Analyzing Astronomical Data with Machine Learning Techniques

Mohammad H. Zhoolideh Haghighi

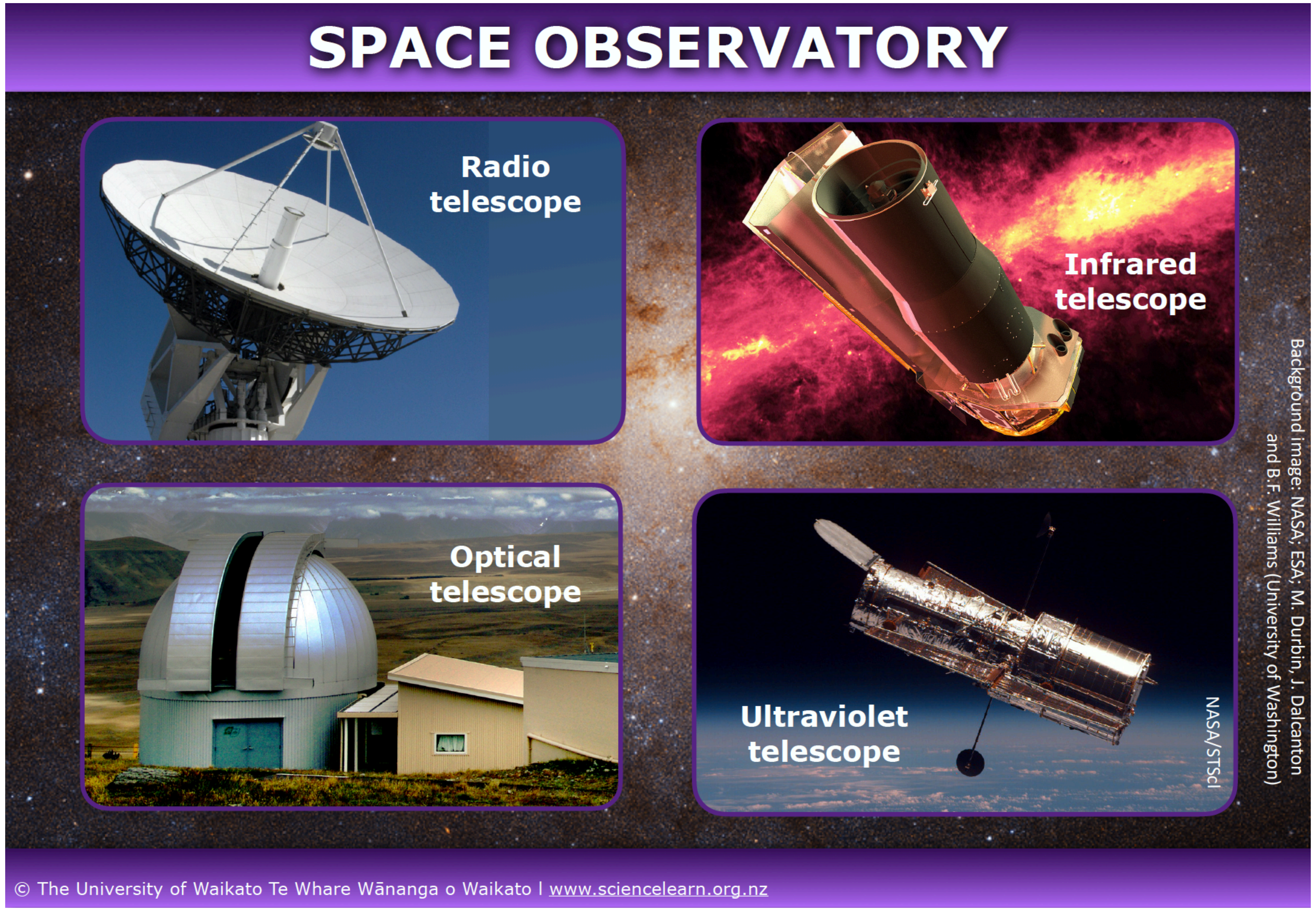
IUT, Nov 2021

Institute for Research in Fundamental Sciences

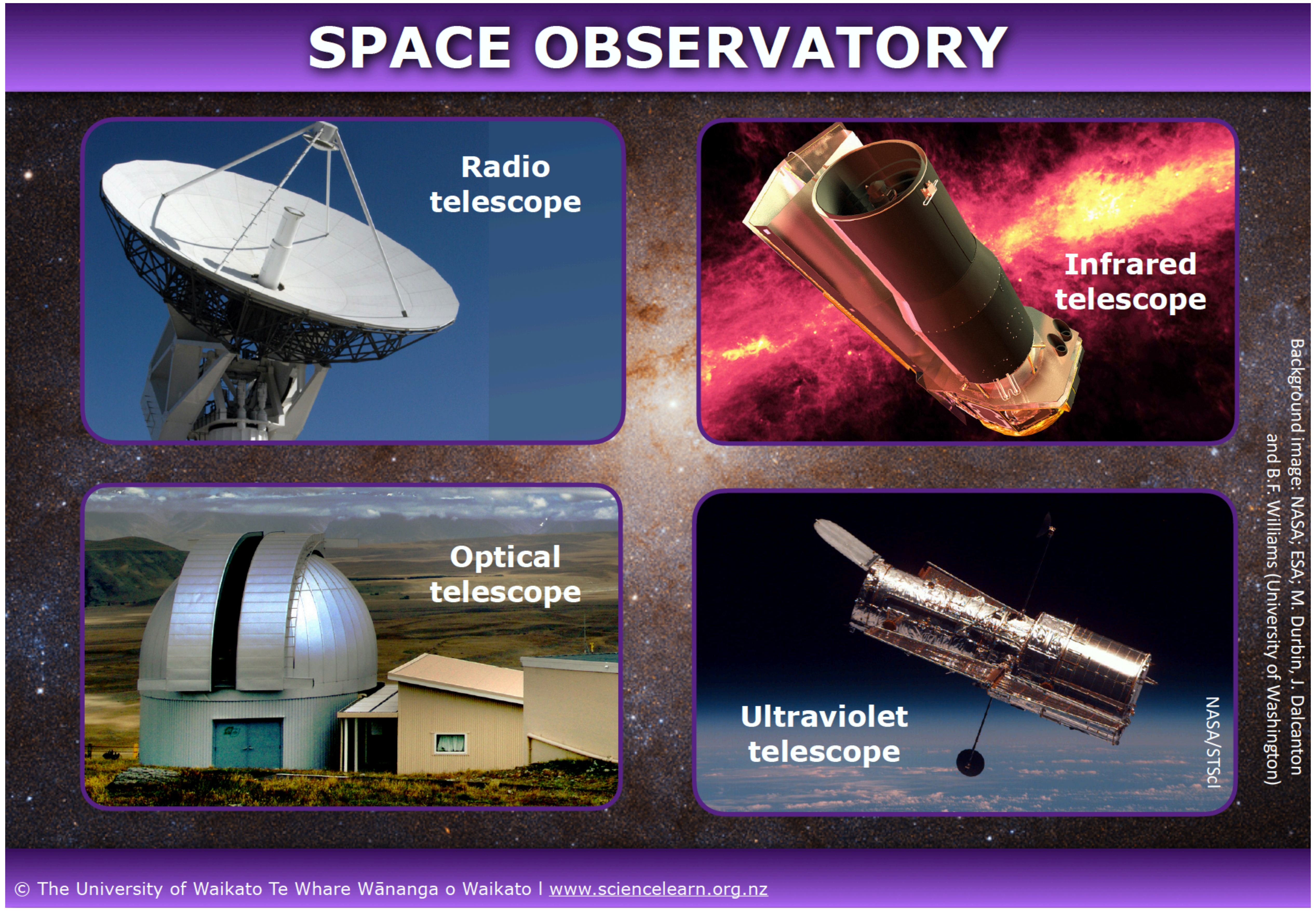
1928
K. N. Toosi
University of Technology

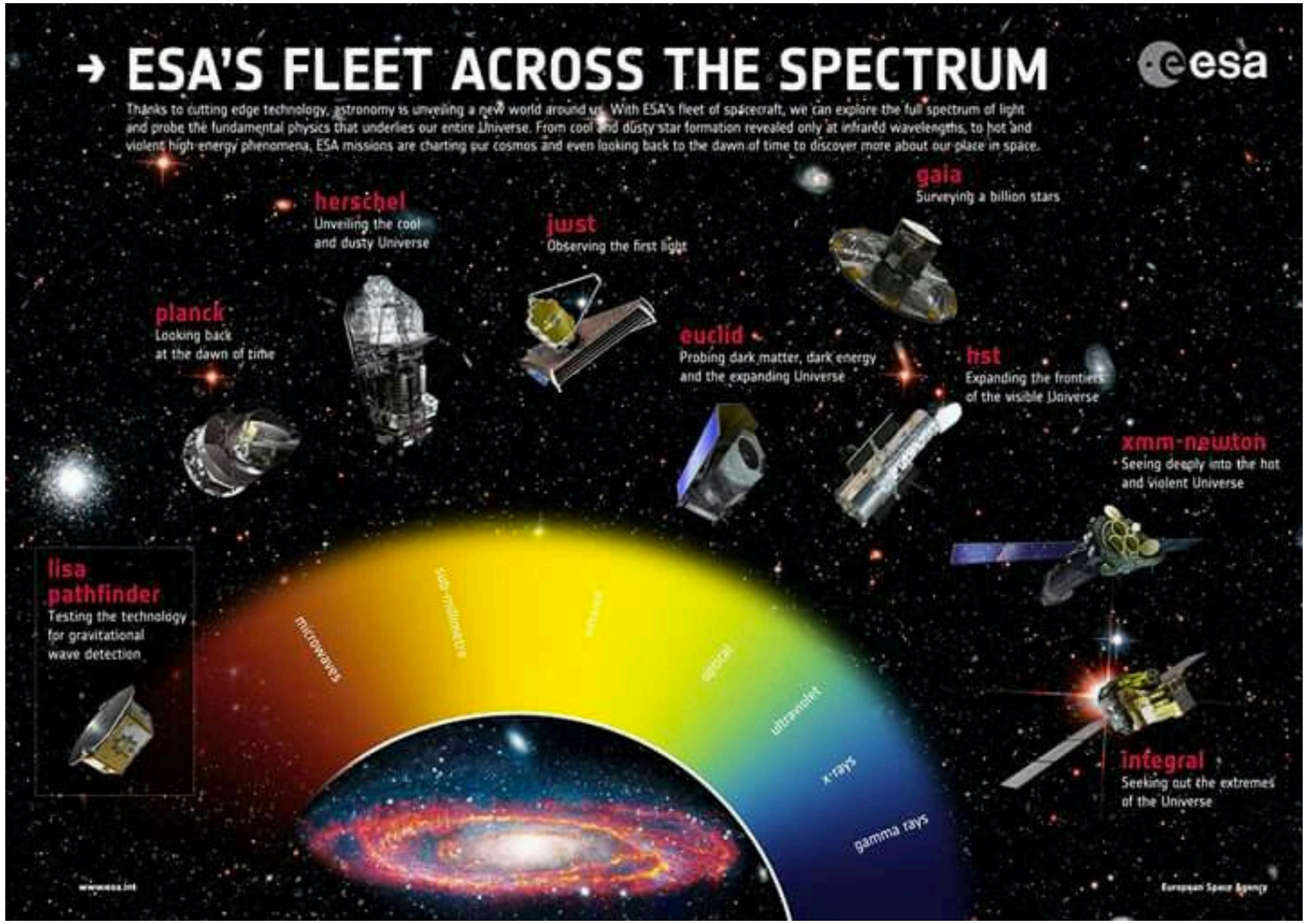
Ground-based and space sky surveys ranging from gamma rays and X-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era.

Ground-based and space sky surveys ranging from gamma rays and X-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era.

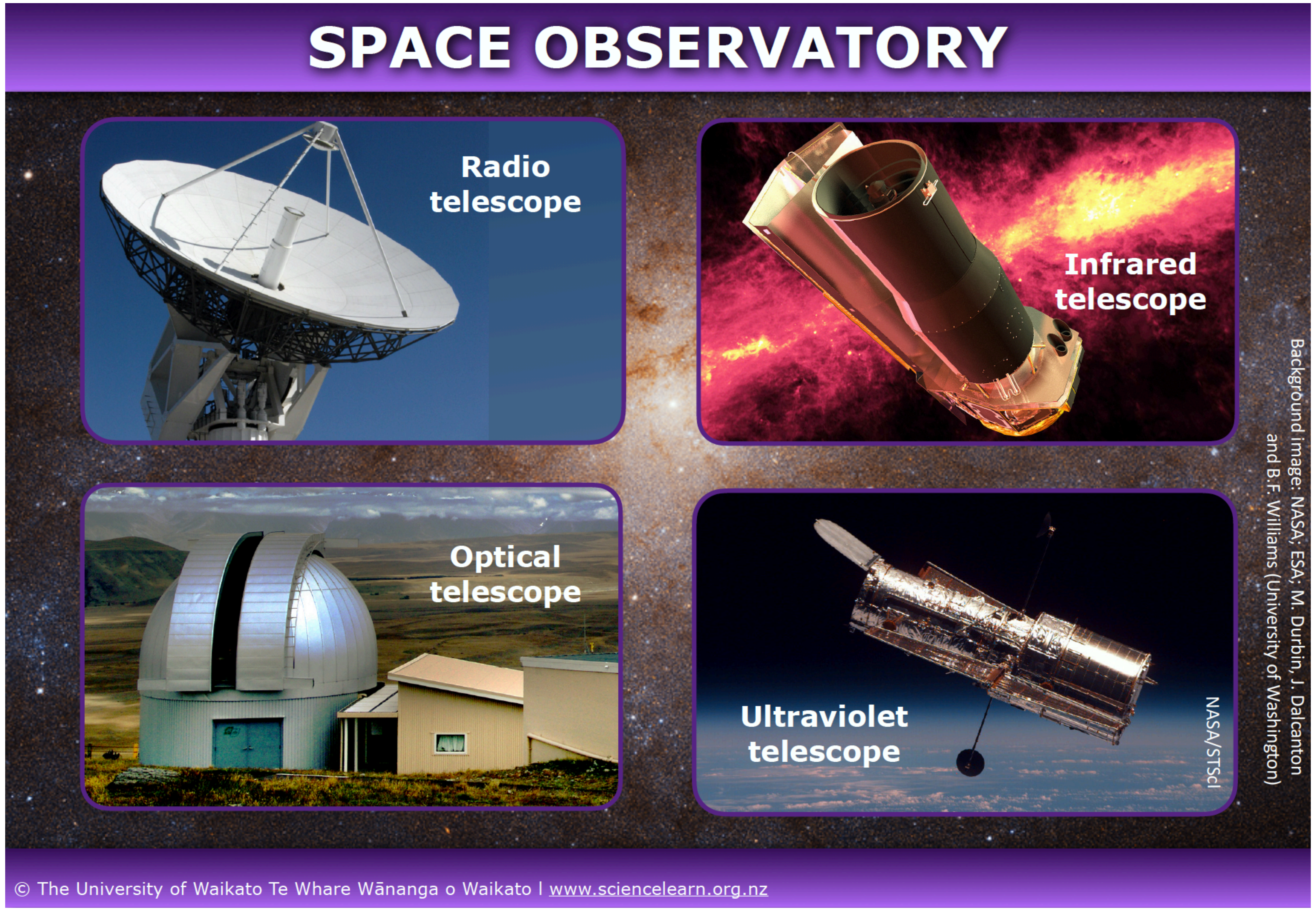


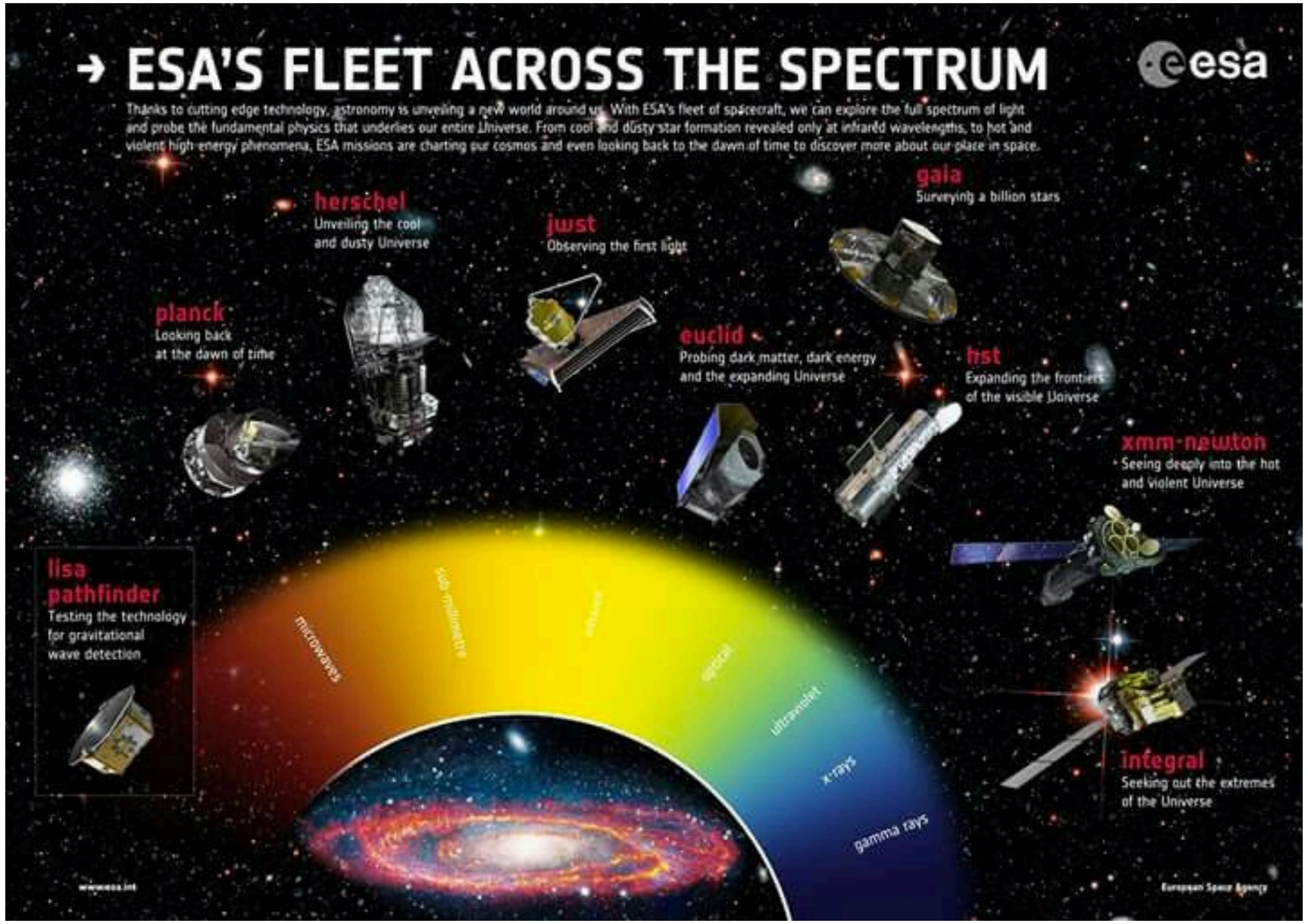
Ground-based and space sky surveys ranging from gamma rays and X-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era.





Ground-based and space sky surveys ranging from gamma rays and X-rays, ultraviolet, optical, and infrared to radio bands is bringing astronomy into the big data era.





You can add data from Gravitational waves and Neutrinos to the above list

Characteristics of big data: **Volume, Variety, and Velocity.**

Characteristics of big data: **Volume, Variety, and Velocity.**

Volume is the amount of data

Characteristics of big data: **Volume, Variety, and Velocity.**

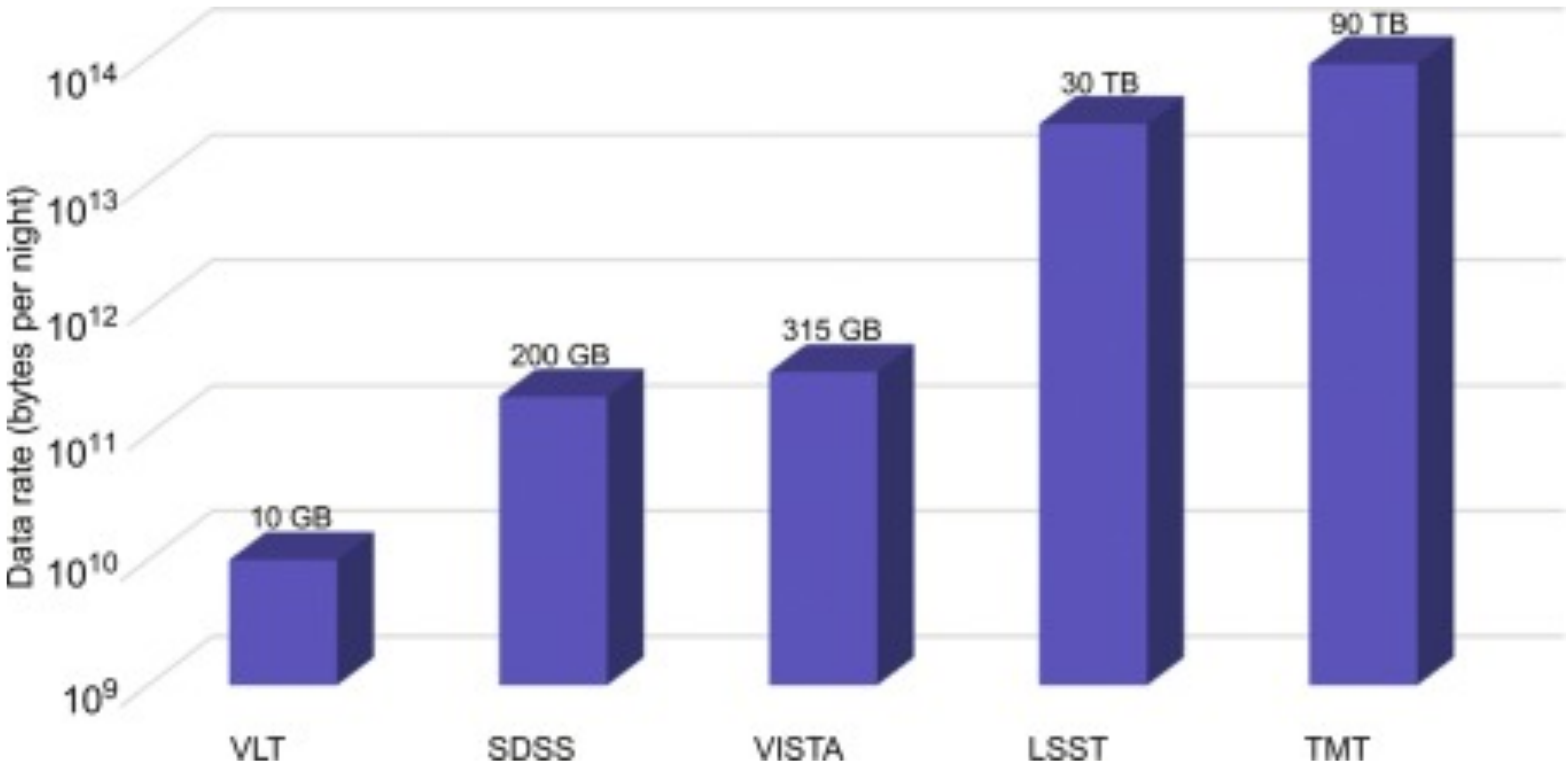
Volume is the amount of data

Big data pose challenges for capture, cleaning, storage, processing, sharing, mining, analysis, and visualization

Characteristics of big data: **Volume, Variety, and Velocity.**

Volume is the amount of data

Big data pose challenges for capture, cleaning, storage, processing, sharing, mining, analysis, and visualization

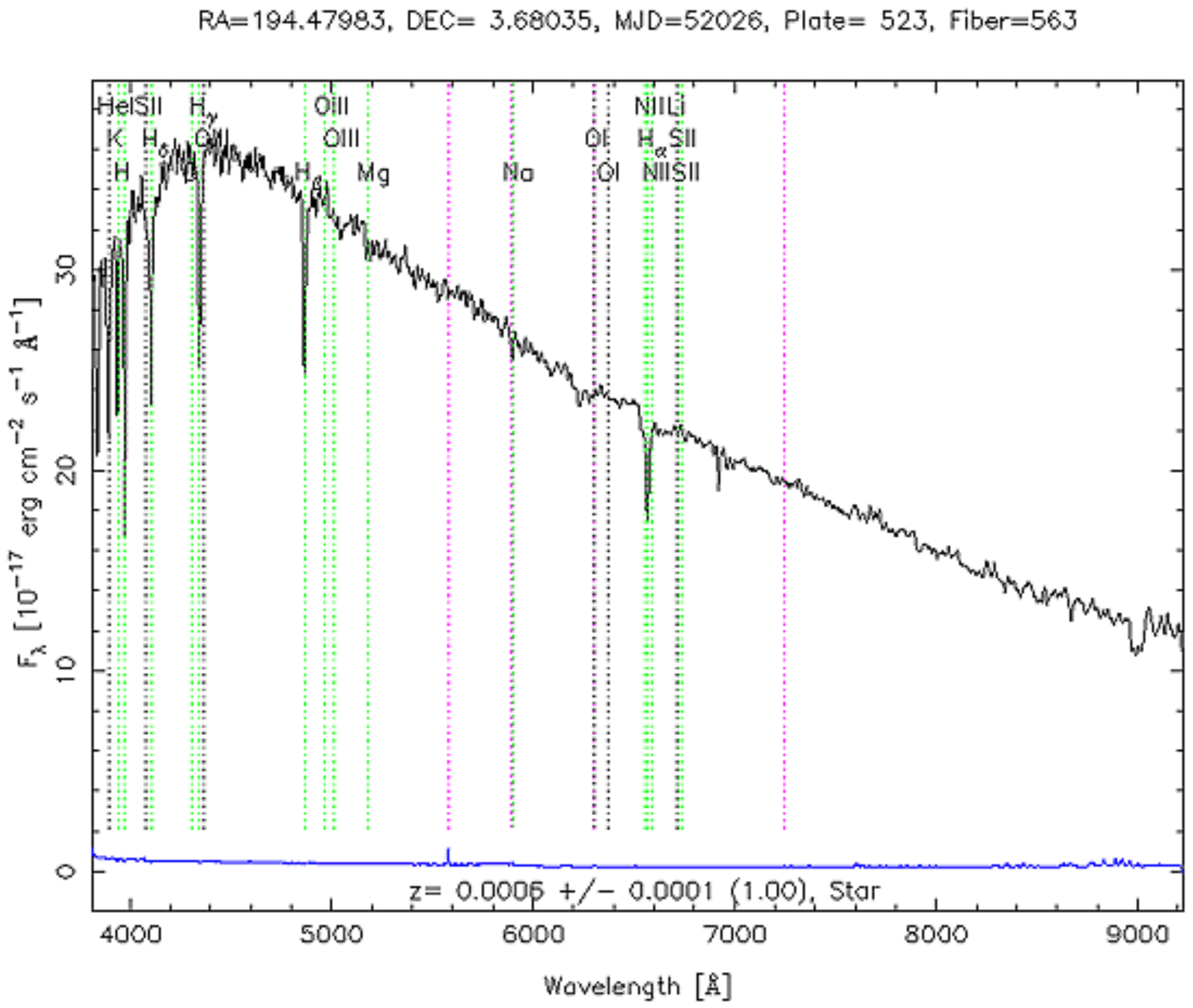


Lei et al. 2020

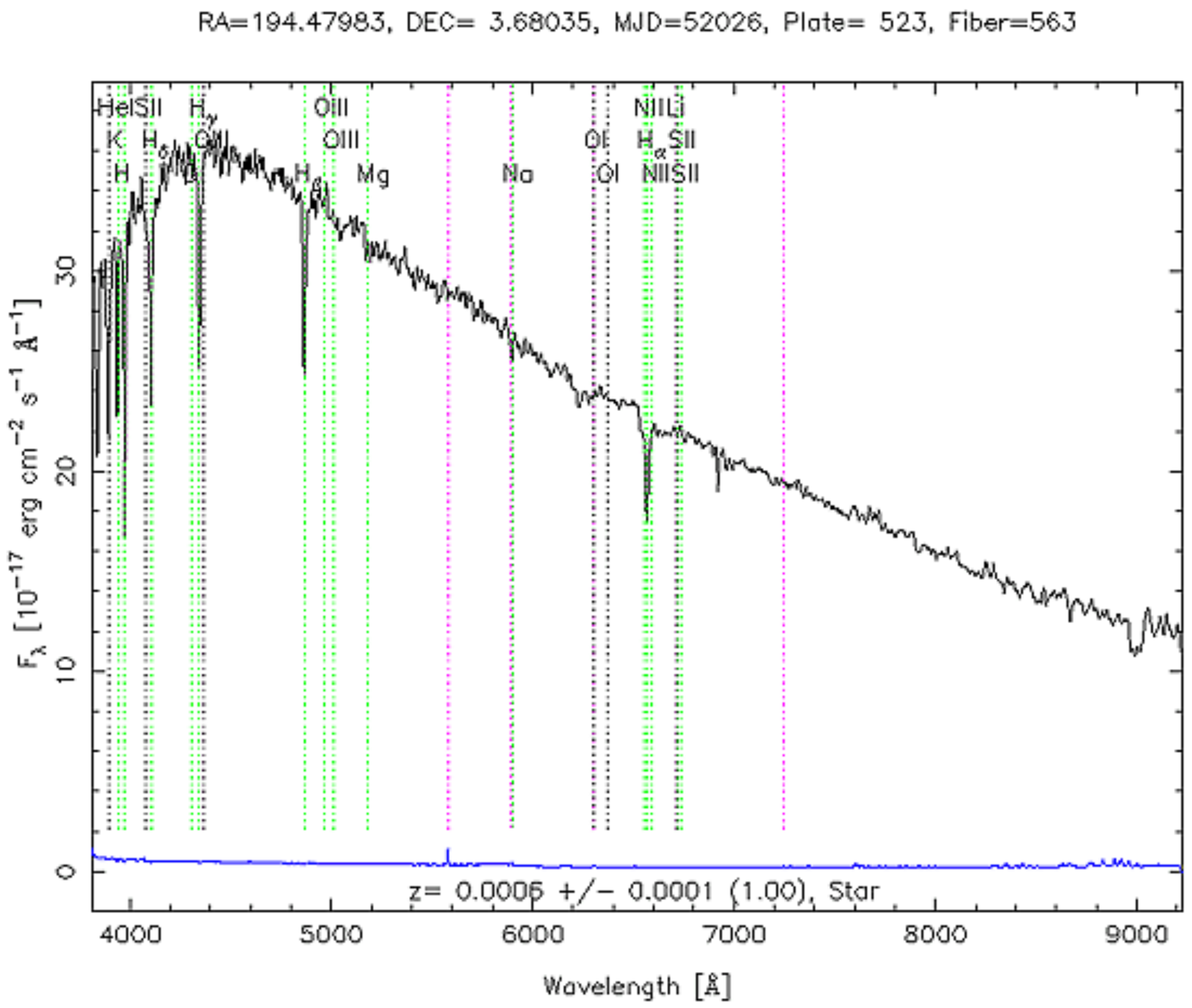
Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.

