

# Astroinformatics and Astrophysics a virtuous synergy in the Big Data era



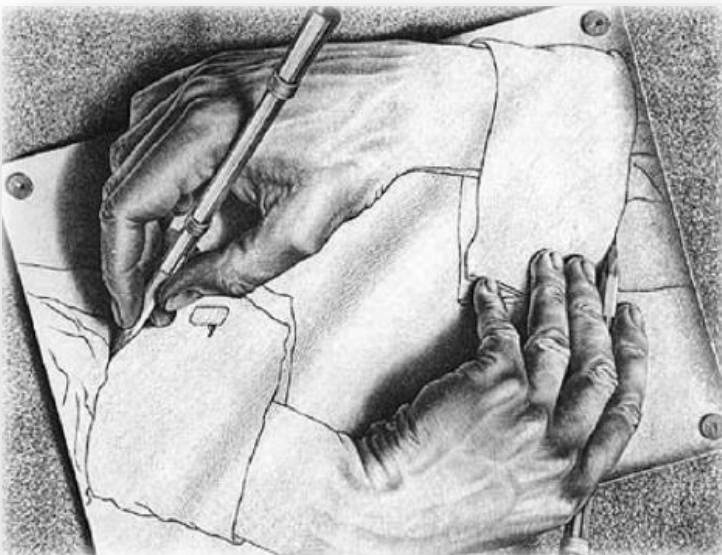
*M. Brescia, G. Riccio, S. Cavuoti, A. Mercurio  
G. Angora, M. Delli Veneri*

# Transformation and Synergy

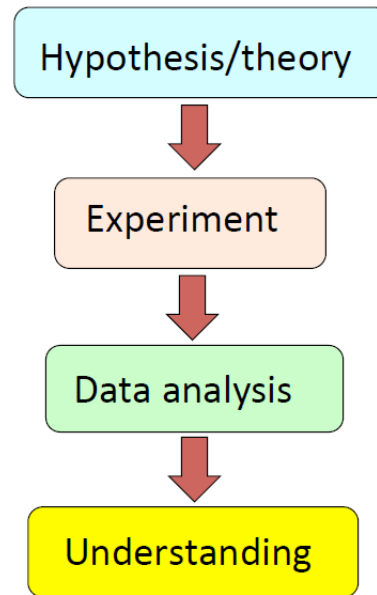
All sciences in the 21<sup>st</sup> century is becoming cyber-science (aka e-science) and with this change comes the need for a new scientific methodology.

The challenge we are tackling:

- management of large, complex, distributed data sets
- effective exploration of such data → new knowledge
- these challenges are universal!
- a virtuous synergy between computationally enabled science and the science-driven IT

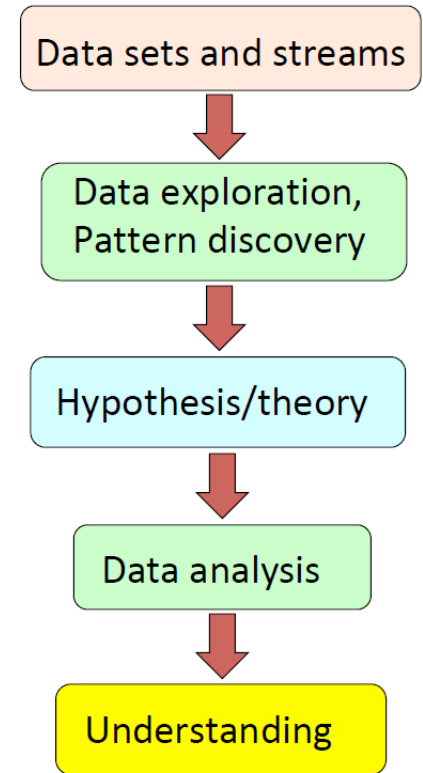


## Hypothesis-driven science



The two approaches are complementary

## Data-driven science



# Why Astrophysics is a Big Data case

Formally, Big Data is a system whose data are characterized by the “3 critical V” rule (**Volume, Velocity, Variety**)



- The **information volumes and rates** grow exponentially

➔ **Most data will never be seen by humans**



- A great increase in the data **information content**

➔ **Data driven vs. hypothesis driven science**

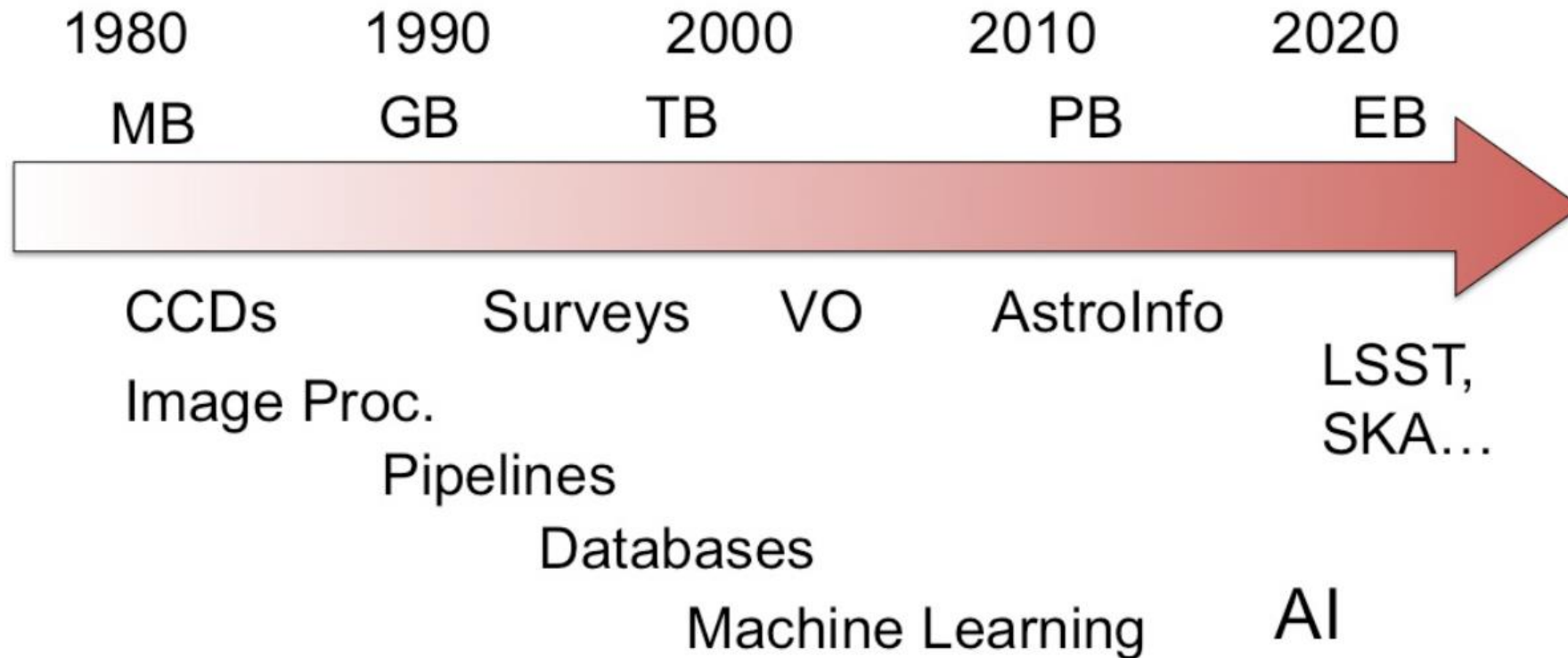
- A great increase in the **information complexity**

➔ **There are patterns in the data that cannot be comprehended by humans directly**



# The evolving data-rich Astronomy

An example of a “Big Data” science driven by the advances in computing/information technology

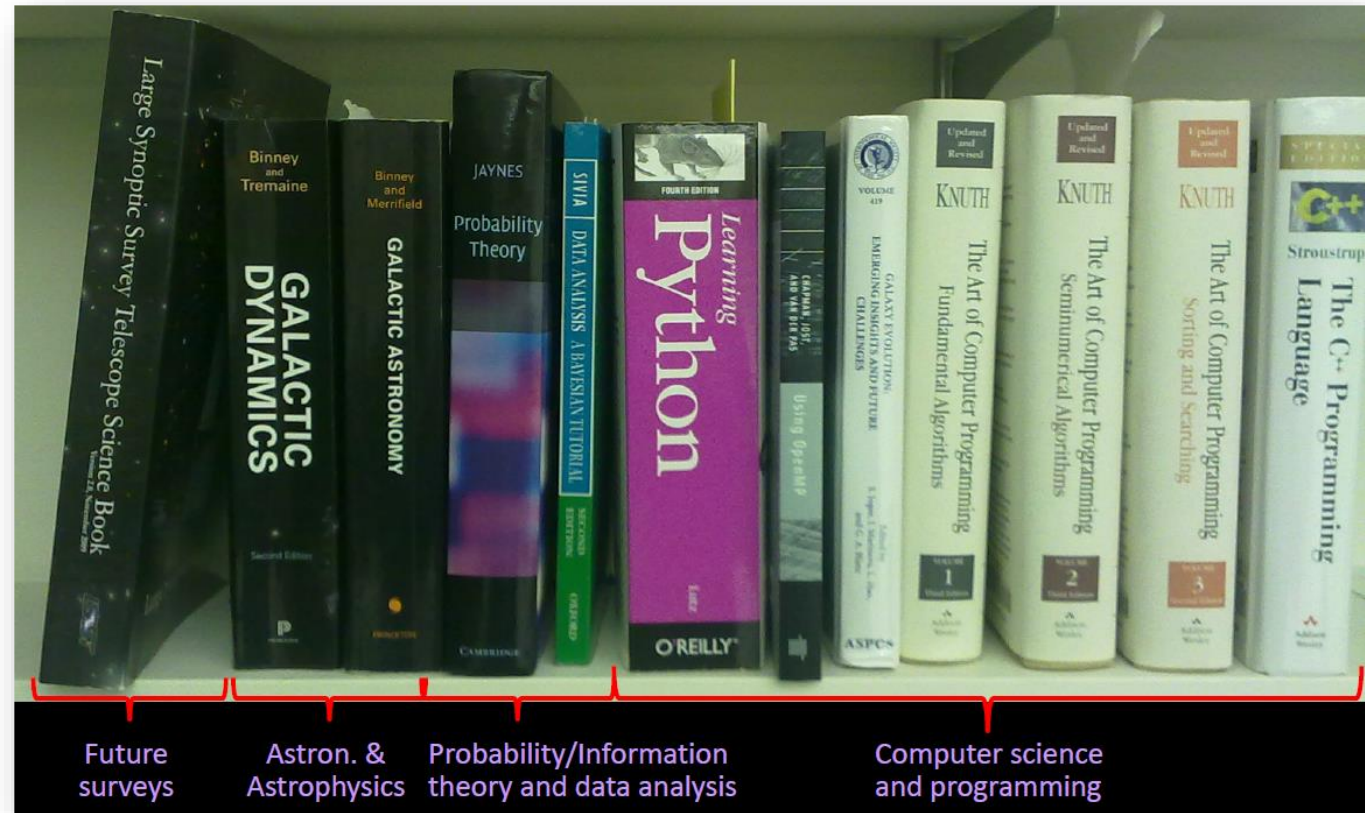


***Key challenges: data heterogeneity and complexity***

# Doing Astronomy in the age of large surveys

Traditionally, Astronomy was a data-starved science. Our approach to research and our analysis methods were shaped by this environment. Surveys are altering it; data is becoming abundant and of unprecedented quality.

Upcoming surveys will cap this transformation. For example LSST will deliver positions, magnitudes and variability information for virtually everything in the southern sky to 24<sup>th</sup> – 27<sup>th</sup> magnitude, with an order of magnitude better controlled systematics than current surveys.

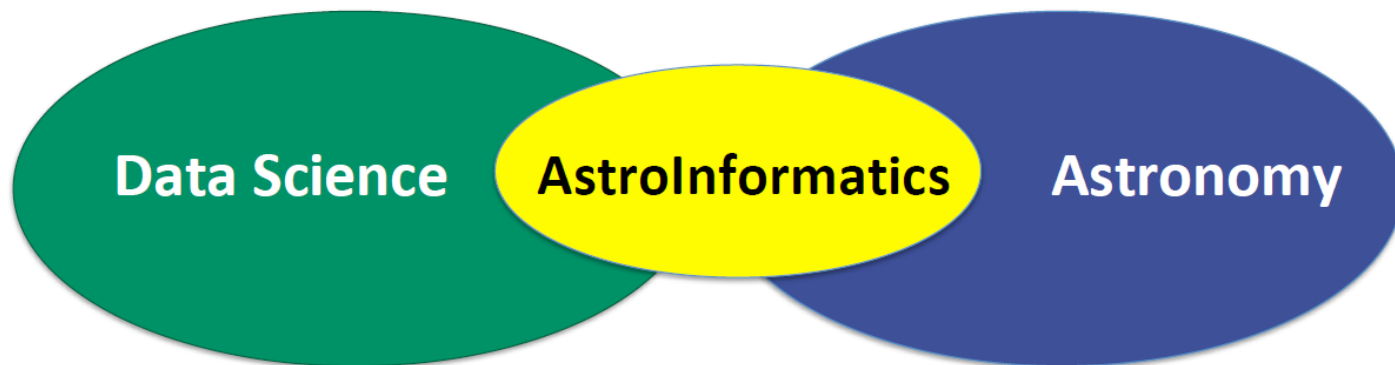


We are entering the age of abundance of high quality data.

