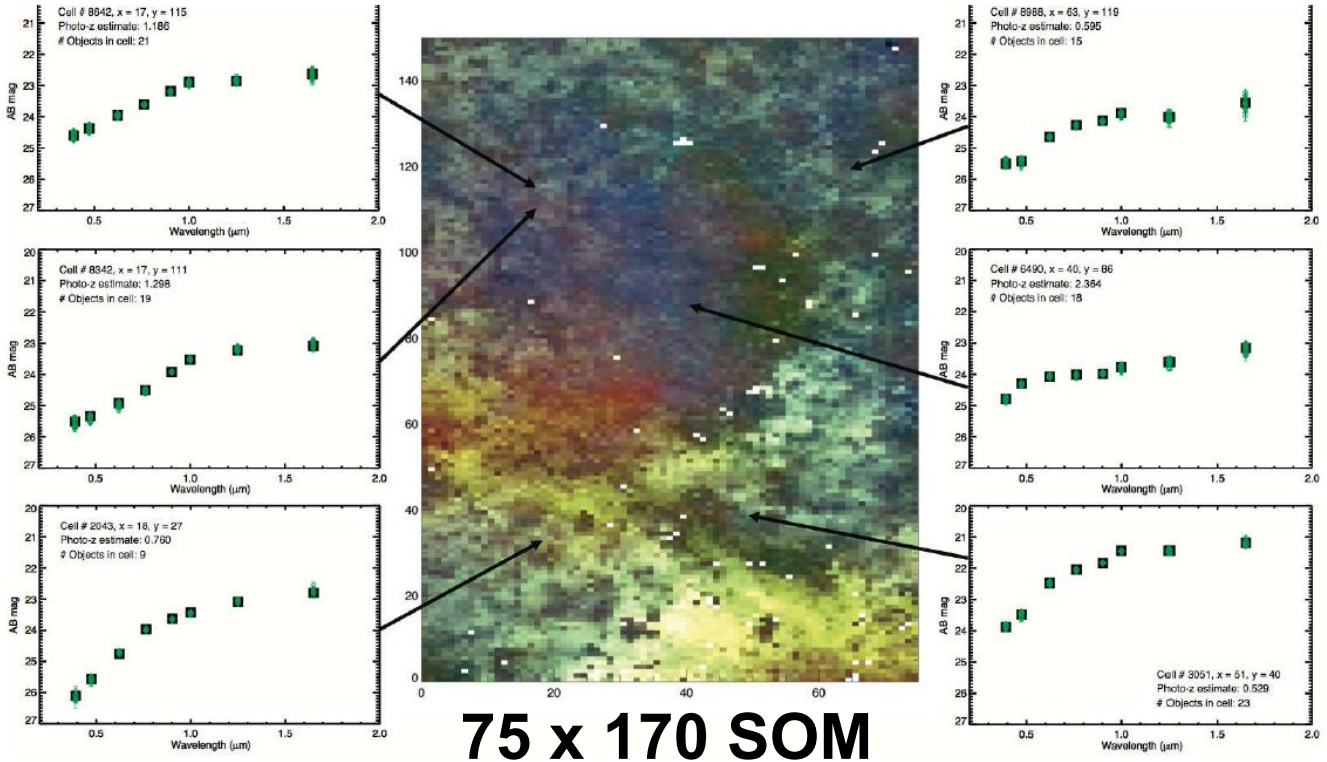
$$min_f \sum_{i=1}^{n} L(f(x)) + \boldsymbol{\lambda} \mathbf{L_{1-norm}}(\mathbf{w})$$

Hara & Maehara 2016, Proceedings of NIPS 2016, Barcelona, Spain

DATA RICH REGIME

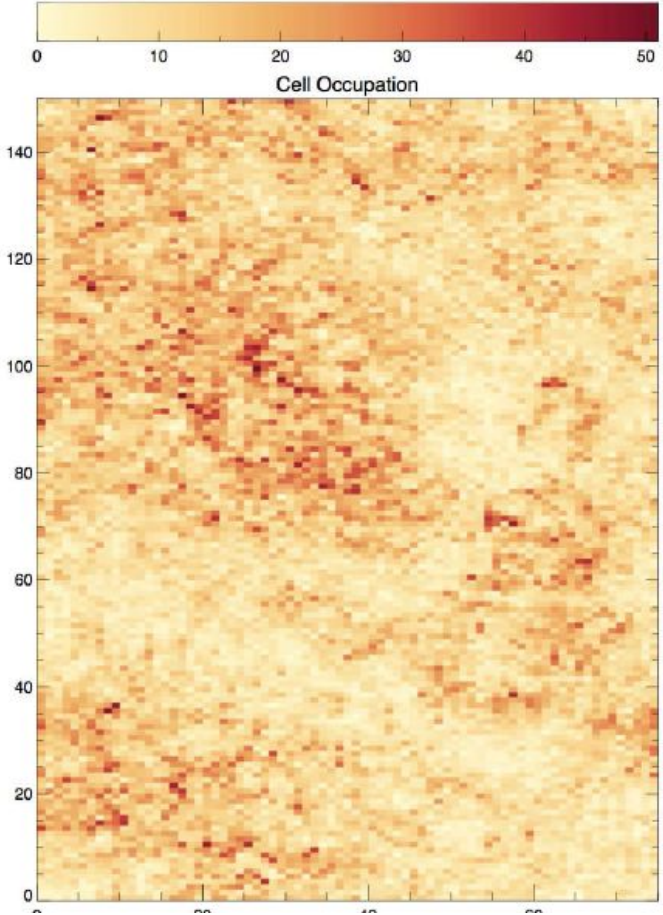
- **Coverage of OPS (Biases in training set)**
  - The OPS is not uniformously covered by the Training set
  - **Do training and test set cover the same OPS?**

