Introduction to Unsupervised Learning

Dalya Baron (Tel Aviv University)

XXX Winter School, November 2018

What is the difference between supervised and unsupervised learning?

What is the difference between supervised and unsupervised learning?

figure of merit

What is the difference between supervised and unsupervised learning?

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.

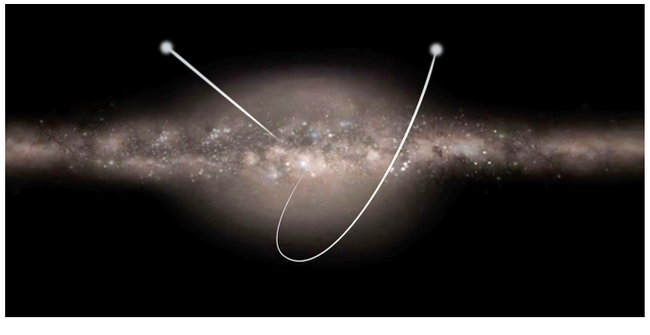
The algorithm detects clusters, complex relations, outliers, or reduces the dimensions of the dataset.

Prior knowledge isn't required!

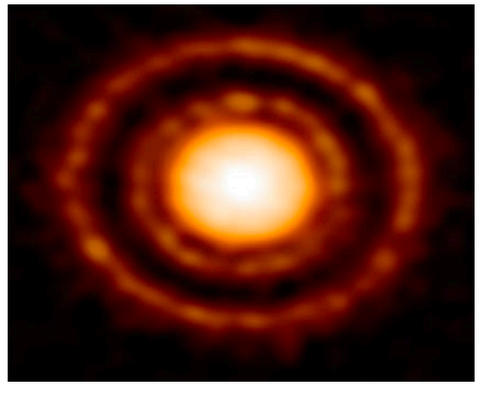
Machine-assisted Knowledge Discovery

Hubble Deep Field

Gaia and runaway stars (ESA)

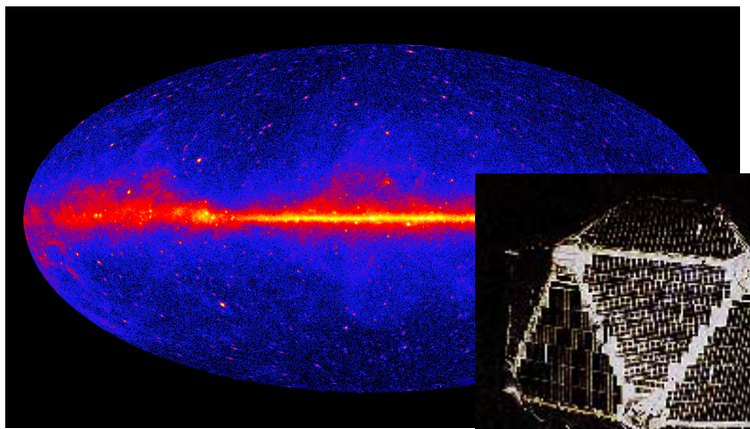


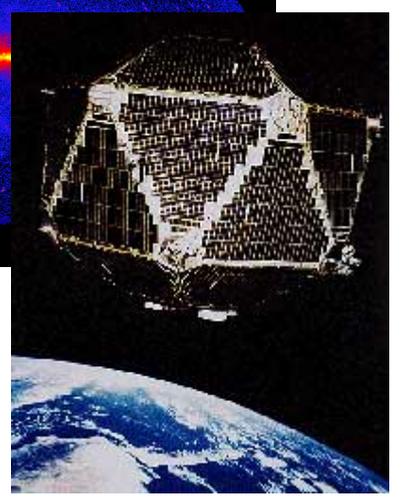
ALMA view of protoplanetary disk, D. Fedele et al.



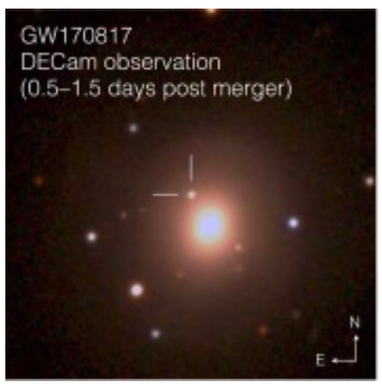
New technology allows us to make new discoveries

The gamma-ray sky by Fermi





The Vela 5B satellite in low-Earth orbit (LANL)



Soares-Santos et al. 2017

Abbott et al. 2017

Anatomy of unsupervised learning algorithms

$$f(\overrightarrow{X}, \{a_1, a_2, ...\}) = \overrightarrow{y}$$

Anatomy of unsupervised learning algorithms

$$f(\overrightarrow{X}, \{a_1, a_2, ...\}) = \overrightarrow{y}$$

Input dataset:

- Raw data (spectra, images, light-curves).
- Extracted features.
- Measured relations between different objects (distances, correlations).