Success in research will depend on the ability to analyze and mine knowledge from that data.

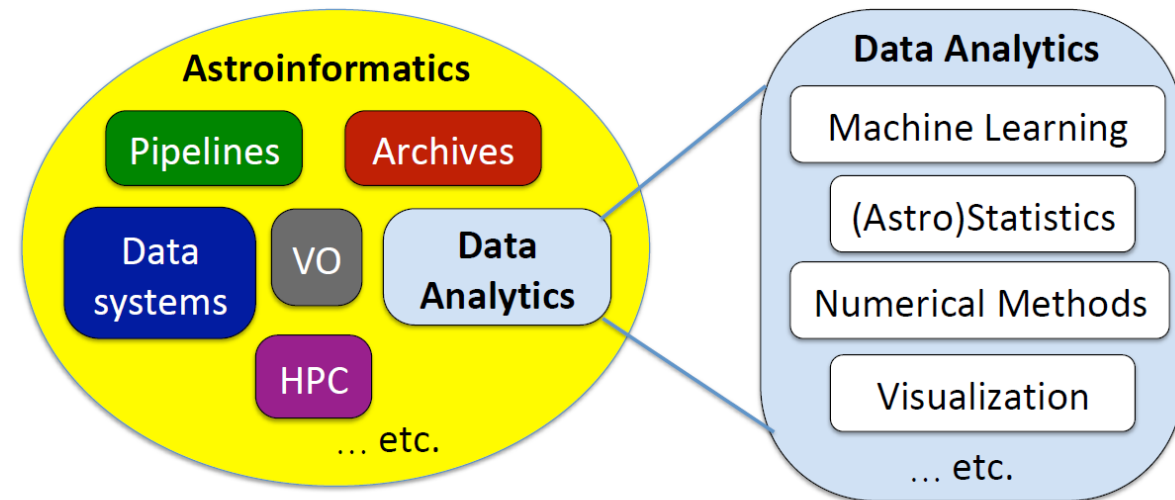
# Astroinformatics

is essentially astronomical applications of Data Science



It contains all of the components of Data Science, in their astronomical applications

- While VO became a global data grid of astronomy, astroinformatics focuses of the **knowledge discovery tools**
- It includes a growing community of scientists, both as contributors and as users
- Like other X-Informatics (X = bio, geo, ...) it is a bridge between astronomy and data science, and for the methodology sharing with other fields.



... and their interconnections

# Astroinformatics – new perspective

Characterize the known

Feature selection, Parameter space analysis

Assign the new from the known

Supervised learning, Regression, classification

Explore the unknown

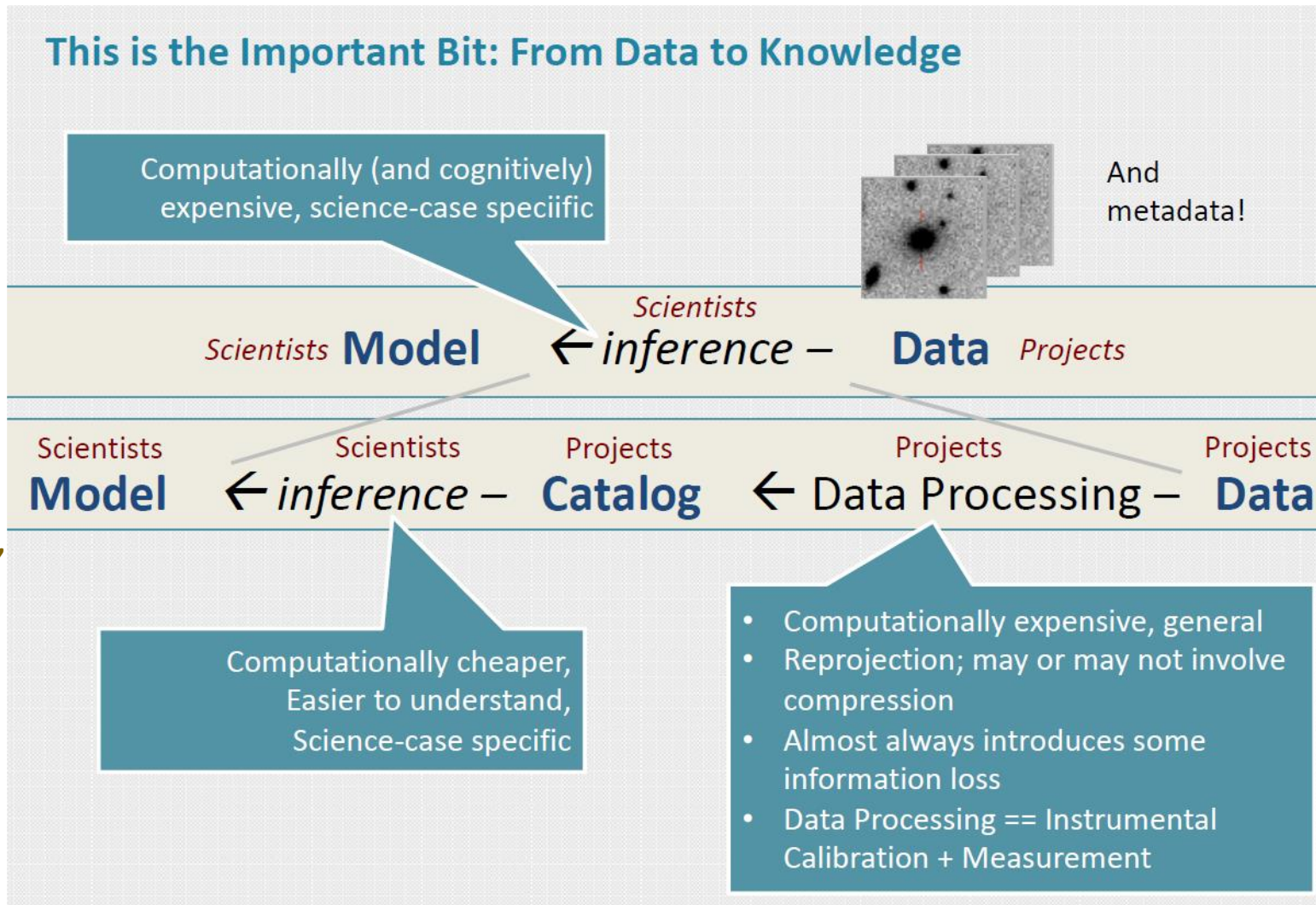
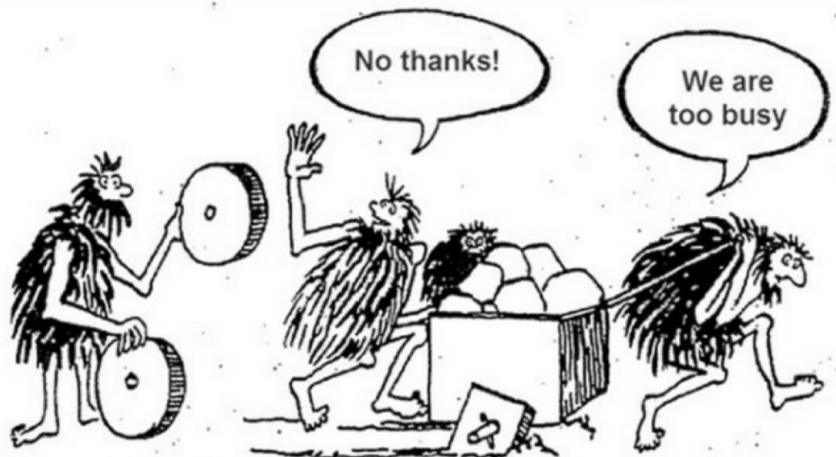
Clustering, unsupervised learning

Discover the unknown

Outlier detection and analytics (serendipity)

Benefits of very large datasets:

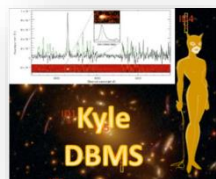
Statistics of “typical” events, cross-correlation, automated search for “rare” events



# Our available methods and fields of interest

<http://dame.fisica.unina.it/>

Database Management System  
(see Giuseppe's talk)



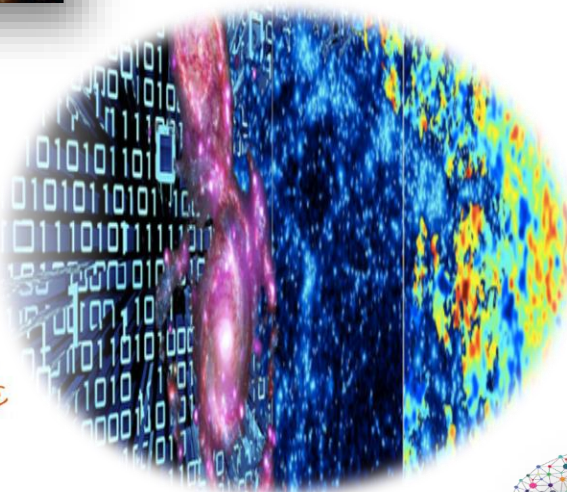
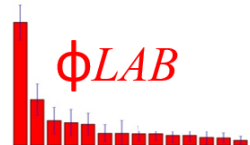
Catalogue cross-matching



Table/image analysis,  
monitoring and statistics  
(see Giuseppe's talk)



Parameter Space exploration,  
(feature selection)



Time series prediction  
and classification (LSTM)



Clustering  
(Growing Neural Gas)



Image/catalogue  
source classification



Deep Learning (CNNs) and data  
augmentation (GAN)



Multi-dimensional data  
visualization (tSNE)



Bayesian Augmenting with  
Gaussian Analytics for Stream  
Classification



Data Mining web app. Includes:

- ❖ MLPQNA neural networks
- ❖ Support Vector Machine
- ❖ Random Forest, K-Means, SOM, genetic algorithms

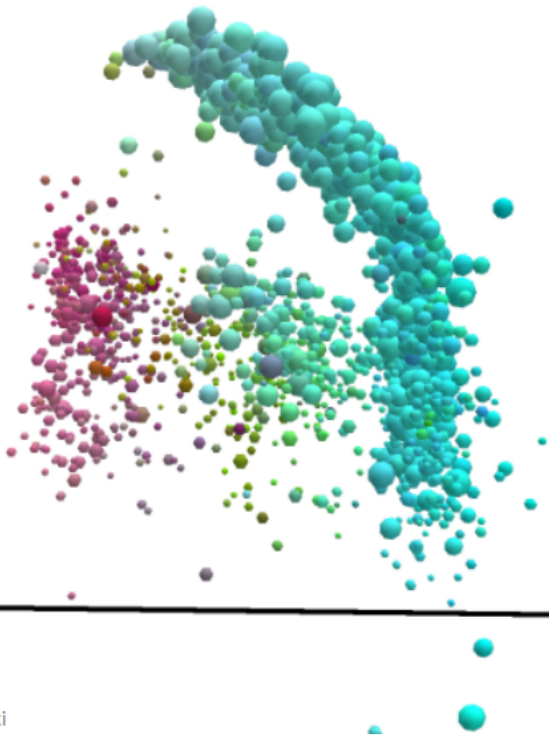


# Parameter Space Exploration

## Exploration of Parameter Spaces is a Central Problem of Data Science

Clustering, classification, correlation and outlier searches, ...

### Machine Learning Is the Key Methodology



### Challenges:

- Algorithm and data model choices
- Data incompleteness
- Feature selection and dimensionality reduction
- Uncertainty estimation
- Scalability
- Visualization
- ... etc.