Masters et al., 2015, APJ

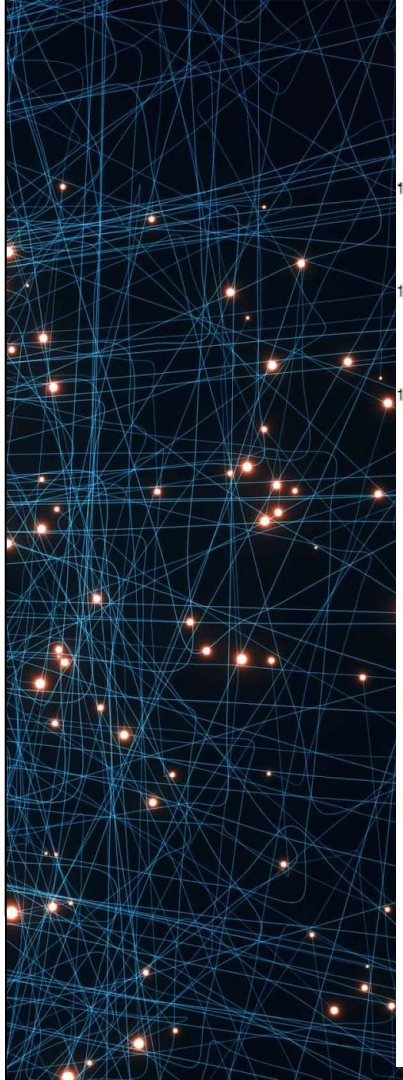


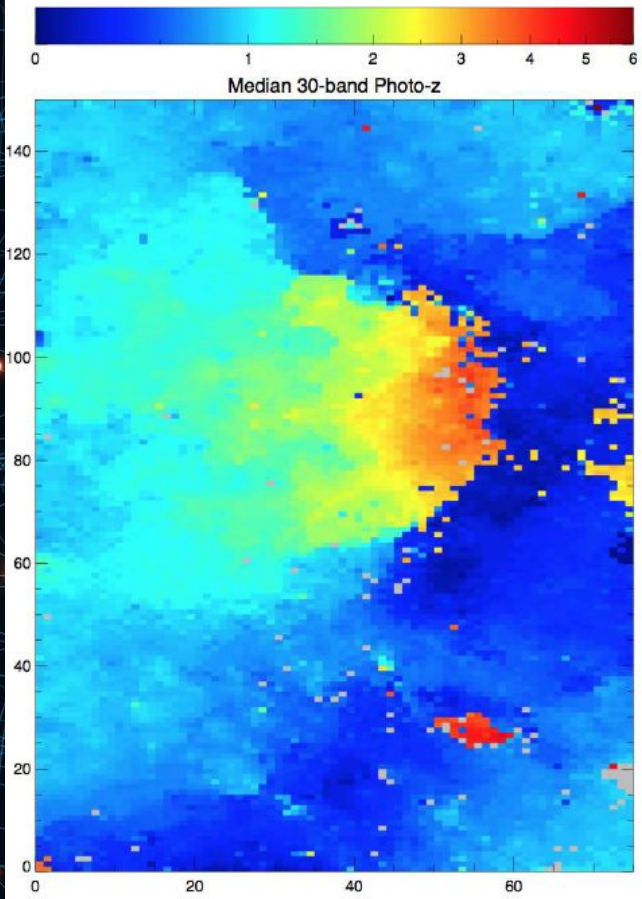
75 x 170 SOM

COSMOS data (EUCLIDISED) and converted to "pseudo-Euclid" photometric system: u,g,r,i,z,Y,J,H; Spectroscopic data from COSMOS master catalogue

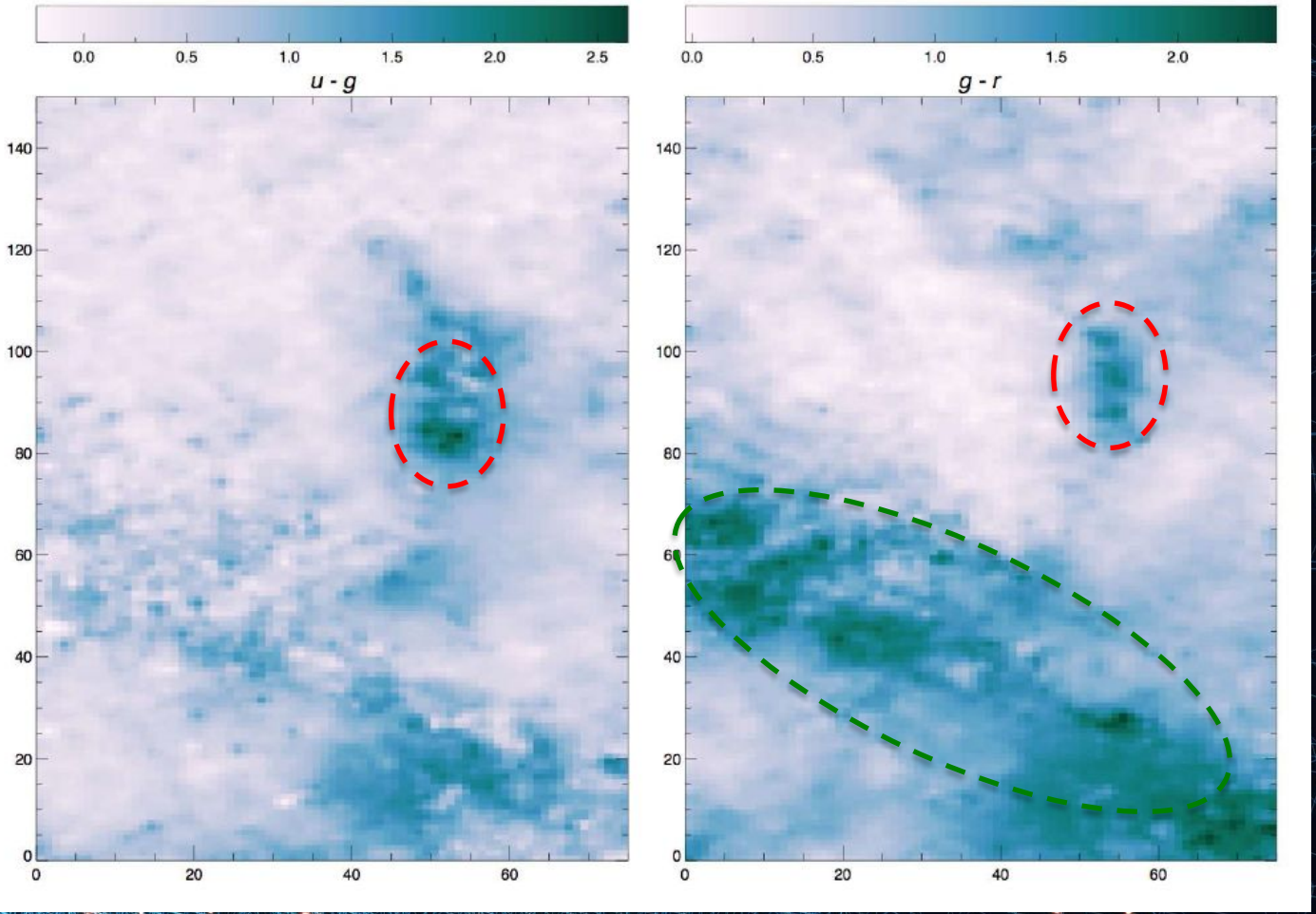


Density of galaxies in the color space (OPS)





Projection of redshift in the OPS

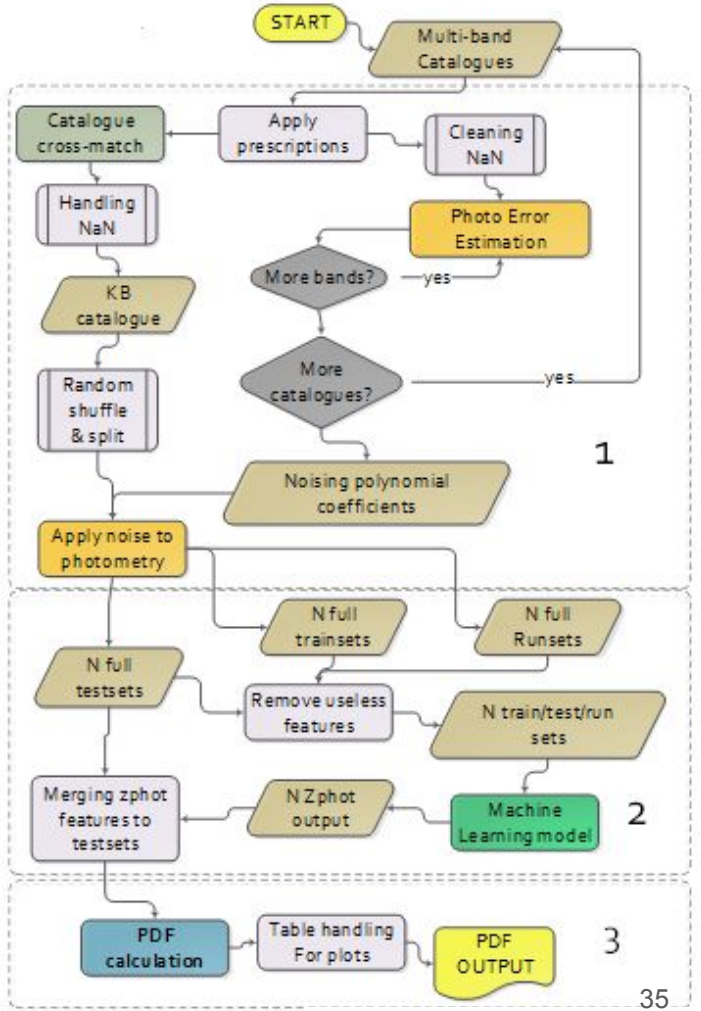


Ly –alpha break
u-g at 2.5<z<3.0
g-r at 3<z<4

Passive and dusty galaxies at low redshift

METAPhoR Workflow

1. INPUT: the KB (train + test sets), the photo-z binning step B (by default 0.01) and the zspec Region of Interest (RoI) $[Z_{min}, Z_{max}]$;
2. Produce N photometric perturbations, thus obtaining N additional test sets;
3. Perform 1 training (or N +1 trainings) and N +1 tests;
4. Derive: number of photo-z bins $(Z_{max}-Z_{min})/B$; N+1 photo-z estimations; the number of photo-z $C_{B,i} \in [Z_i, Z_{i+B}[$;
5. Calculate the probability that a photo-z belongs to all given bins: $PDF(\text{photo-z}) = (P(Z_i \leq \text{photo-z} < Z_{i+B}) = C_{B,i}/N+1)_{[Zmin,Zmax]}$;
6. Calculate and store statistics.