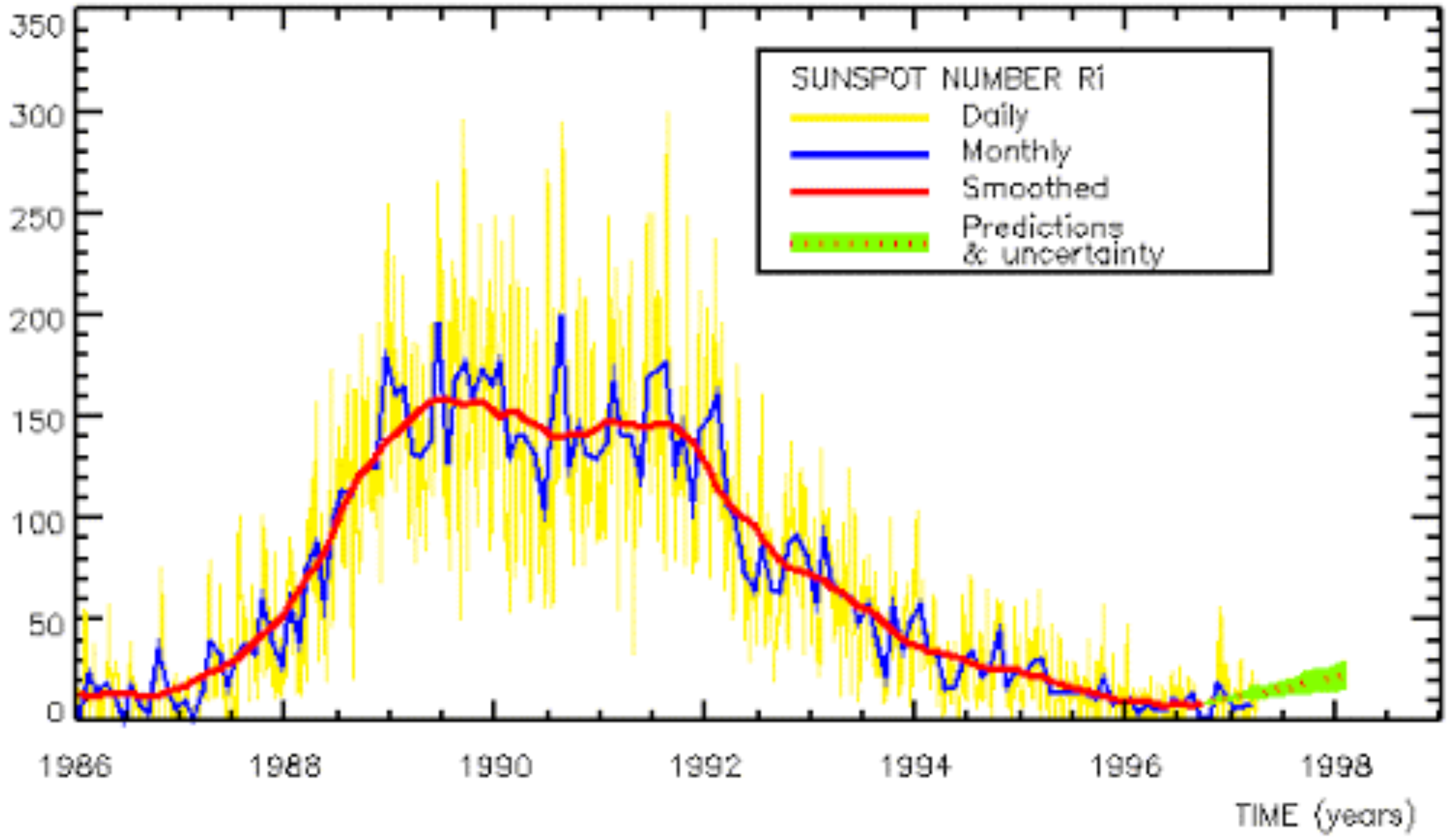
Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.

Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.

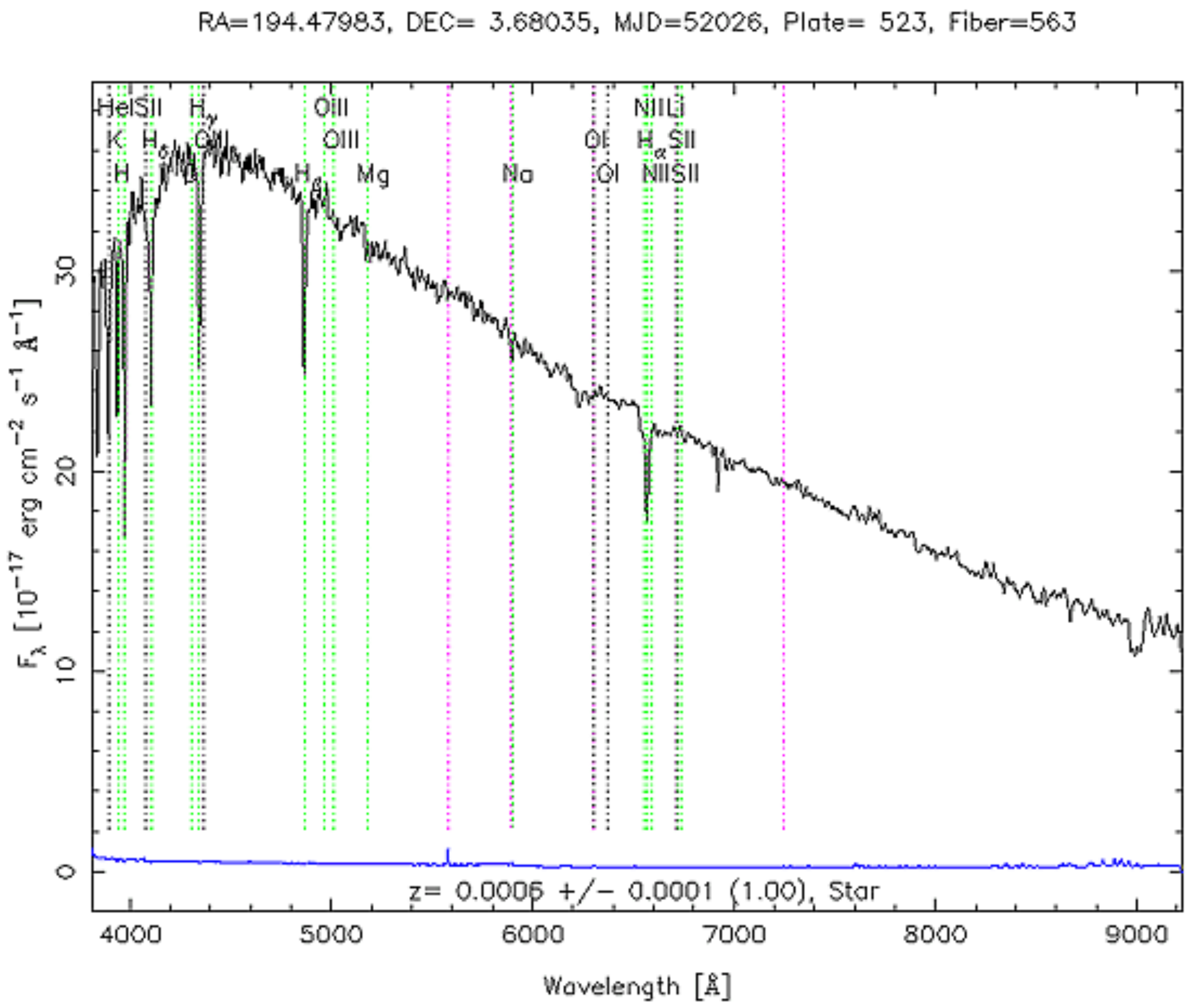


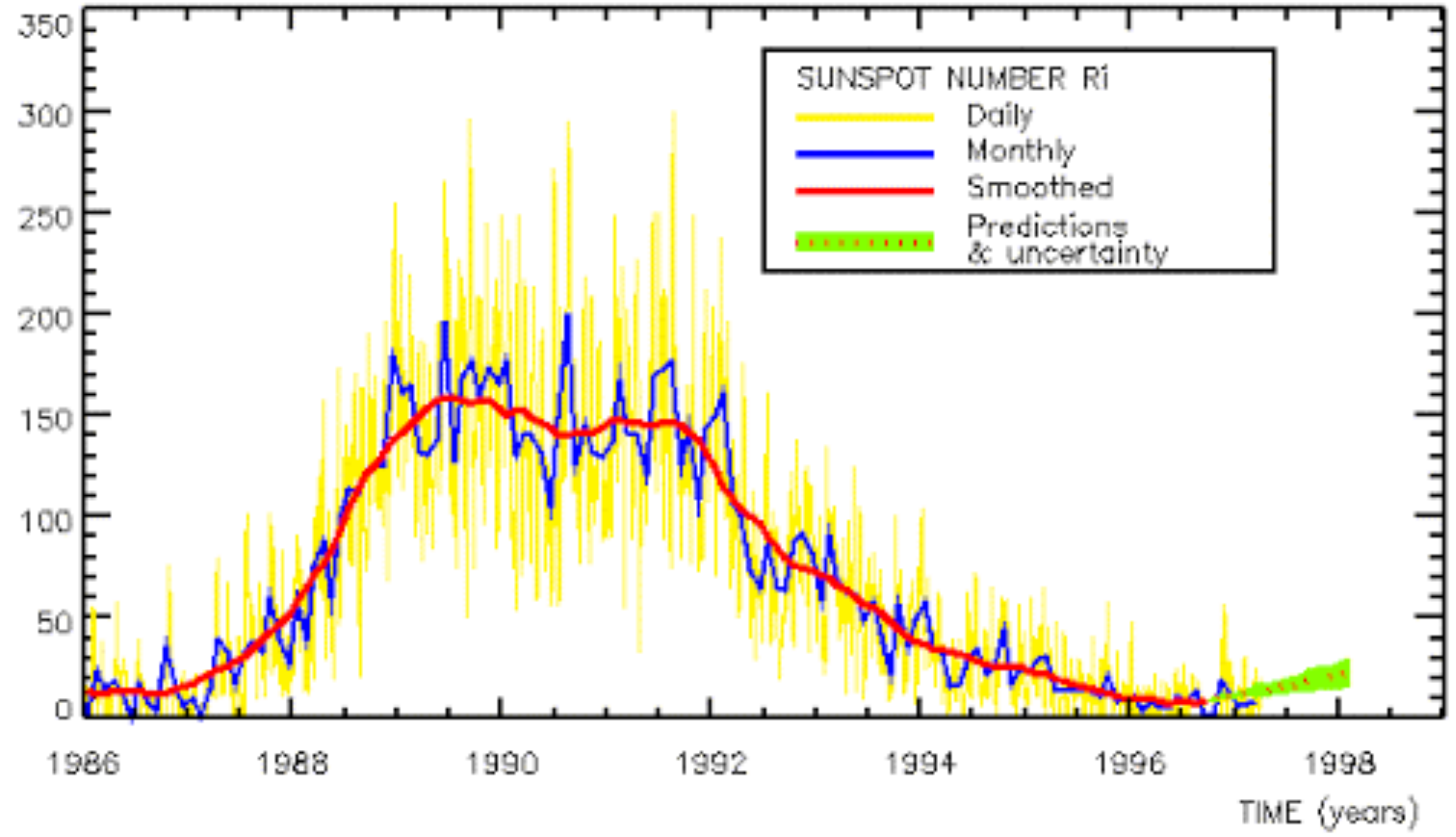
Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.

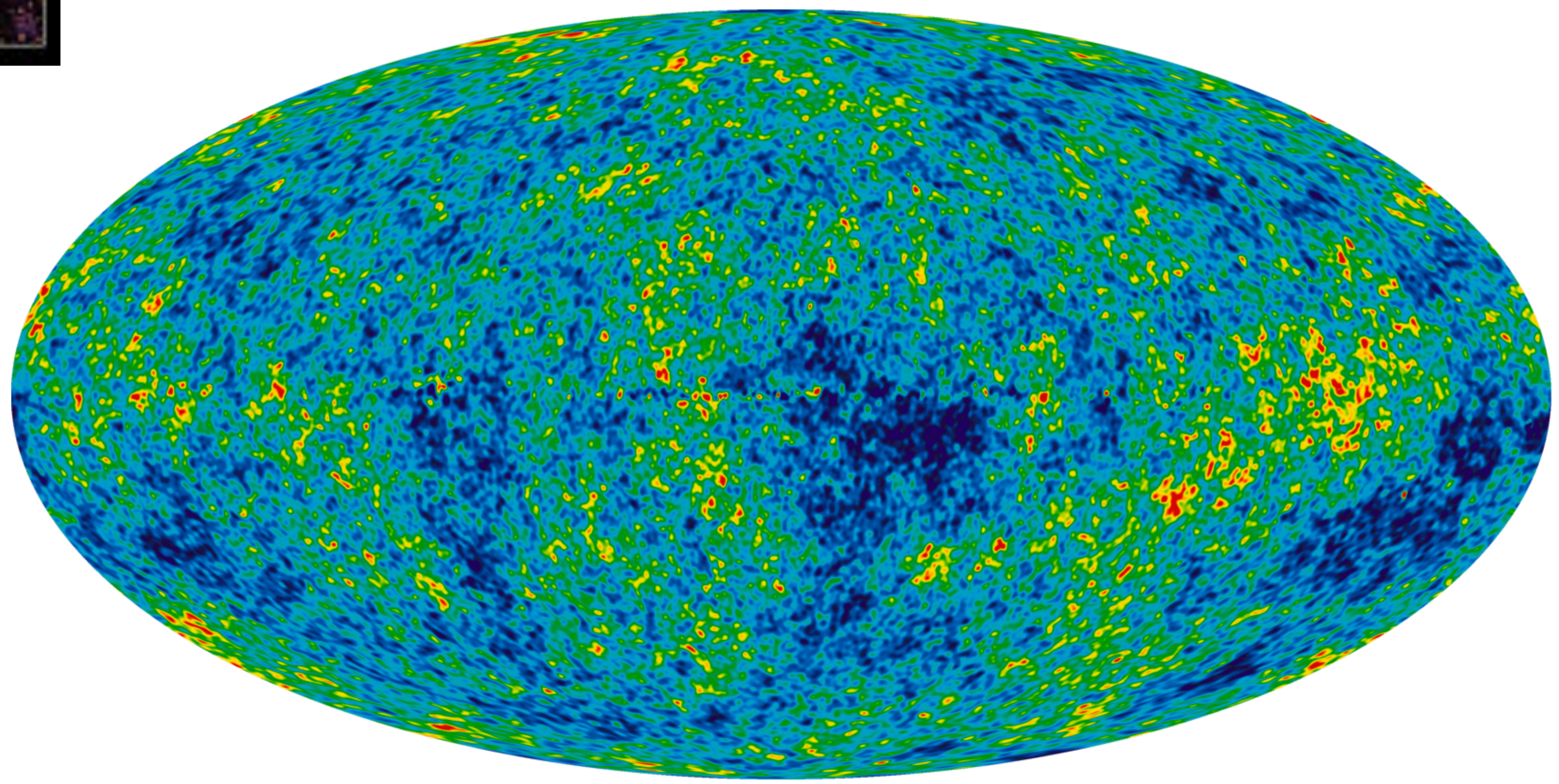




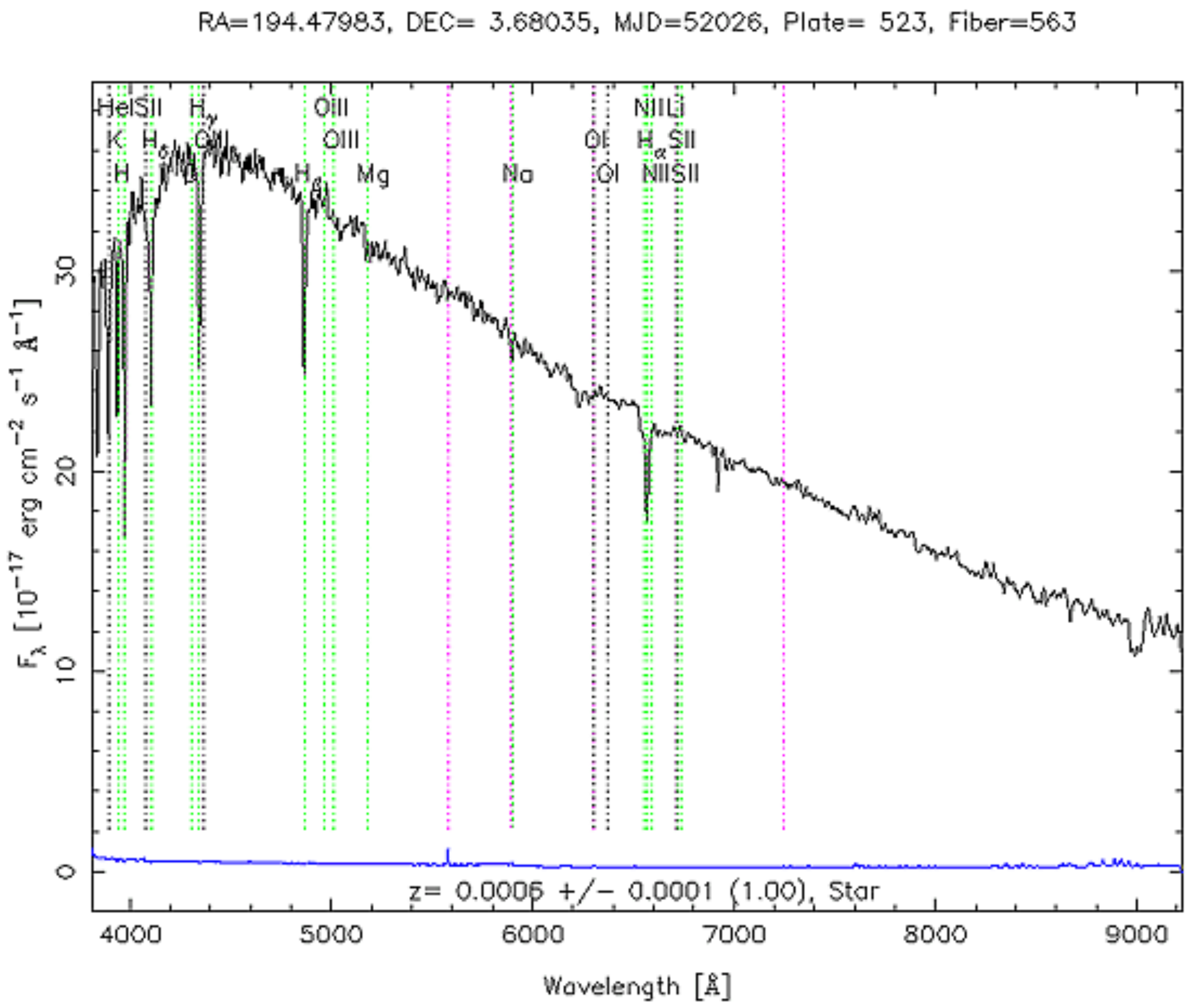
Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.

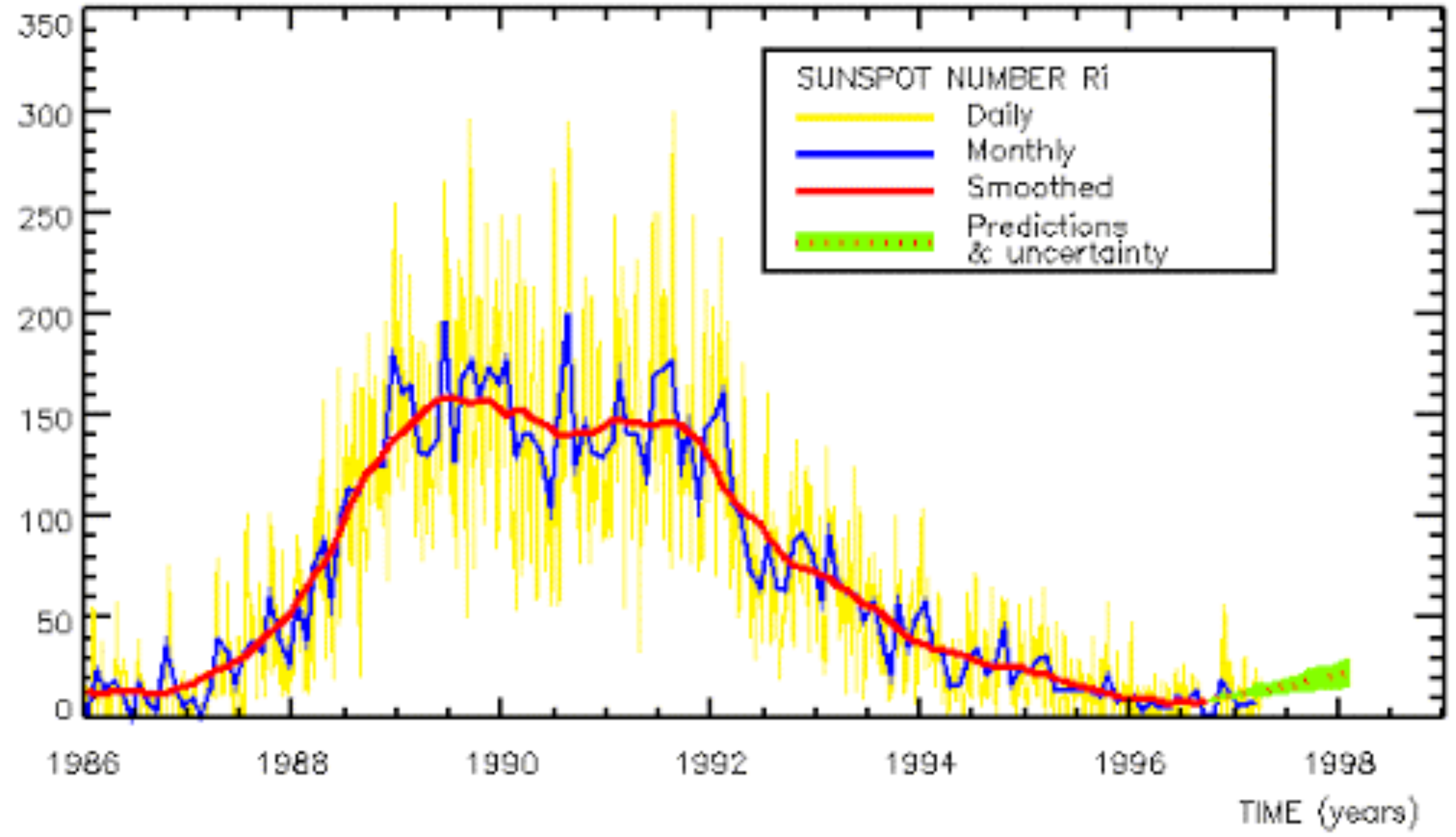


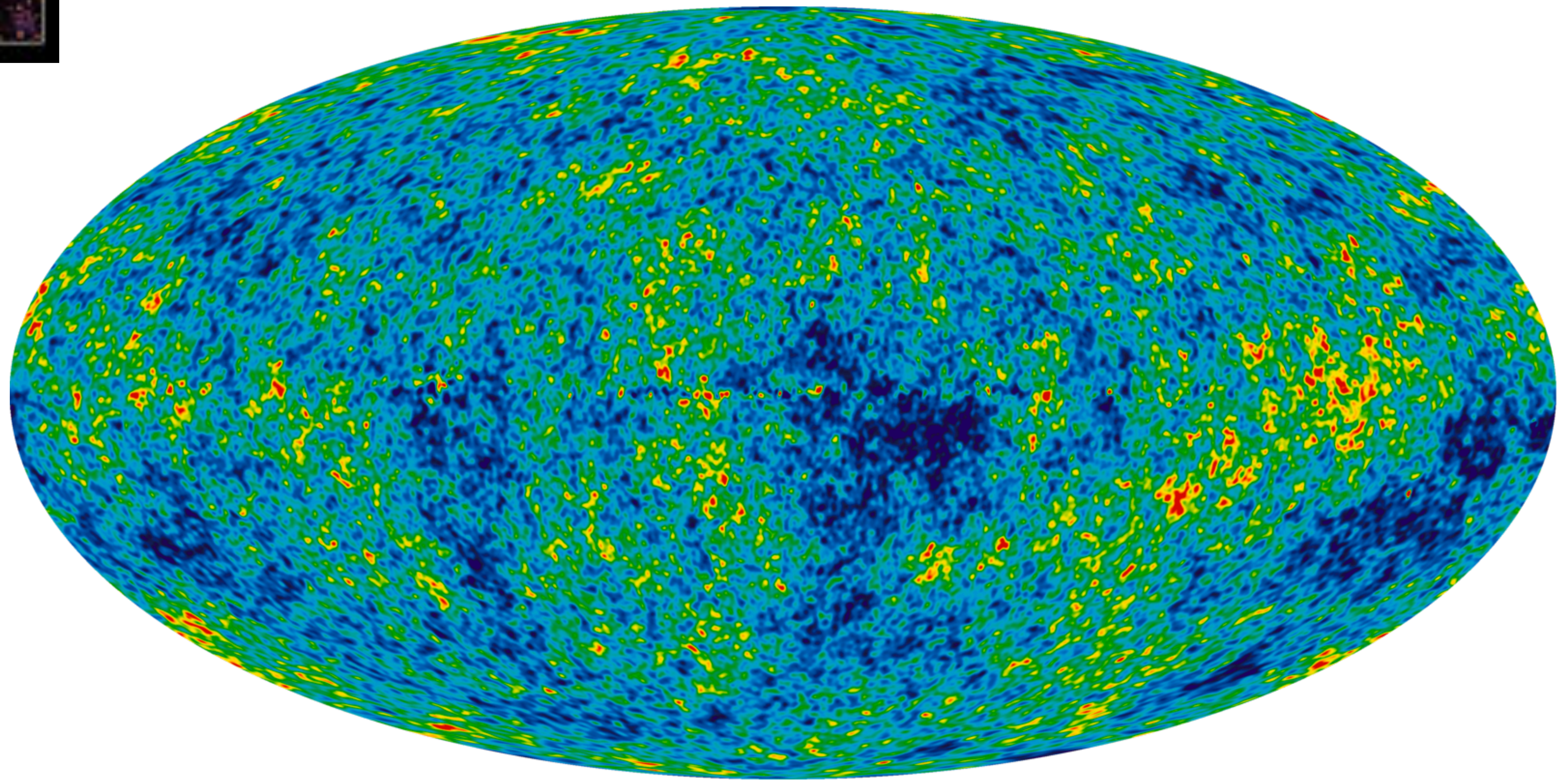


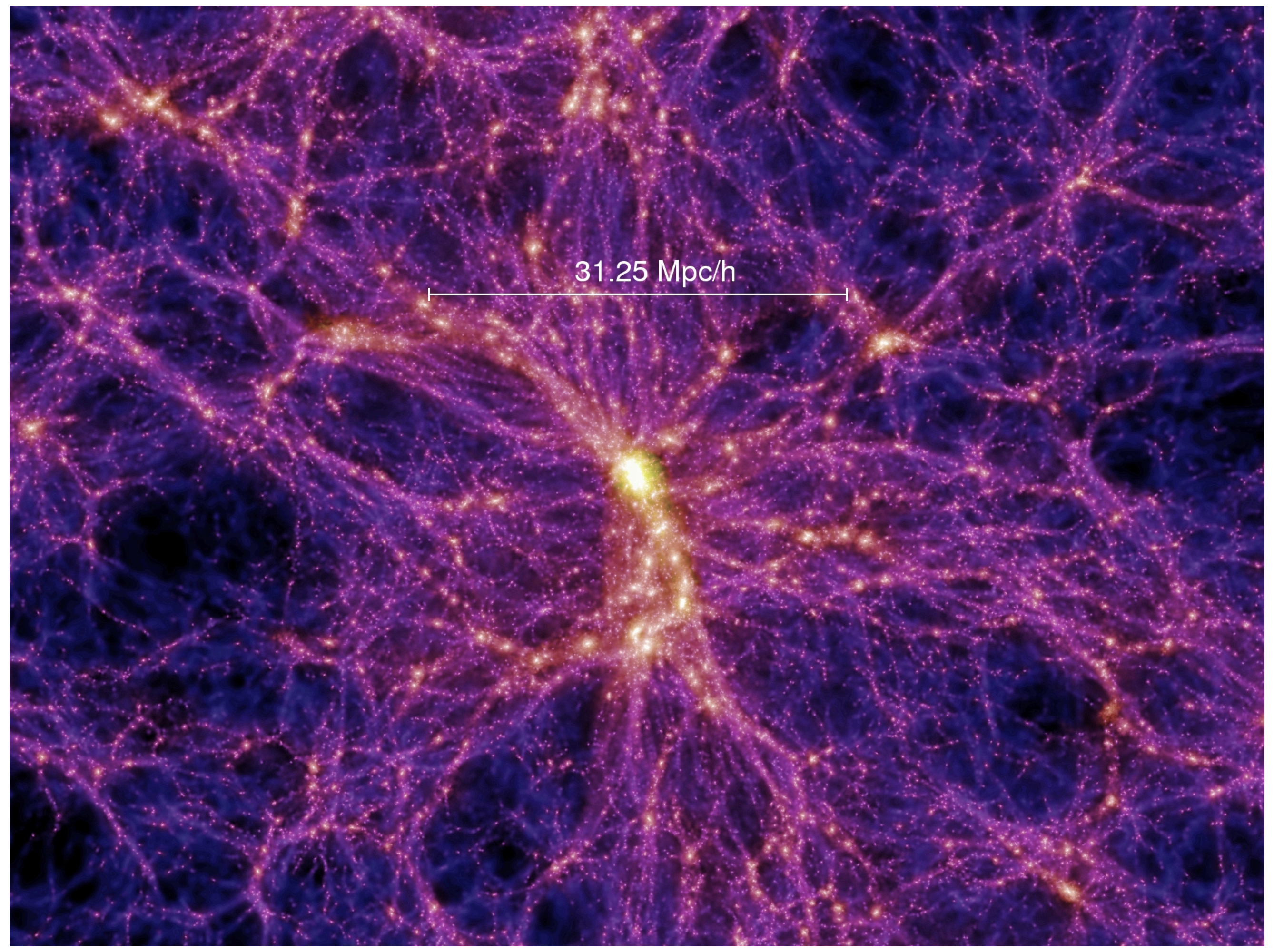


Variety points to data complexity. Astronomical data mainly include images, spectra, time-series data, CMB, and simulation data.