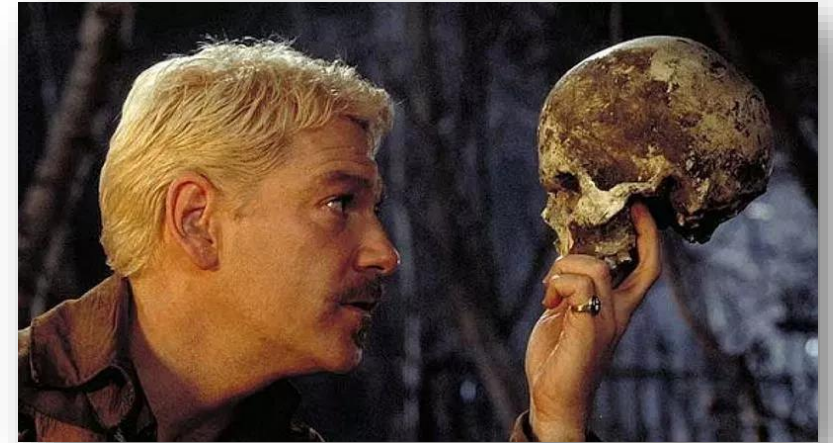
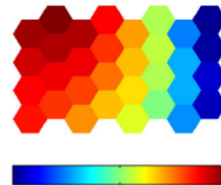
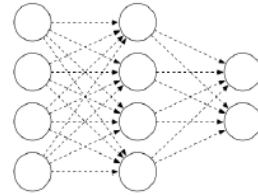
} Especially with the data dimensionality



# Supervised or Unsupervised?...

## Classification, Clustering, and Outliers

- **Supervised learning (classification):** use a known set of objects to train a classifier
  - Hard to find previously unknown things
- **Unsupervised learning (clustering):** let the data tell you how many different kinds of things are there
  - Could find previously unknown types as outliers



### Supervised Algorithms

Neural Networks (MLP)  
 Boltzmann Machines  
 RBM  
 Decision Trees  
 Nearest Neighbor  
 Naive Bayes Classifiers  
 Bayesian Networks  
 Gaussian Processes  
 Regression  
 ...

There is **no** “one size fits all”:  
 different choices  
 for different  
 problems

### Unsupervised Algorithms

K-Means  
 Self-Organizing Maps  
 RDF  
 Fuzzy Clustering  
 CURE  
 ROCK  
 Vector Quantization  
 Probabilistic Principal Surfaces  
 ...

## figure of merit

### Supervised Learning

**Input:** a list of objects with measured properties and **labels**.

The algorithm is optimizing a score (cost function) that depends on the input labels and predicted labels.

**Prior knowledge is required!**

### Unsupervised Learning

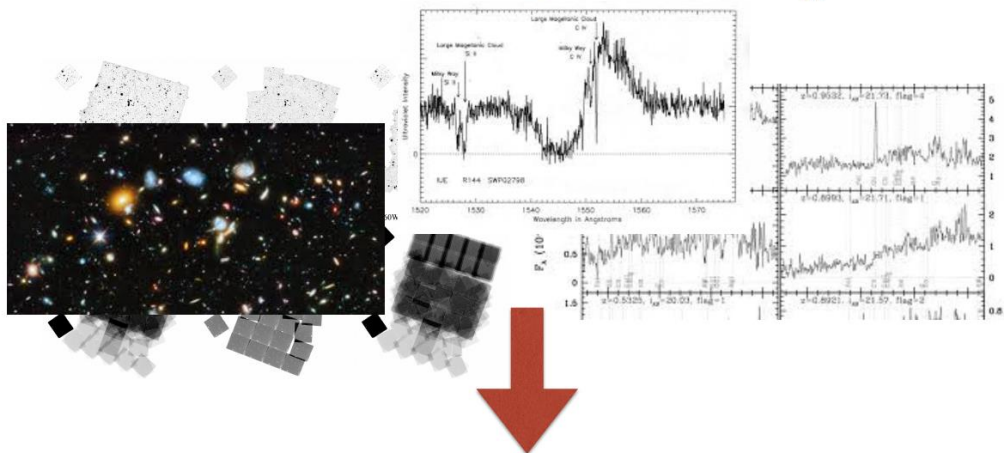
**Input:** a list of objects with measured properties.

The algorithm detects clusters, complex relations, outliers, or reduces the dimensions of the dataset.

**Prior knowledge isn't required!**

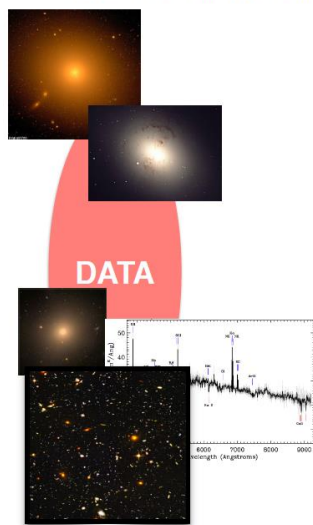
# Deep Learning

What do we put as input?



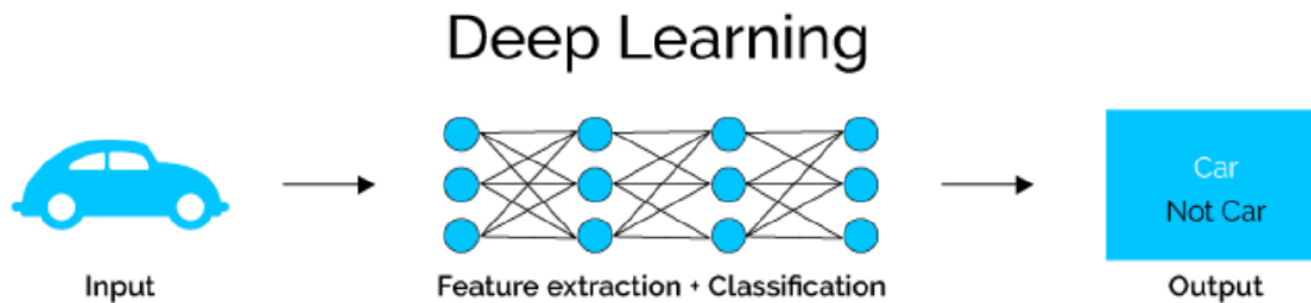
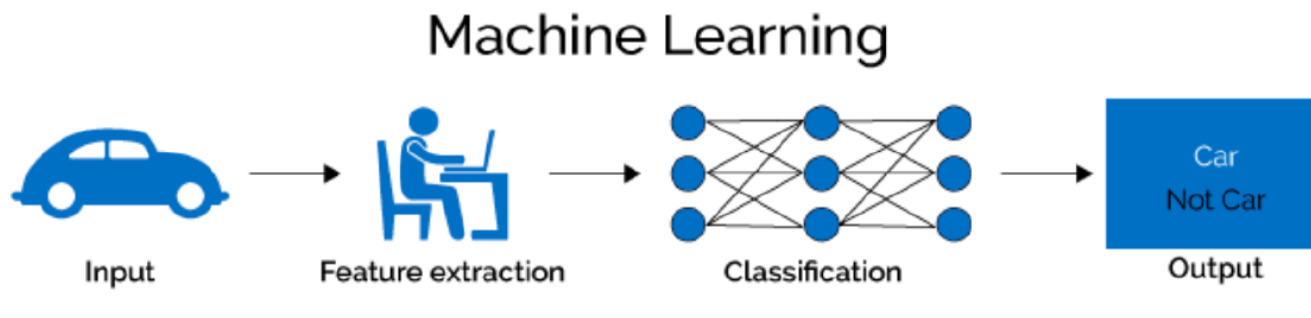
PRE-PROCESS DATA TO EXTRACT MEANINGFUL INFORMATION

THIS IS GENERALLY CALLED **FEATURE EXTRACTION**



- Spiral!
- Emission line!
- Merger!
- AGN!
- Clump!

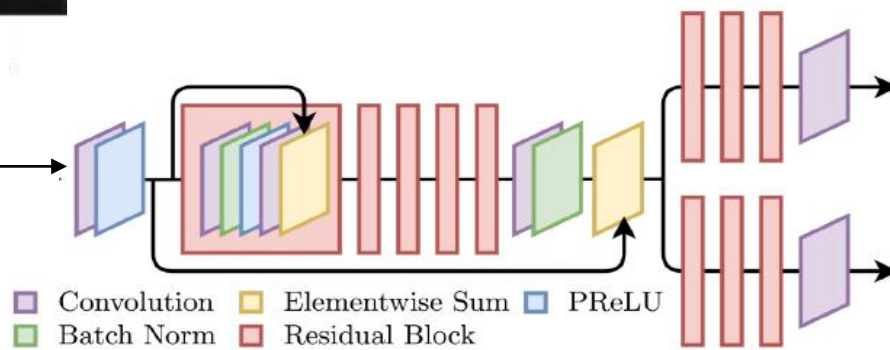
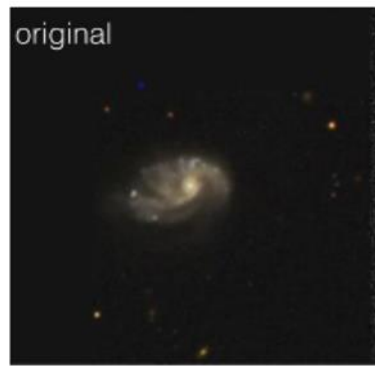
THIS IS A CHANGE OF PARADIGM!



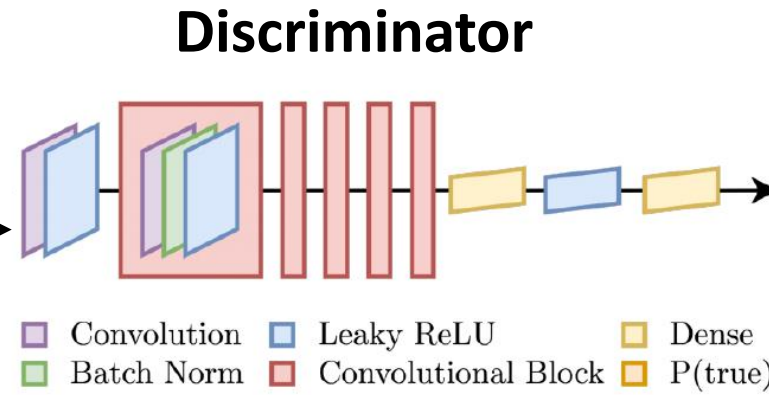
# Generative Adversarial Networks

## Data Augmenting

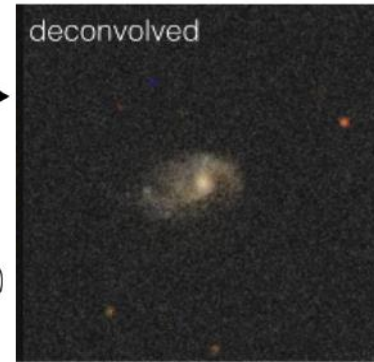
## CNN training



## Generator



## Discriminator



# Use case example: ALMA datacube analysis

**Study of overdensities within high-redshift QSO ( $4.65 < z < 6.67$ ) environments: proto-clusters tracers**

ALMA: interferometer with 66 antennas, reaching an angular resolution  $\approx$  optical telescopes, ideal at mm and sub-mm frequencies

ALMA more effective than HST to detect CII at high-z and able to avoid spectro follow-up to derive spec-z

As higher luminous sources, QSOs are more easily detectable at high-z. Thus they are ideal as proto-cluster tracers

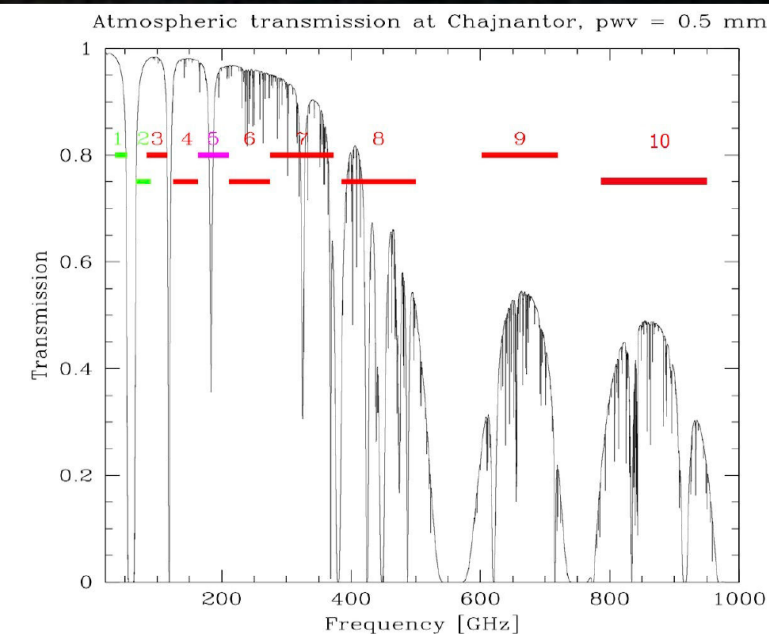
$\text{Ly}\alpha$  cannot be used as cold gas tracer at high-z, mostly due to the ISM obscuring, contamination by sky rows etc.