DATA POOR REGIME

Most astronomical literature deals with

- Optically selected samples
- Large spectroscopic knowledge bases
  - More or less uniform coverage of OPS
- Negligible fraction of missing data

Future panchromatic surveys will deal with

- Non optically selected samples (radio, X ray, etc.)
- Reduced spectroscopic knowledge bases
  - Non uniform and incomplete coverage of parameter space (very sparse)
  - Spectroscopic KB extracted from different regions of the sky (e.g. pencil beam surveys, etc.)
- Huge fraction of missing data

A Comparison of Photometric Redshift Techniques for Large Radio Surveys

Norris, Salvato, Longo, Brescia et al., 2019, ArXiv:1902.05188

The survey **EMU - Evolutionary Map of the Universe**, to be performed with ASKAP will observe Ca. 70 million galaxies

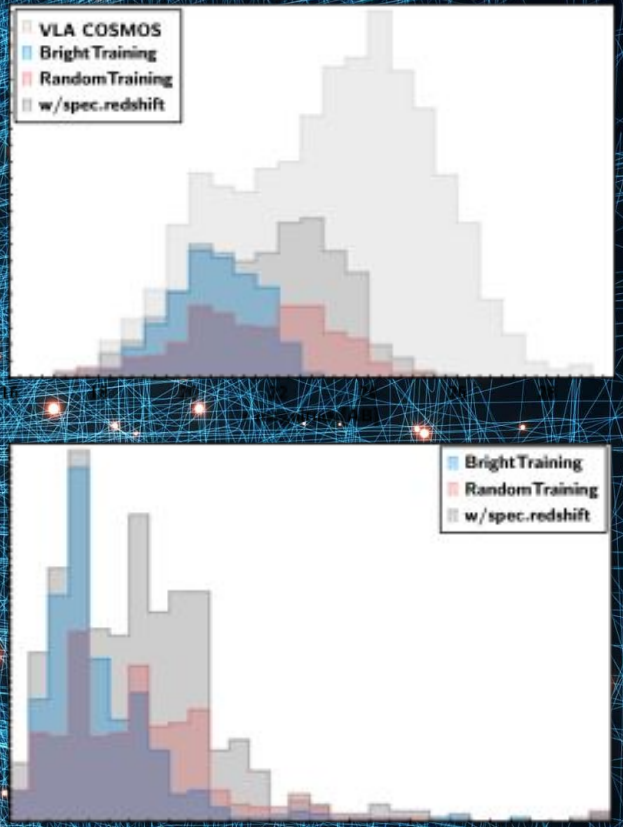
Radio selected samples are dominated (ca 50%) by **starburst** and and **high-z radio loud AGN** (Norris, 2011, 2013). These objects are usually faint and underrepresented in optically selected samples.

The median redshift sample of EMU will be ca z=1.2, while most optically selected samples have median redshift at z=0.5/0.7

Test DATA: VLA-COSMOS 1.4 GHz sample

2242 sources with optical counterparts (Sargent et al. 2010).
757 soTest DATA: VLA-COSMOS 1.4 GHz sample
form the “spectroscopic KB”. (91 (XMM) + 158 (Chandra) X-ray sources).
45 features (photometric measurements)

Small training sets
Poor coverage of OPS
Strongly biased
Incomplete data



16 sets of experiments:

(combinations of…)

1. **Luminosity biases (B or R)**
   Training on shallower sample
   Bright (50%) or **R**andom

2. **Depth (deep or Shallow)**
   Deep: train on deepest data available
   Shallow:: train on data at the same depth of EMU

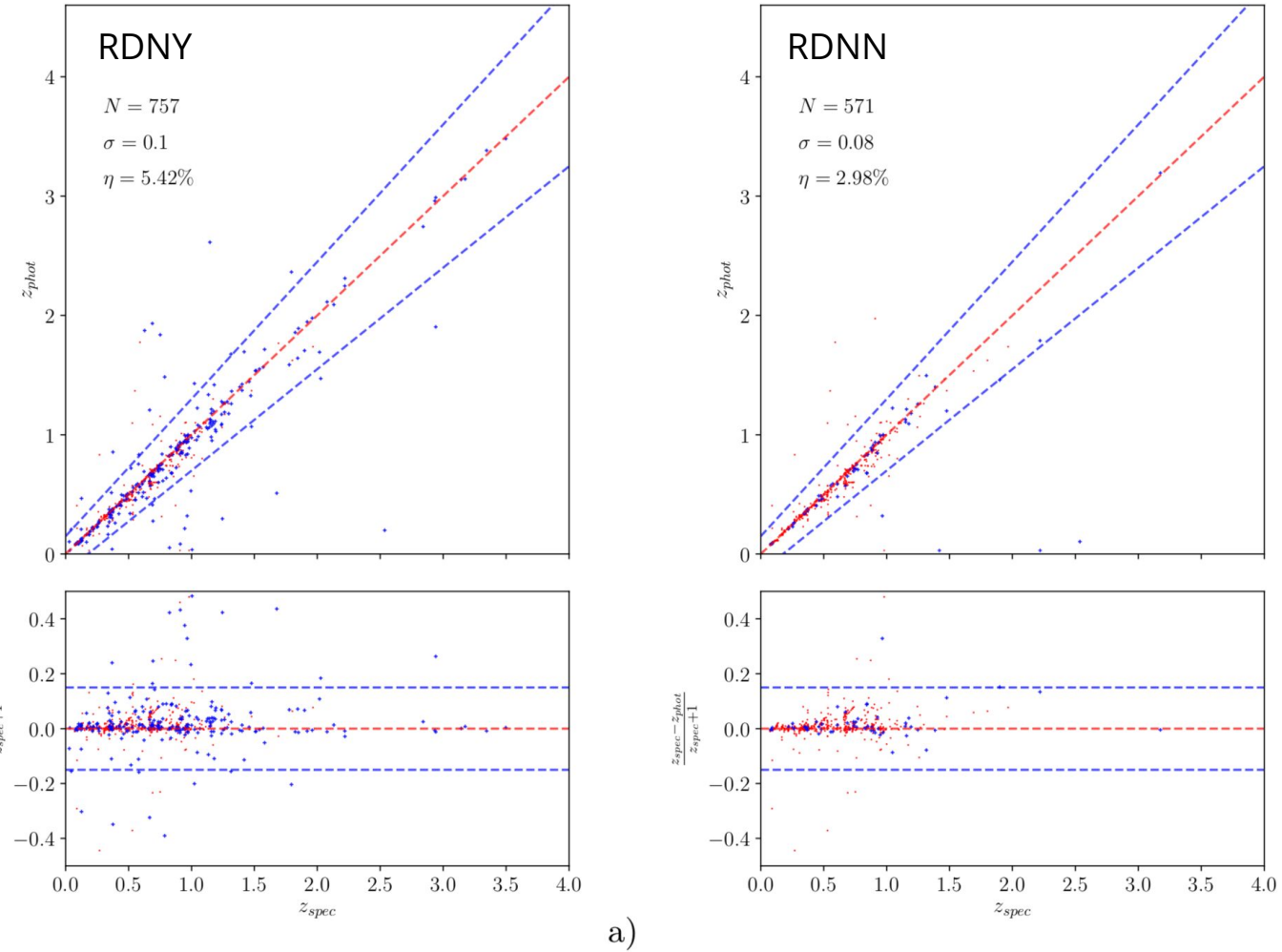
3. **Radio fluxes** (**Y** or **N**)
   Inclusion of the radio fluxes in the OPS