Anatomy of unsupervised learning algorithms

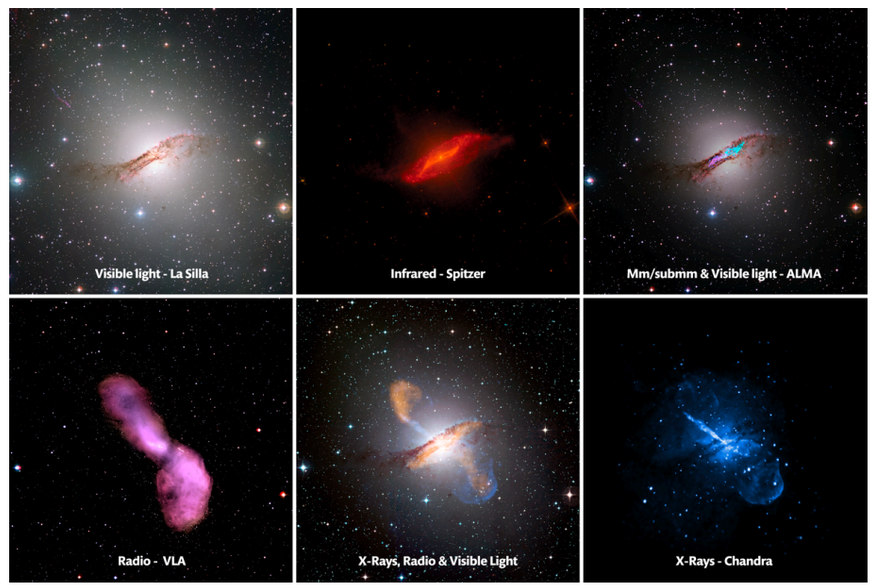
$$f(\overrightarrow{X}, \{a_1, a_2, ...\}) = \overrightarrow{y}$$

Input dataset:

- Raw data (spectra, images, light-curves).
- Extracted features.
- Measured relations between different objects (distances, correlations).

Hyperparameters

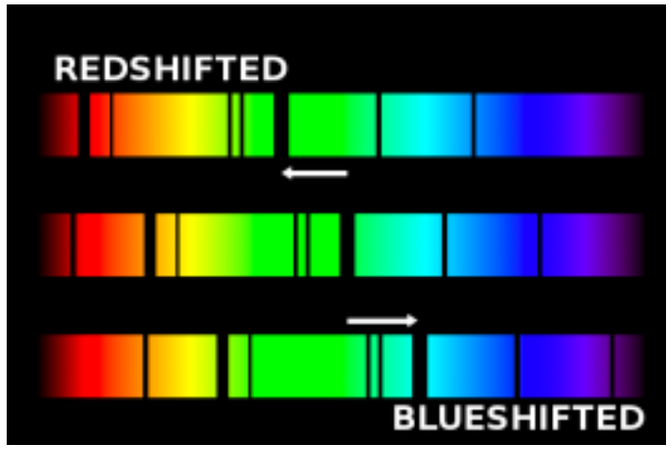
- Tuning parameters of the algorithm.
- Can strongly affect the result.
- Traditionally, cannot be optimized for.



Anatomy of unsupervised learning algorithms

Internal choices and/or internal cost function

- Usually, we cannot control these.
- Strongly affect the result, and define the range of possible outputs.



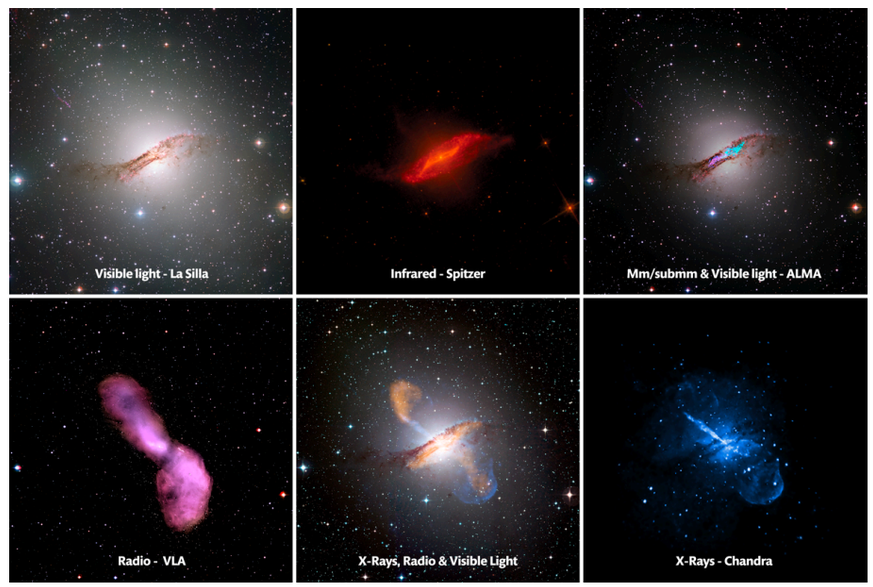
$$f(\vec{X}, \{a_1, a_2, \ldots\}) = \vec{y}$$

Input dataset:

- Raw data (spectra, images, light-curves).
- Extracted features.
- Measured relations between different objects (distances, correlations).