A valid alternative is CII in the FIR at  $\lambda=158$  and  $z > 6$ , as the dominant cooler of ISM in star-forming galaxies ( $\sim 0.1\% - 1\%$  of contribution to the FIR galaxy luminosity) and with a sufficient luminosity to derive a precise spec-z



Workshop Laboratorio Spettroscopia INAF – Roma, 10-11 Giugno, 2019

Band	frequency (GHz)/(mm)
3	84.0-116.0/2.59-3.57
4	125.0-163.0/1.84-2.40
5	163.0-211.0/1.84-2.40
6	211.0-275.0/1.09-1.42
7	275.0-373.0/0.80-1.09
8	385.0-500.0/0.60-0.78
9	602.0-720.0/0.42-0.50
10	787.0-950.0/ 0.32-0.38

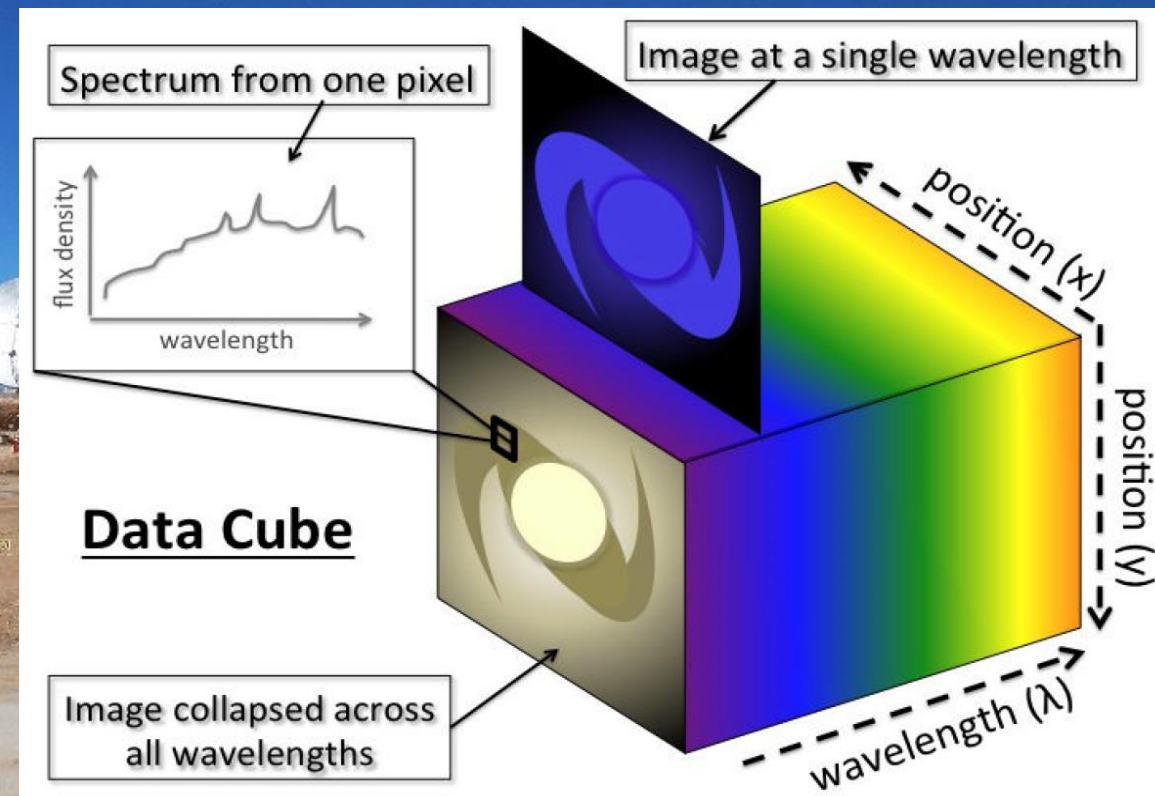
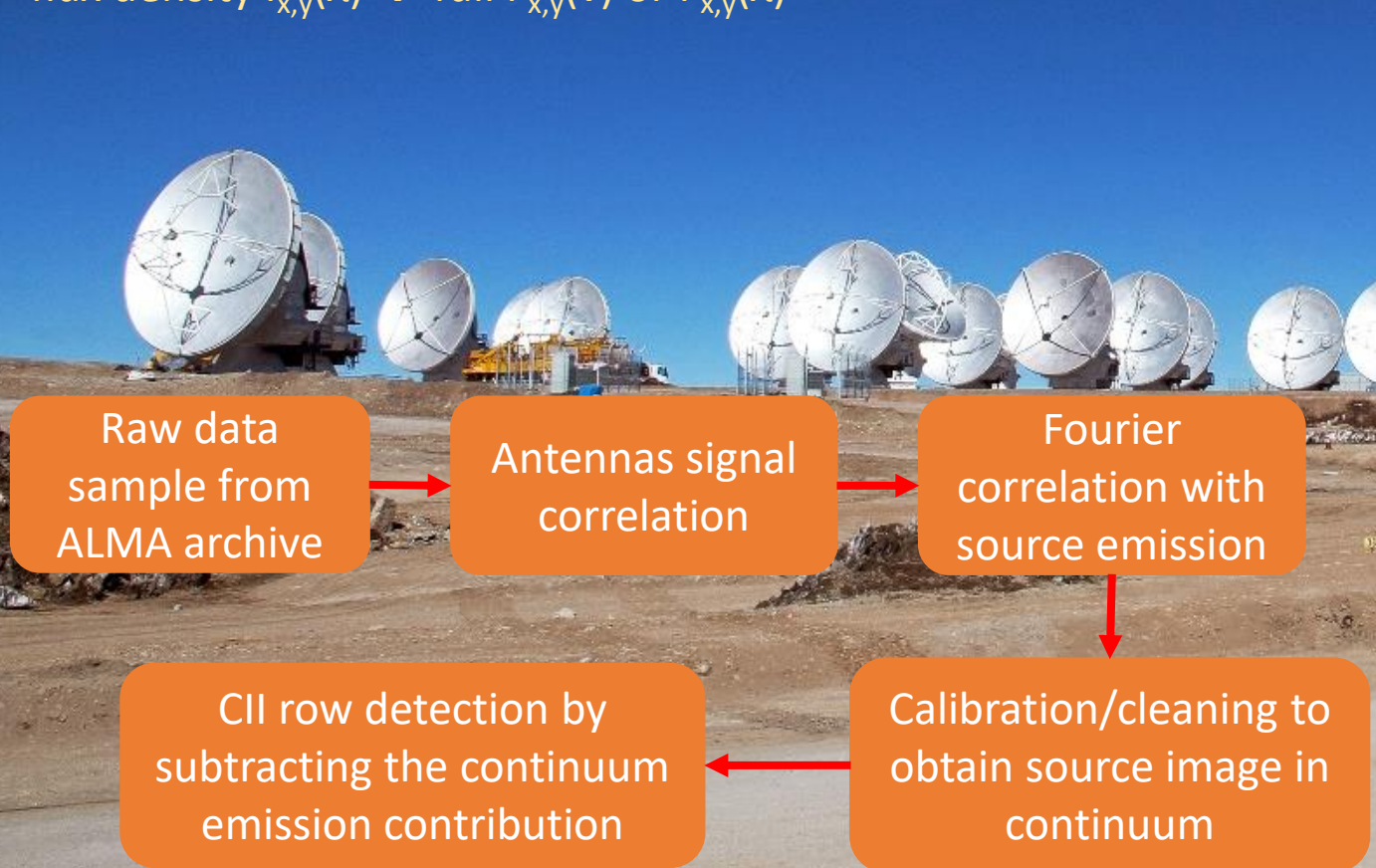


# ALMA datacube analysis

Each datacube section shows the source surface brightness  $I_{x,y}(\nu)$  for each sky pixel  $(x,y)$

Source spectrum for each pixel showing  $I_{x,y}(\nu)$  or flux density  $I_{x,y}(\lambda) \rightarrow$  full  $F_{x,y}(\nu)$  or  $F_{x,y}(\lambda)$

Data sample: 21 sources in  $4.65 < z < 6.5$ , CII at  $\lambda = 0.158 \mu\text{m}$

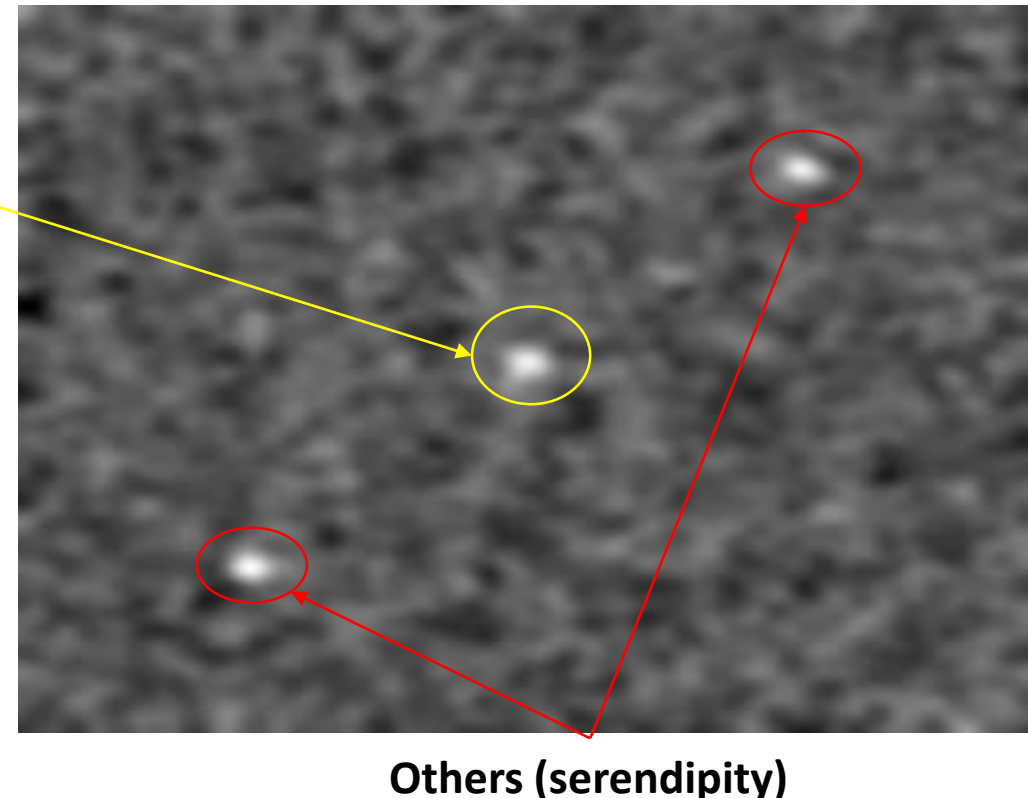
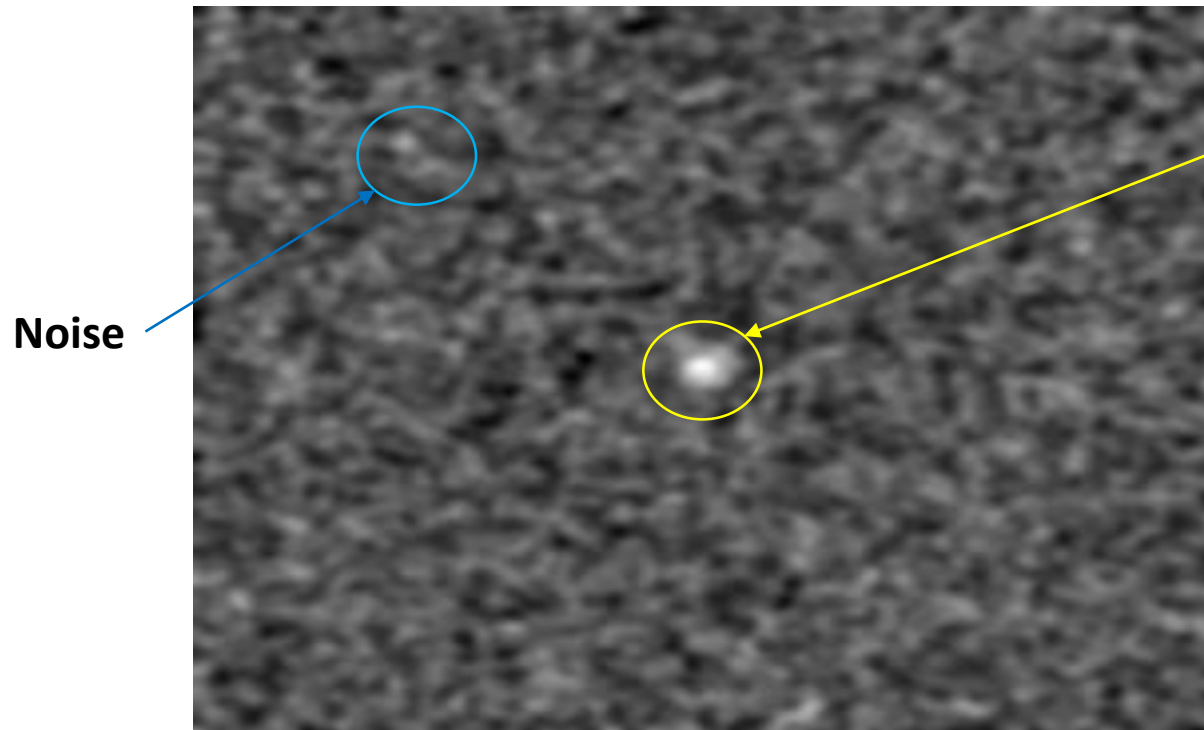