4. **X-ray AGN** (**Y** or **N**)
   Included (not) in the training set

Experiment		A1	B1	C1	D1	E1	F1	G1	H1	A2	B2	C2	D2	E2	F2	G2	H2
Code		BDNY	BDYY	BDNN	BDYN	BSNY	BSYY	BSNN	BSYN	RDNY	RDYY	RDNN	RDYN	RSNY	RSYY	RSNN	RSYN
Training set size		391	391	302	302	391	391	302	302	343	343	278	278	343	343	278	278
Max test set size		366	366	457	457	366	366	457	457	416	416	481	481	416	416	481	481
kNN	N=	366	366	293	293	366	366	293	438	414	414	322	322	414	414	322	322
	NMAD=	0.15	0.15	0.13	0.14	0.1	0.48	0.1	err	0.05	0.05	0.05	0.04	0.23	0.24	0.22	0.22
	η=	56	58	58	59	31	95	28	95	18	18	11	11	49	52	49	52
	β=	44	42	27	26	69	5	46	5	82	82	60	60	51	48	34	32
RF-JHU	N=	366	366	438	438	366	366		438	414	414	467	467	414	414	467	467
	NMAD=	0.11	0.12	0.12	0.12	43	0.45		err	0.07	0.07	0.07	0.07	0.09	0.09	0.1	0.1
	η=	28	27	28	30	95	95		95	15	15	16	16	20	19	21	19
	β=	72	73	69	67	5	5		5	85	85	82	82	80	81	77	79
RF-NA	N=	366	366	293	293	366	366	293	293	414	414	322	322	414	414	322	322
	NMAD=	0.13	0.12	0.16	0.17	0.11	0.09	0.12	0.12	0.07	0.07	0.06	0.06	0.13	0.13	0.11	0.1
	η=	33	25	86	83	28	22	35	33	14	15	8	7	36	36	28	25
	β=	67	75	9	11	72	78	42	43	86	85	62	62	64	64	48	50
MLPQNA	N=	366	366	293	293	366	366	293	293	414	414	322	322	414	414	322	322
	NMAD=	0.2	0.25	0.15	0.14	0.13	0.12	0.08	0.09	0.06	0.06	0.05	0.05	0.12	0.14	0.11	0.12
	η=	80	88	36	31	40	40	22	27	17	19	14	13	36	38	27	32
	β=	20	12	41	44	60	60	50	47	83	81	58	58	64	62	49	46
Le Phare	N=	757		571		509		549		757		571		509		549	
	NMAD=	0.02		0.01		0.08		0.08		0.02		0.01		0.08		0.08	
	η=	5		3		22		23		5		3		22		23	
	β=	95		73		52		56		95		73		52		56	

Table 3. Results of the 16 experiments. Line 2 of the header gives the code as described in §3: Bias (Bright/Random), IR Depth (Deep/Shallow), Radio (Y/N), X-ray (Y/N). Column 1: method name; column 2: metric: N=number of redshifts estimated, σ=standard deviation of estimated-true, η=percentage of outliers, β= overall success rate, expressed as a percentage, as defined in the text.

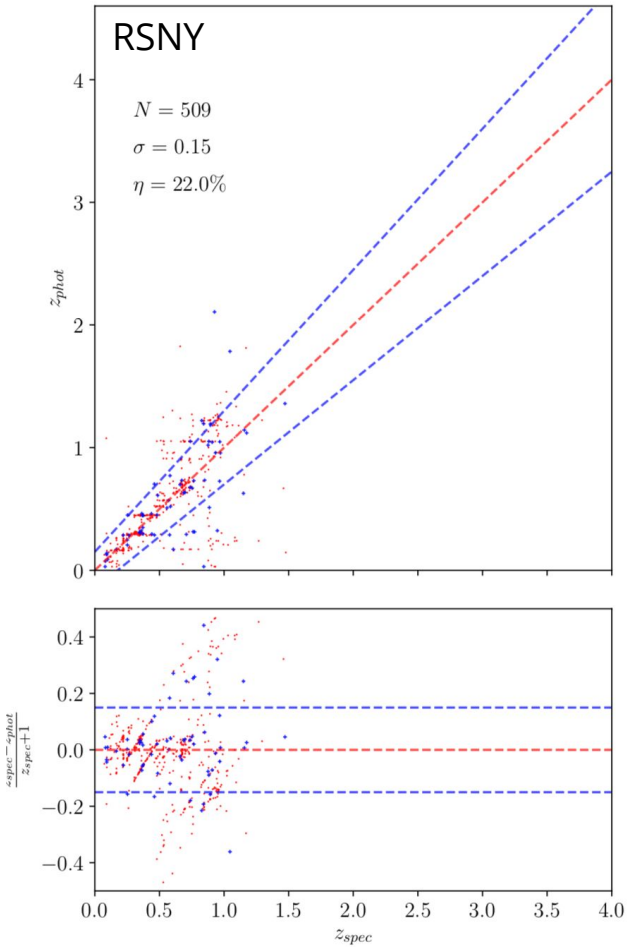
Random Forest (2 implementations), MLPQNA, LE-Phare (SED), BPZ (hybrid), K-NN



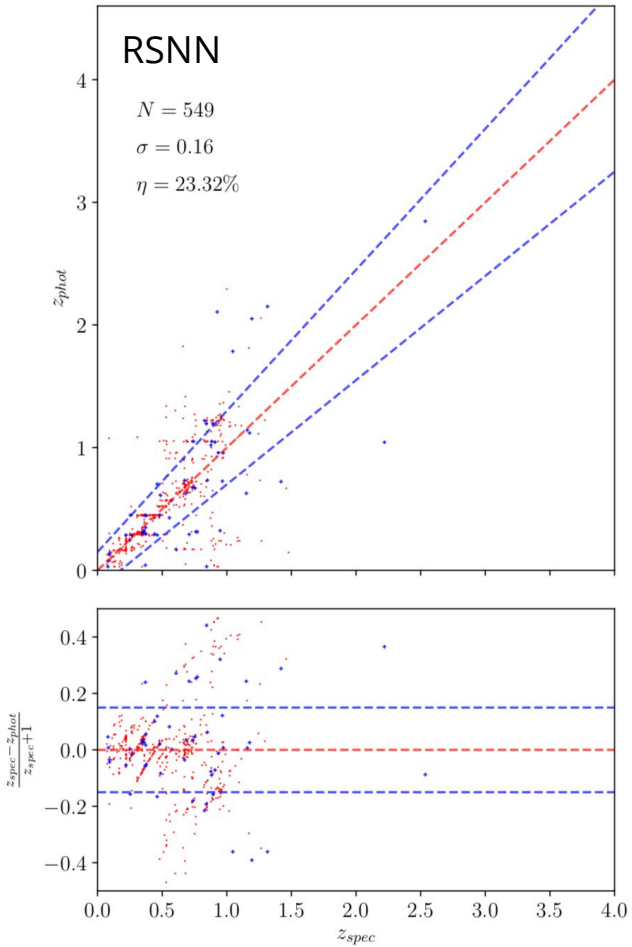
Le Phare: SED fitting

Blu: AGN
Red: non-AGN

Makes use of full COSMOS wavelength coverage



c)