Velocity means the speed of producing, transmitting, and analyzing data. Speaking of data volume, LSST will generate one SDSS each night for 10 years.

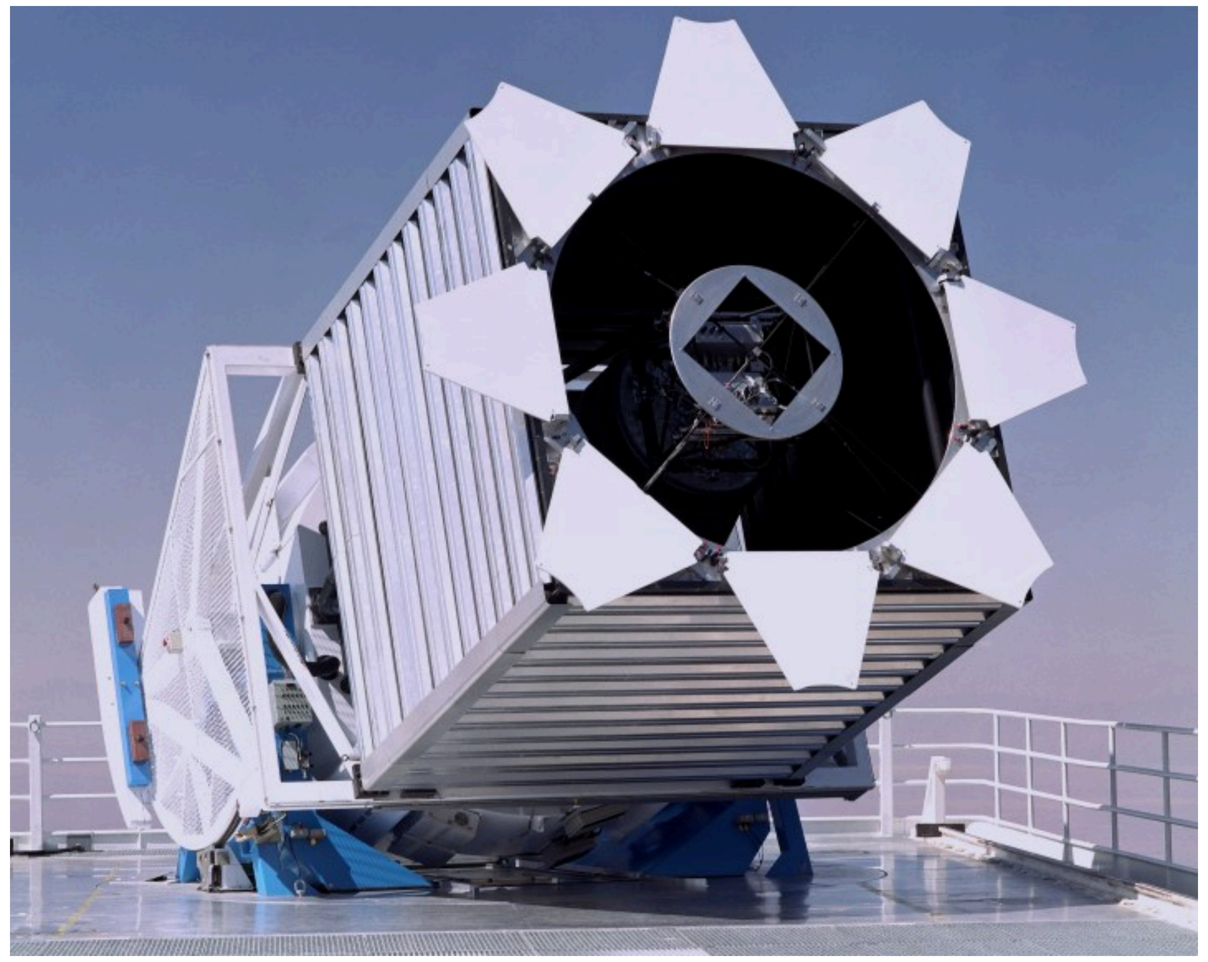
Velocity means the speed of producing, transmitting, and analyzing data. Speaking of data volume, LSST will generate one SDSS each night for 10 years.

LSST

Velocity means the speed of producing, transmitting, and analyzing data. Speaking of data volume, LSST will generate one SDSS each night for 10 years.

LSST

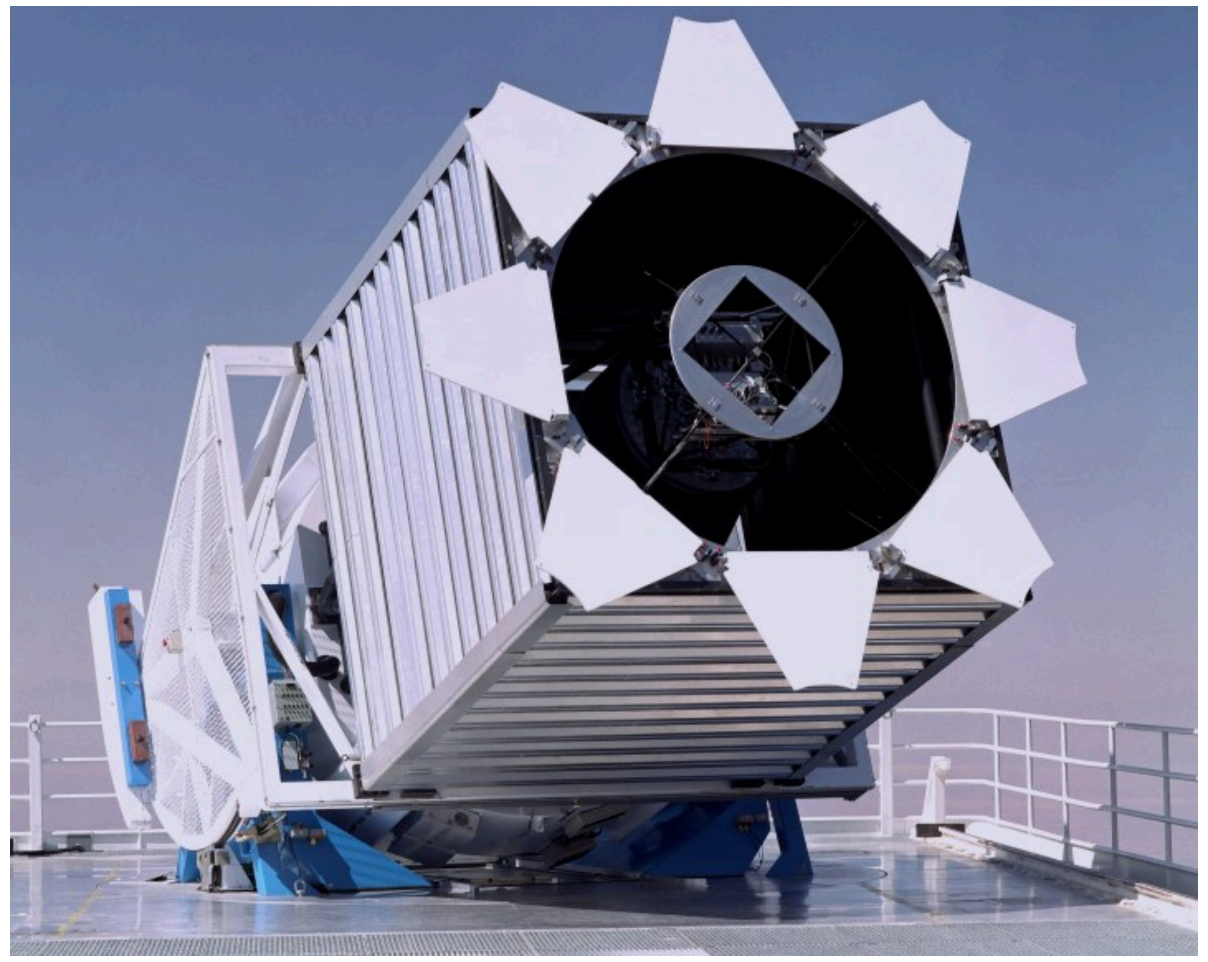
SDSS



Velocity means the speed of producing, transmitting, and analyzing data. Speaking of data volume, LSST will generate one SDSS each night for 10 years.

LSST

SDSS



LSST expects to find 1000 new supernovae each night for 10 years. How to mine, classify, and target the supernovae candidates and make follow-up observations in 10 years time is a huge challenge for astronomers.

It is a golden era for Astrophysics to deal with the challenges and opportunities generated by the massive data volume, rates, and complexity from next-generation telescopes.

It is a golden era for Astrophysics to deal with the challenges and opportunities generated by the massive data volume, rates, and complexity from next-generation telescopes.

Data mining is of great importance in the big data era. It helps physicists to discover potential and useful information from the large amounts of data.

It is a golden era for Astrophysics to deal with the challenges and opportunities generated by the massive data volume, rates, and complexity from next-generation telescopes.

Data mining is of great importance in the big data era. It helps physicists to discover potential and useful information from the large amounts of data.

Data mining tasks mainly consist of summarization, classification, regression, clustering, and outlier/anomaly detection.

What is Machine Learning?

What is Machine Learning?

"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

What is Machine Learning?

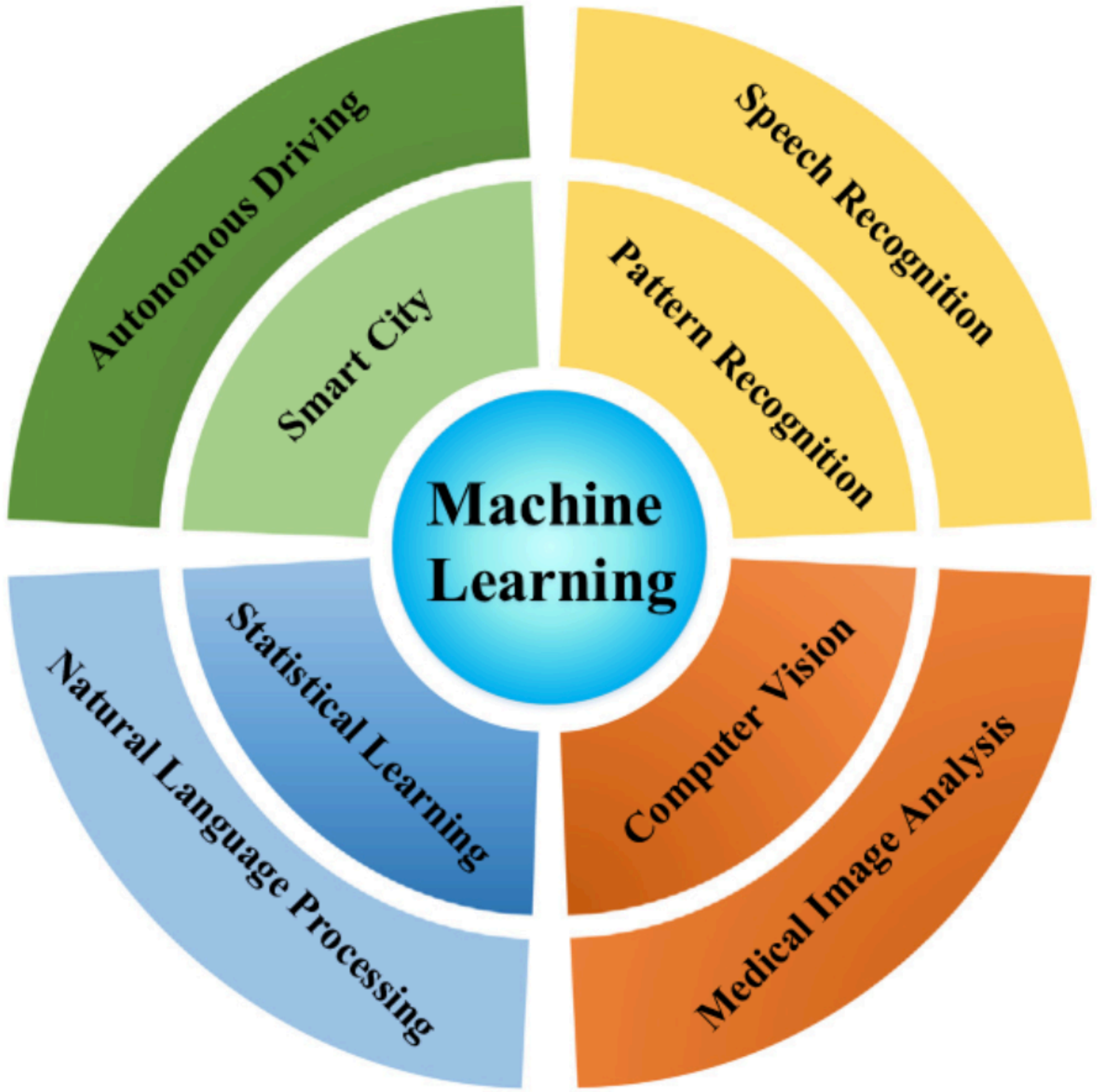
“Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

Where can Machine Learning be used?

What is Machine Learning?

"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

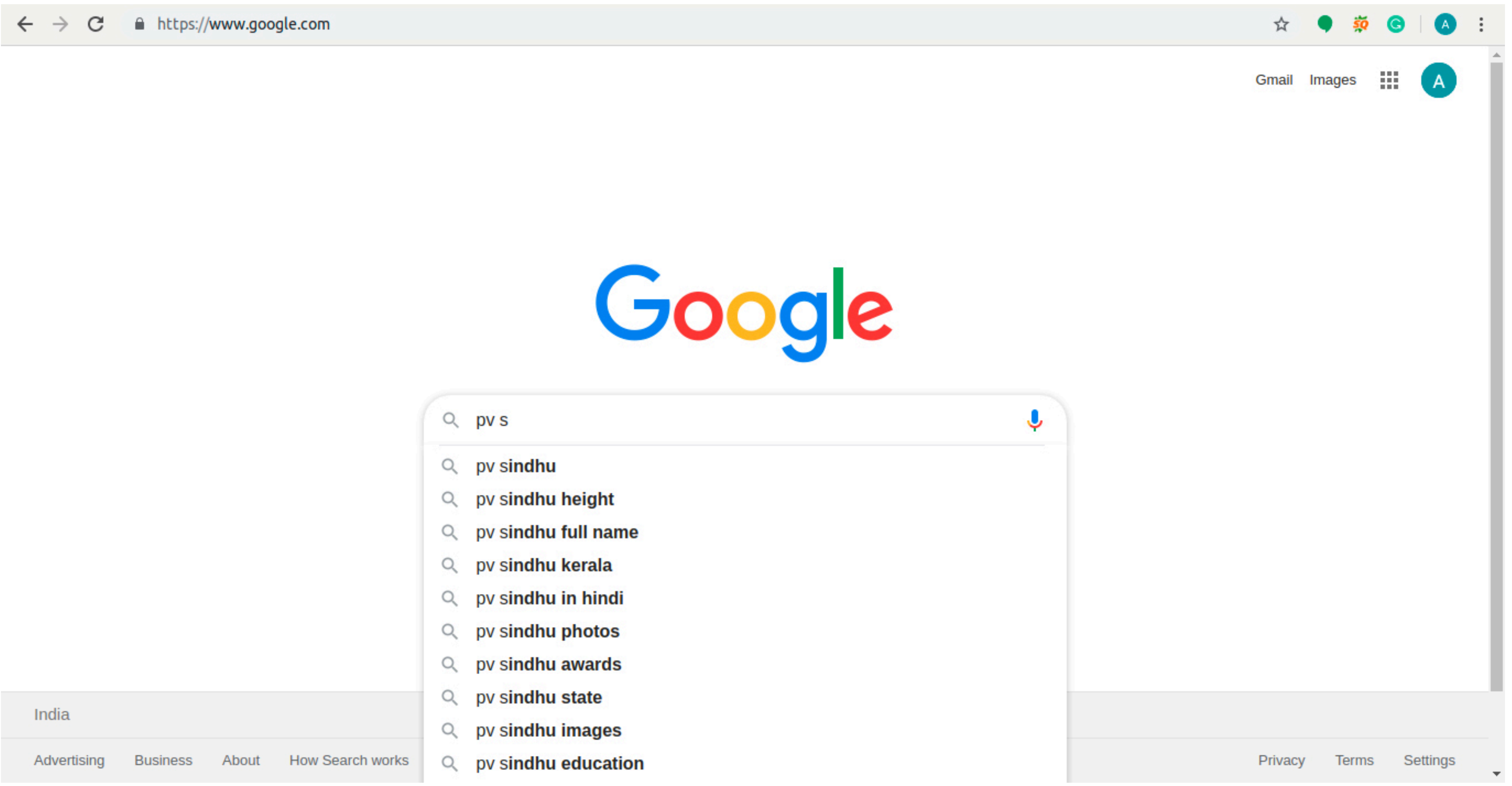
Where can Machine Learning be used?

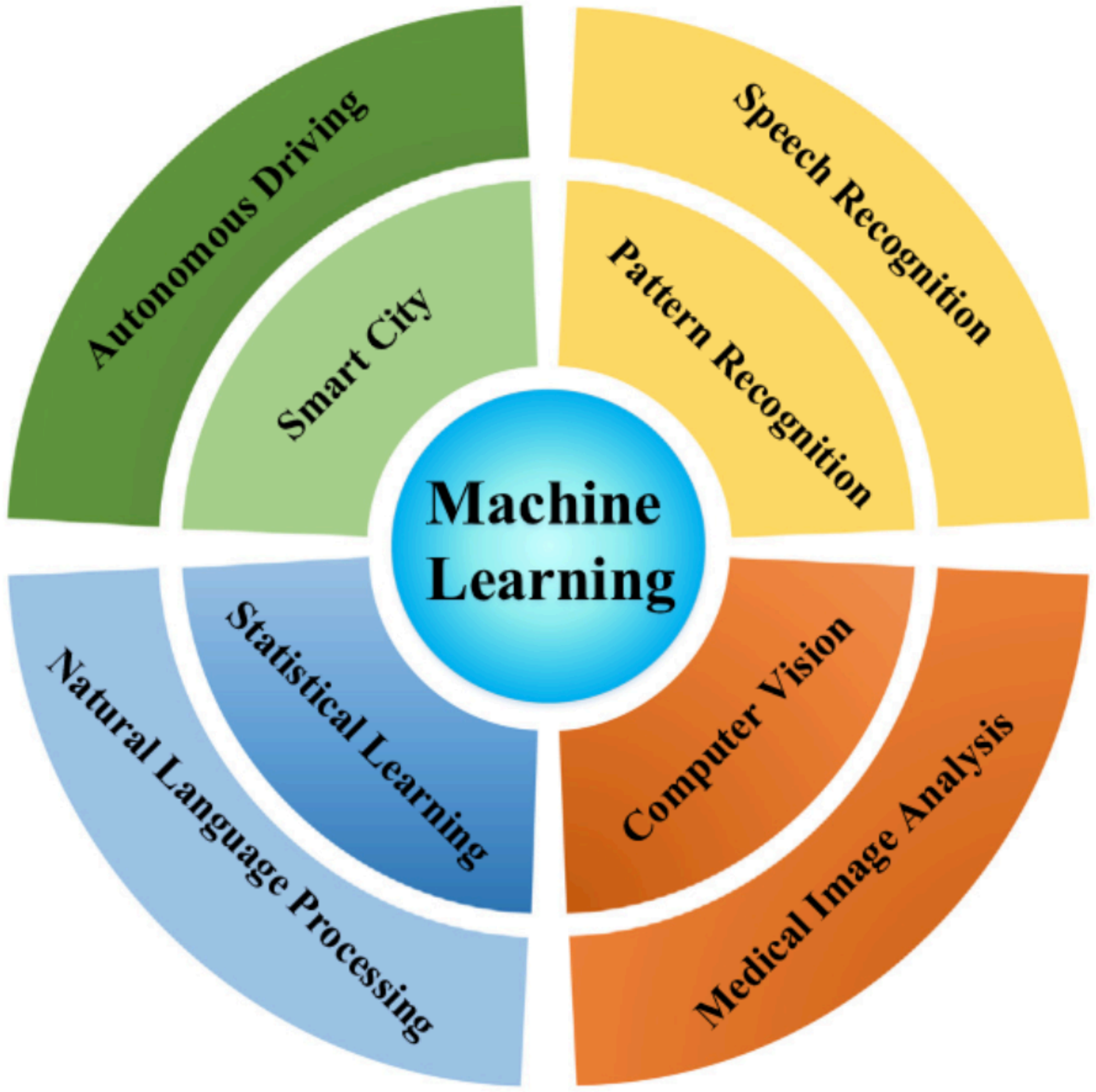


What is Machine Learning?

"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

Where can Machine Learning be used?

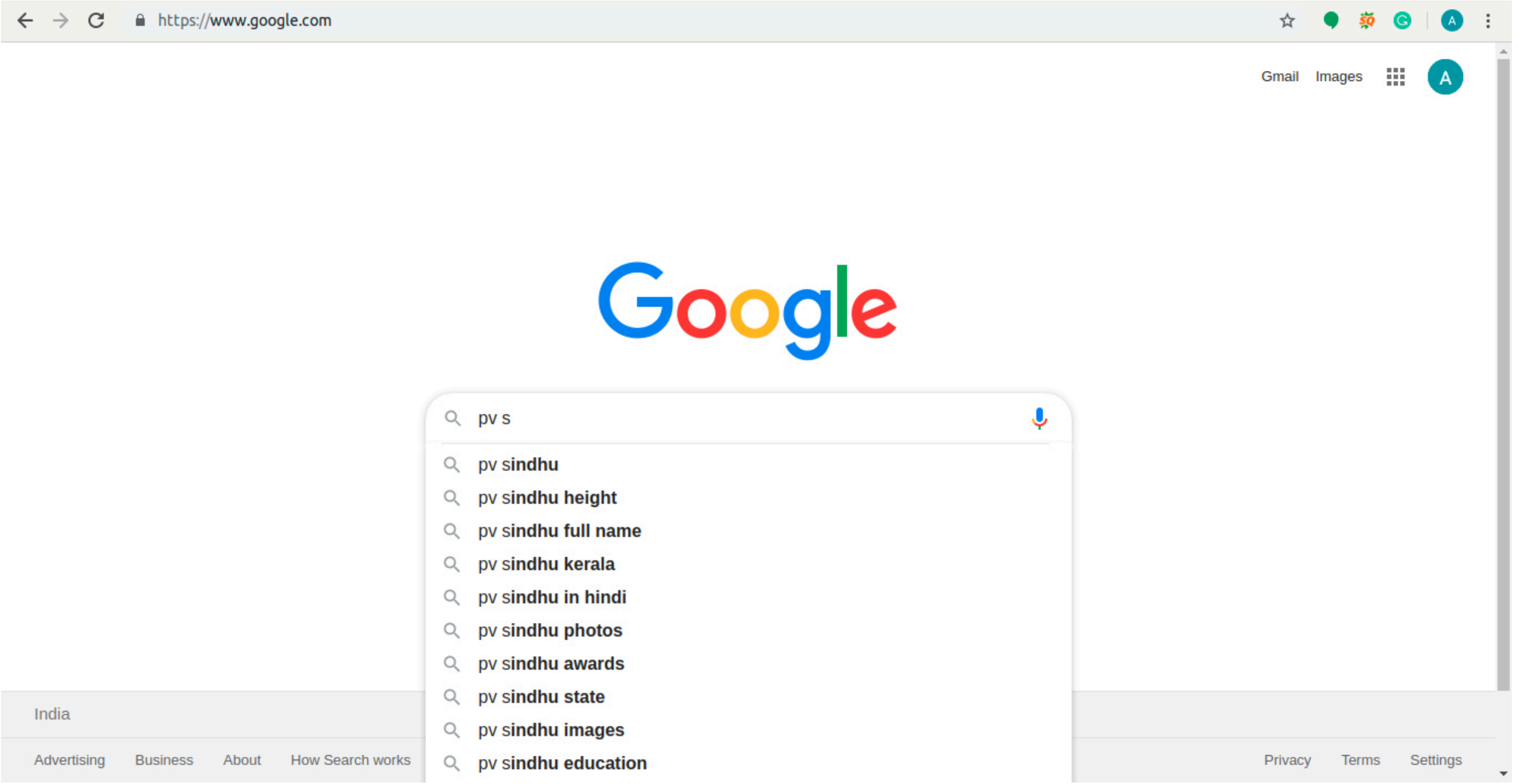


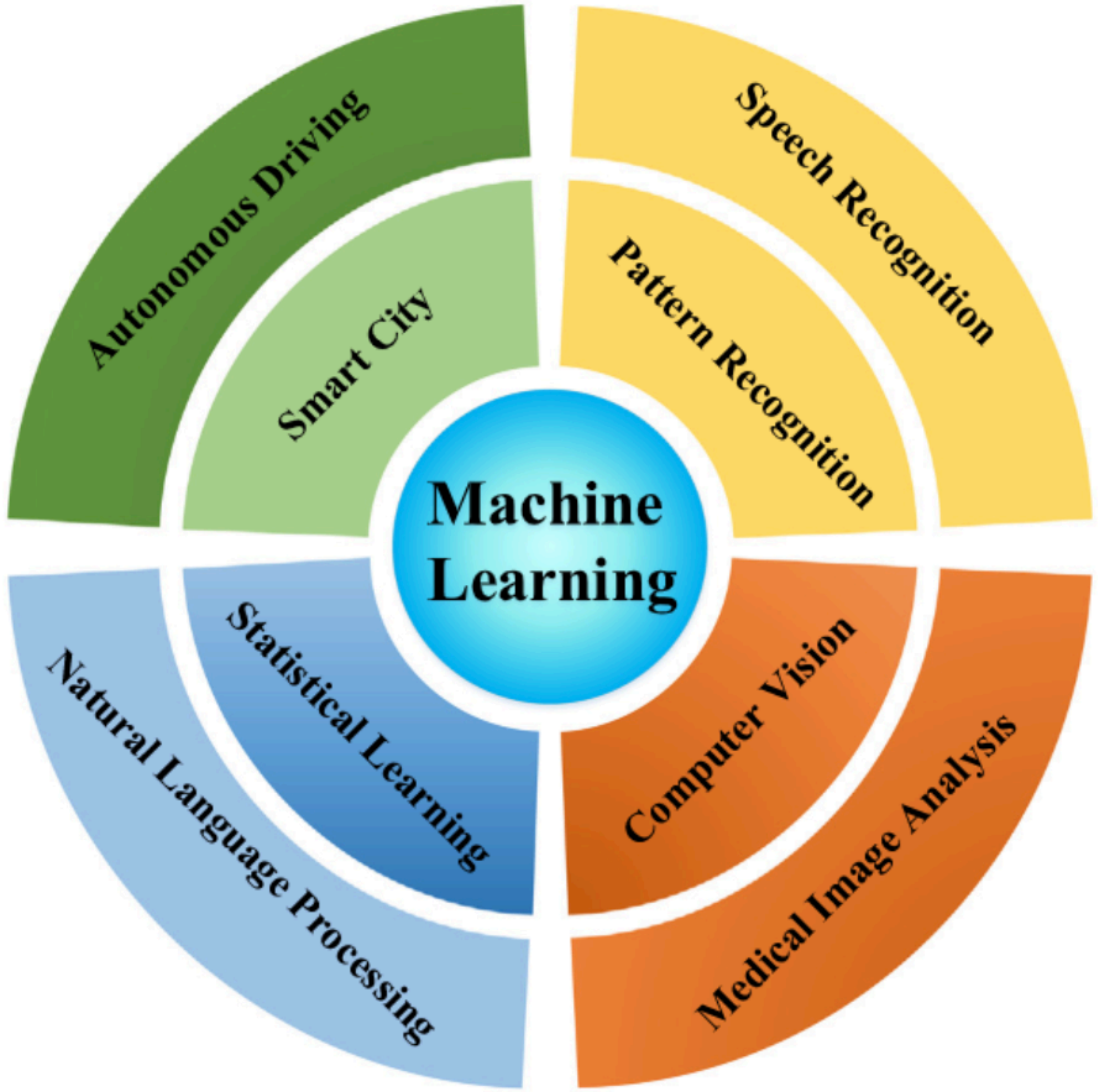


What is Machine Learning?

"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

Where can Machine Learning be used?