# ALMA datacube analysis

**Goal: to detect all source candidates in each datacube**

**Method for datacube analysis**

<b>Input datacube</b>	<b>Detection threshold <math>S/N &gt; 3</math></b>	<b>Re-binning of channels in frequency (to remove noise and detect signal peaks)</b>	<b>From “neighbor peaks” in frequency, got peak with highest S/N</b>
-----------------------	--	--	--



# ALMA datacube analysis

Simulated datacubes (with Gaussian noise)  
provided to verify the method accuracy

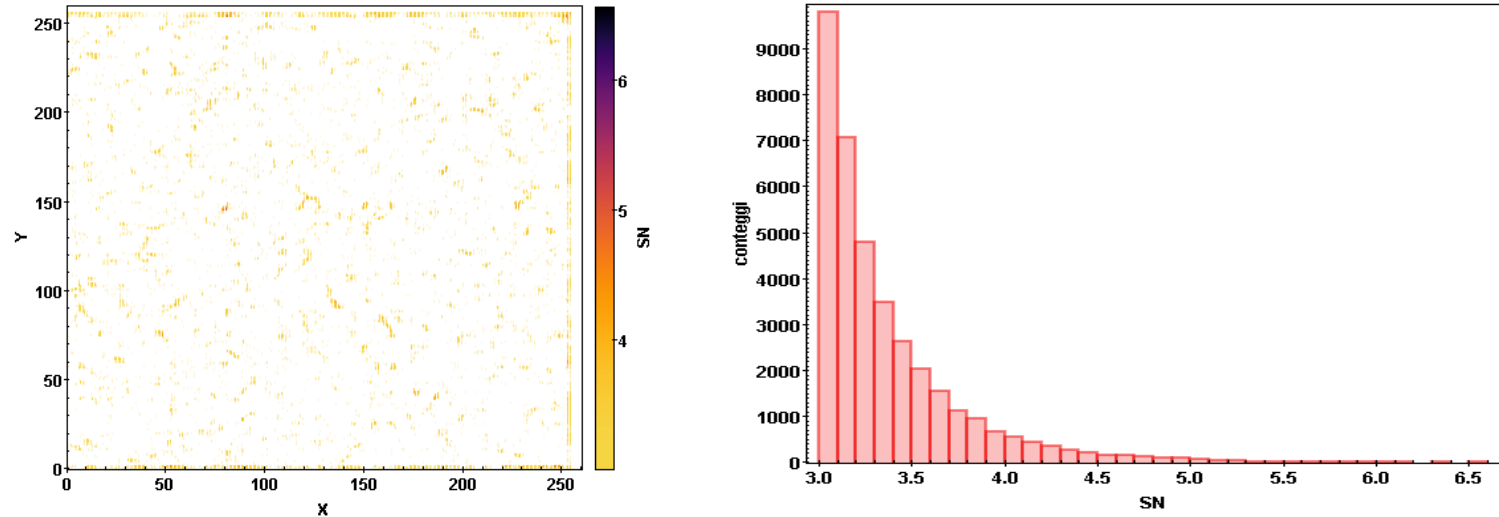
$$L_{line} = 1.04 \times 10^{-3} \times S_{line} \Delta v D_L^2 v_{obs} L_{\odot}$$

*Carilli & Walter 2013*

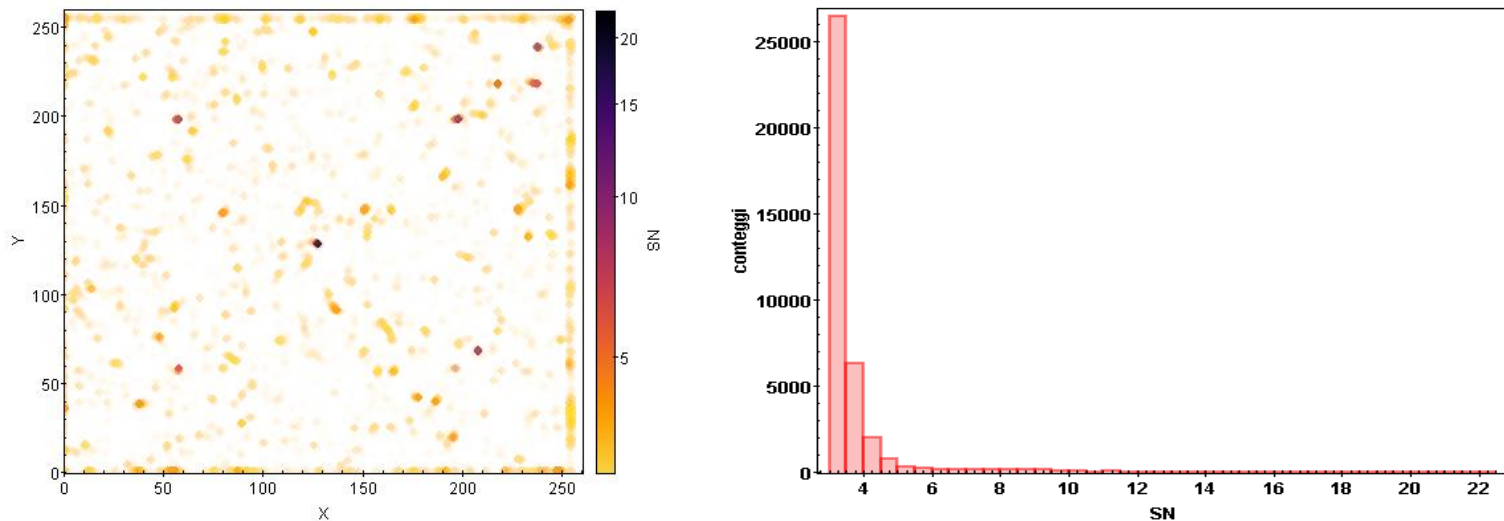
Generation of source catalogue  
with: position, flux, S/N and luminosity ( $L_{SOL}$ )

TARGET	X	Y	FLUSSO	S/N	L ( $L_{SOL}$ )
QSO J0842-1218	167	197	0.025	8.4	1.4E+09
SDSS J092303.53+024739.5	194	46	0.003	6.1	4.7E+07
	201	51	0.018	11	3E+08
	206	48	0.029	14	5.1E+08
QSO J1319+0950	122	140	0.003	6.1	6E+07
	137	129	0.011	9.4	2.3E+08
	126	145	0.007	6.1	1.5E+08
SDSS J132853.66-022441.6	234	42	0.016	10	3.2E+08
CFHQS J210054-171522	9	131	0.007	6.5	3.3E+08
	9	134	0.035	6.6	1.6E+09
PJ065-26	123	126	0.038	6.2	1.6E+09
PSO J167.6415-13.4960	121	122	0.021	7.8	4.7E+08
PJ231-20	123	128	0.006	6.5	1.3E+08
	124	111	0.048	12	1.1E+09
J308-21	77	120	0.009	6	2E+08
	127	136	0.014	8.5	2.8E+08
	157	116	0.016	6.9	3.4E+08
	70	121	0.013	6.5	2.8E+08
[WMH2013] 5	129	127	0.007	6.8	1.4E+08
QSO J1509-1749	131	133	0.005	6.2	2E+08
QSO J1306+0356	121	126	0.022	8.1	5.5E+08
	118	124	0.022	6.6	5.4E+08

Datacube with Gaussian noise only



Datacube with Gaussian noise and sources





# ALMA datacube analysis

Determination of Luminosity Density Function, by taking into account:

- ✓ Noise within ALMA datacubes
- ✓ For each source, total volume of datacubes in which the source is detectable
- ✓ Datacube volume ( $\text{Mpc}^3$ ), where 3<sup>rd</sup> dimension is the frequency
- ✓ Amount of sources per  $\text{Mpc}^3$  over a certain luminosity

Comparison with literature

- ✓ General agreement
- ✓ Higher luminosity depth ( $\sim 1$  order of growth)
- ✓ Larger data sample
- ✓ Overestimate w.r.t. theoretical models (*de Looze+2014*)
- ✓ Overdensity of emitters [CII] in QSO high-z fields

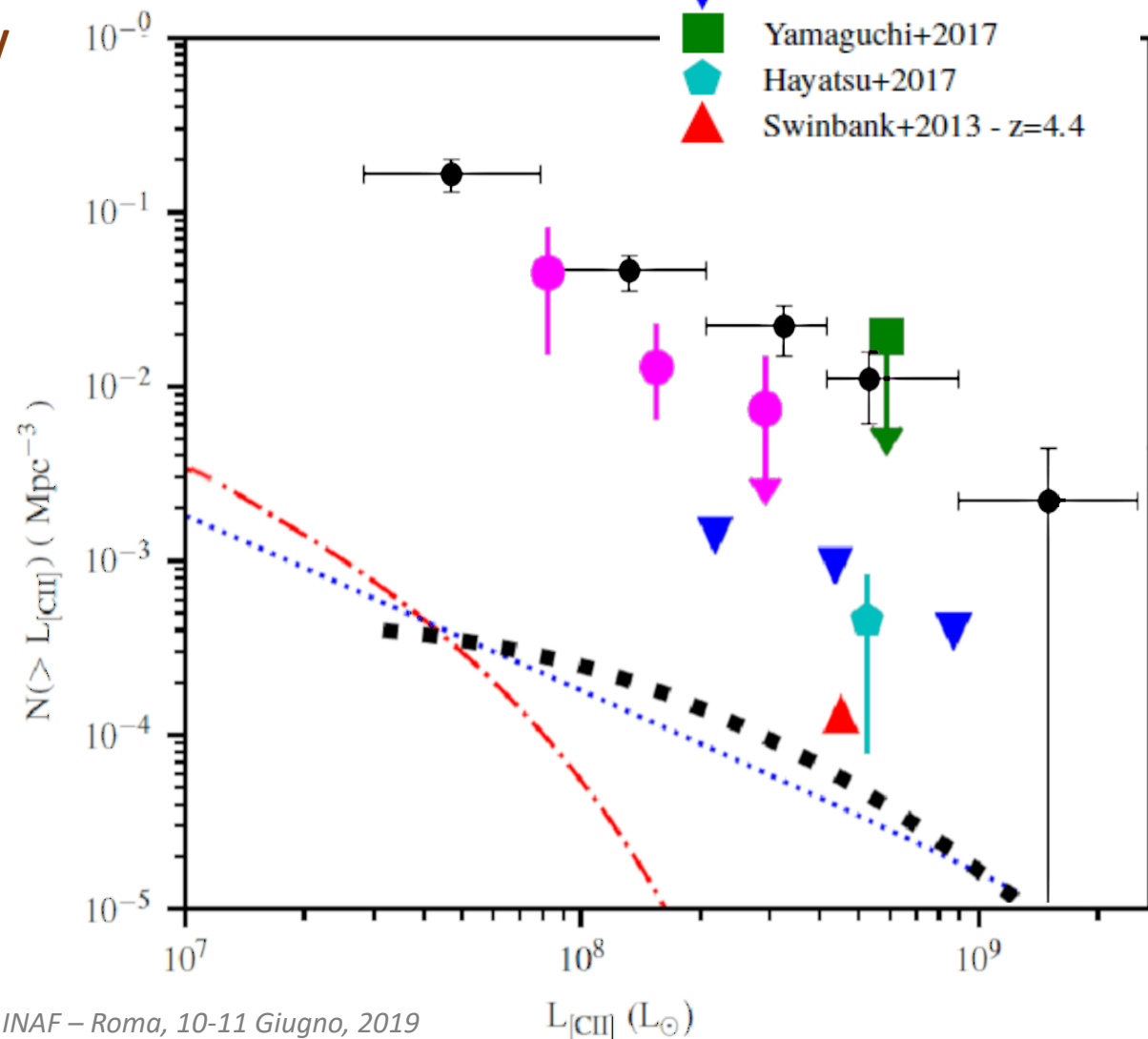
**NEXT TASK:**

**Alternative approach to detect candidates with Deep Neural Networks (CNN+GAN)**



**WG: A. Marconi, M. Brescia, S. Carniani, G. Angora, G. Longo, R. Ragusa**

- This work - Candidates
- Miller+2018
- Popping+2016
- - - Lagache+2017
- Hayward, Behroozi+2013
- ▼ Aravena+2016
- Yamaguchi+2017
- Hayatsu+2017
- ▲ Swinbank+2013 - z=4.4



# Cluster members identification

Identification of Cluster Members (CM) from other source types.