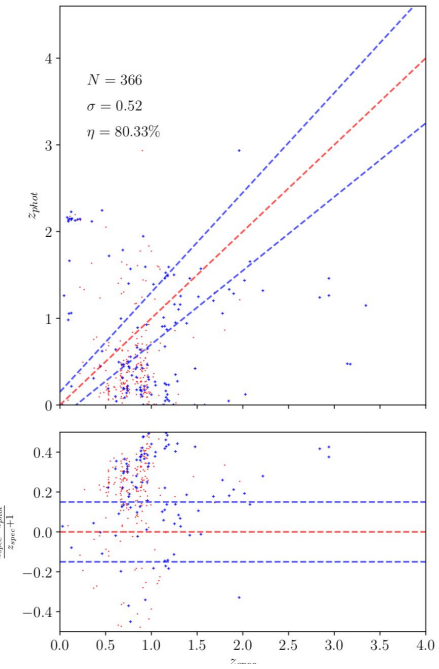


d)

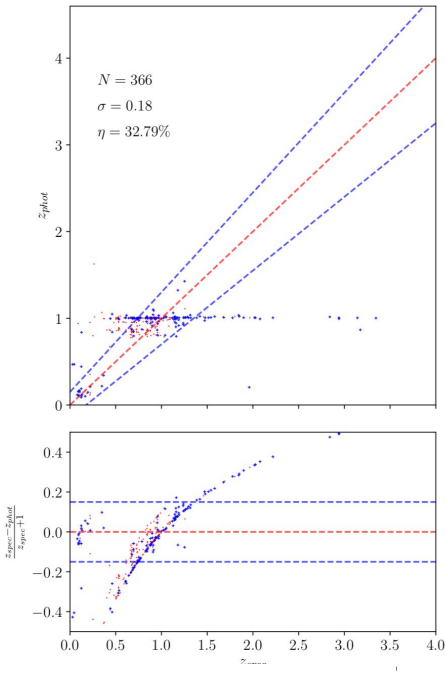
Le Phare

Blu: AGN
Red: non-AGN

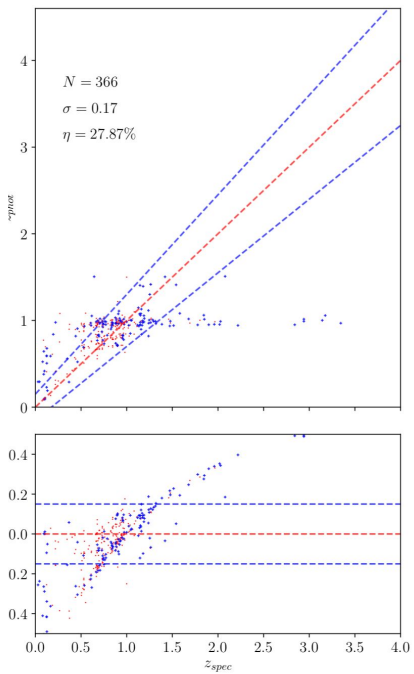
Blu: AGN
Red: non-AGN



MLPQNA

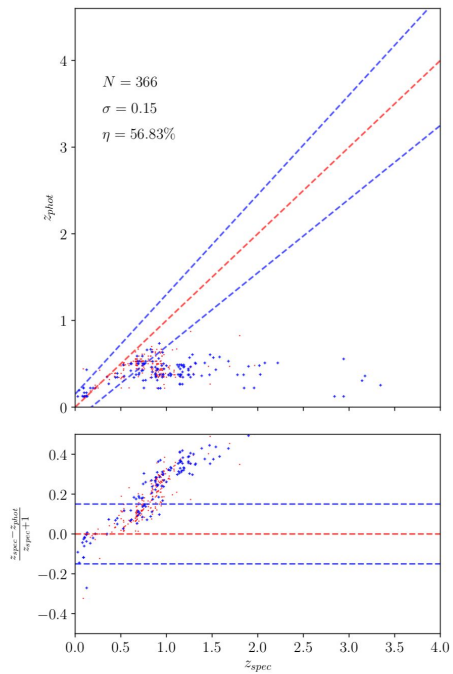


RF-NA



c)

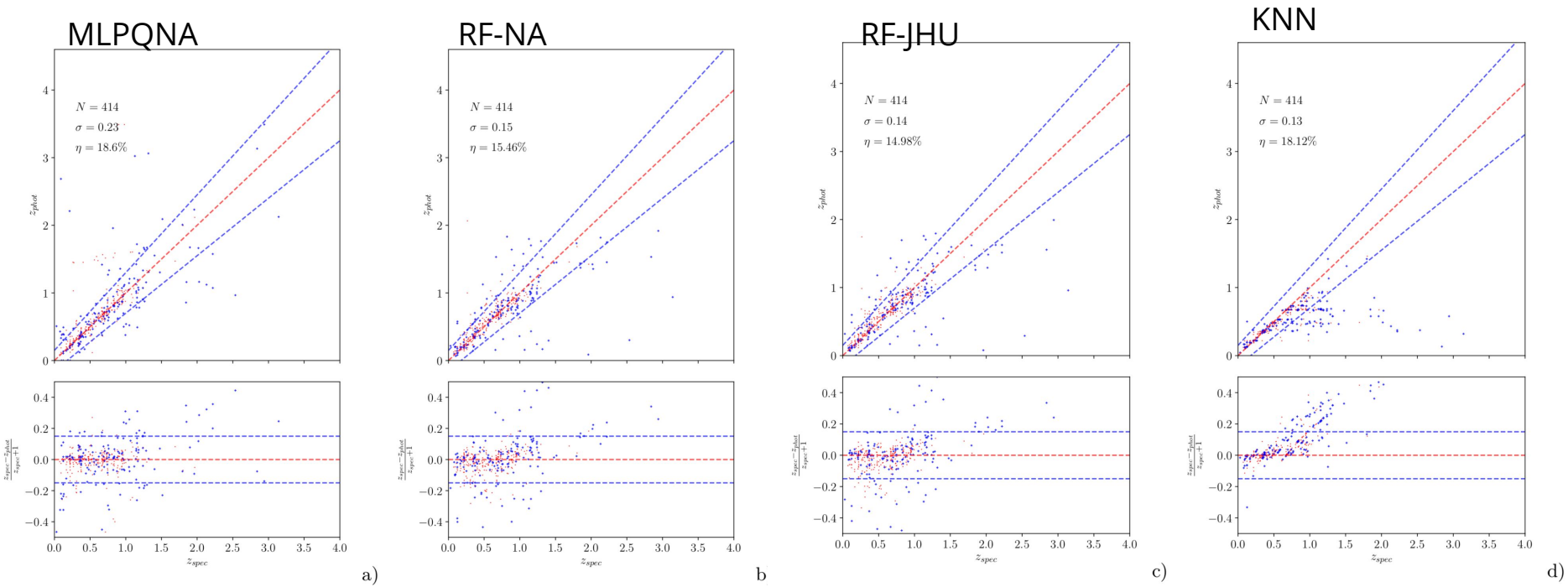
RF-JHU



d)

KNN

Exp. A1/BDNY:
most realistic for radio surveys (trained on bright 50%)



Exp. B2/RDYY

(random training, deep sample, radio fluxes used, conf. AGN in the training)

Data overabundance vs annotated data scarcity

Common to many (most) domains

...different strategies to cope with it

but no clear cut, unique solution....