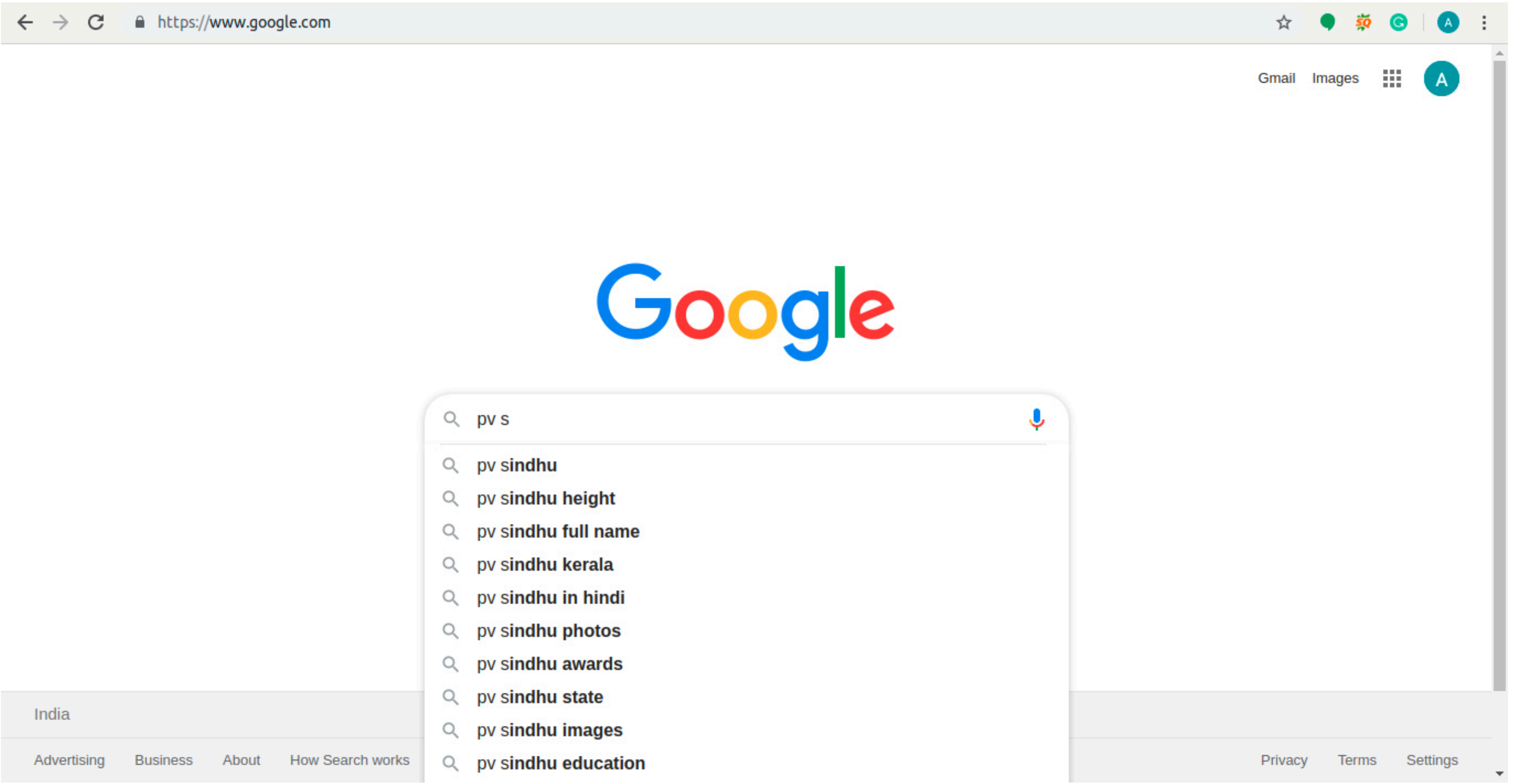


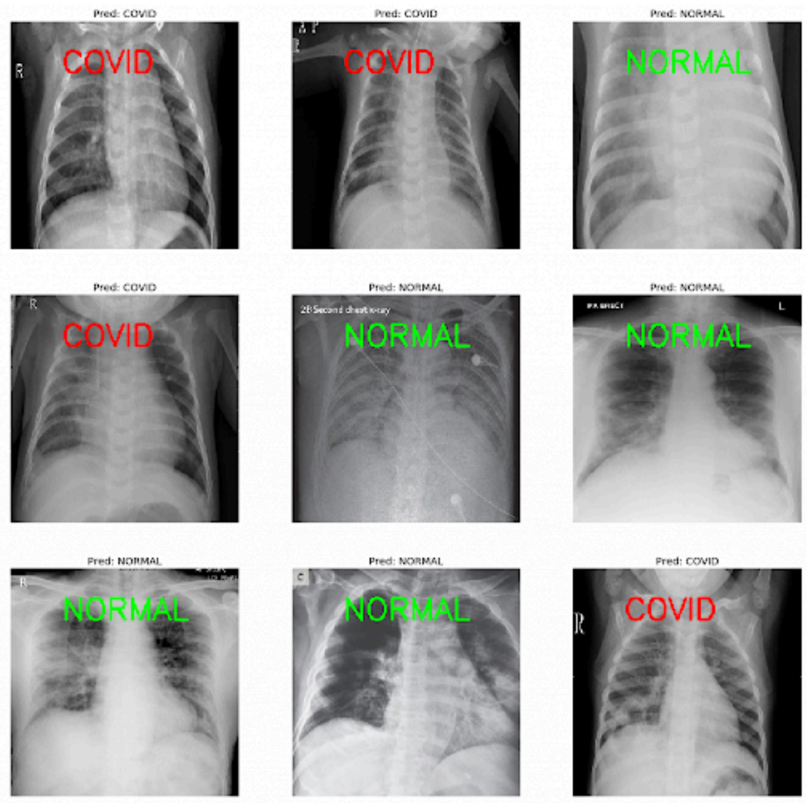


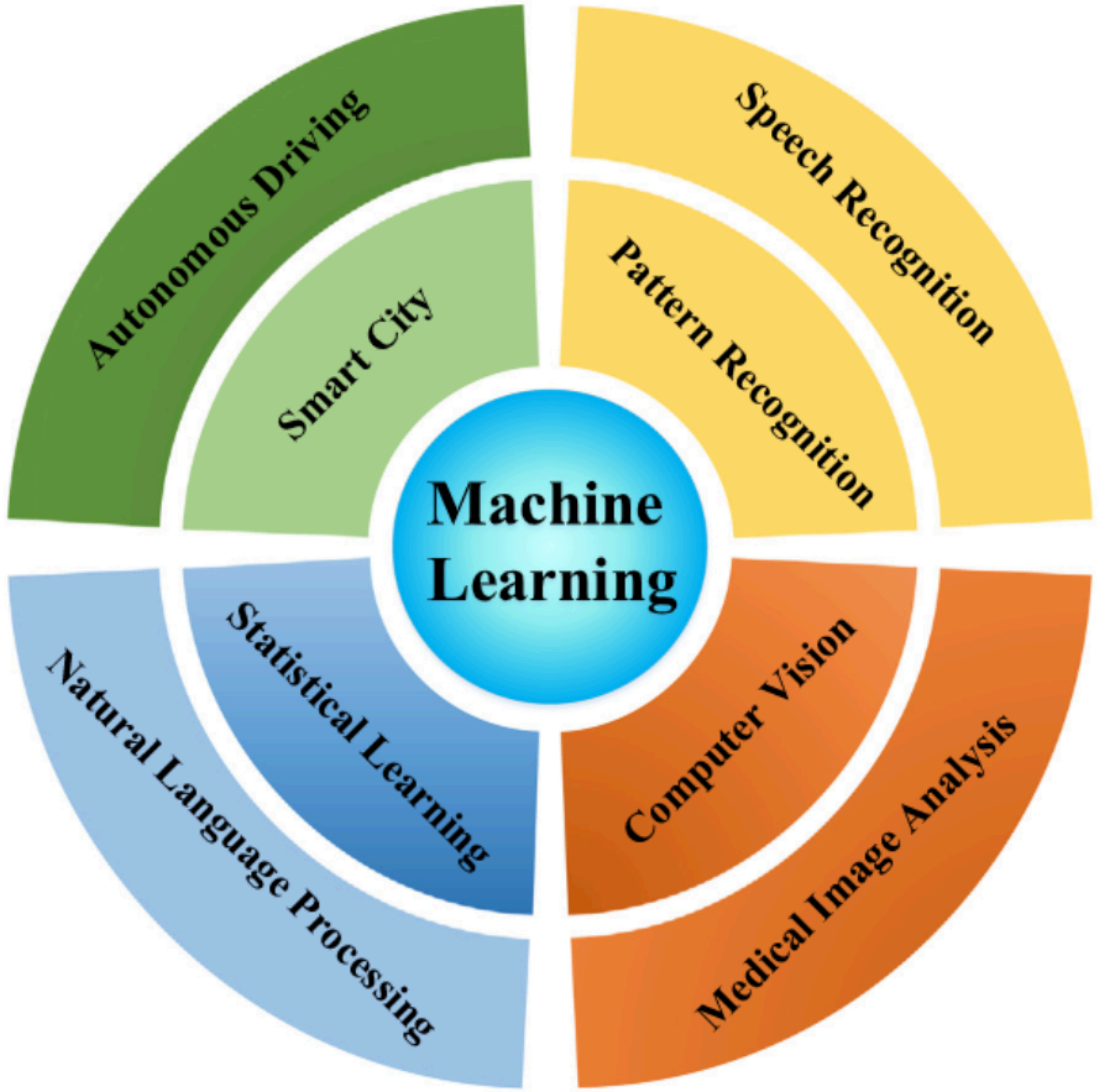
What is Machine Learning?

"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

Where can Machine Learning be used?

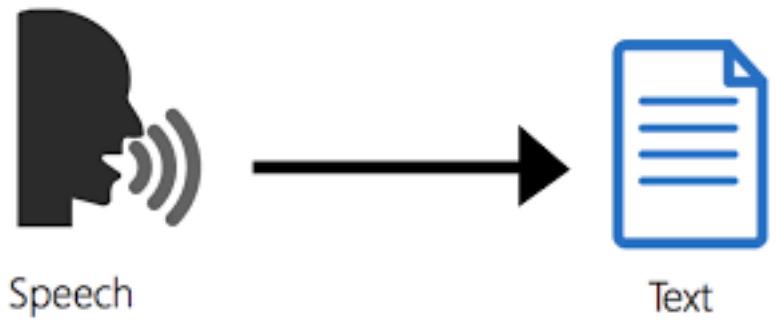




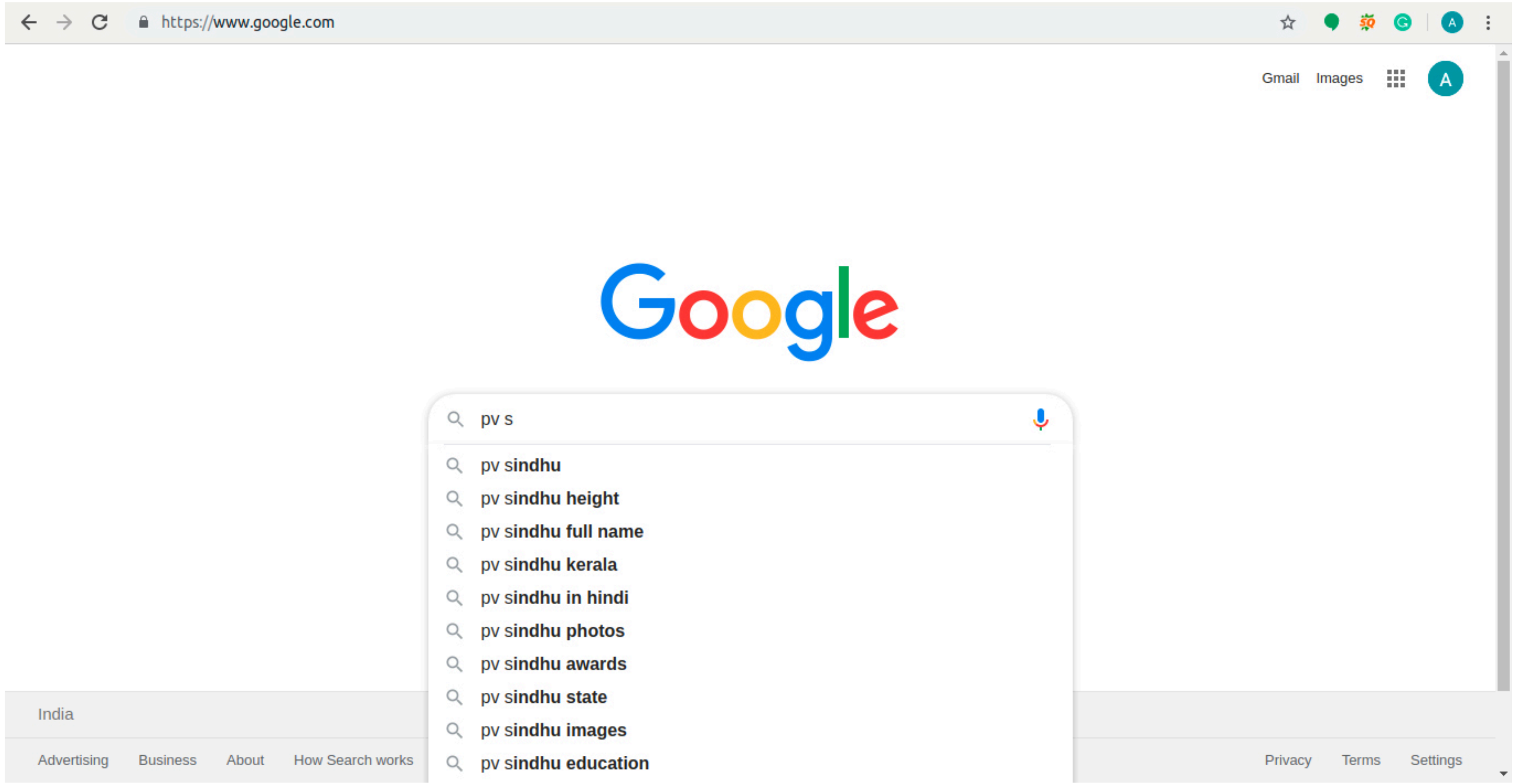


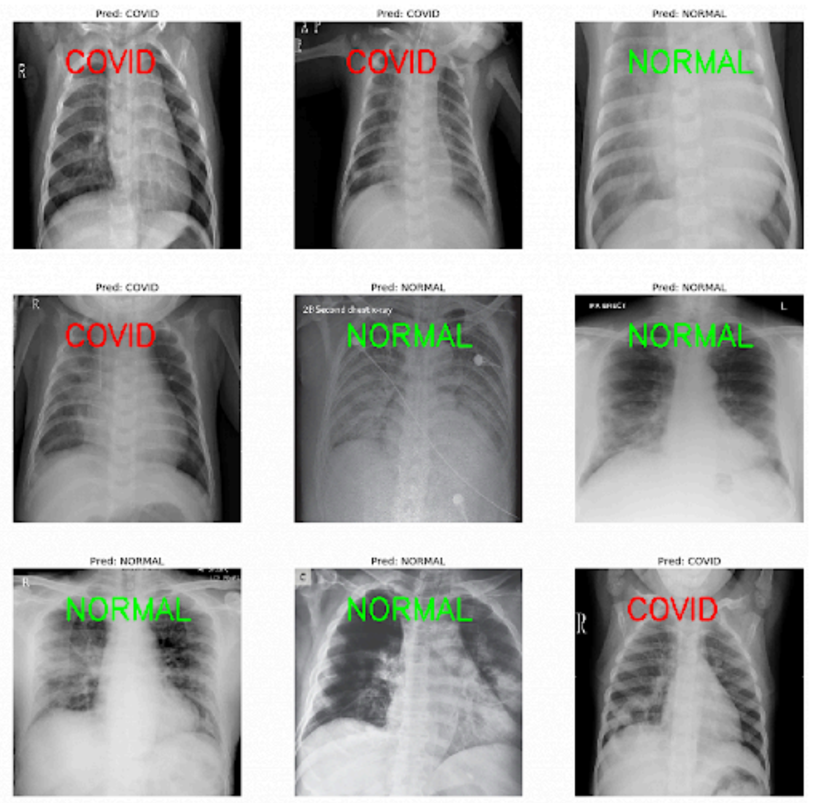
What is Machine Learning?

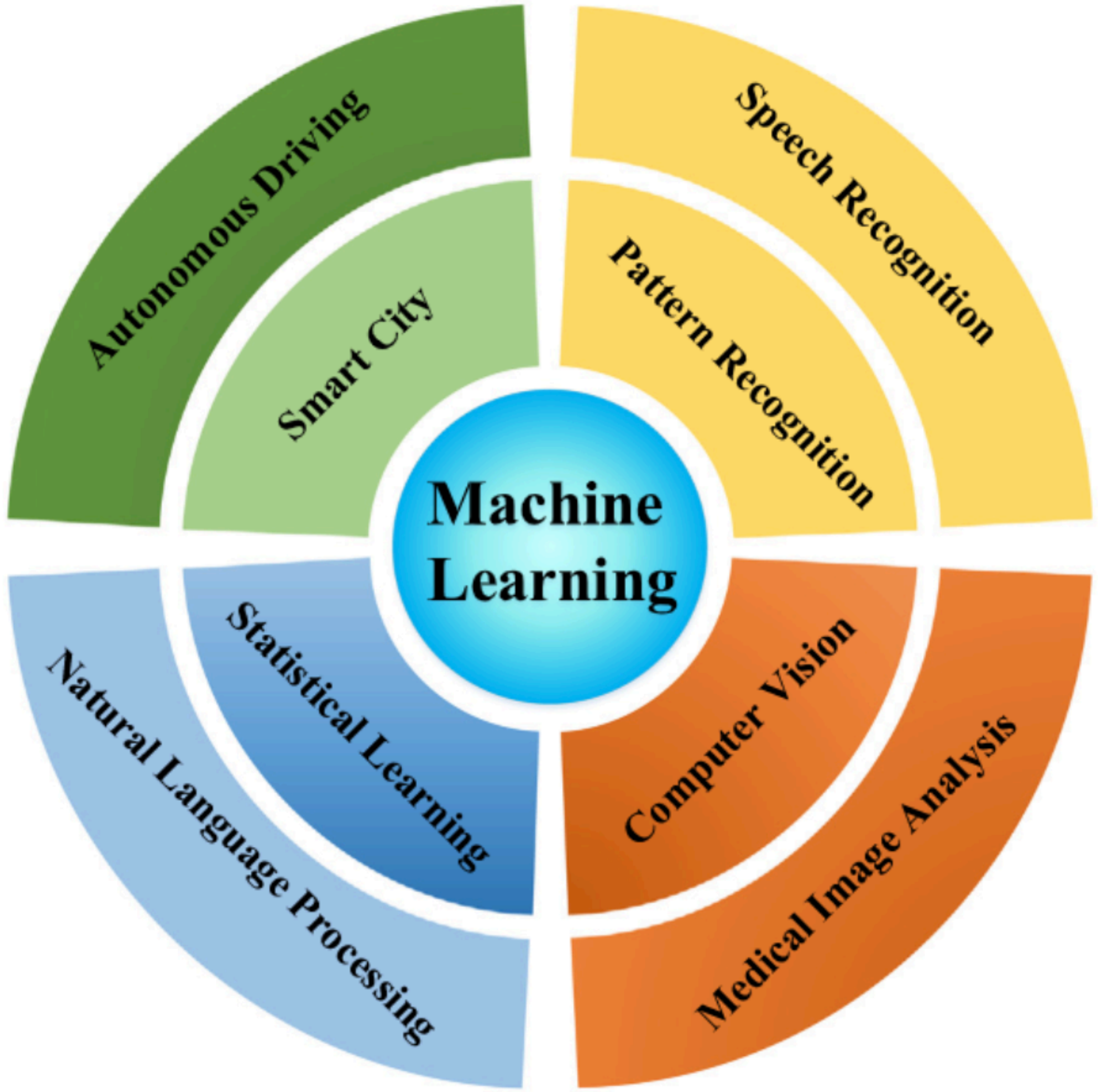
"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.



Where can Machine Learning be used?

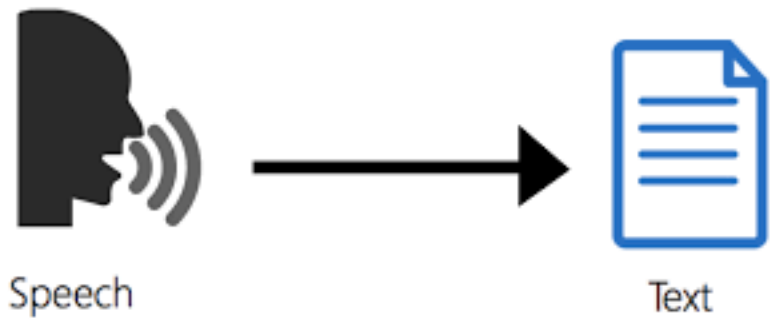




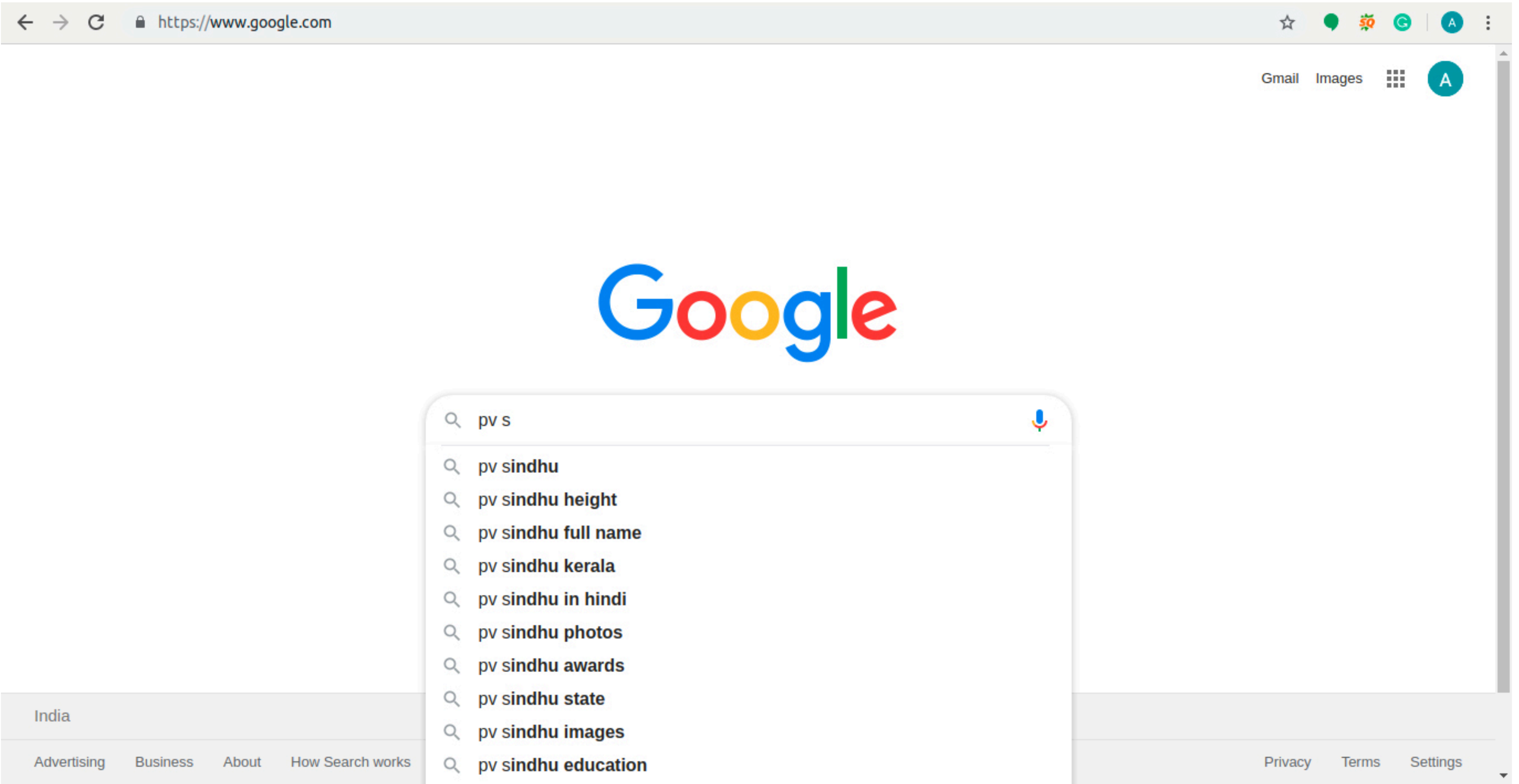


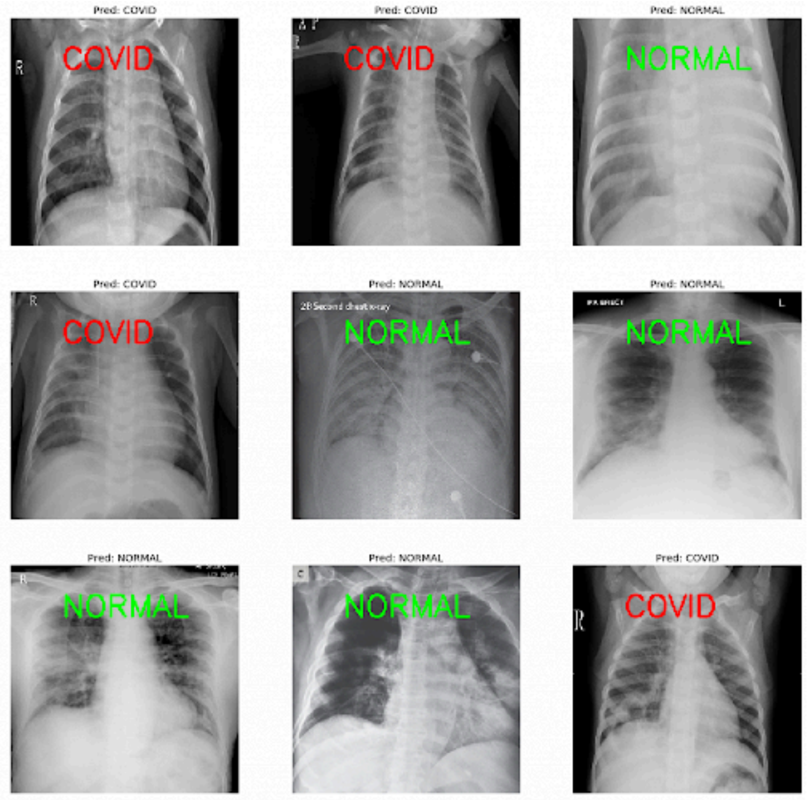
What is Machine Learning?

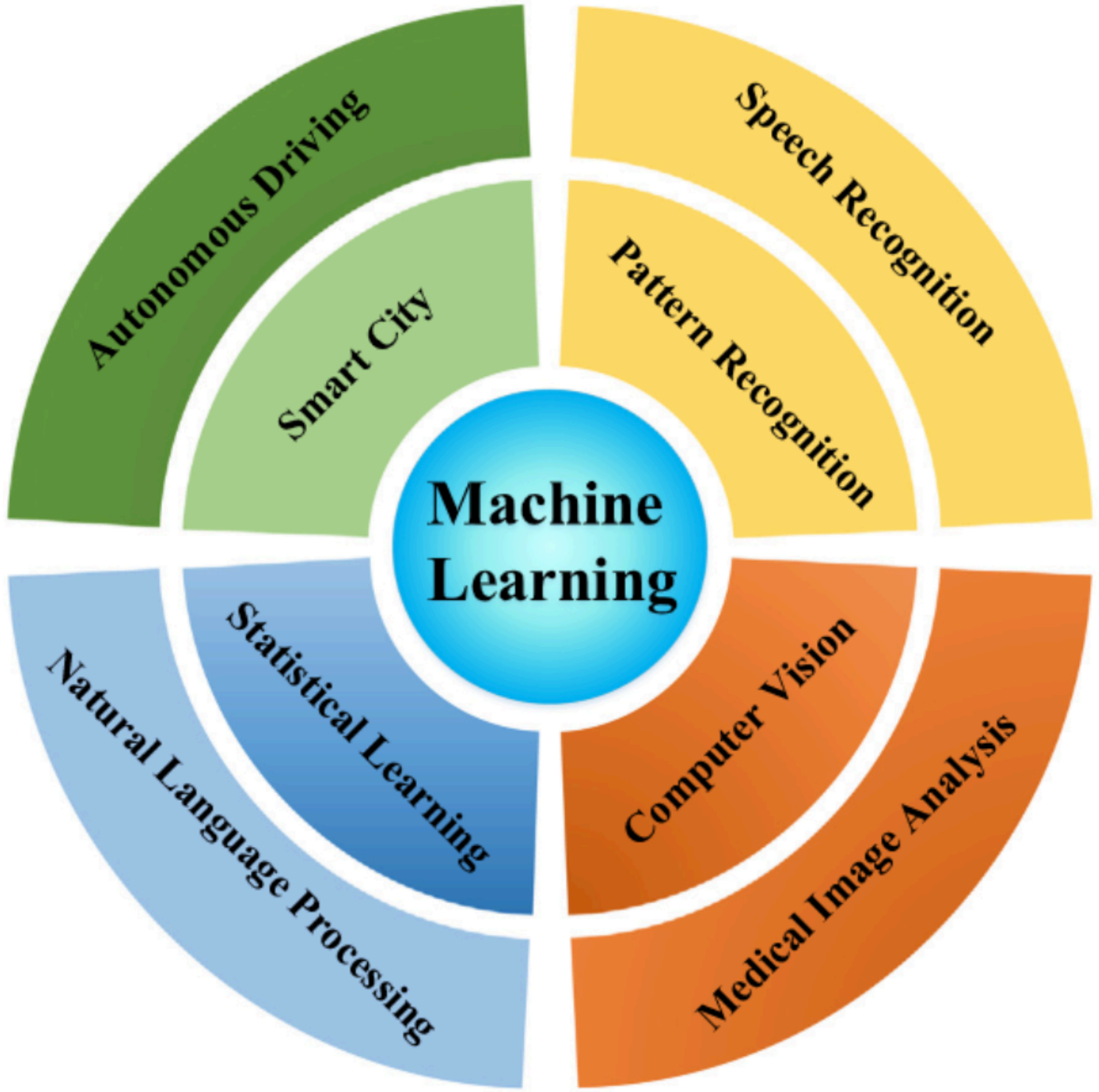
"Machine Learning is the science of getting computers to realize a task without being explicitly programmed.