HST ACS/WFC3 images, KB: spectro sources, assuming as CM a galaxy with separation from cluster  $< 3000$  Km/s (rest frame)

*Grillo+ 2015, Caminha+ 2016*

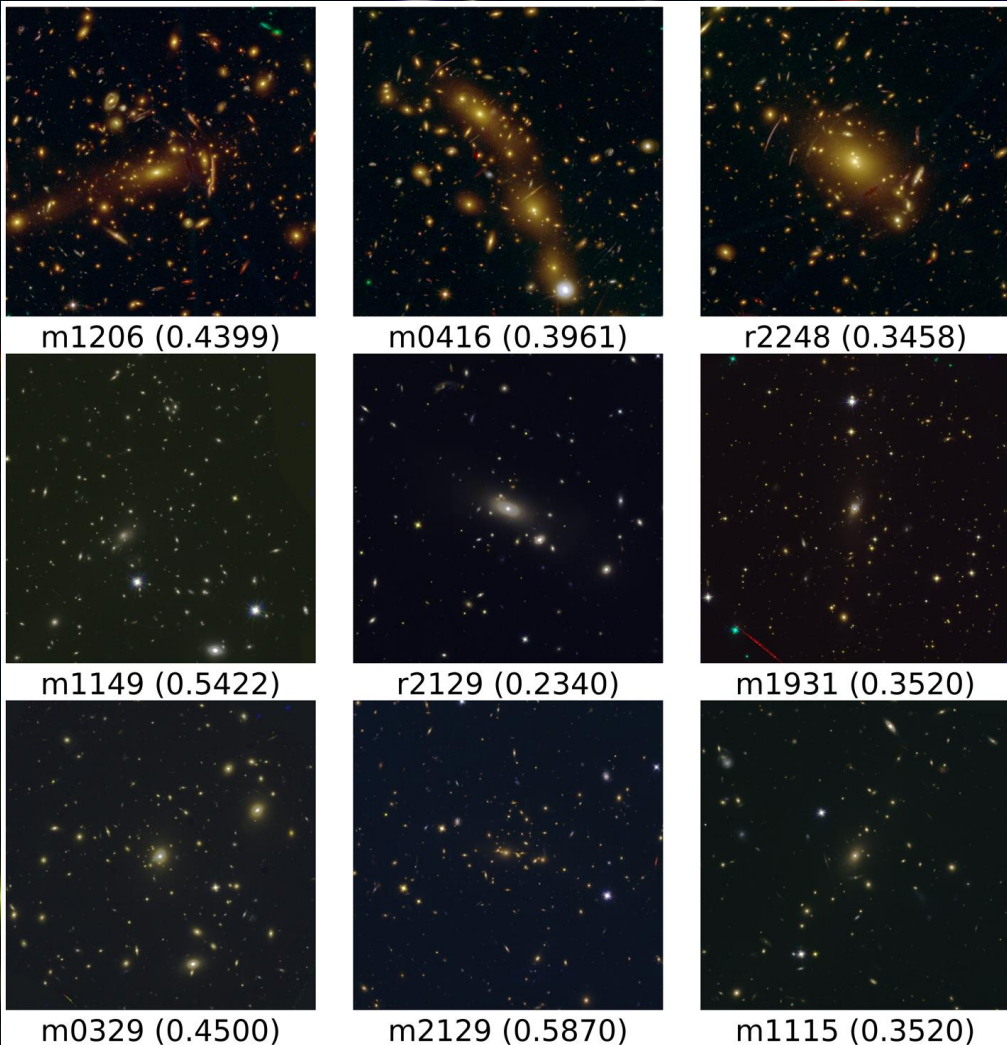
For each source we extracted from HST images a squared thumbnail with a side of  $\sim 3.8''$ , centered on the source position

Given the limited number of sources (about 100 CMs for each cluster), all the experiments involve a data augmentation based on rotations and flips.

This pre-process makes the network invariant to the performed transformations.

In order to avoid the introduction of strong correlations within the dataset, we constrain the pre-processing, by applying the transformation to the 15% of the sample, implying an augmentation factor of 2.05.

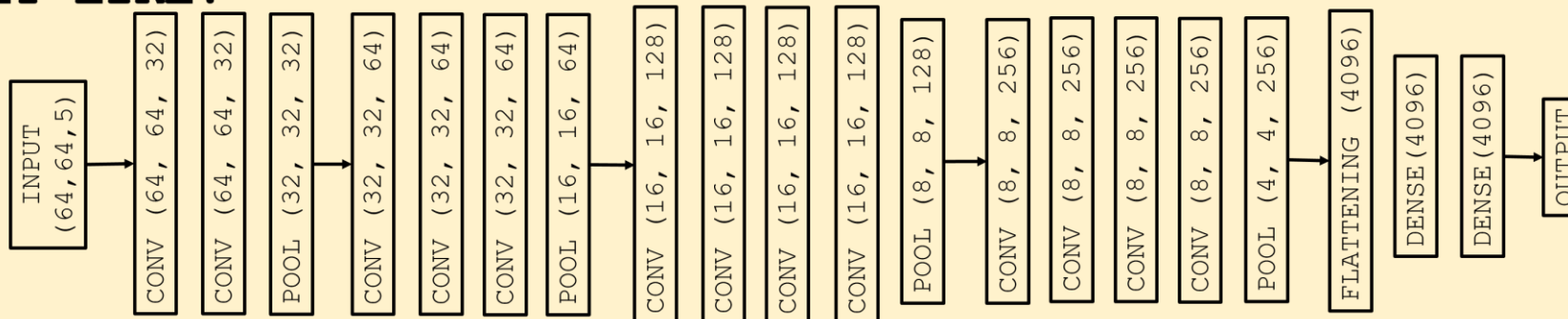
***G. Angora, P. Rosati, M. Meneghetti, A. Mercurio, M. Brescia***



# Cluster members identification

We approached the problem with Deep Learning: two Convolutional Neural Networks (CNN)

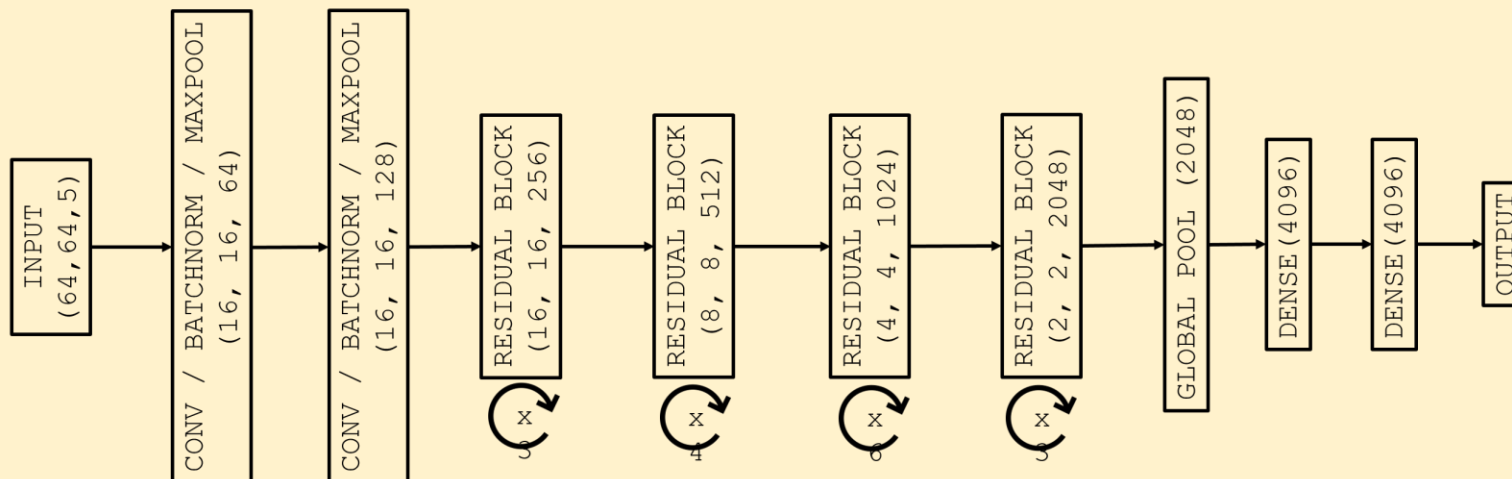
## VGGNET LIKE:



VGGNET: a canonical CNN, based on a chain of convolution + pooling layers and cross-entropy as cost function

RESNET: a more complex CNN, including residual blocks, i.e. additional blocks implementing the identity mapping (adding original input to the output). It helps to build more complex networks, avoiding the known problem of “evanescing error gradient”

## RESNET LIKE:



# Cluster members identification

We performed 2 different classification experiments:

- ❑ First training set is obtained by **stacking several clusters**, forcing the networks to extract features able to recognize CMs at different depths (i.e. showing different photometric and morphologic properties);
- ❑ Second training set is built with a stack of low-redshift clusters, testing the **dependence on redshift** and the possibility to predict CM at different depths.

**stacking**

**Z dependence**

VGGNET		[1]	[2]	[3]	[4]	[5]	[6]	[7]
	AE	87.1	90.5	90.9	89.5	90.1	90.7	90.9
CM	Pur	86.8	91.6	91.6	90.5	89.7	90.2	92.6
	Comp	85.5	91.9	90.4	87.8	90.7	87.0	85.9
	F1	86.1	91.7	91.0	89.1	90.2	88.6	89.2

RESNET		[1]	[2]	[3]	[4]	[5]	[6]	[7]
	AE	91.8	91.3	90.7	89.7	89.8	91.1	90.6
CM	Pur	91.7	93.0	89.7	85.4	84.5	88.0	85.7
	Comp	88.0	89.0	90.5	95.4	94.6	94.0	95.8
	F1	89.8	90.9	90.1	90.2	89.7	91.9	90.1

VGGNET		Valid	Test1	Test2
	AE	87.9	81.2	81.5
CM	Pur	91.7	82.4	78.7
	Comp	86.6	80.9	95.1
	F1	89.1	81.2	86.2

RESNET		Valid	Test1	Test2
	AE	90.0	80.9	81.9
CM	Pur	93.4	85.7	89.1
	Comp	86.9	79.1	79.9
	F1	89.9	82.4	84.2

[1] = m1206 + m0416 + r2248; [2] = [1] + m1149;  
 [3] = [2] + r2129; [4] = [3] + m1931;  
 [5] = [4] + m0329; [6] = [5] + m2129;  
 [7] = [6] + m1115

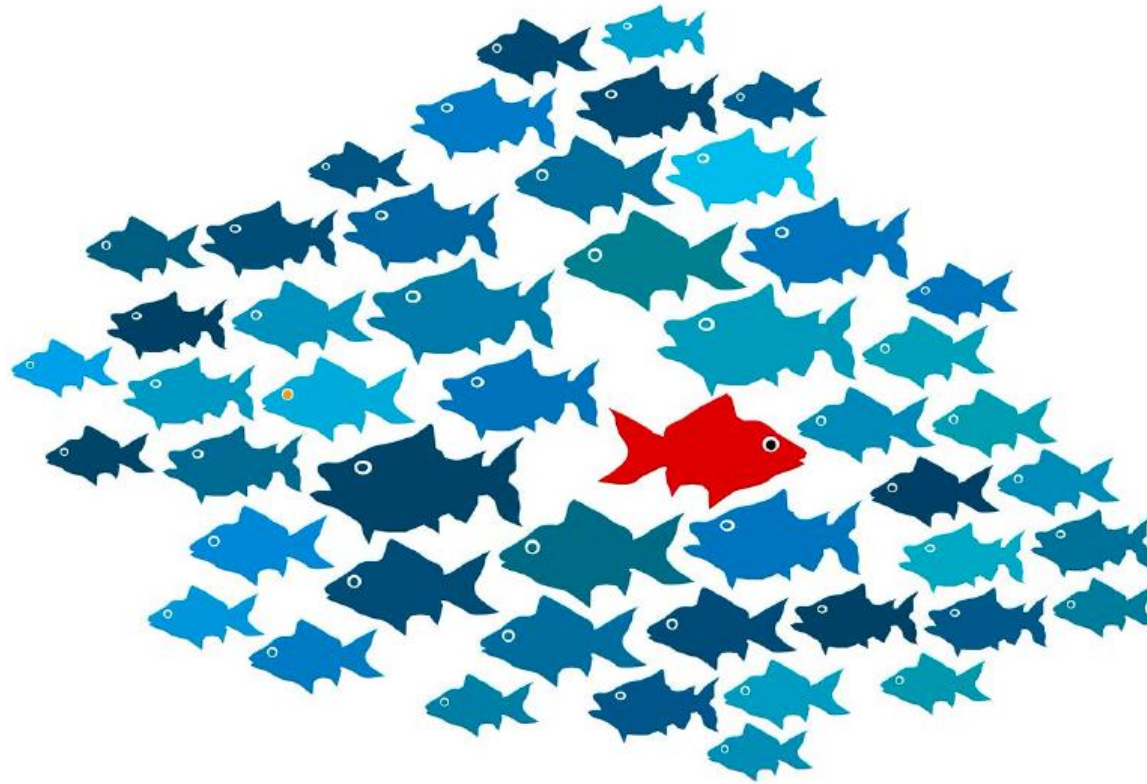
(Blind test set ~30% of KB)

**Train/Valid**(20%) = low-z clusters: m1206 + m0416 + r2248 + m1931 + m0329 + m1115;  
**Test1** = higher-z clusters: m1149 + m2126;  
**Test2** = lower-z clusters: r2129

# Outlier identification

## How do we find outliers?

Supervised learning-based outlier detection will uncover the outliers that “shout the strongest”.