Crowdsourcing

Semi-supervised learning

Generative adversarial networks

Active Learning

Domain adaptation/transfer learning

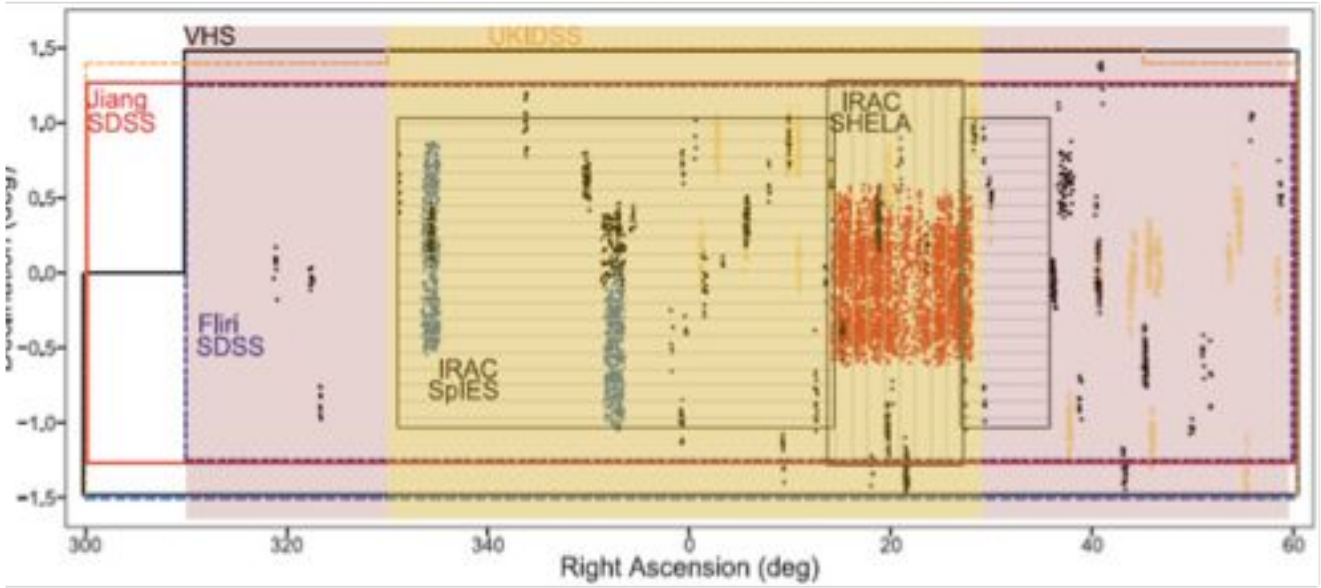
SImulations (basically useless to uncover new science)

Domain knowledge and structure

Photometric Redshifts for X-ray selected Active Galactic Nuclei in the eROSITA era

M. Brescia[1]*, M. Salvato[2]†, S. Cavuoti[1,3,4]‡, T. T. Ananna[5,6], G. Riccio[1], S. M. LaMassa[7], C. M. Urry[6] and G. Longo[3,4]

Sample composed by ca. 7.000 sources in Stripe 82 with X ray counterpart (La Massa et al. 2017)



igure 1. Map of the original multi-wavelength coverage of Stripe 82X area discussed in A17. The total area extends for ~ 2.5° in eclination and 120° in Right Ascension. The dots represent X-ray sources, respectively, from XMM-Newton AO13 (red), AO10 (blue), chival XMM-Newton sources (yellow) and Chandra sources (black). While standard photo-z are generated for the entire area (in red), e selection of the best features discussed in the first part of the paper is obtained considering only the sources in the yellow area.

Filter	BAND DEPTH								
	NOMINAL	BEST	SDSS	SDSS & VHS	SDSS & IRAC	SDSS & WISE	SDSS VHS & IRAC	SDSS VHS & WISE	SDSS VHS IRAC & WISE
FUV	21.99	—	—	—	—	—	—	—	—
NUV	21.99	—	—	—	—	—	—	—	—
u	31.22	28.54	28.54	28.54	28.54	28.54	28.54	28.54	28.54
g	28.77	24.20	24.39	24.20	24.39	24.39	24.20	24.20	24.20
r	27.13	23.25	23.43	23.25	23.43	23.43	23.25	23.25	23.25
i	27.21	22.35	23.49	22.64	23.49	22.45	22.64	22.35	22.35
z	30.46	22.42	23.35	22.46	22.99	22.42	22.46	22.42	22.08
J	24.74	21.64	—	24.64	—	—	21.64	21.64	21.51
H	24.15	22.87	—	22.87	—	—	21.61	22.87	21.61
K	22.60	21.63	—	21.63	—	—	21.63	21.63	21.63
Juk	23.44	—	—	—	—	—	—	—	—
Huk	22.69	—	—	—	—	—	—	—	—
Kuk	22.41	—	—	—	—	—	—	—	—
CH1_SPIES	24.27	20.82†	—	—	21.64†	—	21.06†	—	20.49†
CH1_SHELA	22.80		—	—		—		—	
CH2_SPIES	22.88	20.49†	—	—	21.41†	—	21.07†	—	20.22†
CH2_SHELA	23.88		—	—		—		—	
W1	21.16	20.71	—	—	—	20.71	—	20.71	20.61
W2	20.74	20.59	—	—	—	20.63	—	20.63	20.59
W3	18.20	18.04	—	—	—	18.11	—	18.11	18.04
W4	16.15	16.06	—	—	—	16.13	—	16.13	15.94
N. of sources	5990	2290	4855	3218	2293	3291	1620	2696	1380
N. of sources w/ z_{spec}	2933	1686	2793	2218	1596	2160	1279	1935	1121
N. of sources w/ $F_X > 10^{-14}$	2351	1249	2025	1649	1051	1619	888	1445	793
N. of sources w/ $F_X > 10^{-14}$ and z_{spec}	1550	1025	1483	1309	857	1256	758	1174	683

Table 1. Summary table for depth, amount of sources and redshift coverage. The first column refers to the nominal depth of the entire sample of reliable counterparts in Stripe 82X, as presented in A17. The following columns refer to the magnitudes reached in the various experiments, i.e., the faintest magnitude reported in the Stripe 82X catalogue for the various sub-samples for which the photo-z have been computed. The values in the column ***BEST*** represent the faintest magnitudes of the sub-sample of sources in the yellow area of Fig. 1, used for the features analysis performed with ΦLAB, (Sec. 3.1). The bands marked with a — symbol have been discarded from that specific experiment.

- FUV and NUV magnitudes and corresponding errors from GALEX all-sky survey (Martin et al. 2005);
- u, g, r, i, z SDSS AUTO magnitudes and corresponding errors from Fliri & Trujillo (2016);
- J, H, K from VISTA (Irwin et al. 2004). As shown in A17 additional data in J_{UK},H_{UK},K_{UK} data from UKIDSS (Lawrence et al. 2007) are available for the same area but were not used in this paper;
- 3.6 and 4.5 μm magnitudes and corresponding errors from IRAC. Here two complementary surveys are used: SPIES (Timlin et al. 2016) and SHELA (Papovich et al. 2016). Given the similarity of the two surveys, we do not differentiate sources belonging to one or another;
- W1, W2, W3, W4 magnitudes and corresponding errors from AllWISE (Wright et al. 2010).

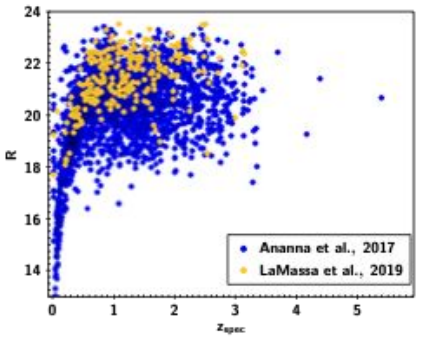


Figure 2. Redshift and magnitude distribution for the sources with spectroscopic redshift. The blue sources were presented in A17 and have been used in this work as training and blind test samples. The 258 yellow sources are on average fainter and were recently presented in LaMassa et al. (2019). They are used as additional blind test sample.