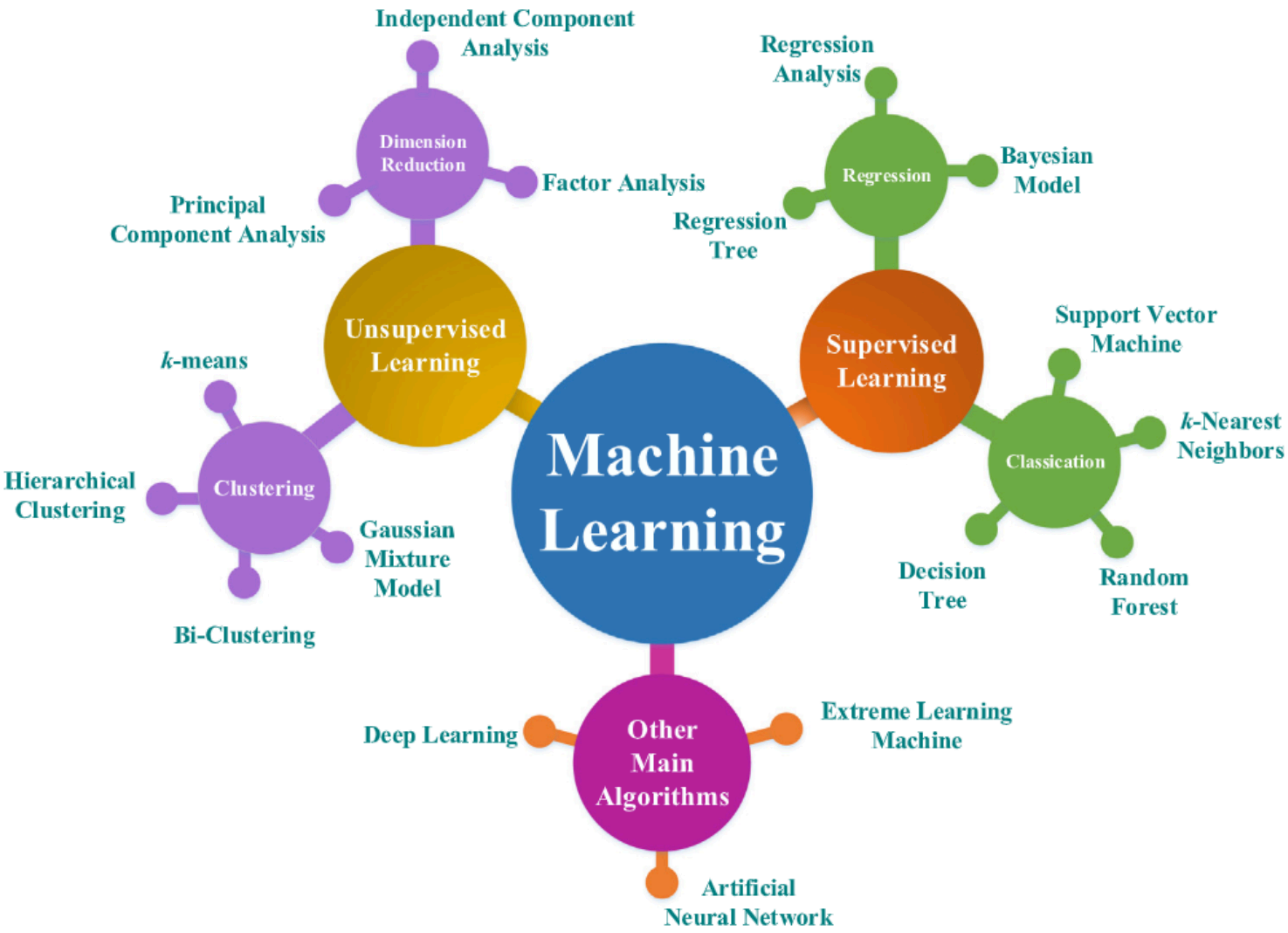


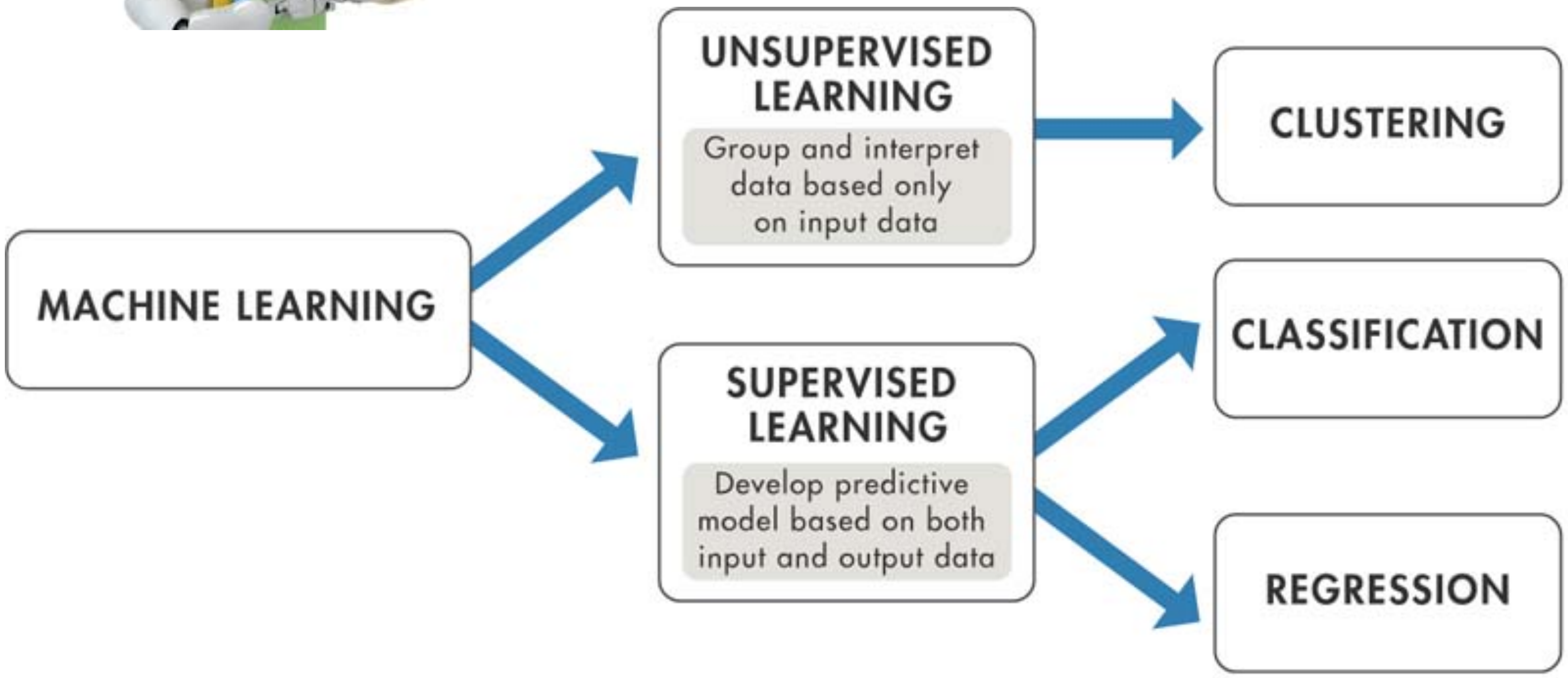
Where can Machine Learning be used?








Machine Learning
Unsupervised Learning
Dimension Reduction
Independent Component Analysis
Factor Analysis
Principal Component Analysis
Clustering
k-means
Hierarchical Clustering
Gaussian Mixture Model
Bi-Clustering
Supervised Learning
Regression
Regression Analysis
Bayesian Model
Regression Tree
Classication
Support Vector Machine
k-Nearest Neighbors
Decision Tree
Random Forest
Other Main Algorithms
Deep Learning
Extreme Learning Machine
Artificial Neural Network


MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Supervised learning

Supervised learning

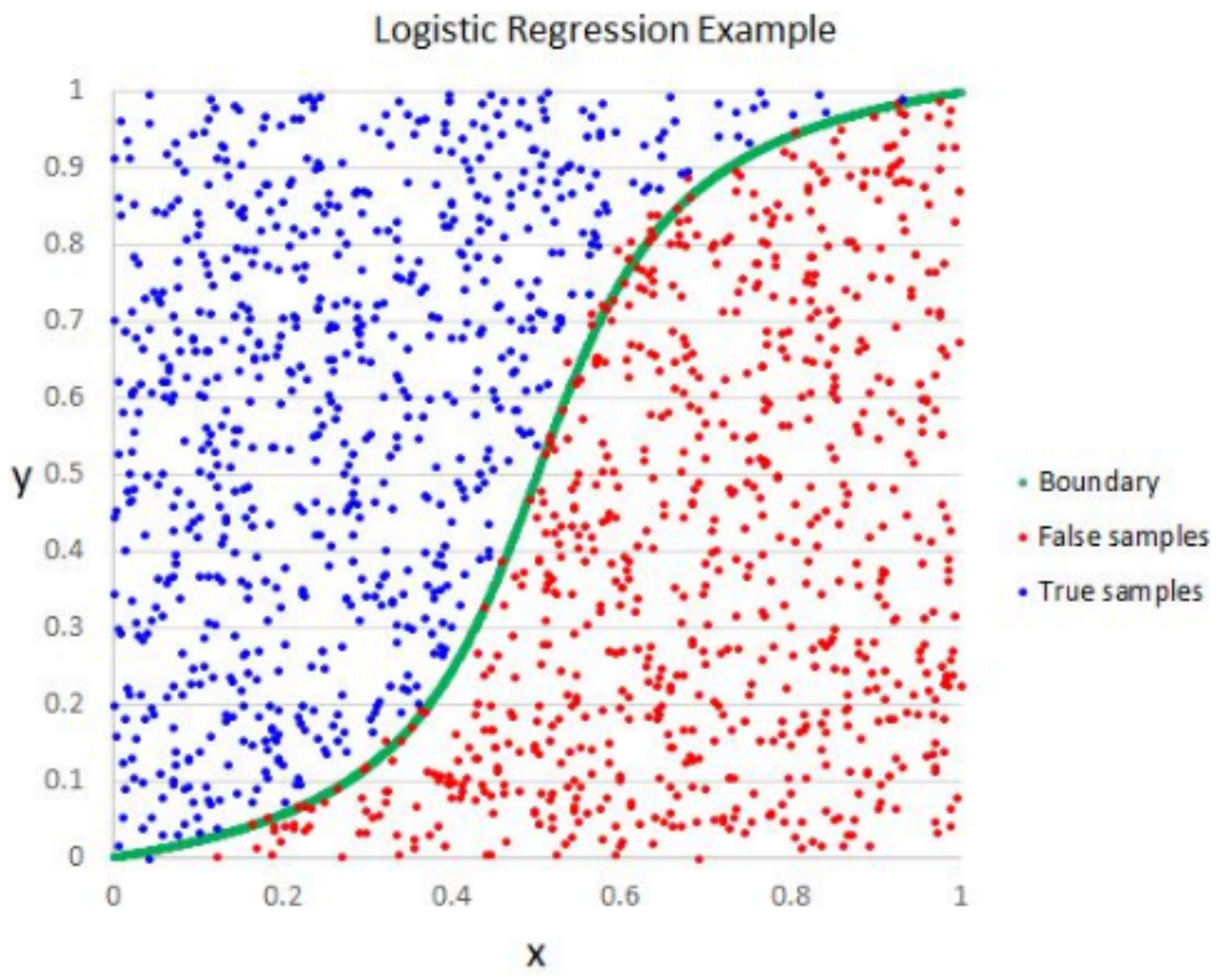
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Supervised learning



Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

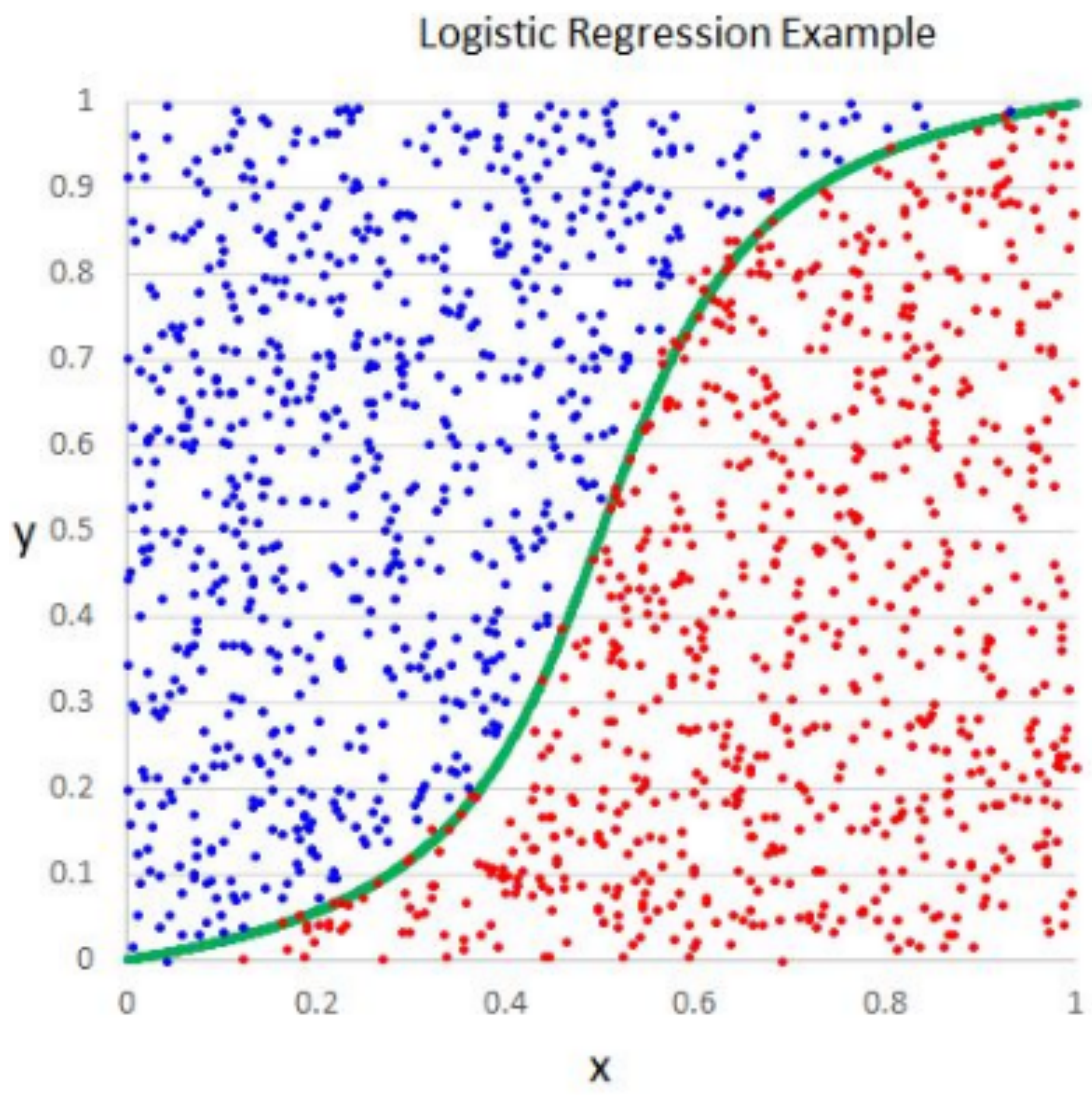
Supervised learning

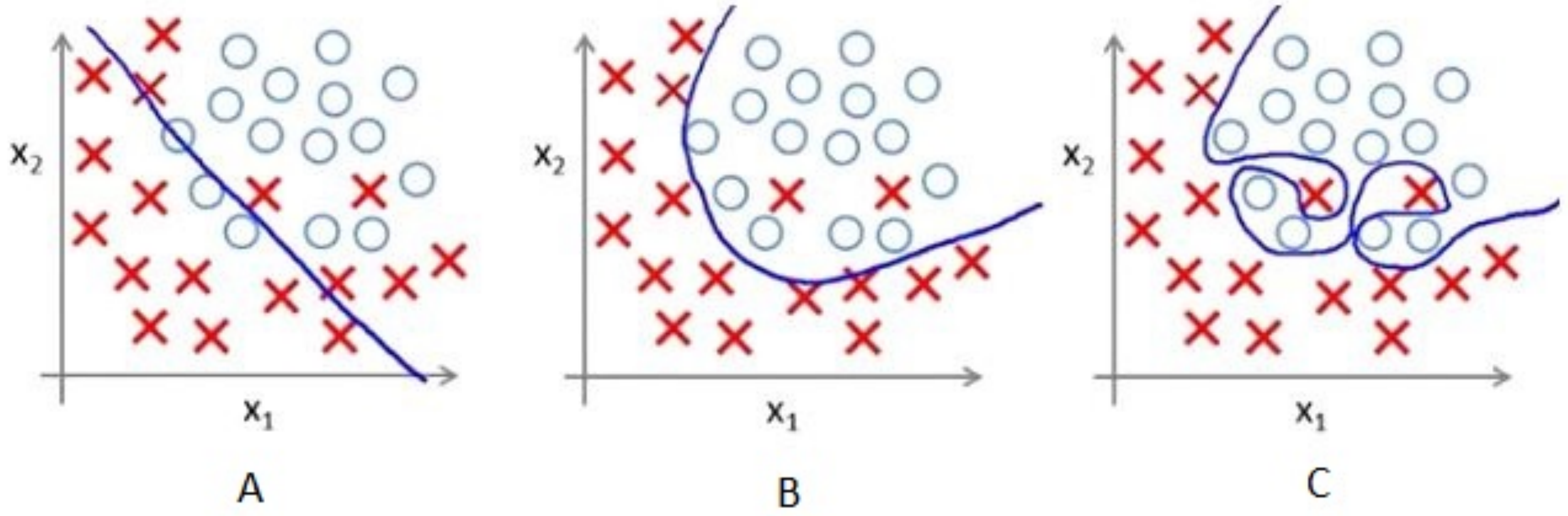
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

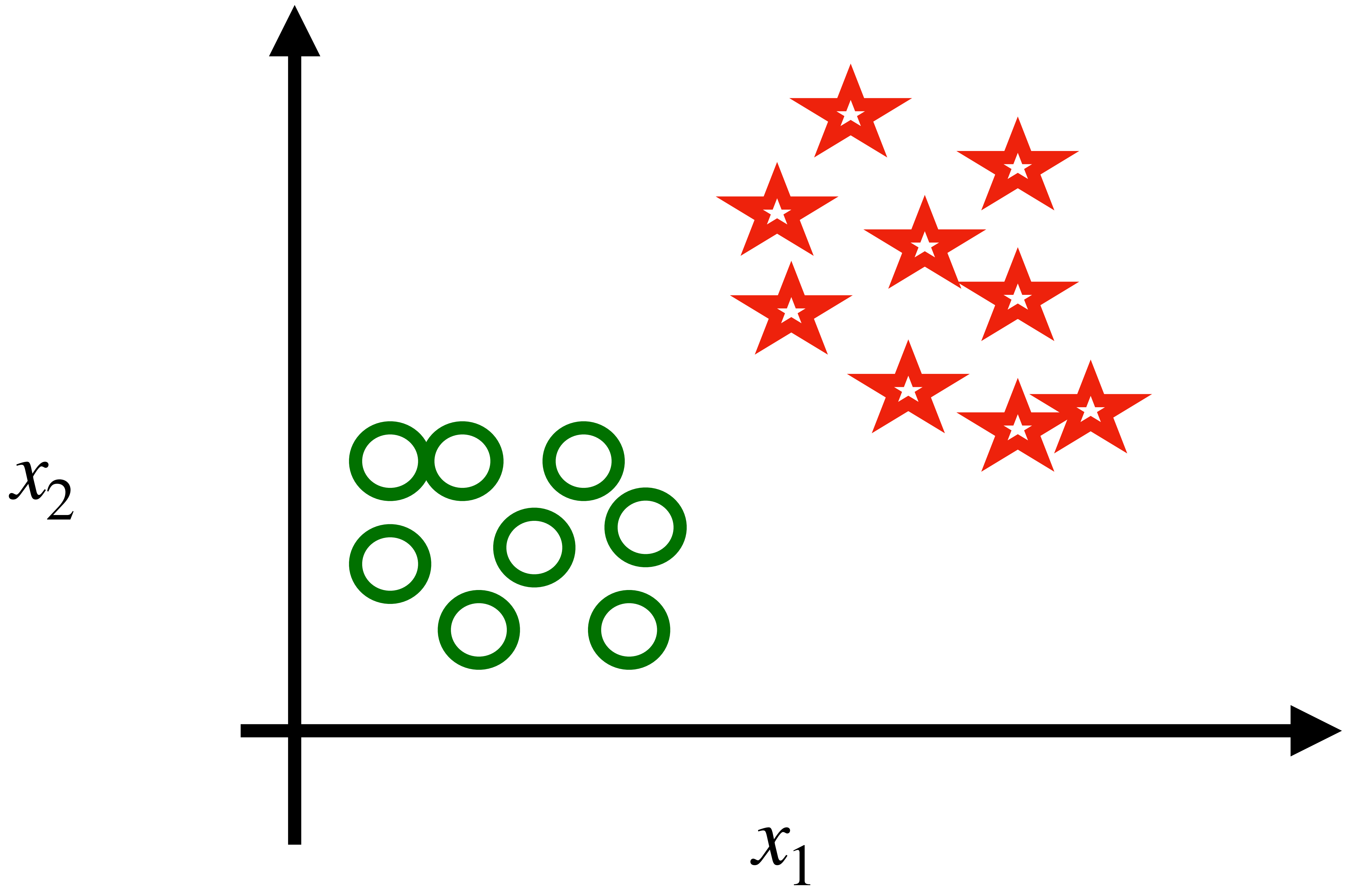
Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$





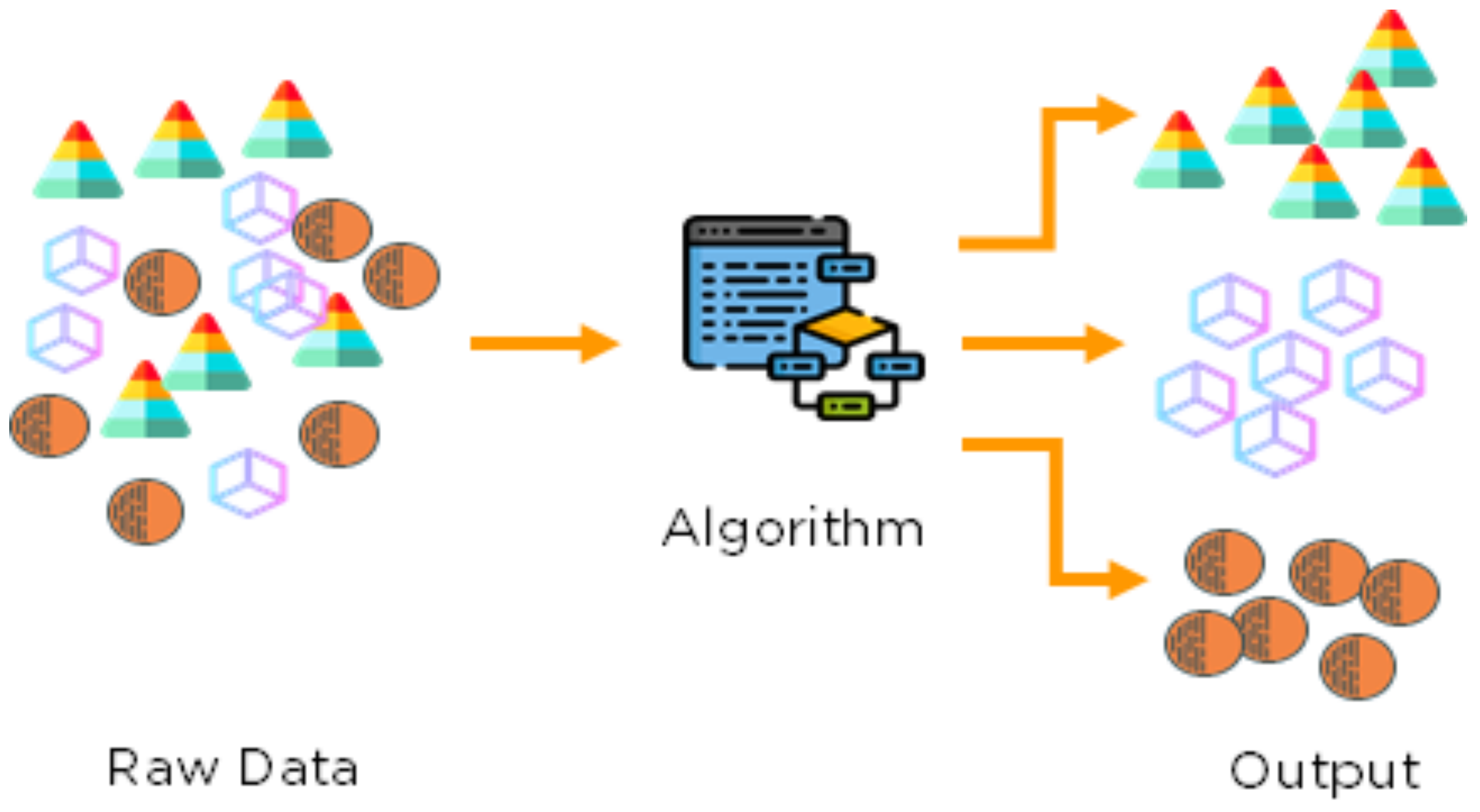
Supervised Learning



Unsupervised learning

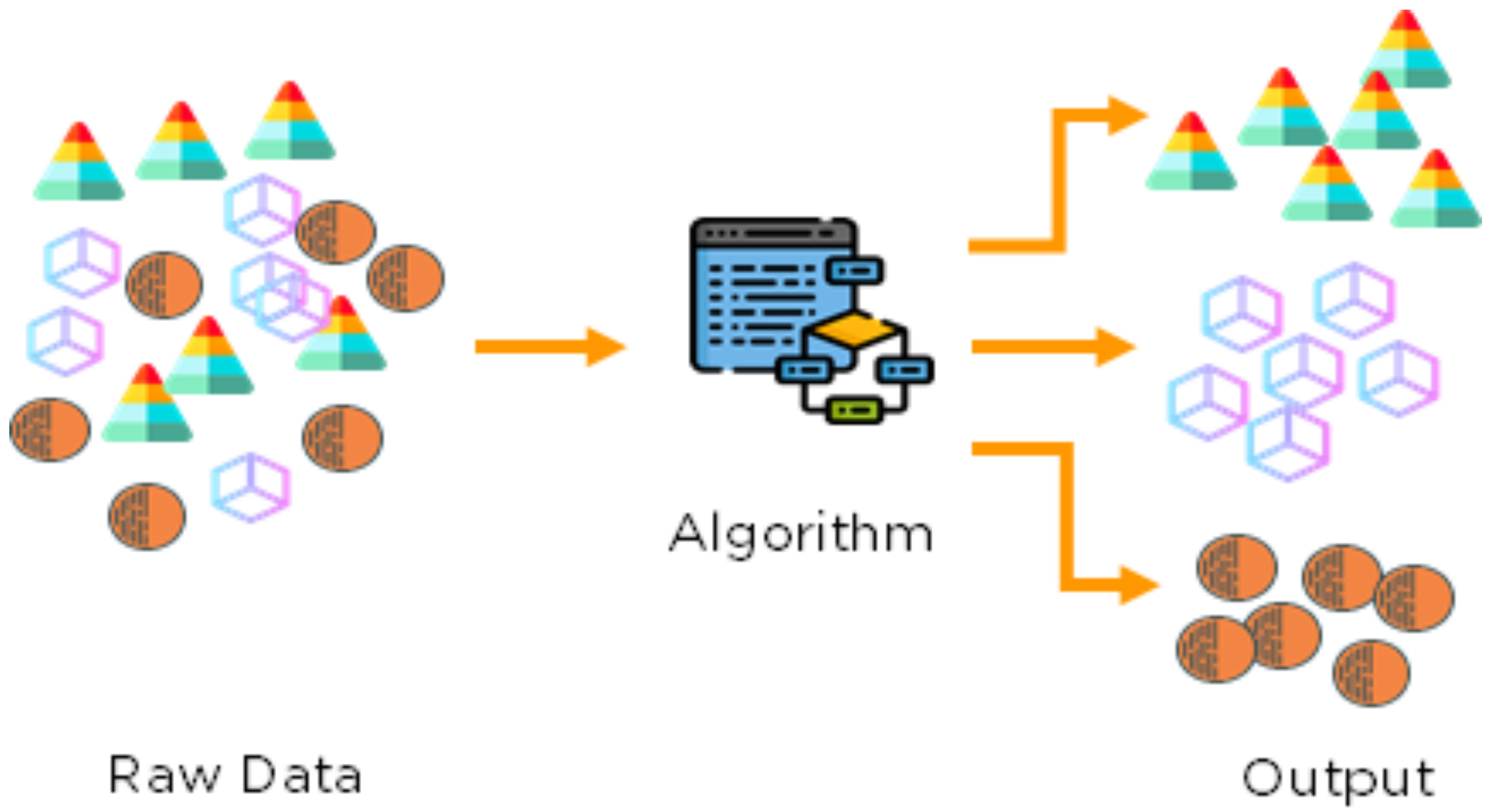
Unsupervised learning

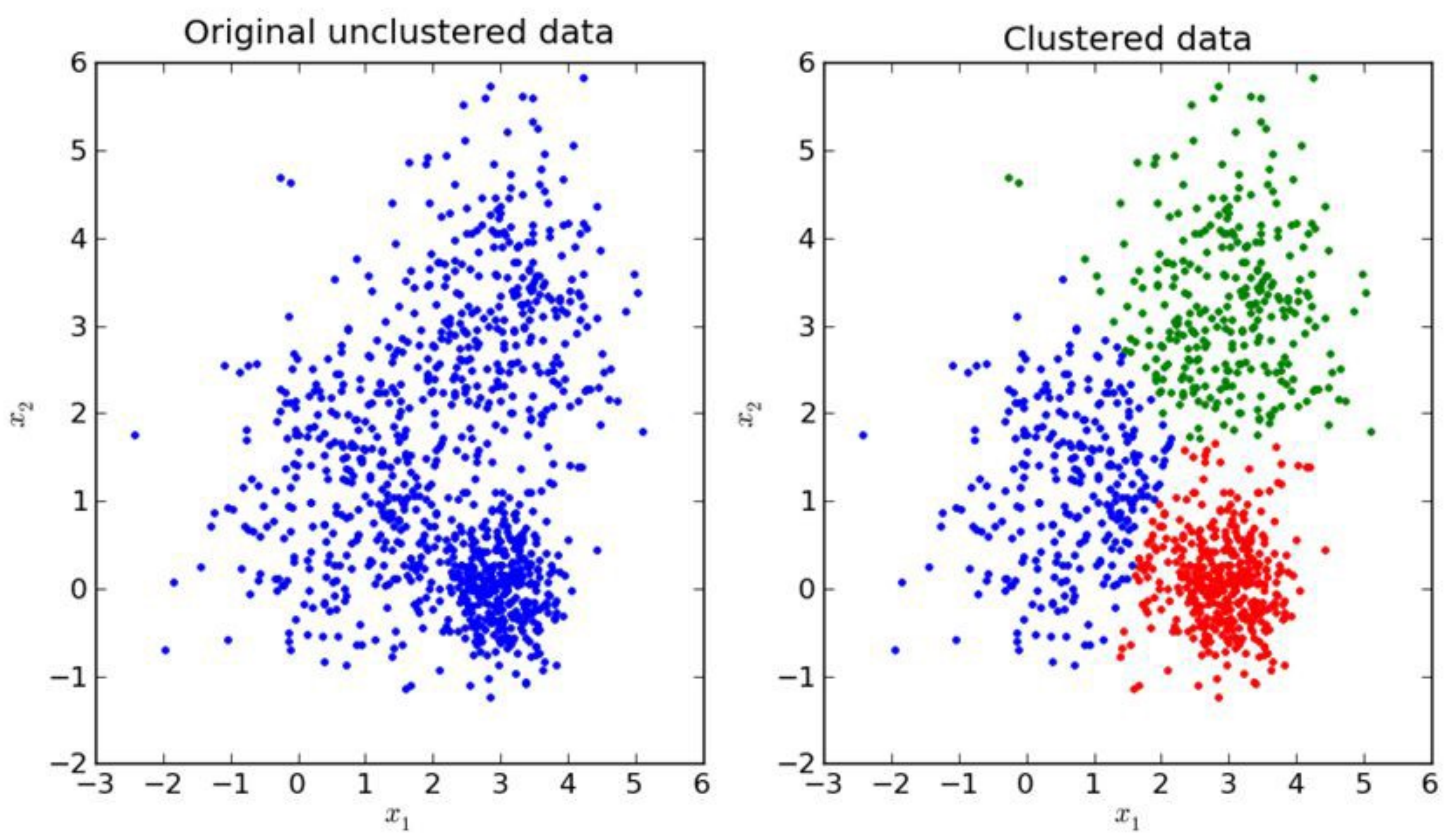
No label provided. Machine is supposed to find some structures in the raw data.