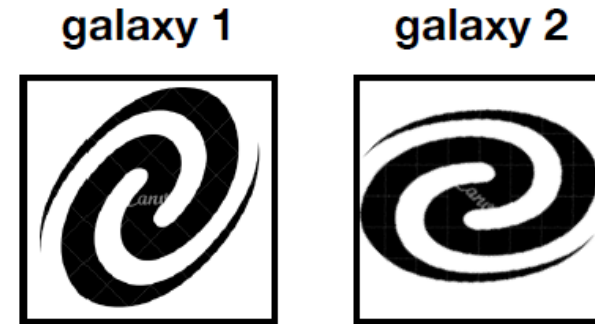
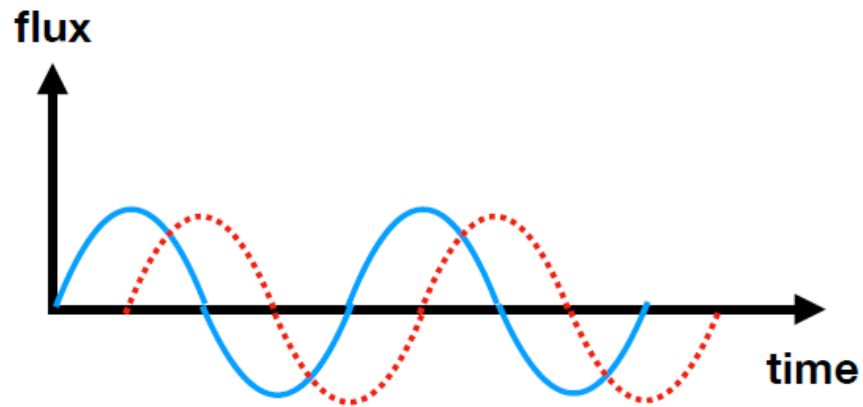


# Unsupervised Random Forest

## Outlier detection on **other datasets** using unsupervised RF?

The unsupervised Random Forest assumes a regular grid, and thus will work for spectra or extracted features.

It will not work for images or time series, because it does not have translational and rotational symmetry!

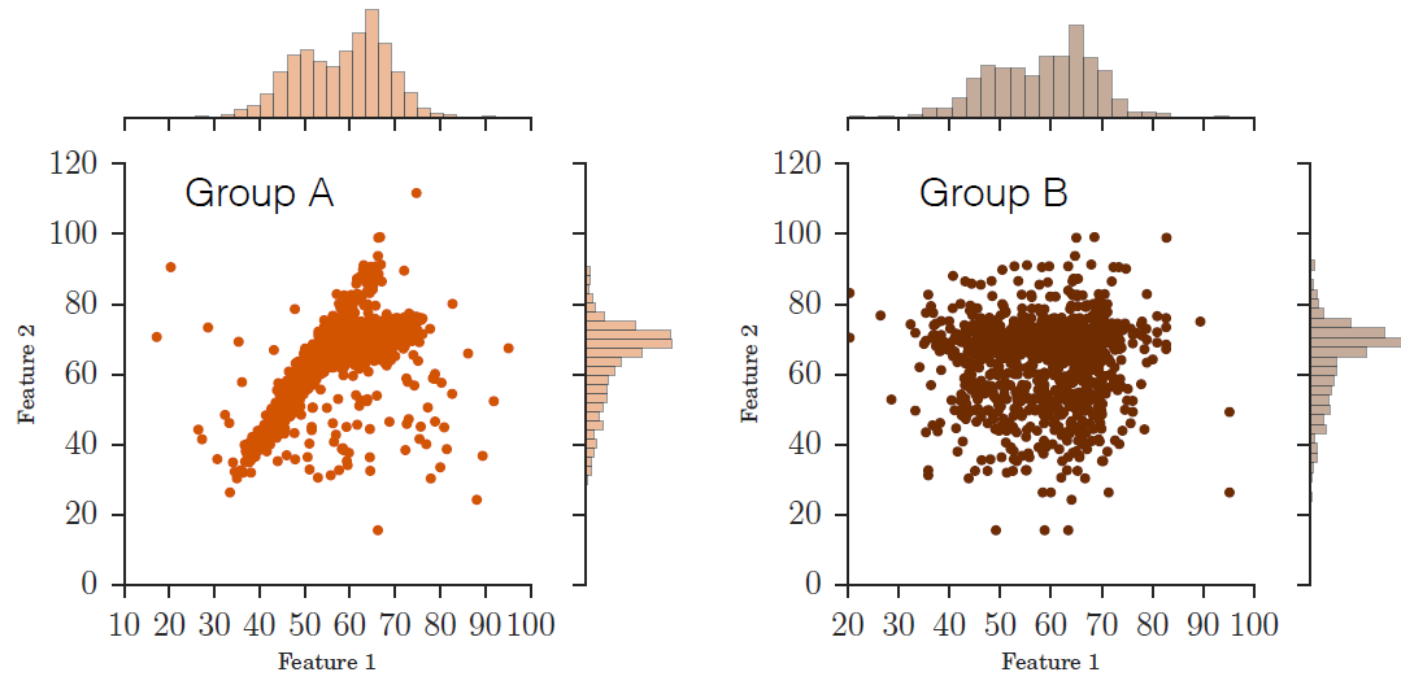
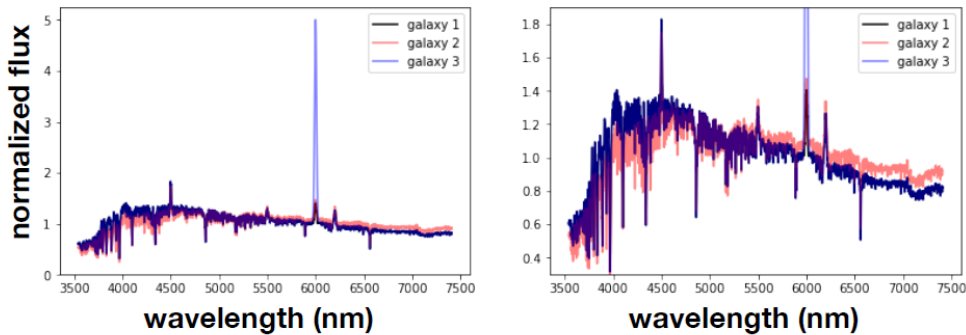
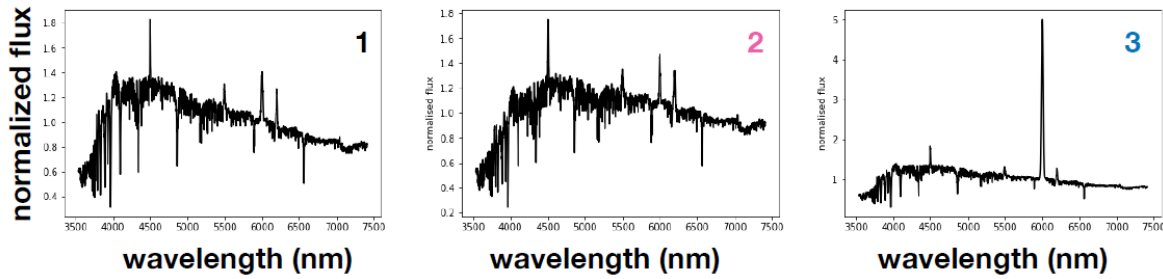


**possible solution:** find a representation of the signal on a regular grid (e.g., FFT of time series).

# Unsupervised RF

**Random Forest** can be used as an unsupervised algorithm, to produce pair-wise similarity for the objects in our sample.

For the problem of finding outliers we do not have a training set. Instead we take our entire dataset - label it as class "real", and generate a synthetic dataset of similar size, and similar marginal distributions in all the features, but without covariance between the features



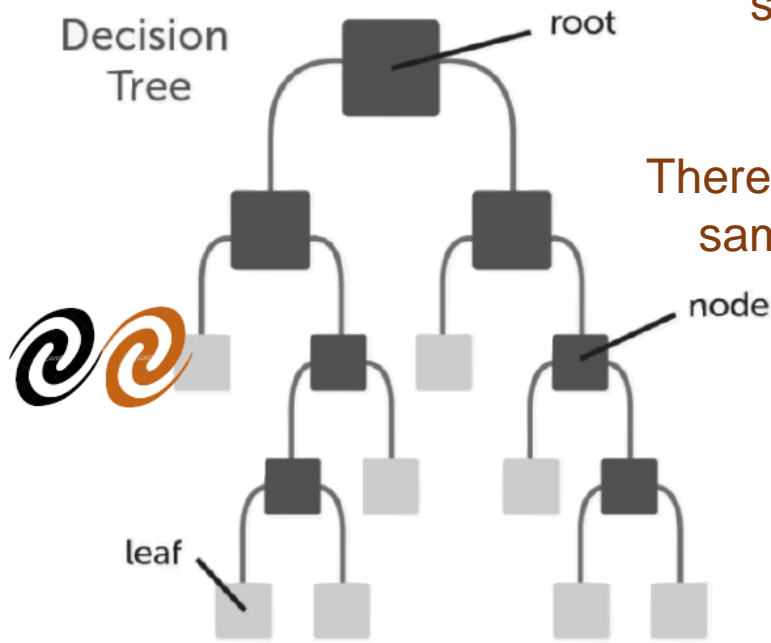
# Unsupervised RF

Training the RF on these sets teaches it to recognize objects that have covariance. Now, given two objects that we pass through all the trees, we can ask how often they ended up in the same "real" terminal leaf.

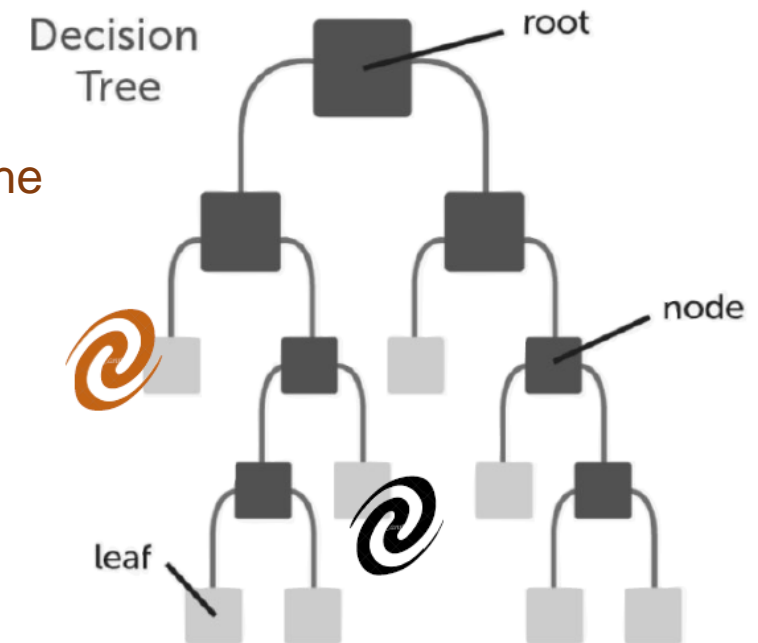
Two completely identical objects will have the exact same features and always end up together.  
Two very dissimilar objects will never do.

Therefore, counting how often two objects land in the same leaf is a measure of similarity, or distance, which was our purpose.

The process is repeated for all the trees in the forest.  
Therefore, the similarity ranges from 0 to N, the number of trees in the forest.



**similarity += 1**



**similarity += 0**



# Dimensionality Reduction

## Why do we need dimensionality reduction?

- **“Practical”:**

- Improve performance of supervised learning algorithms: original features can be correlated and redundant, most algorithms cannot handle thousands of features.
- Compressing data (e.g., SKA).

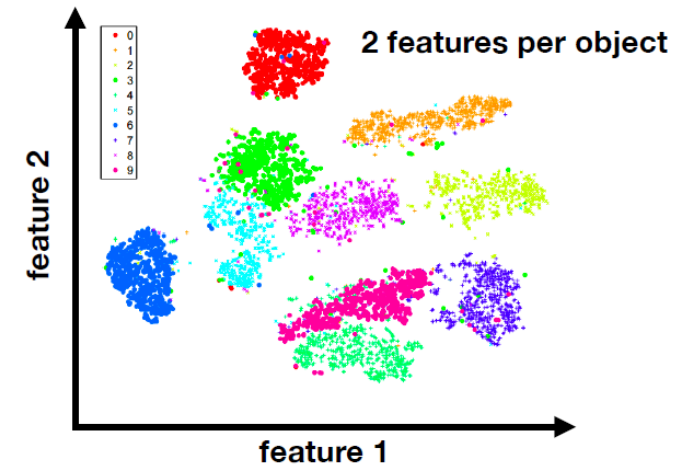
- **“Artistic”:**

- Data visualization and interpretation.
- Uncover complex trends.
- Look for “unknown unknowns”.