Spectroscopic KB

FS with PhiLab

feature	importance	feature	importance
R-Z	14.51%	J-K	0.40%
G-I	12.44%	U-CH1	0.35%
CH1-CH2	7.50%	H-CH1	0.34%
U-G	6.00%	R-CH2	0.33%
Z-W1	5.84%	U-I	0.33%
Z-CH1	4.24%	R-W2	0.33%
G-R	4.03%	K-CH1	0.33%
K	3.14%	R-W1	0.31%
G-Z	3.03%	U	0.30%
I-W1	2.00%	U-J	0.30%
I-CH2	1.94%	G-W2	0.27%
H	1.81%	G-CH2	0.24%
R-I	1.67%	I-J	0.23%
I-CH1	1.51%	CH1-W2	0.23%
J	1.45%	G	0.22%
H-K	1.34%	J-CH2	0.21%
R	1.21%	G-CH1	0.21%
I	1.21%	G-K	0.20%
W1	1.18%	J-W1	0.20%
I-Z	1.08%	H-W2	0.20%
Z	0.99%	K-CH2	0.17%
H-W1	0.97%	K-W2	0.16%
K-W1	0.83%	U-W1	0.16%
Z-W2	0.83%	Z-J	0.16%
CH2-W1	0.77%	U-K	0.15%
Z-CH2	0.68%	R-CH1	0.14%
U-R	0.68%	H-CH2	0.13%
U-Z	0.62%	CH1	0.13%
G-W1	0.56%	Z-H	0.13%
J-CH1	0.54%	U-H	0.12%
Z-K	0.52%	J-W2	0.09%
I-W2	0.50%	I-K	0.08%
J-H	0.46%	R-K	0.07%
W1-W2	0.45%	—	—

Table 2. Results of the feature analysis (percentages of estimated feature importance) performed with ΦLAB in the case of the parameter space composed by considering all magnitudes and colours available.

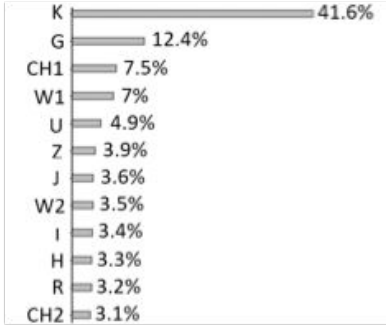


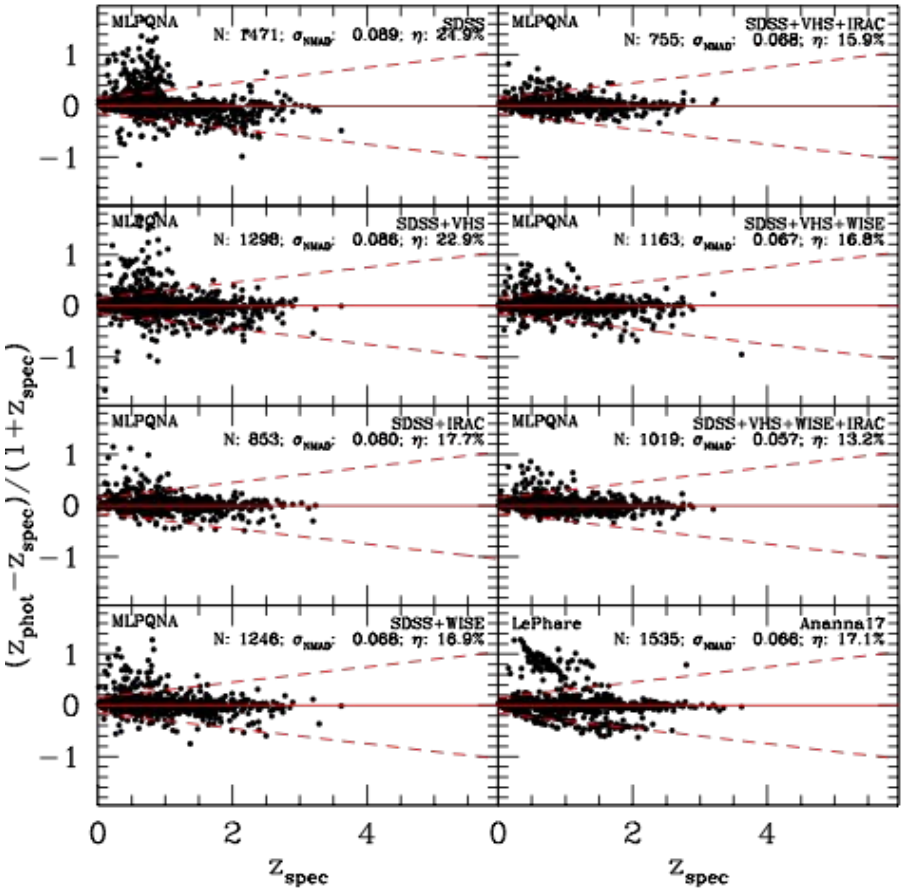
Figure 3. Results of the feature analysis performed with ΦLA
The importance of each feature is estimated for the case in whi
only magnitudes are considered for the sample *BESTmagopt*.

Due to different depths …. need to handle missing data

(i) SDSS, VHS, WISE & IRAC (*sdssVWI*)
(ii) SDSS, VHS & WISE (*sdssVW*);
(iii) SDSS, VHS & IRAC (*sdssVI*);
(iv) SDSS & WISE (*sdssW*);
(v) SDSS & IRAC (*sdssI*);
(vi) SDSS & VHS (*sdssV*);
(vii) SDSS.

	Number of sources	$\|bias\|$	σ	σ_{68}	$\sigma_{\rm NMAD}$	η
A17	258	0.0066	0.292	0.129	0.089	27.07
sdss	227	0.0037	0.367	0.158	0.129	33.48
sdssV	135	0.0357	0.322	0.211	0.149	41.48
sdssW	144	0.0073	0.288	0.173	0.137	34.03
sdssI	110	0.0119	0.202	0.184	0.163	40.91
sdssVW	111	0.0459	0.272	0.167	0.143	33.33
sdssVI	58	0.0343	0.255	0.161	0.116	32.76
sdssVWI	25	0.0298	0.151	0.152	0.104	32.00
$\rm MLPQNA_{merged}$	229	0.0182	0.270	0.192	0.154	38.43

f all statistical results for the new sample of 258 spectroscopic redshifts presented in LaMassa et al. 2019.



igure 5. Comparison between spectroscopic redshift and photo-z for the sources cut at the eROSITA flux and divided on the basis o vailable photometric points. For comparison, the result from A17 is reported in the lower right panel of the figure. By comparing the ccuracy and the fraction of outliers in every panel with the corresponding row in Table 8, we see that computing photo-z using only DSS for bright X-ray sources is not recommended.

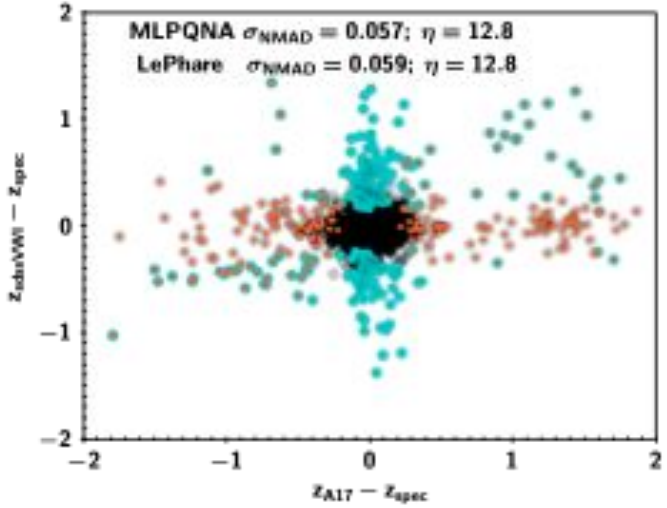


Figure 6. Difference between spectroscopic redshift and photo-z computed via MLPQNA and LePhare for the sub-sample of 1679 sources with SDSS, VHS, WISE and IRAC photometry. Sources that are outliers for MLPQNA (LePhare) are plot in cyan (orange). For this sub-sample the accuracy and fraction of outliers are very similar for the two methods. However, the majority of the outliers are such only for one of the two algorithms.