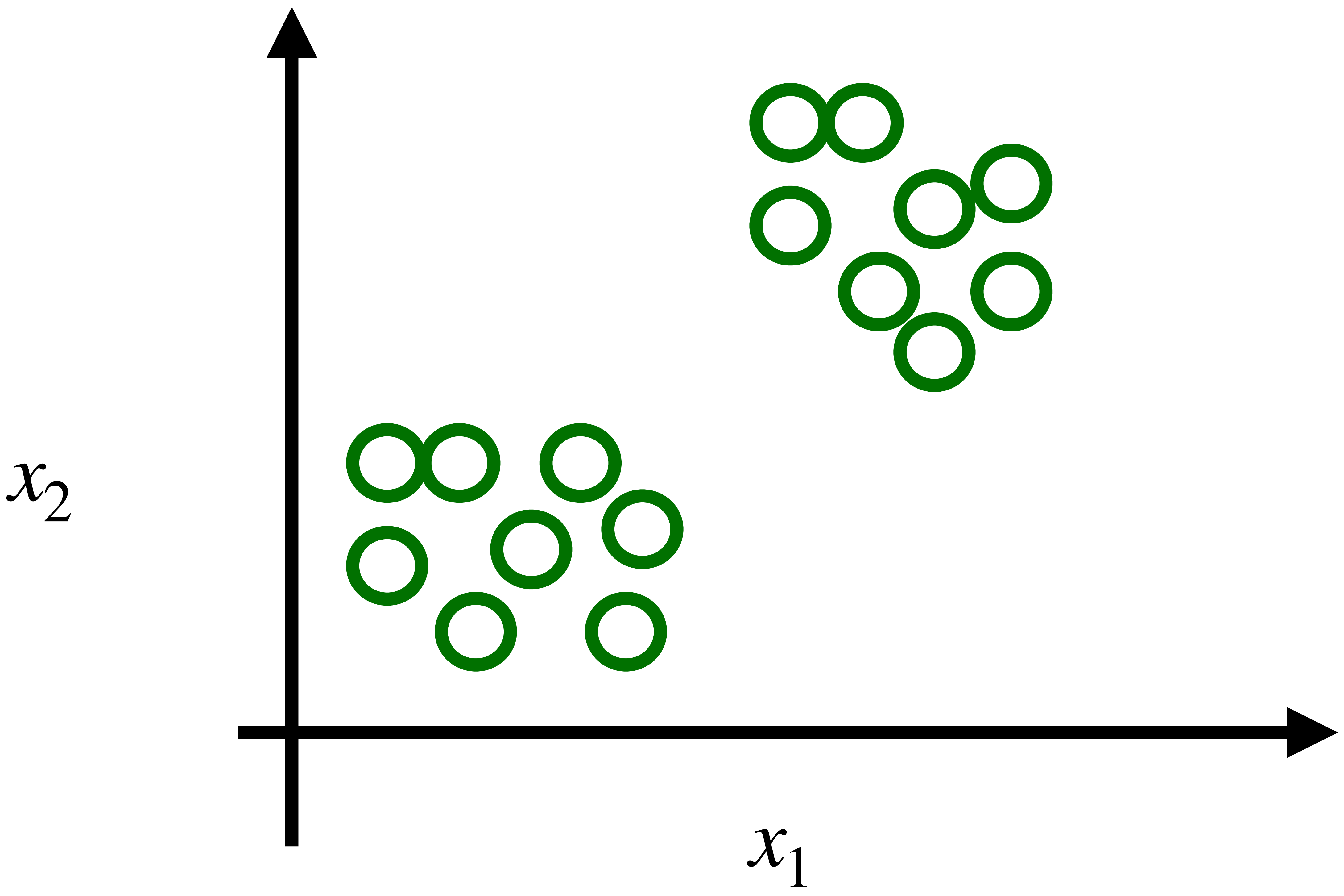
Unsupervised learning

No label provided. Machine is supposed to find some structures in the raw data.





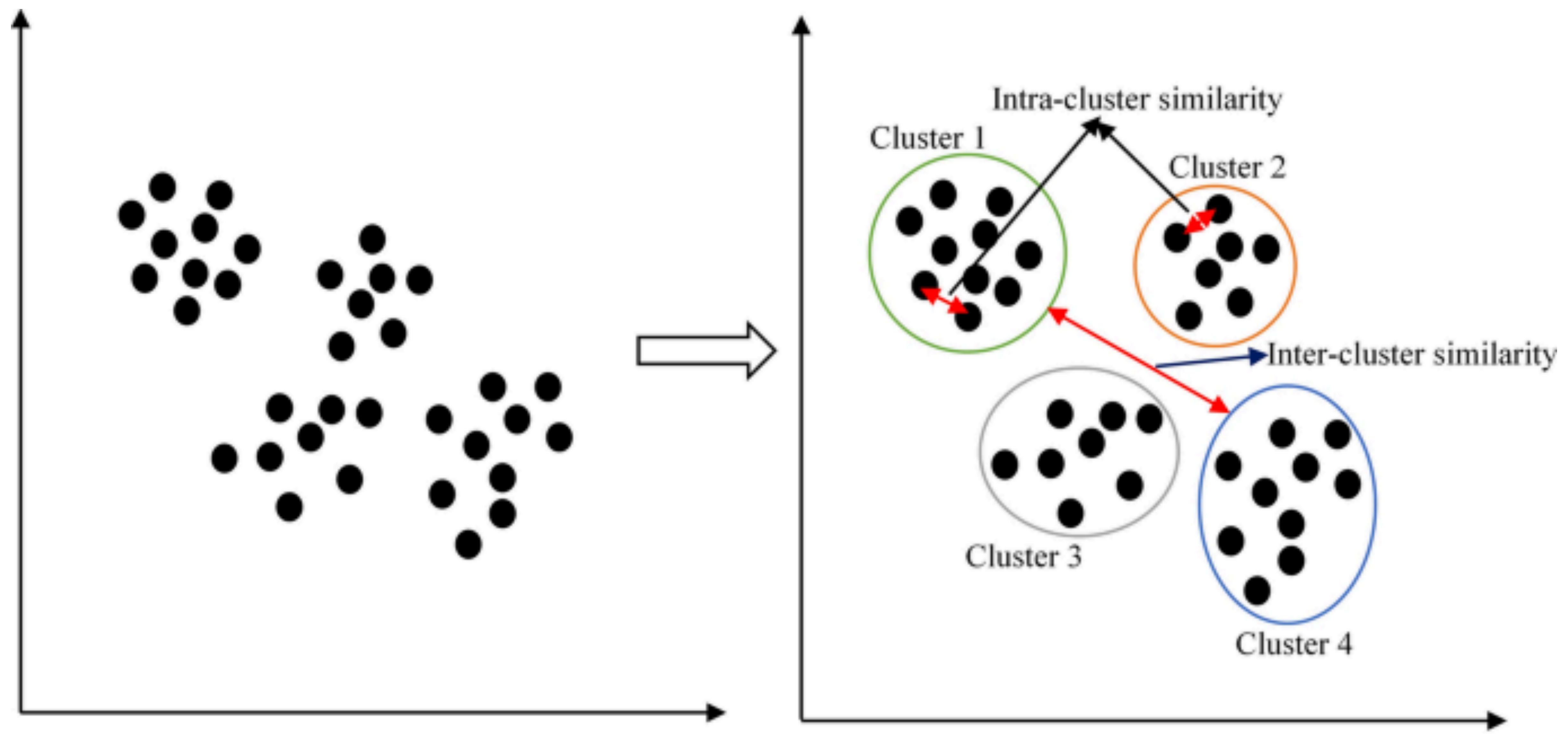
Unsupervised Learning



ML in Astrophysics

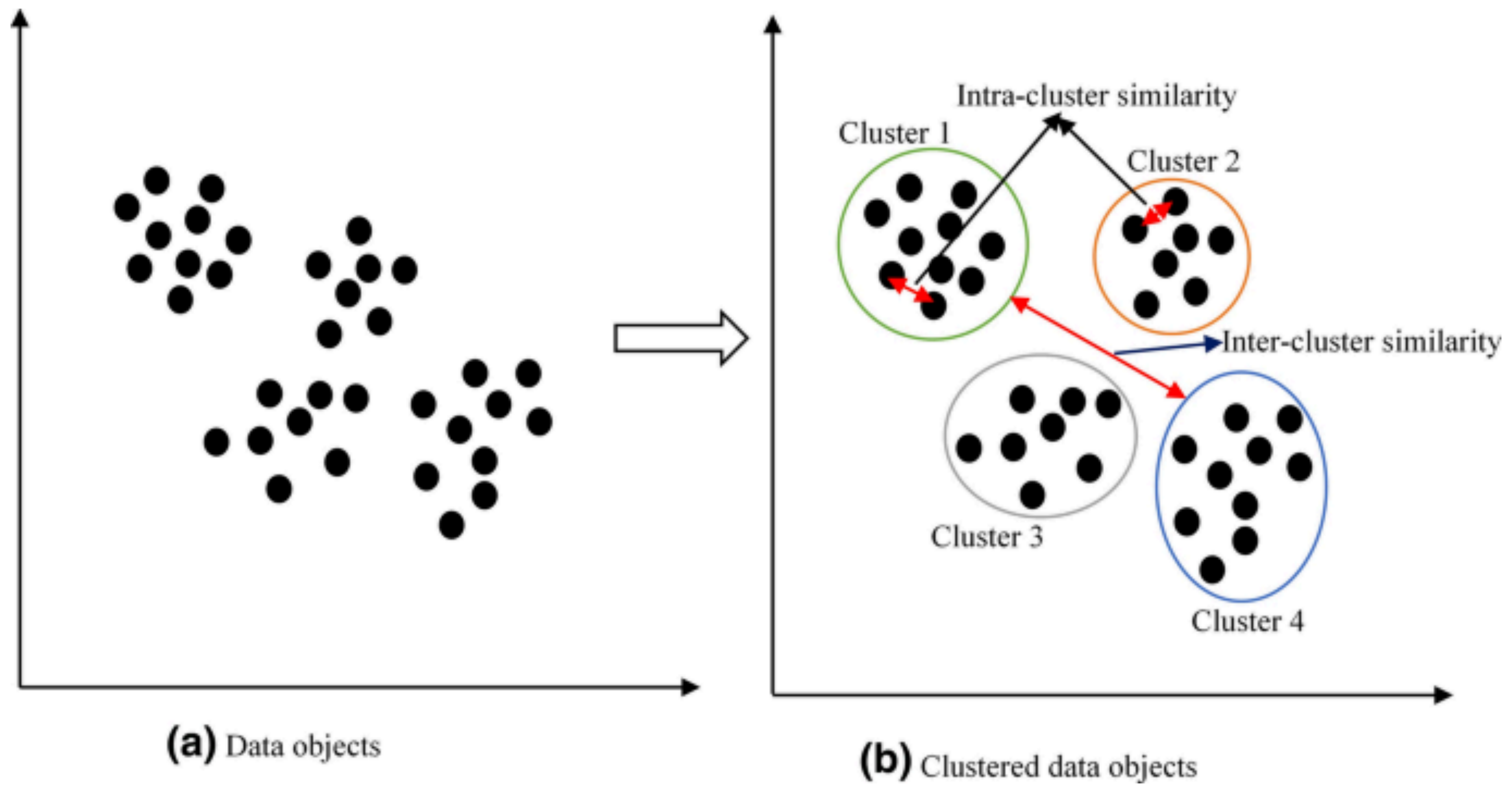
- Spatial clustering
- Source classification with images
- Object detection
- Data cleaning
- Inferring stellar parameters from spectra
- Signal detection
- ...

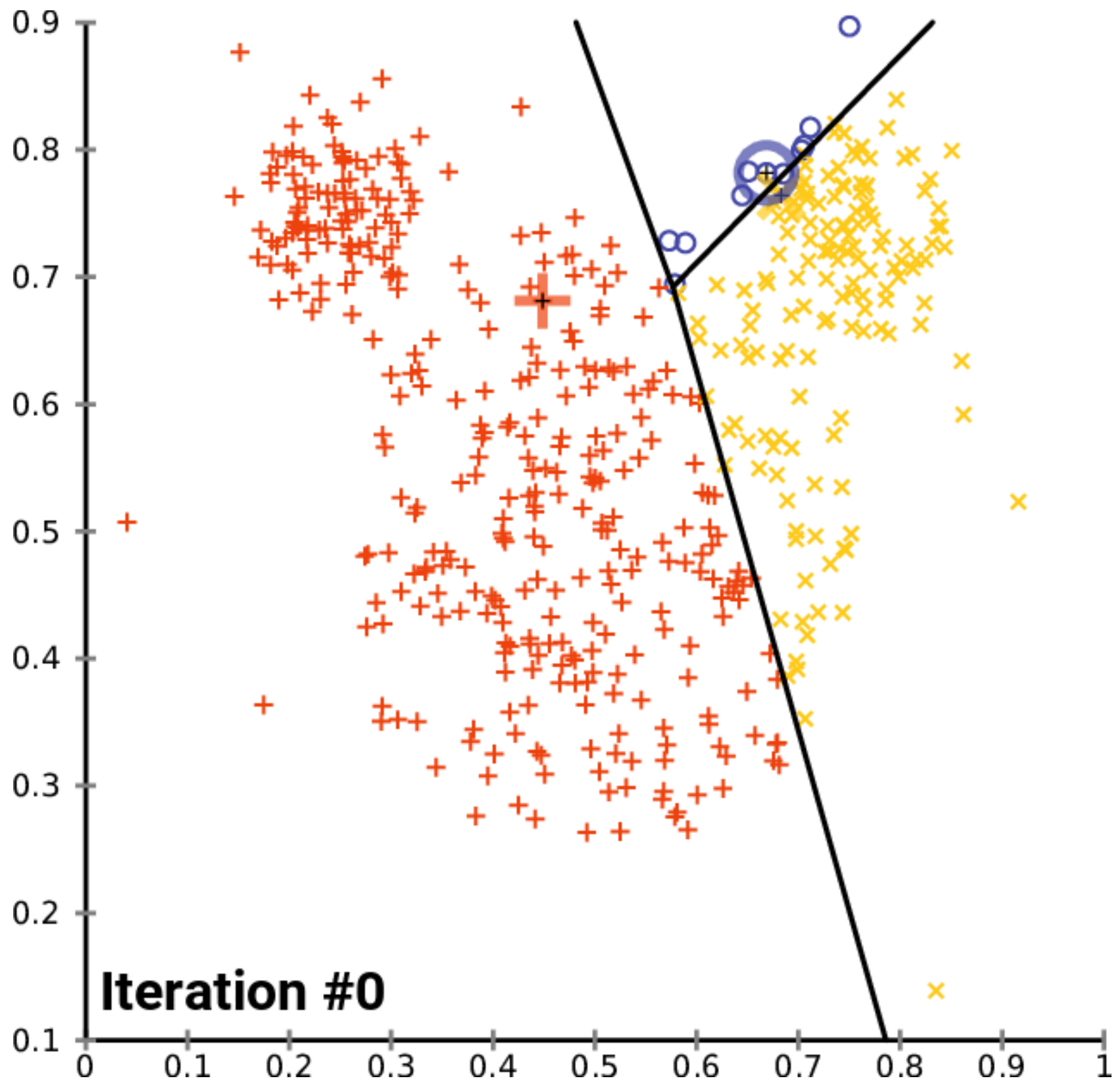
k-means clustering



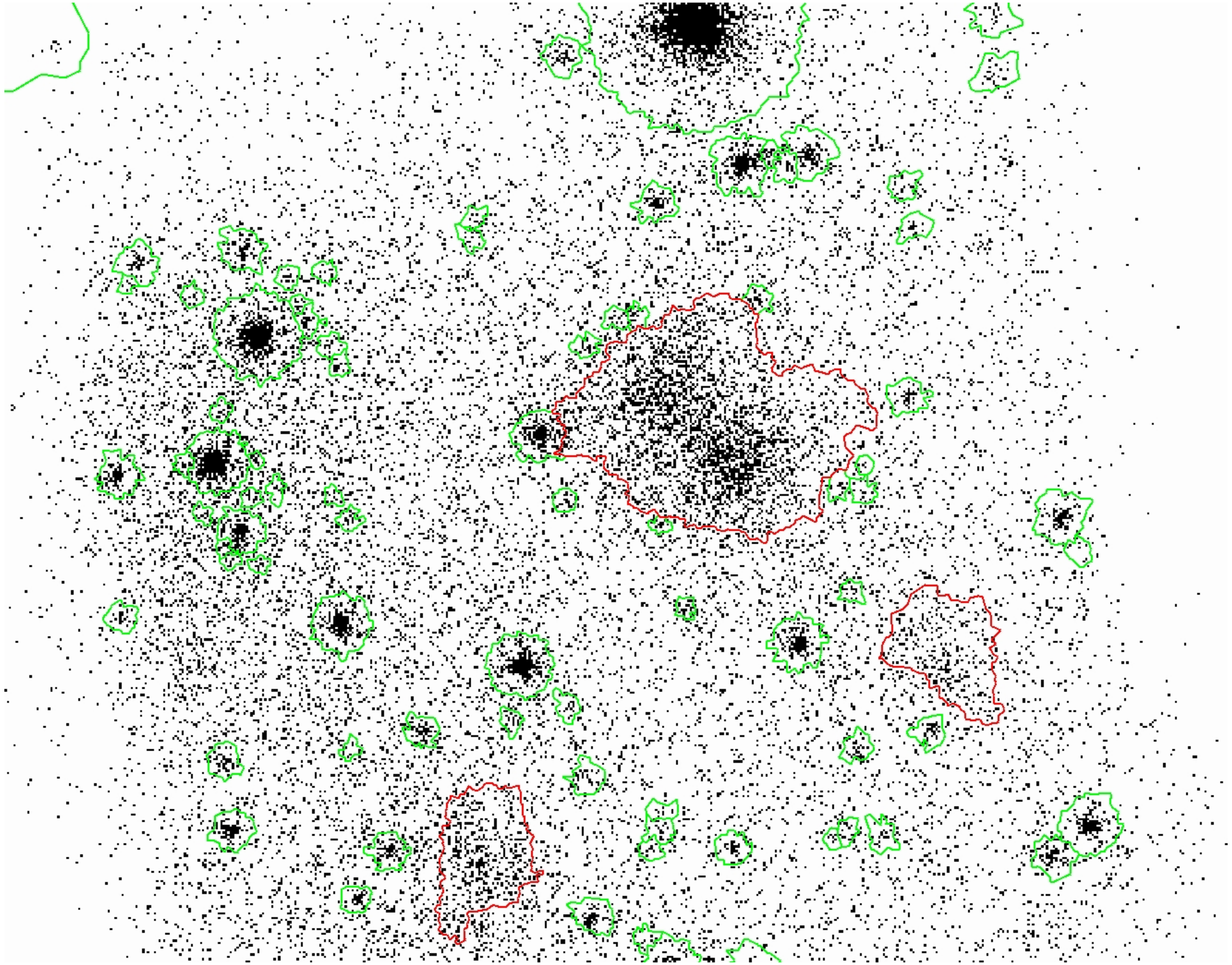
(a) Data objects (b) Clustered data objects