28 x 28 features per object

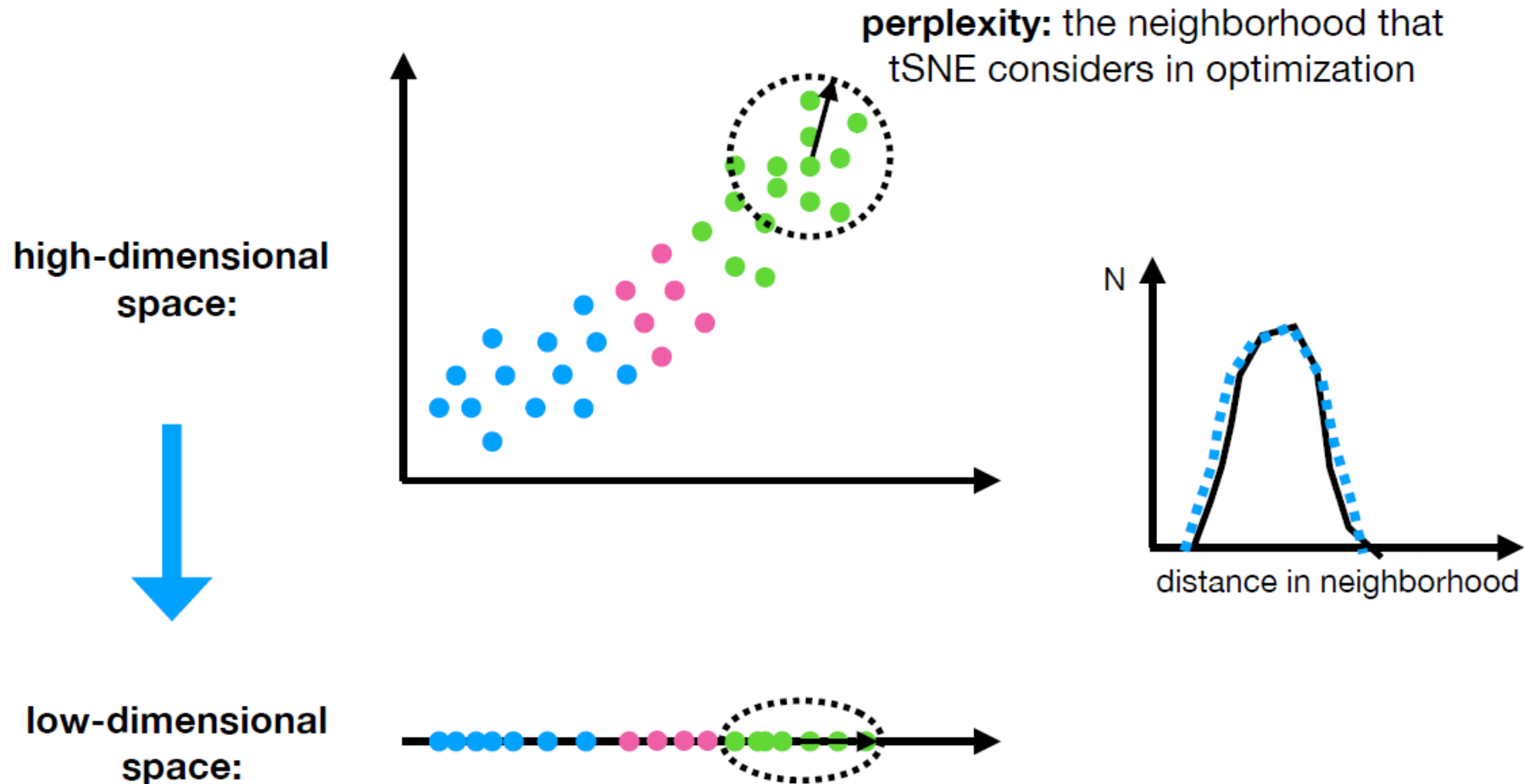
Dimensionality Reduction algorithm



2 features per object

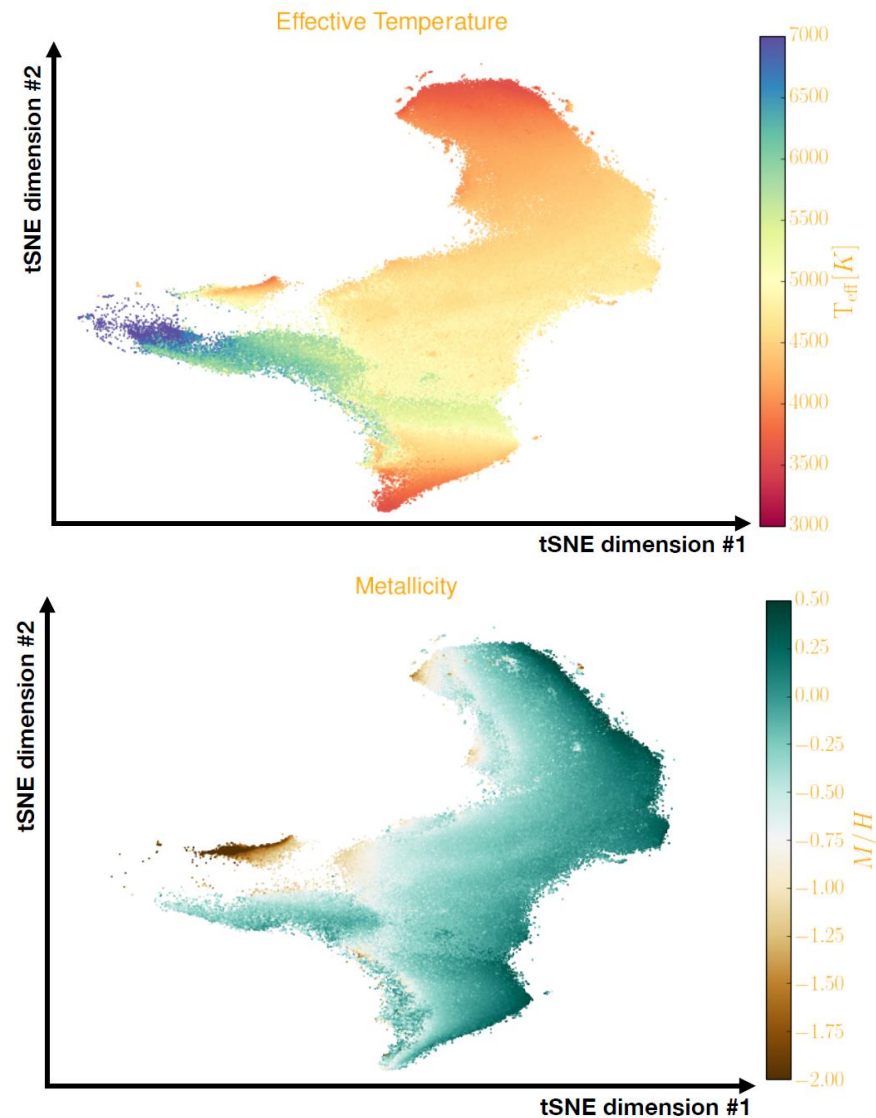
# tSNE

**Intuition:** tSNE tries to find a low-dimensional embedding that preserves, as much as possible, the **distribution of distances** between different objects.

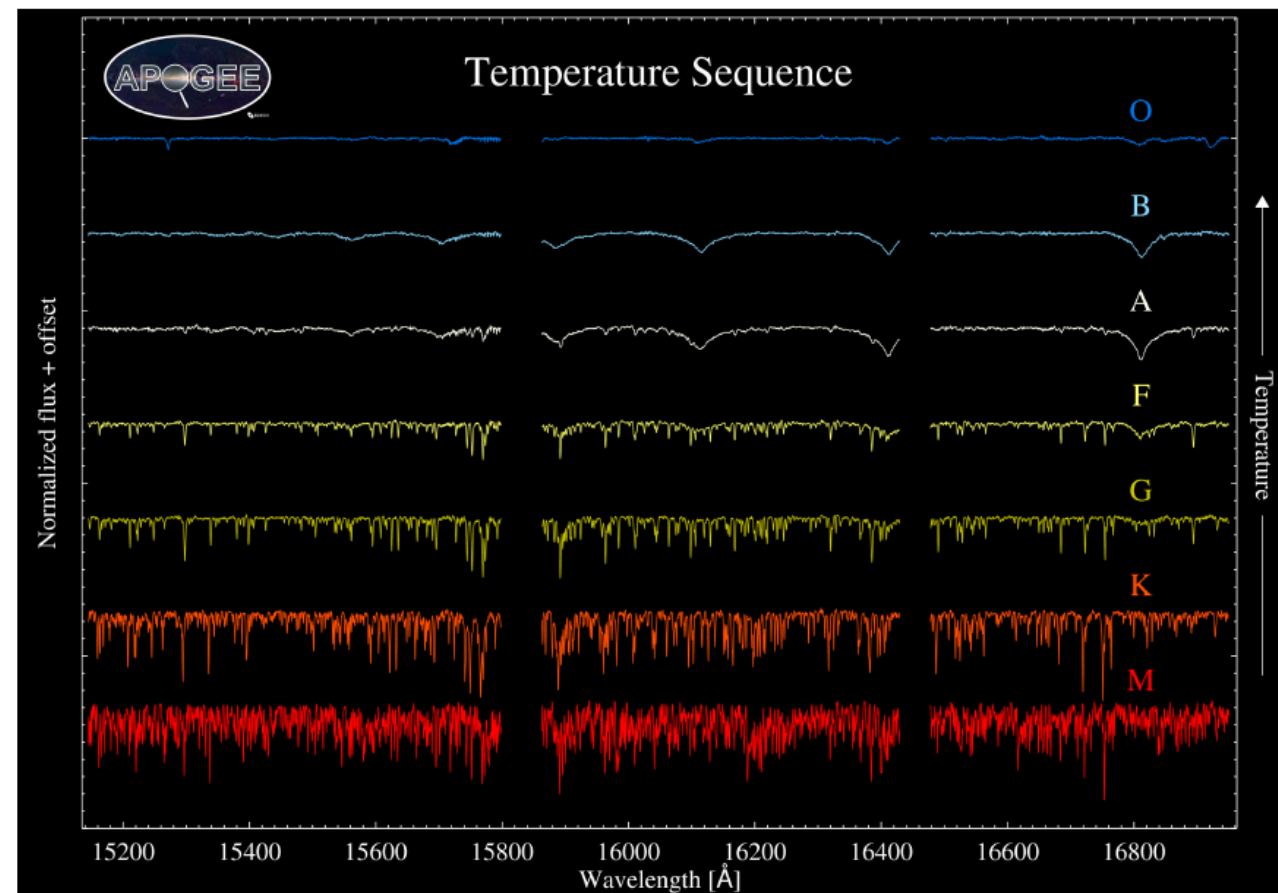


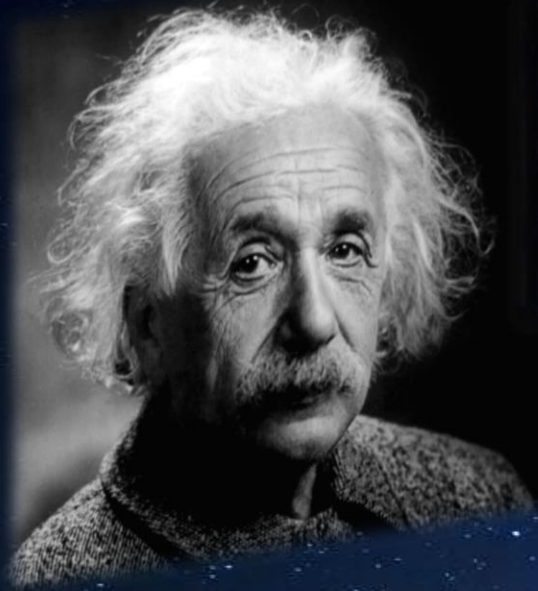
# Example on APOGEE dataset

Colours points according tabulated parameters (e.g. SDSS)



**APOGEE stars:** infrared spectra of  $\sim 250K$  stars.  
 Calculate **Random Forest** distance matrix  $\rightarrow$  Apply **tSNE** for dimensionality reduction.  
 See Reis+17.






"Chi dice che è  
impossibile...  
Non dovrebbe  
disturbare chi ce  
la sta facendo"  
Albert Einstein



“



...what we want is a  
machine that can learn  
from experience.

Alan Turing, 1947

”

# Astroinformatics...