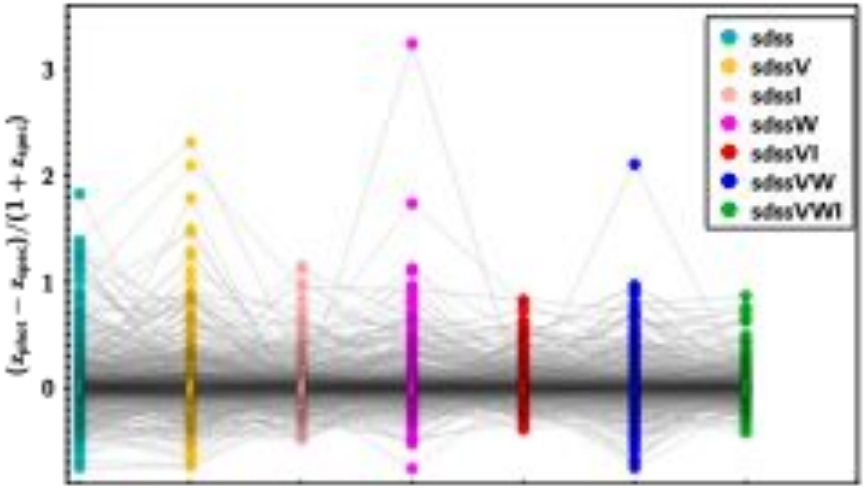
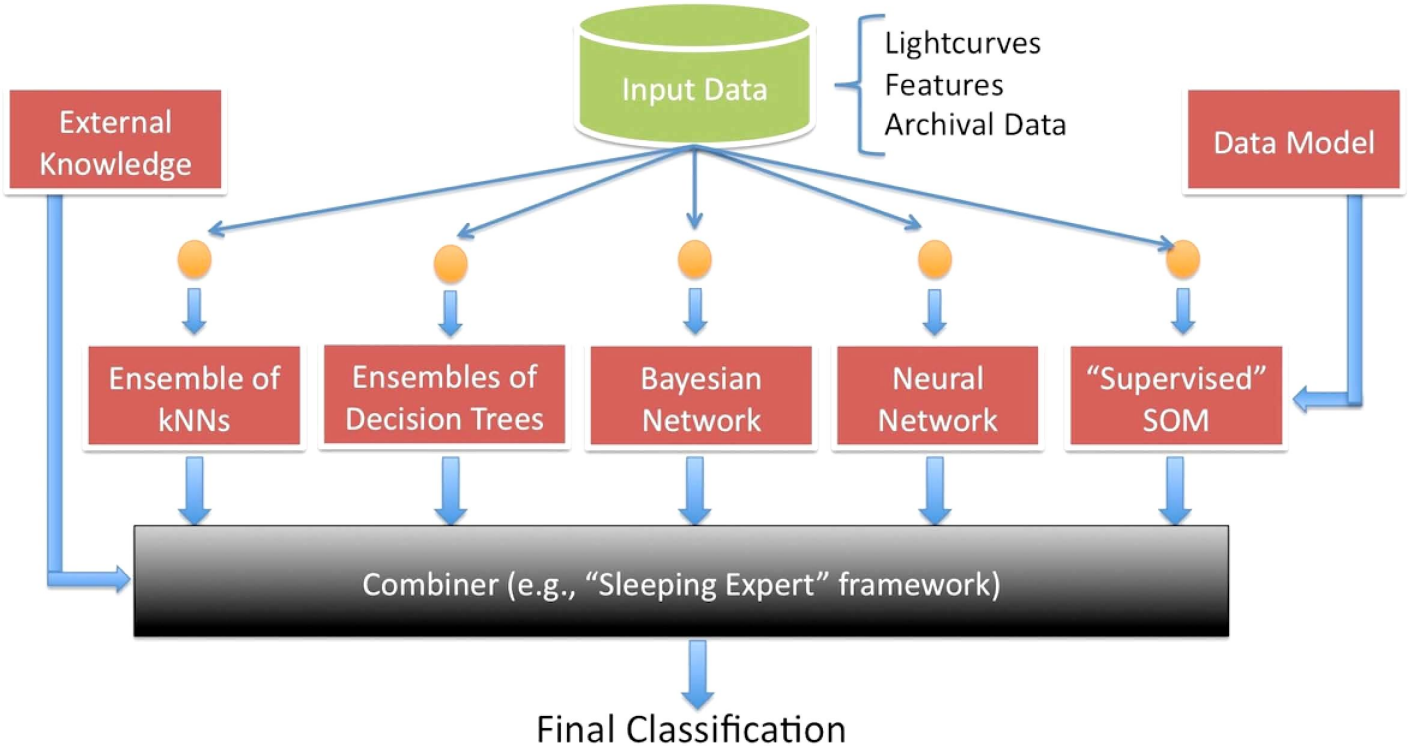


Figure 7. One-to-one comparison of accuracy for photo-z computed via MLPQNA with different combinations of photometry. For this plot only sources present in all the subsamples have ben used.

Metaclassification:

optimal combining of classifiers



Exploring a variety of techniques for an optimal classification fusion:
Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit,
Sleeping Expert…

Some conclusions on upervised methods

- **If large annotated, reliable data sets are available, all methods are substantially equivalent (DL, RF, MLPQNA, K-NN, etc.)**
  - Need for extensive feature selection (different approaches substantially equivalent)
  - Differences are in the range of a few % which are usually negligible when errors are properly taken into account

- **If data are heterogeneous (depth, coverage, etc.) or biased... methods matter**
  - DL substantially useless, RF or KNN outperformed by normal MLP's (better at generalising ?)
  - Handling biases and understanding results becomes the crucial part.
  - Lots of work remains to be done to be able to apply these methods to future surveys

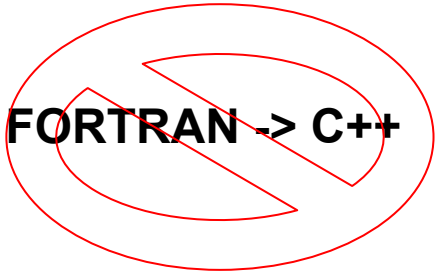
- **The scientific exploitation of future large survey projects requires better "annotated data"**

Data Science is a science and not a tool.

- ML packages are mainly written for the analysis of data of different nature and DO NOT MATCH (BY ANY MEANS) THE NEEDS OF PHYSICS
- ML based tools cannot be used off the shelf by a domain expert . (or: **you cannot take a Physics PhD student, let him download a piece of software from a library and expect that he produces physically meaningful results**).
- **You can put a Physics PhD student with another one from DS and possibly one from Statistics and only so you can obtain something significant**
- PHYSICIST's superiority complex does not work here: DATA SCIENCE IS A SCIENCE AND NOT A TOOL: hundreds of papers out every day, hundreds of methods, etc... specific knowledge anmd training required
- **The proper solution of a problem (even a simple one) may take to a medium sized DS team up to 5-10 years to be properly solved**

It is not a tool to do usual staff

I.e. **FORTRAN -> C++**



Other communities (bioinformatics, geo-informatics, economic, etc) are in the same situation.

Domain experts NEED TO REMAIN DOMAIN EXPERTS CAPABLE TO DEAL WITH Experts

DEEP LEARNING (CNN, AE, GAN; etc) is likely NOT THE SOLUTION TO ANY OF THE ABOVE PROBLEMS (errors, computing time, EAI, etc...) fashionable but...

Big data is like teenage sex:

everyone talks about it,

nobody really knows how to do it,

everyone thinks everyone else is doing it,

so everyone claims they are doing it.

Dan Ariely

Very few do it properly

Thanks for the attention