k-means clustering



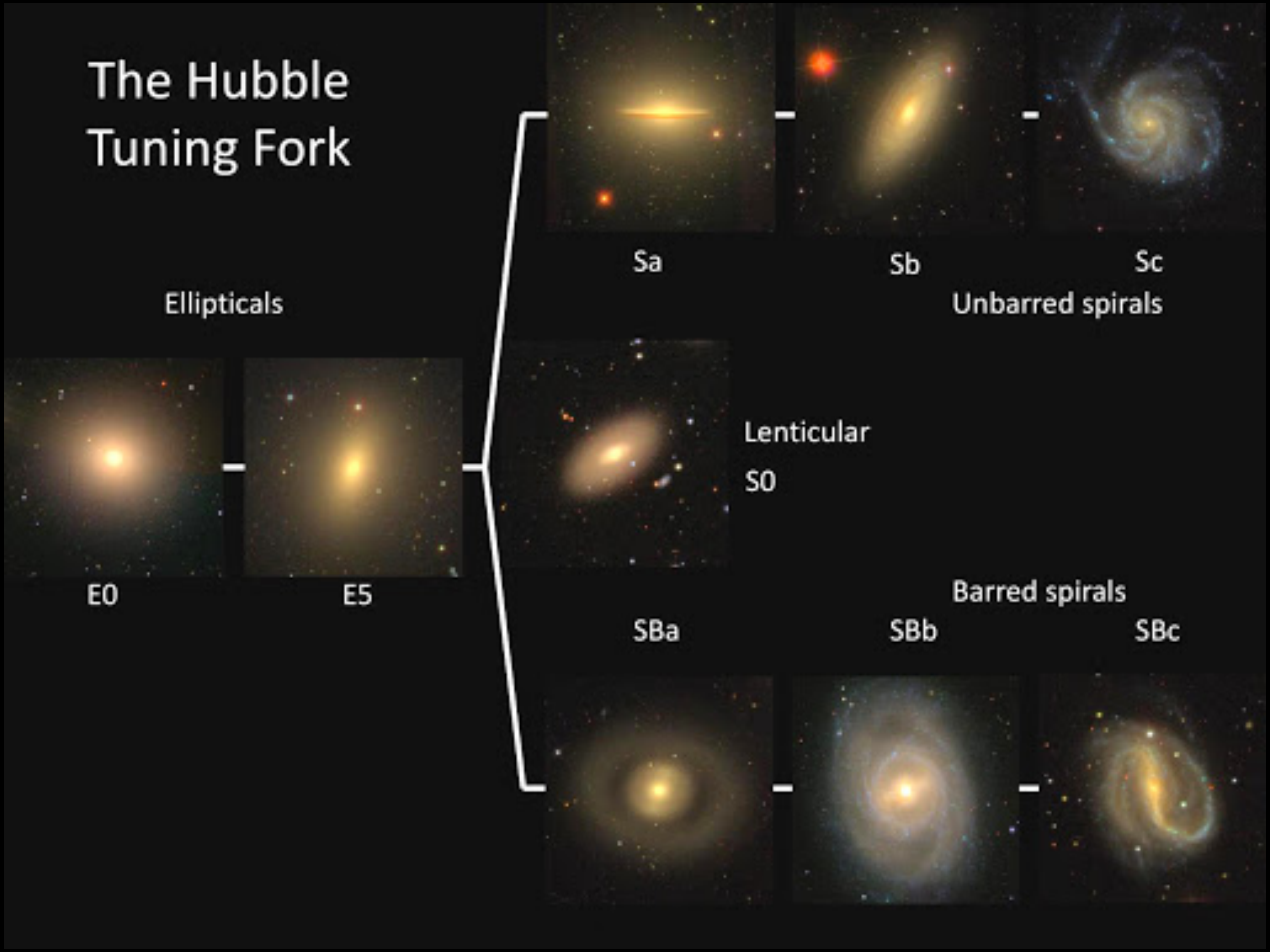


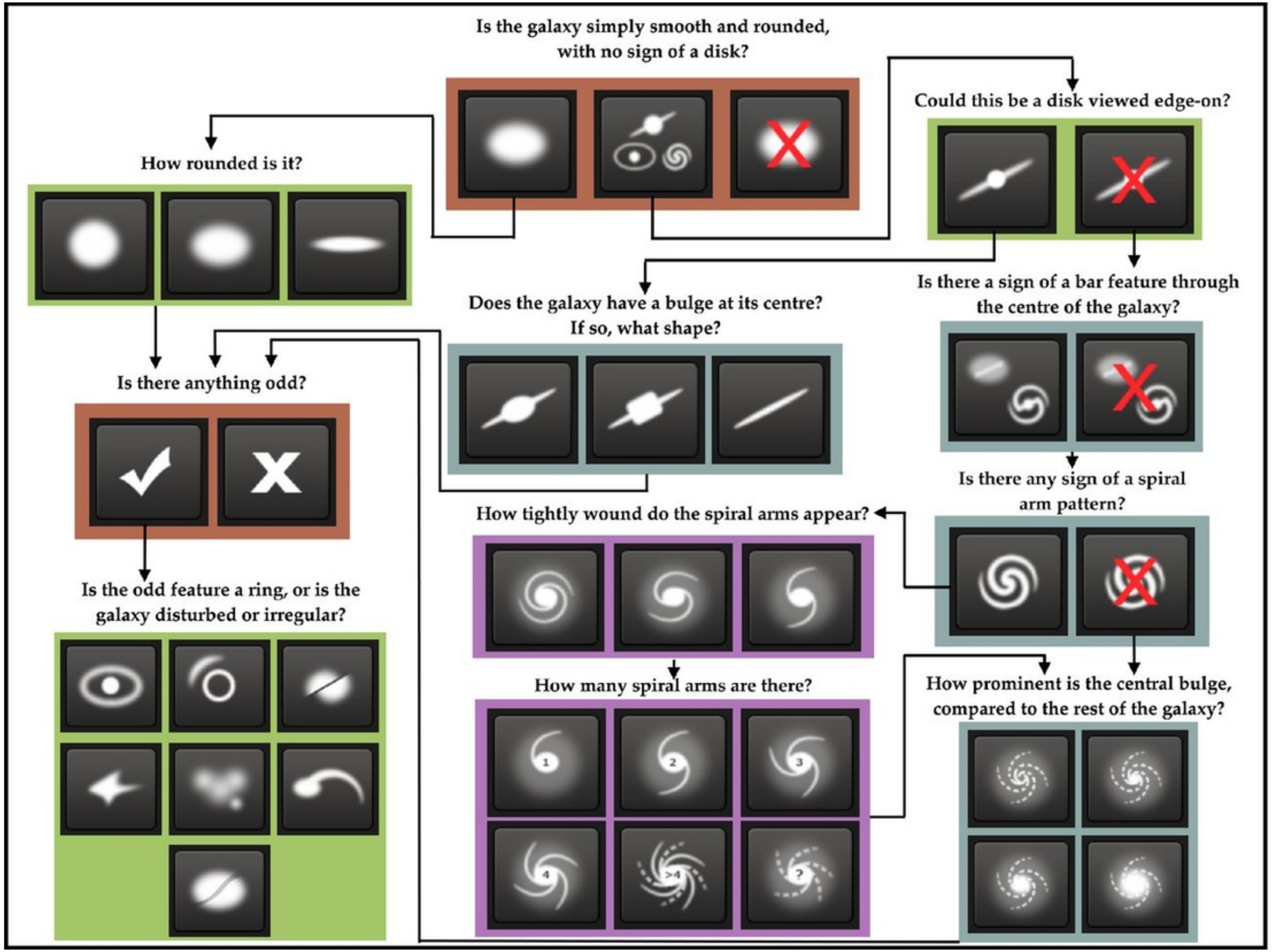
Spatial clustering



Tundo et al. 2012

Source classification with images

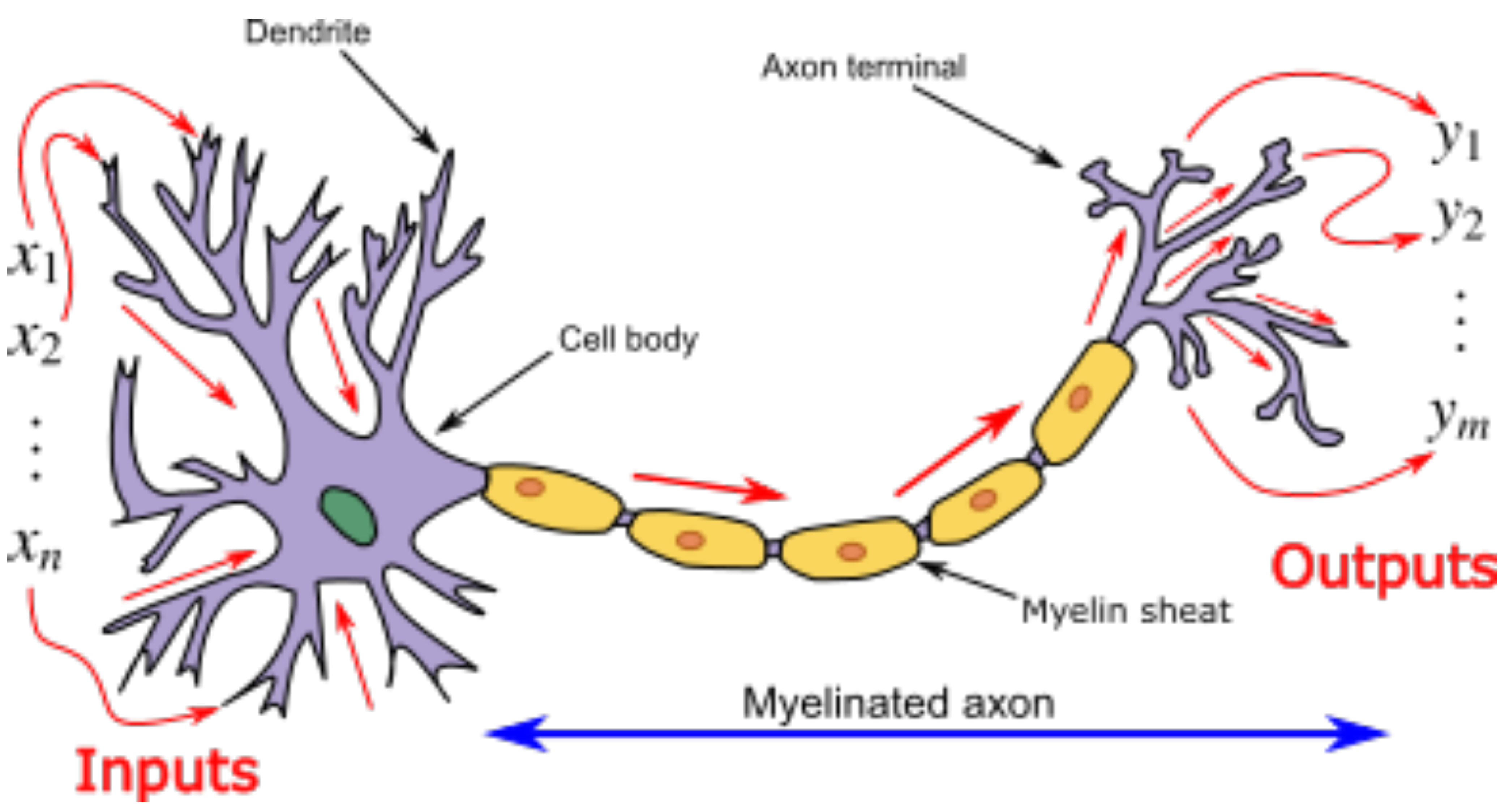


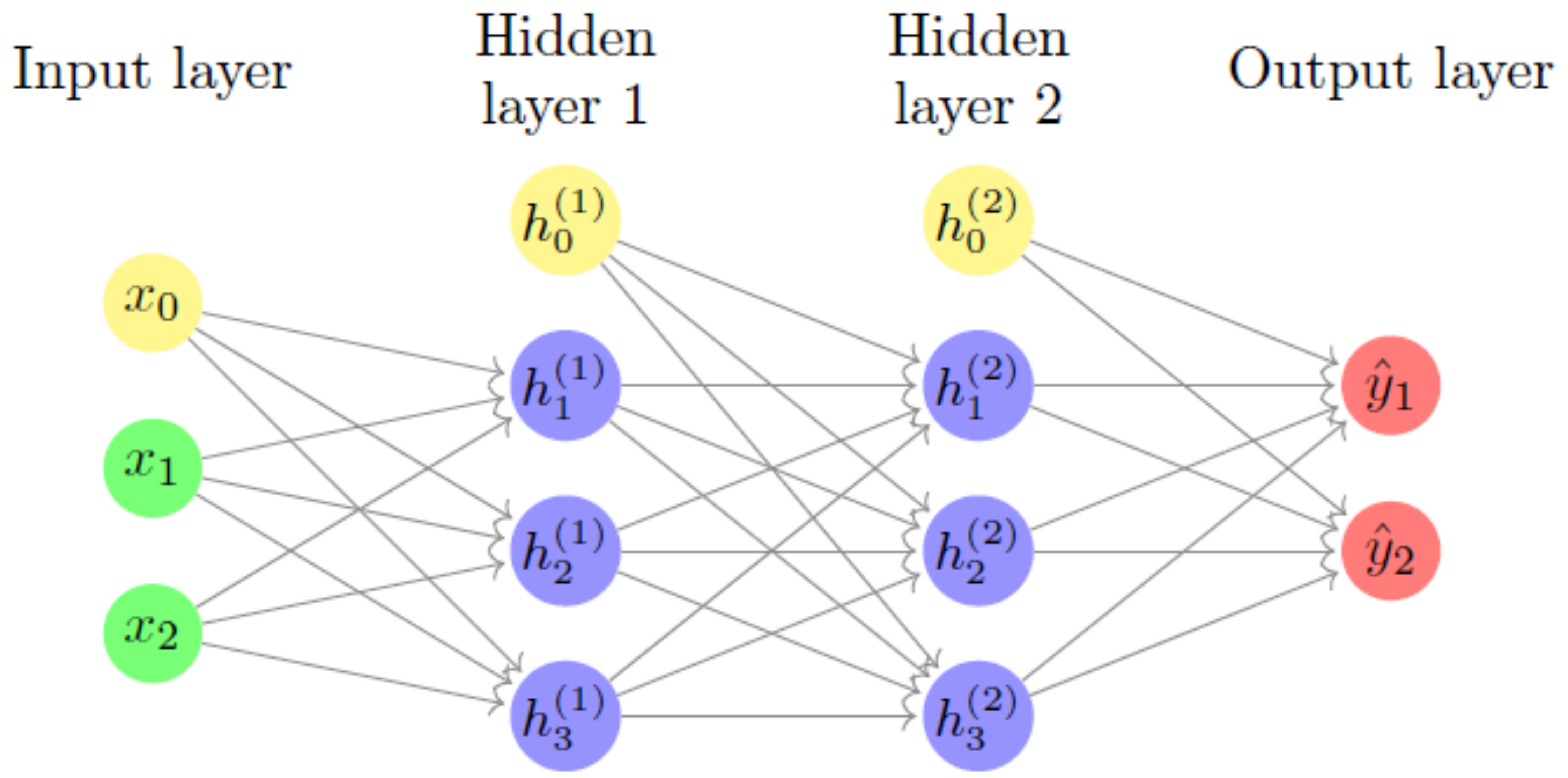


Kyle W. Willett et al. 2013

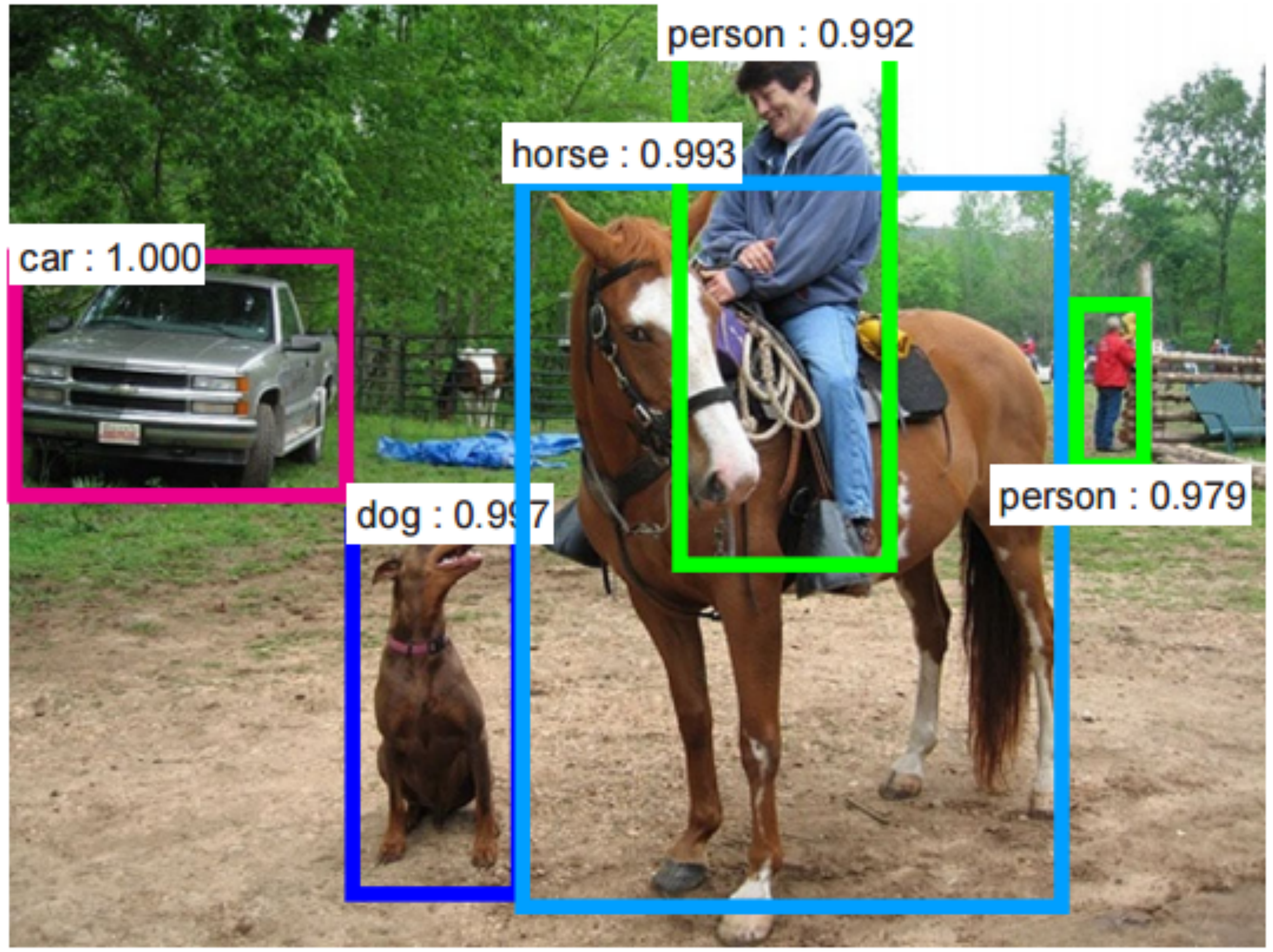
https://www.galaxyzoo.org/

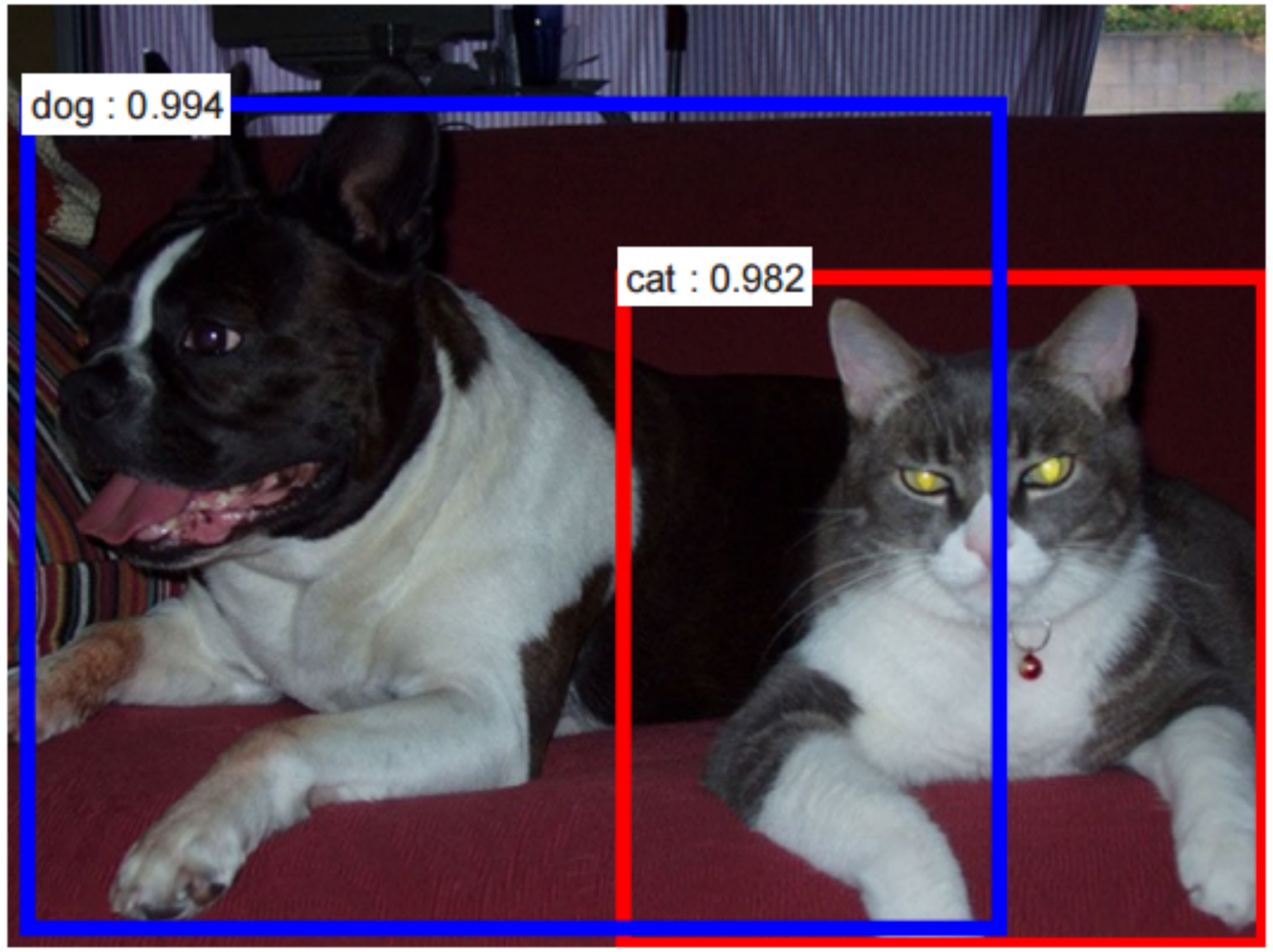
Neural Networks (NNs)



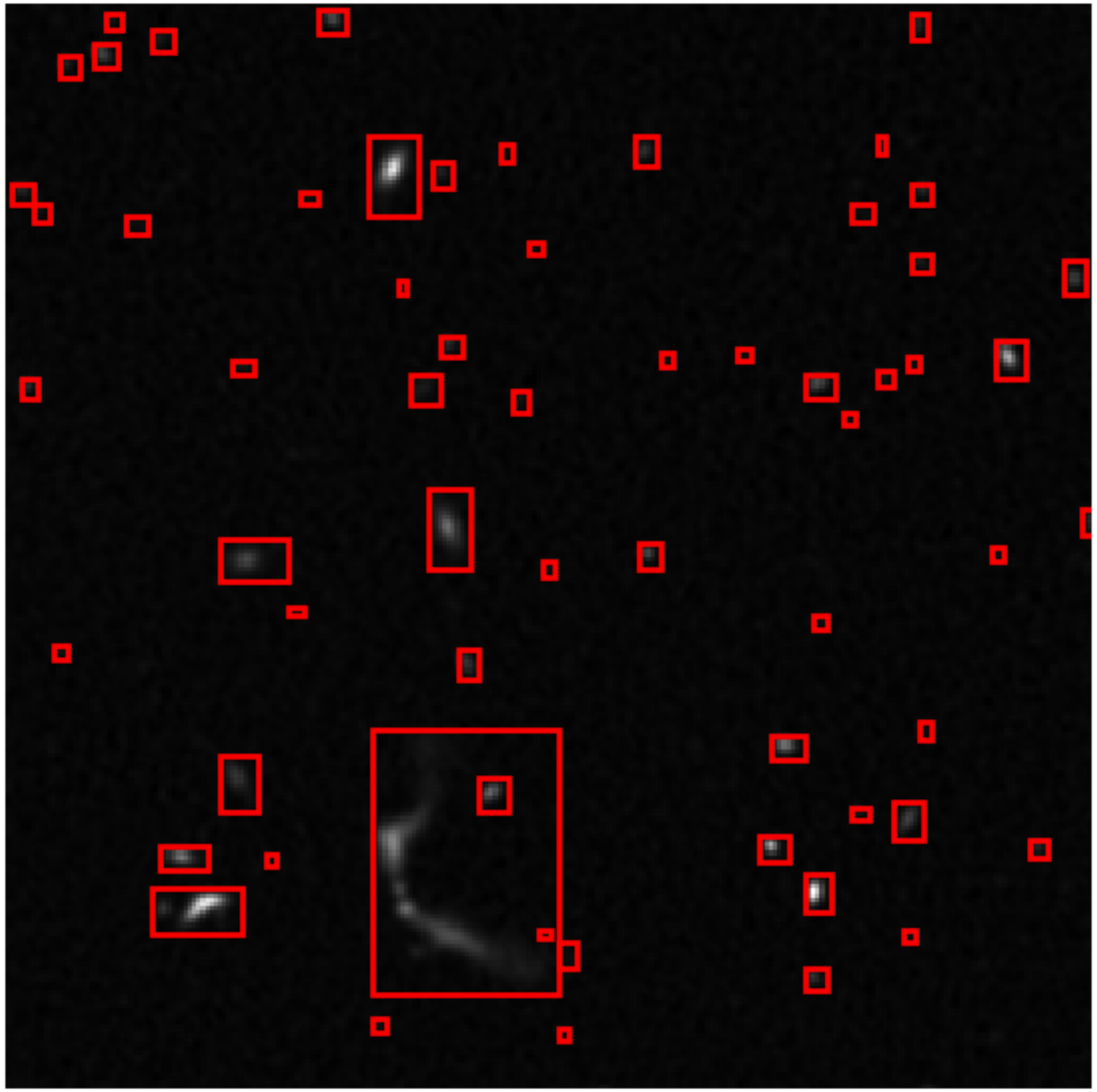


Object detection in images

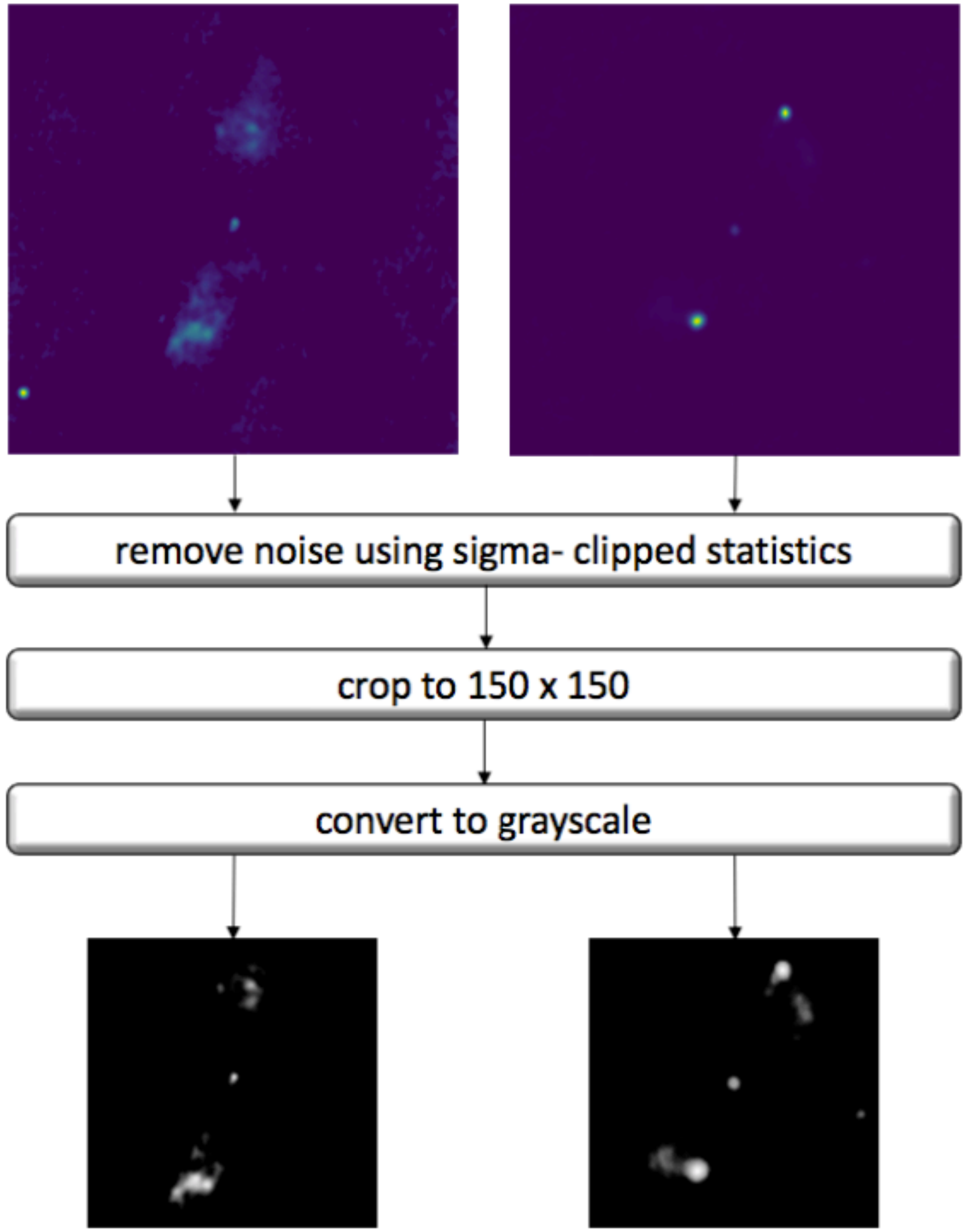




SKA Data challenge #1



Bonaldi+ …. + Zhoolideh +…2020



Morphological classification of compact and extended radio galaxies using convolutional neural networks and data augmentation techniques

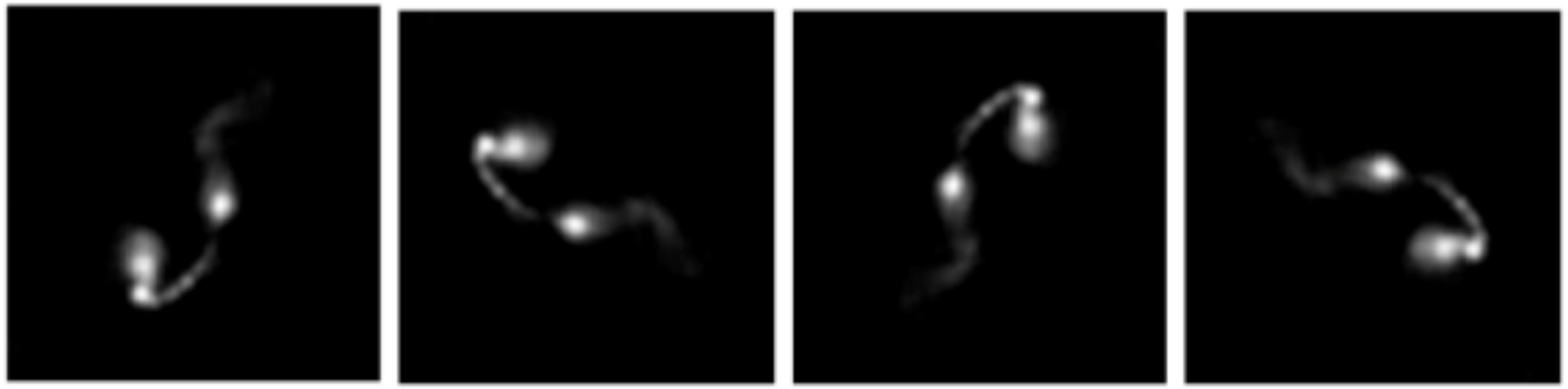
Viera Maslej-Krešňáková[1], Khadija El Bouchefry[2] and Peter Butka[1]⋆
[1]Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Computer Science, Technical University of Košice, Slovakia
[2]South African Radio Astronomy Observatory, Johannesburg, South Africa

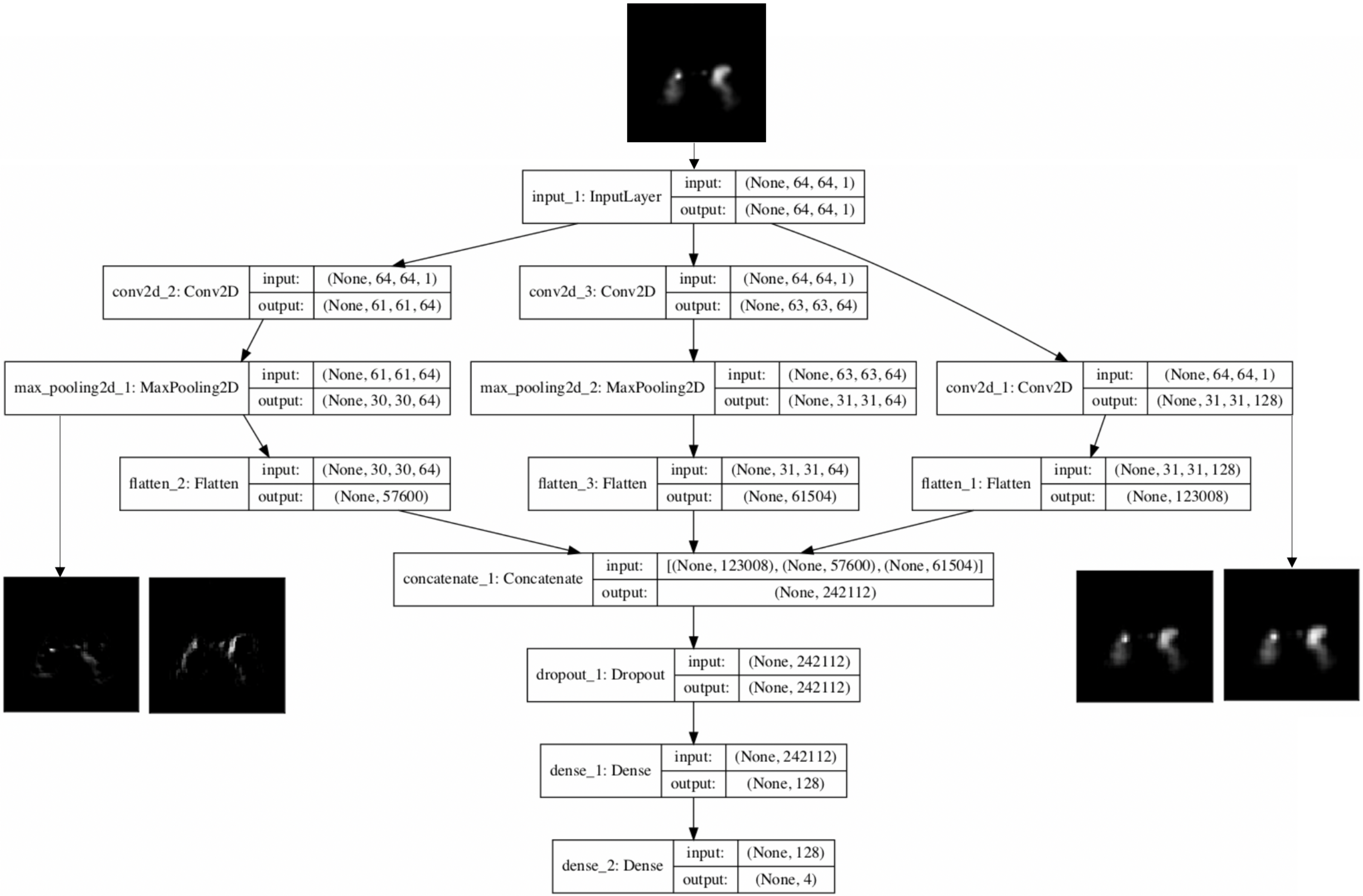
Accepted 2021 May 12. Received 2021 May 11; in original form 2020 December 10

ABSTRACT

Machine learning techniques have been increasingly used in astronomical applications and have proven to successfully classify objects in image data with high accuracy. The current work uses archival data from the Faint Images of the Radio Sky at Twenty Centimeters (FIRST) to classify radio galaxies into four classes: Fanaroff-Riley Class I (FRI), Fanaroff-Riley Class II (FRII), Bent-Tailed (BENT), and Compact (COMPT). The model presented in this work is based on Convolutional Neural Networks (CNNs). The proposed architecture comprises three parallel blocks of convolutional layers combined and processed for final classification by two feed-forward layers. Our model classified selected classes of radio galaxy sources on an independent testing subset with an average of 96% for precision, recall, and F1 score. The best selected augmentation techniques were rotations, horizontal or vertical flips, and increase of brightness. Shifts, zoom and decrease of brightness worsened the performance of the model. The current results show that model developed in this work is able to identify different morphological classes of radio galaxies with a high efficiency and performance.

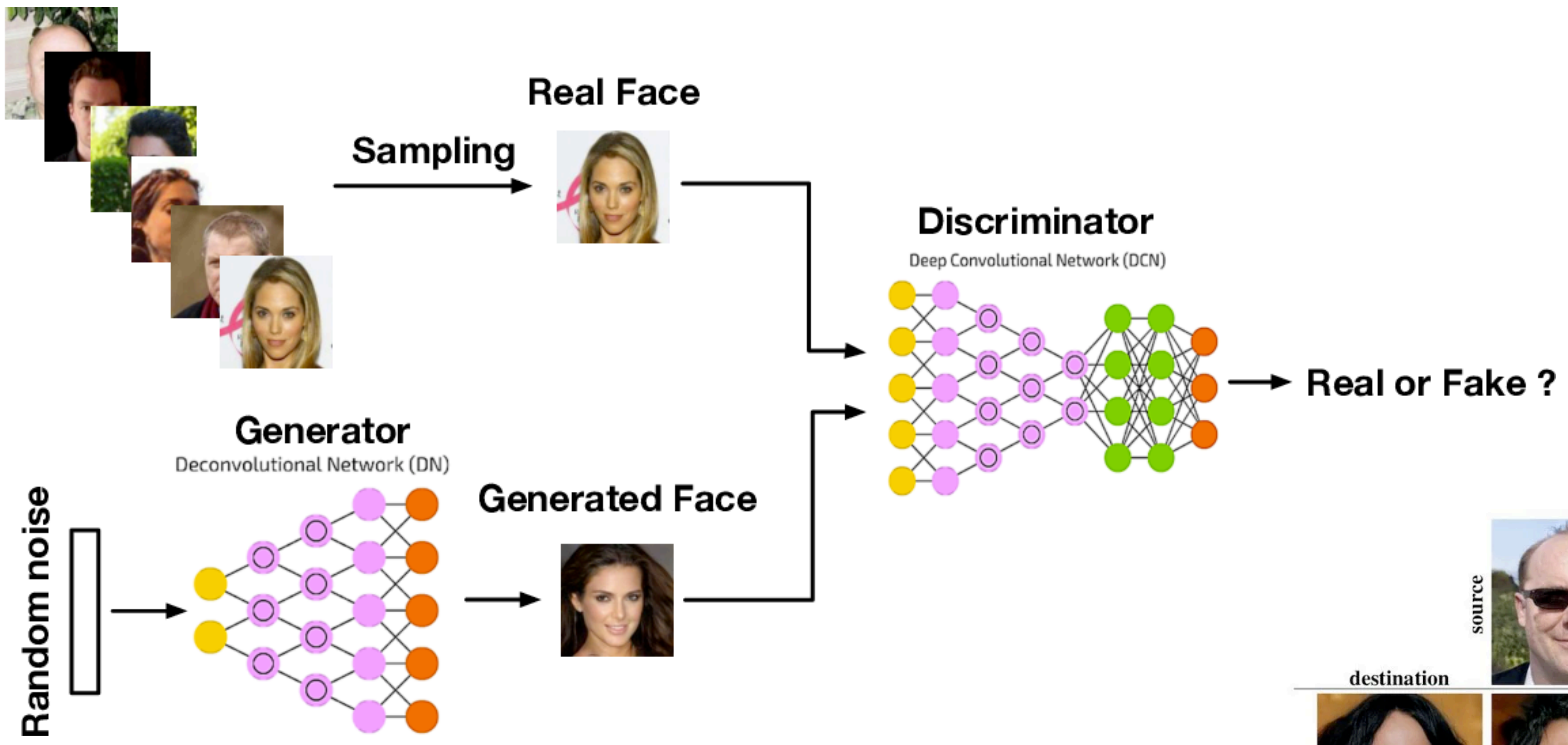
Key words: methods: data analysis – methods: statistical– software: data analysis – radio continuum: galaxies





input_1: InputLayer
input: (None, 64, 64, 1)
output: (None, 64, 64, 1)
conv2d_2: Conv2D
input: (None, 64, 64, 1)
output: (None, 61, 61, 64)
conv2d_3: Conv2D
input: (None, 64, 64, 1)
output: (None, 63, 63, 64)
max_pooling2d_1: MaxPooling2D
input: (None, 61, 61, 64)
output: (None, 30, 30, 64)
max_pooling2d_2: MaxPooling2D
input: (None, 63, 63, 64)
output: (None, 31, 31, 64)
conv2d_1: Conv2D
input: (None, 64, 64, 1)
output: (None, 31, 31, 128)
flatten_2: Flatten
input: (None, 30, 30, 64)
output: (None, 57600)
flatten_3: Flatten
input: (None, 31, 31, 64)
output: (None, 61504)
flatten_1: Flatten
input: (None, 31, 31, 128)
output: (None, 123008)
concatenate_1: Concatenate
input: [(None, 123008), (None, 57600), (None, 61504)]
output: (None, 242112)
dropout_1: Dropout
input: (None, 242112)
output: (None, 242112)
dense_1: Dense
input: (None, 242112)
output: (None, 128)
dense_2: Dense
input: (None, 128)
output: (None, 4)

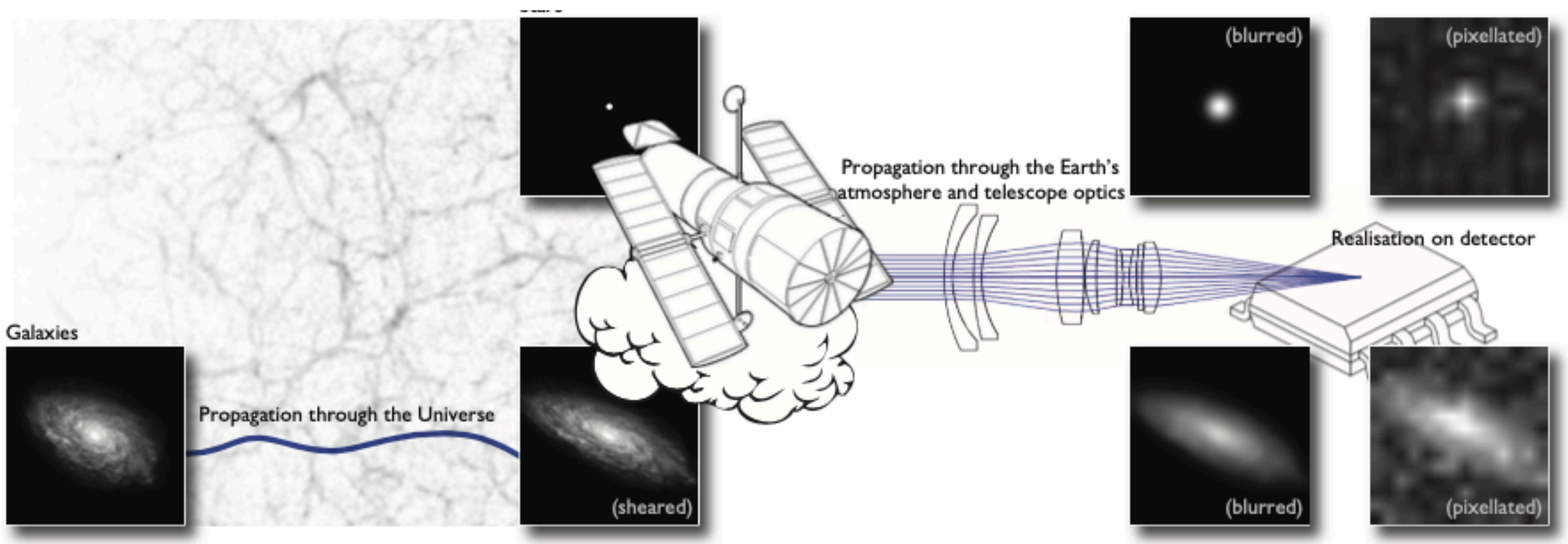
GAN (Generative Adversarial Network)

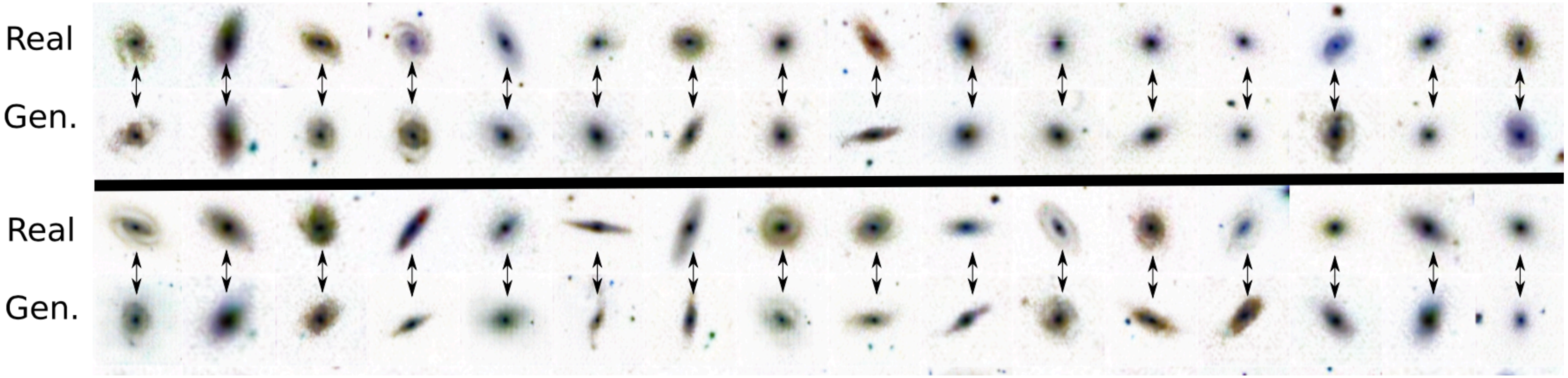


Enabling Dark Energy Science with Deep Generative Models of Galaxy Images

Siamak Ravanbakhsh[1], François Lanusse[2], Rachel Mandelbaum[2], Jeff Schneider[1], and Barnabás Póczos[1]

[1]*School of Computer Science, Carnegie Mellon University*
[2]*McWilliams Center for Cosmology, Carnegie Mellon University*



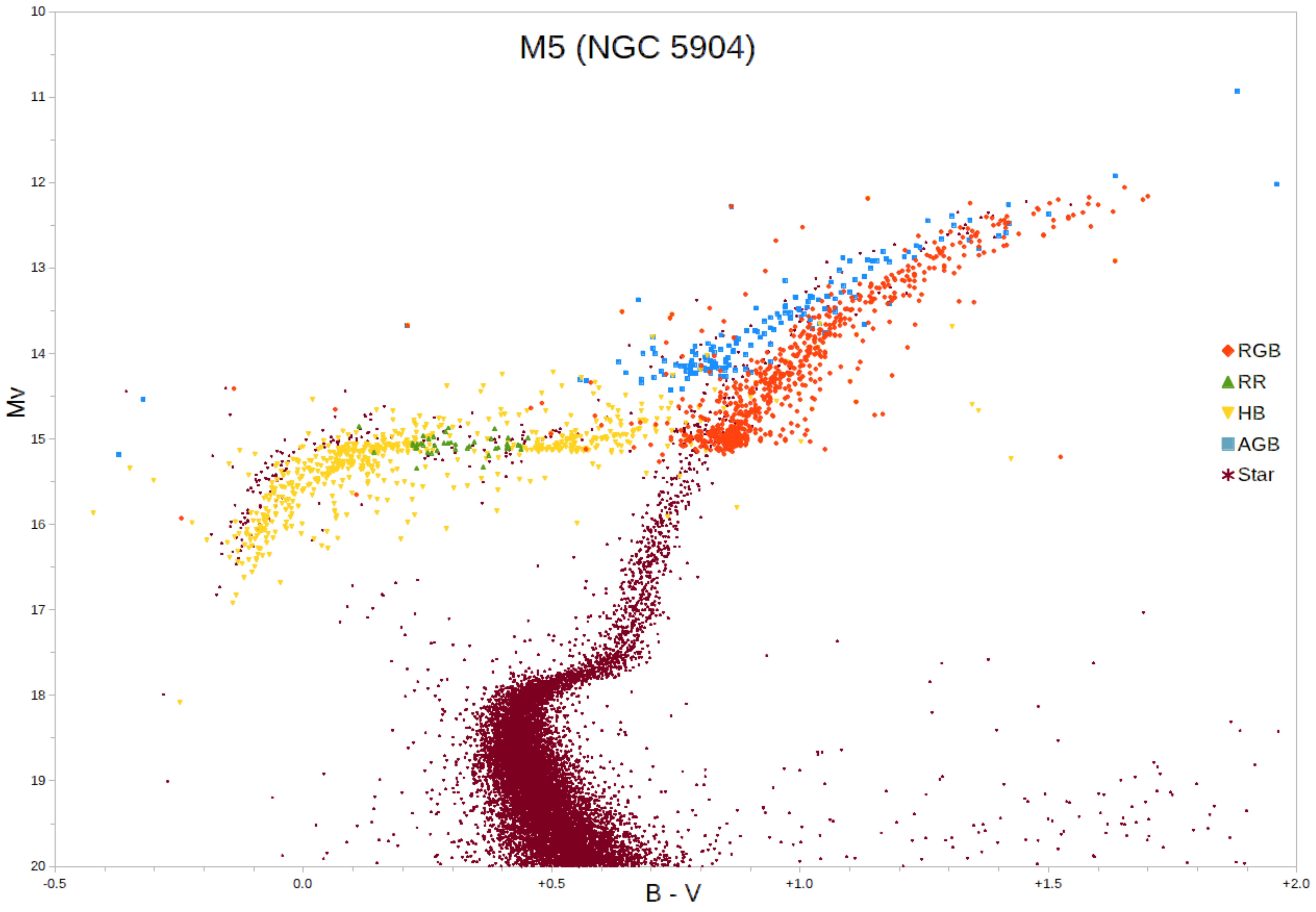


Hands on part: Classifying RR Lyrae variables and main sequence stars

RR Lyrae variables are pulsating horizontal branch stars of spectral class A or F, with a mass of around half the Sun's.

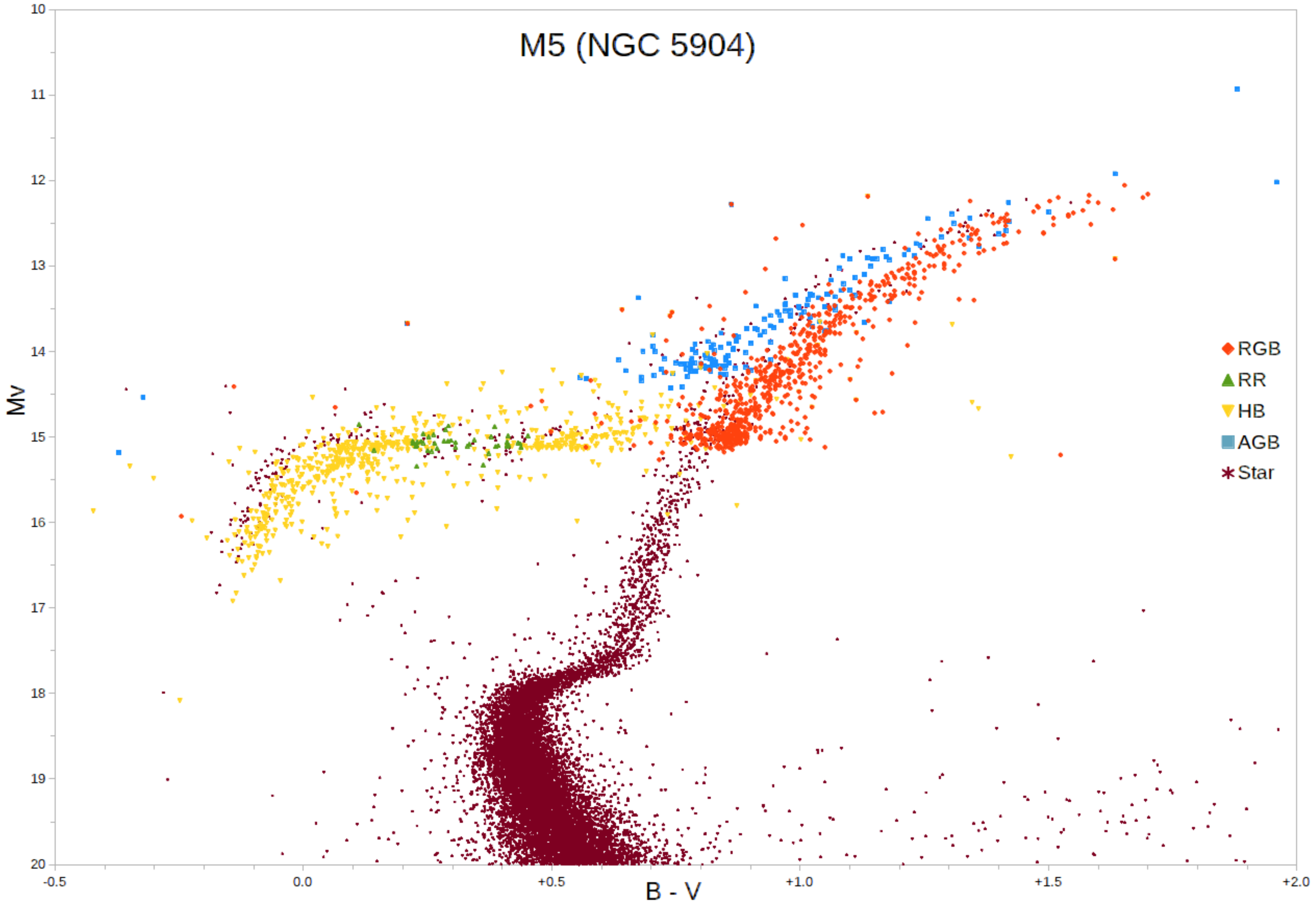
Hands on part: Classifying RR Lyrae variables and main sequence stars

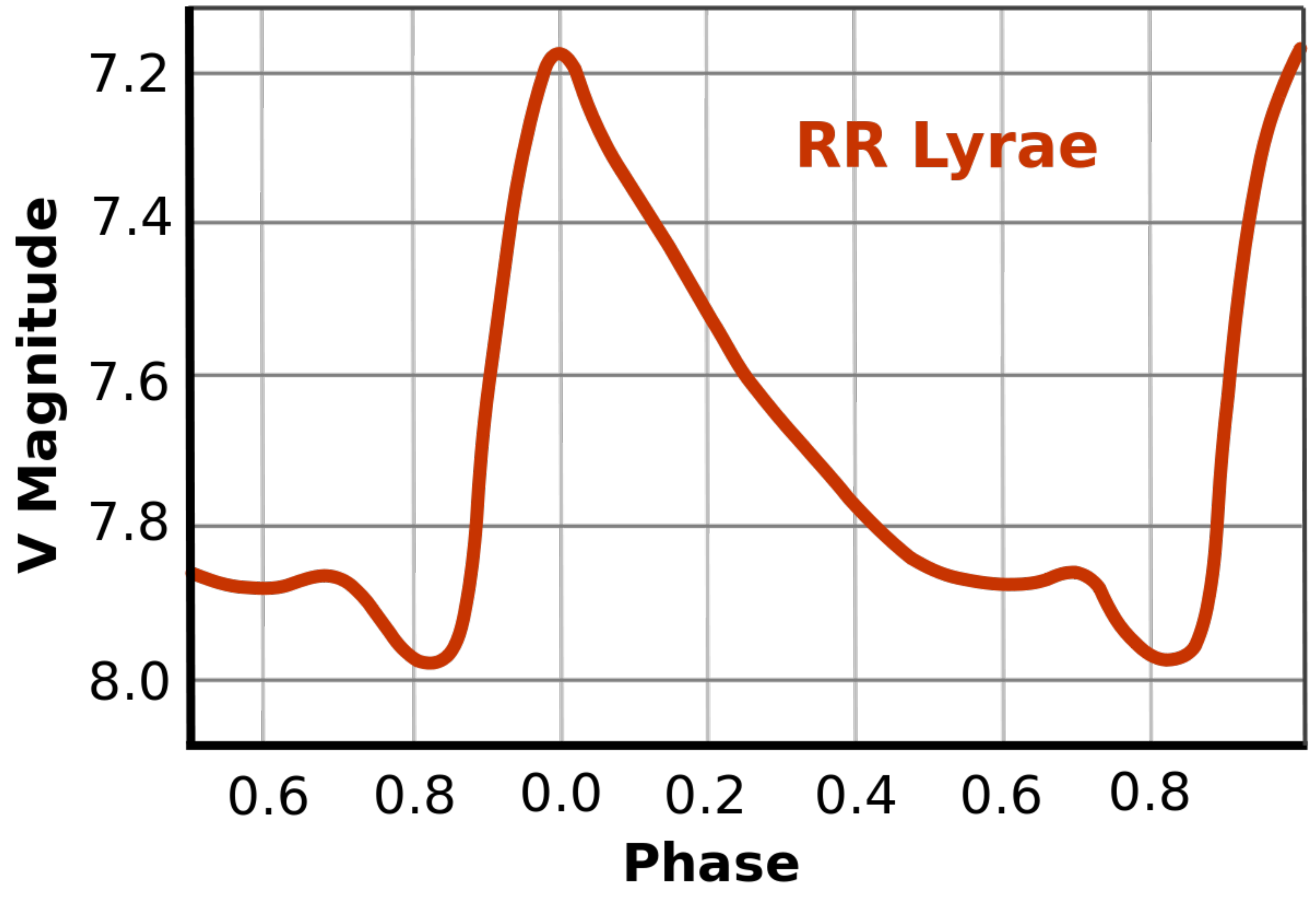
RR Lyrae variables are pulsating horizontal branch stars of spectral class A or F, with a mass of around half the Sun's.



Hands on part: Classifying RR Lyrae variables and main sequence stars

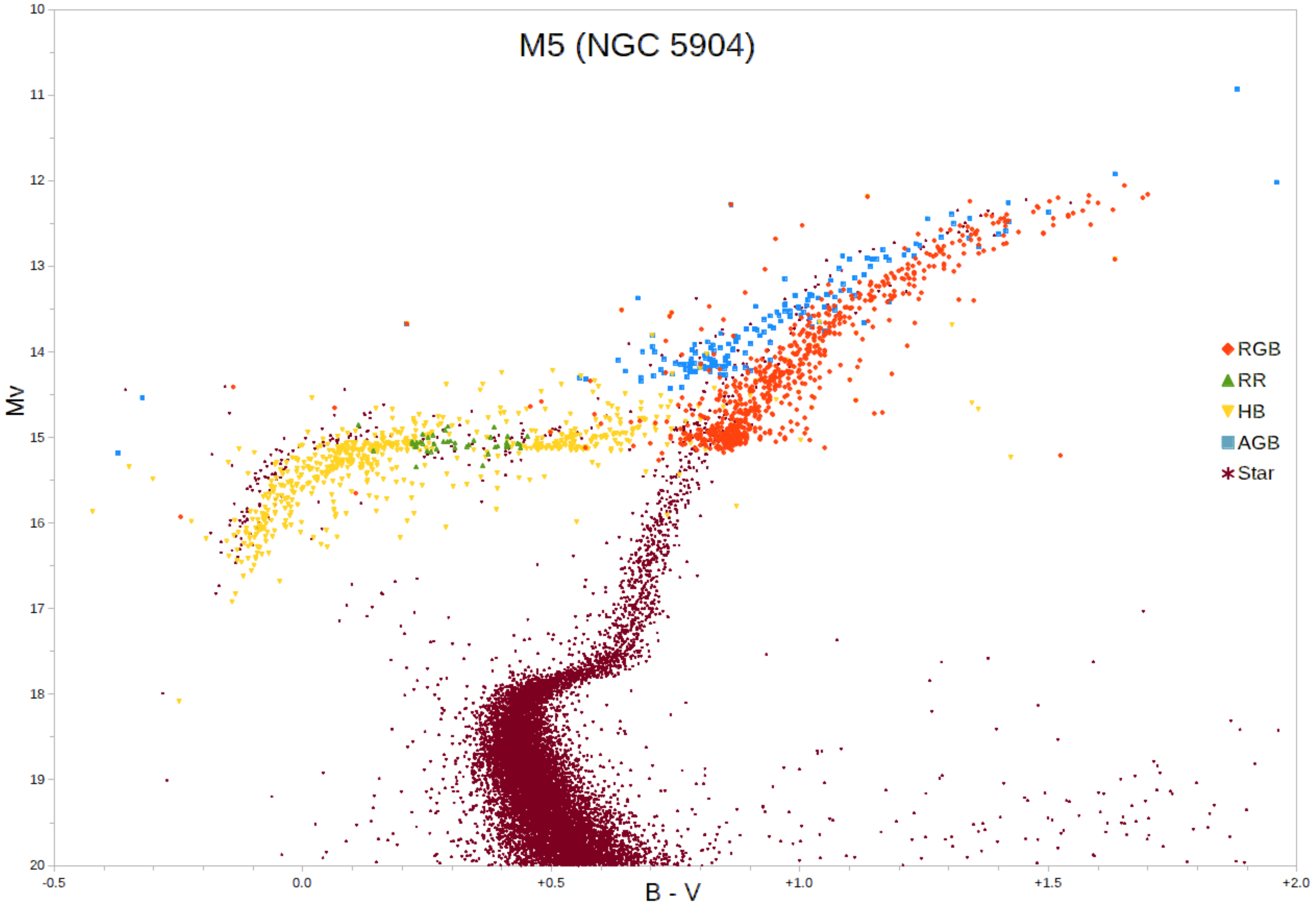
RR Lyrae variables are pulsating horizontal branch stars of spectral class A or F, with a mass of around half the Sun's.

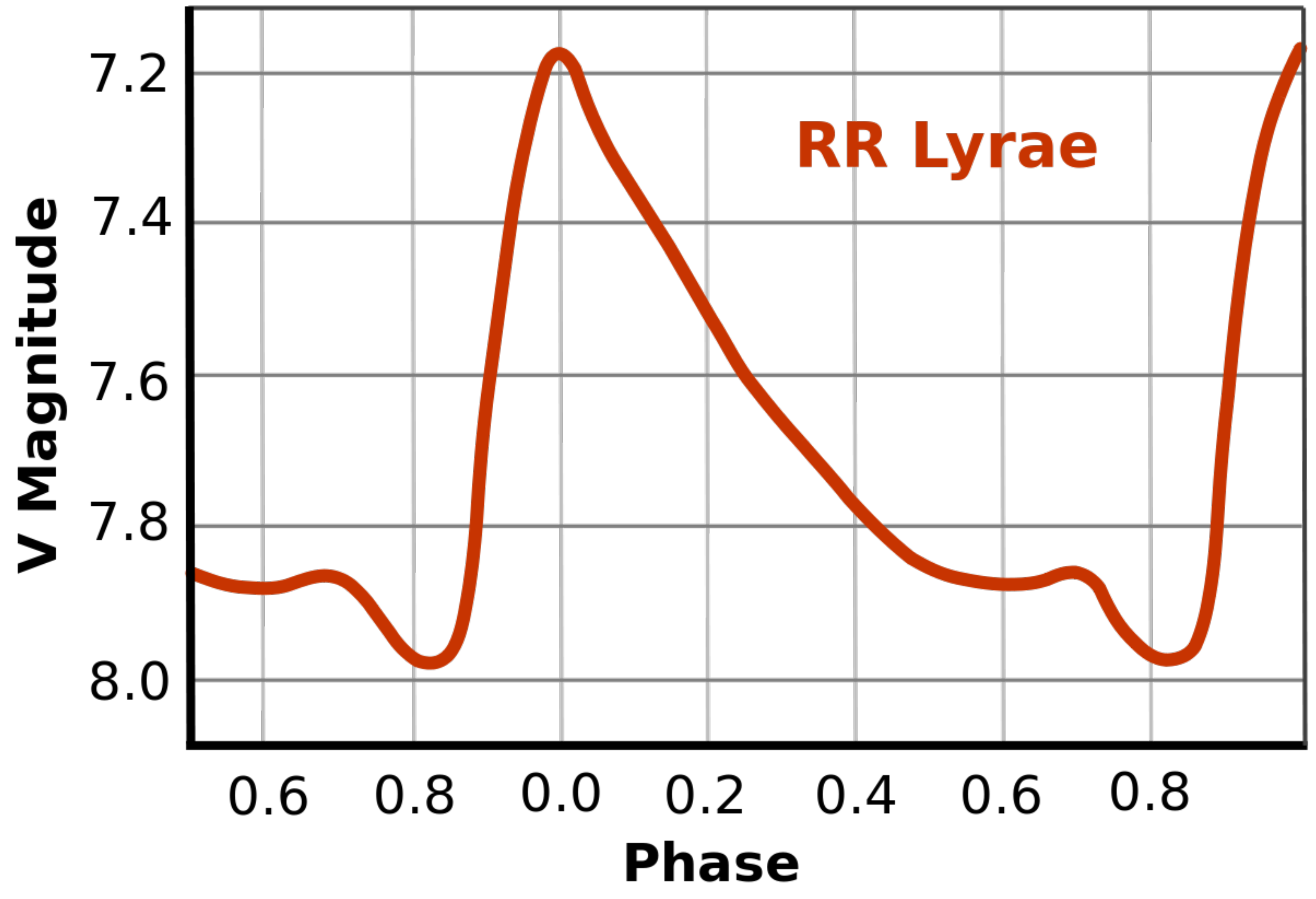


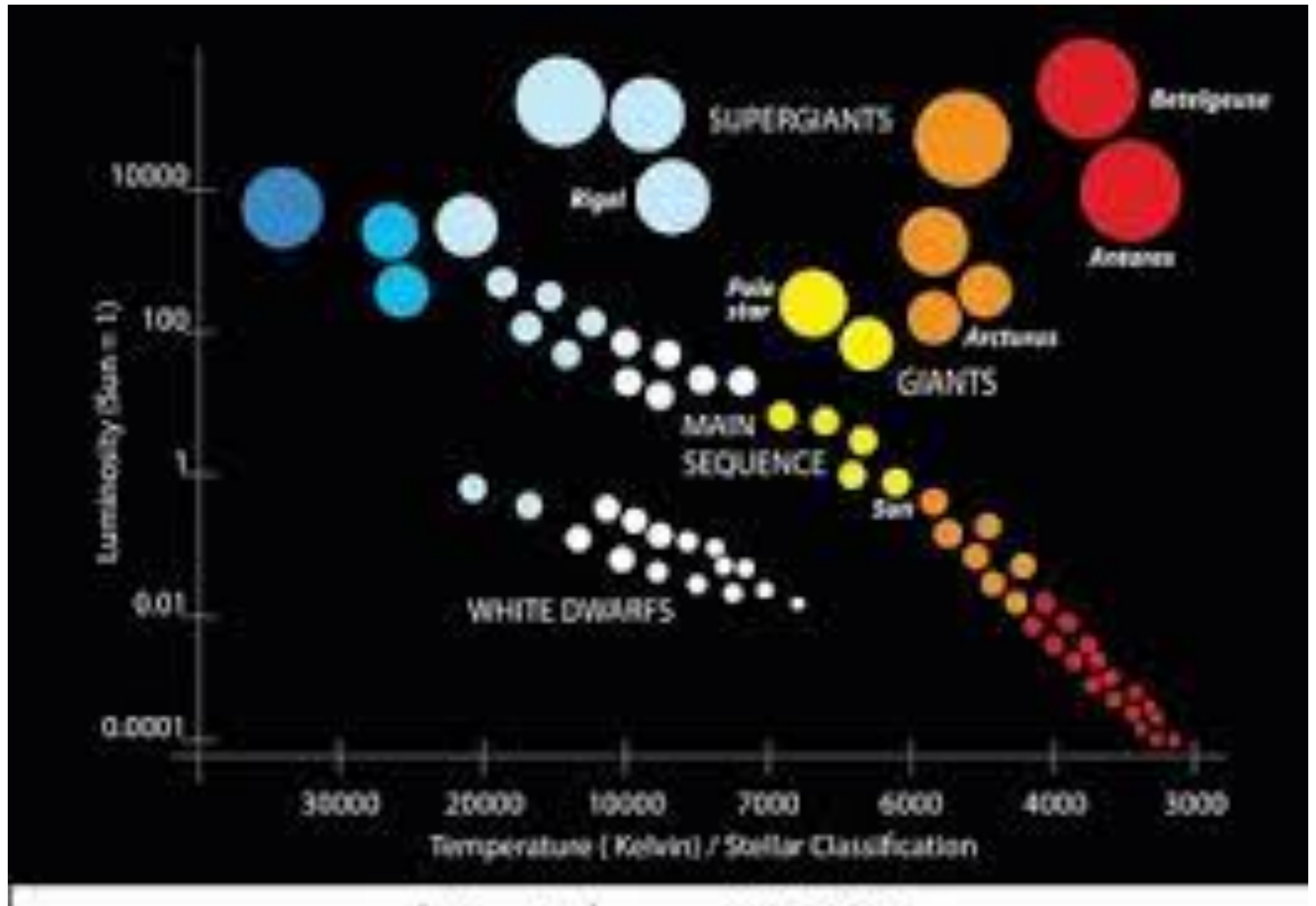


Hands on part: Classifying RR Lyrae variables and main sequence stars

RR Lyrae variables are pulsating horizontal branch stars of spectral class A or F, with a mass of around half the Sun's.







Predicted Label		Real Label: Positive	Real Label: Negative
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$