Hyperparameters

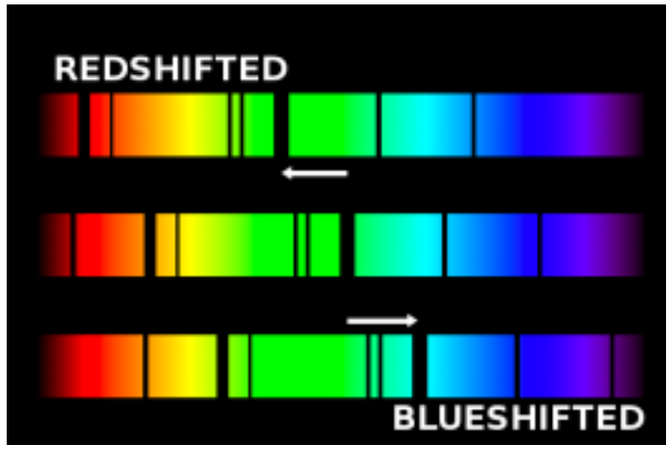
- Tuning parameters of the algorithm.
- Can strongly affect the result.
- Traditionally, cannot be optimized for.



Anatomy of unsupervised learning algorithms

Internal choices and/or internal cost function

- Usually, we cannot control these.
- Strongly affect the result, and define the range of possible outputs.



Algorithm output: clusters, high-dimensional relations, outliers, sparse representation.

Our goal:
exploration or **inference**

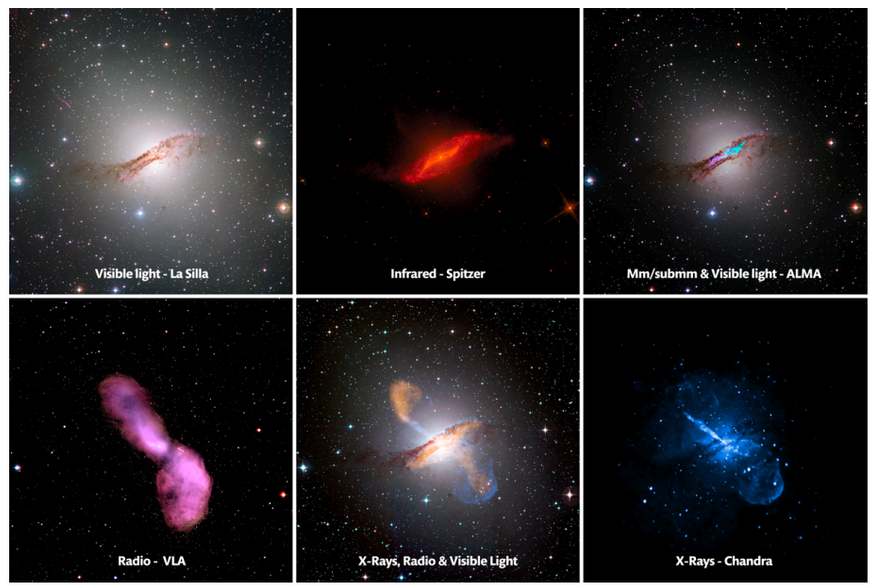
$$f(\vec{X}, \{a_1, a_2, ...\}) = \vec{y}$$

Input dataset:

- Raw data (spectra, images, light-curves).
- Extracted features.
- Measured relations between different objects (distances, correlations).

Hyperparameters

- Tuning parameters of the algorithm.
- Can strongly affect the result.
- Traditionally, cannot be optimized for.



Good Practices

- **Start simple:**
  - Simulate simple low-dimensional dataset, without noise, where the output can be anticipated.
  - Compare the output of the algorithm for different data representations and different choices of hyper-parameters.
- **Gradually complicate the model:**
  - Add more dimensions (some of them should be uninformative).
  - Add noise.
  - Compare the output for different representations and hyper-parameters.
- **Physically-motivated model:**
  - Simulate a physically-motivated dataset.
  - Experiment with different noise properties, different representations, and hyper-parameters.
- **Try to break the algorithm.**

Topics

Clustering algorithms:

- K-means
- Hierarchical Clustering
- Gaussian Mixture model

Ensemble methods:

- Supervised decision trees and random forests
- Unsupervised random forests

Dimensionality Reduction algorithms:

- PCA and variants
- tSNE and UMAP
- Autoencoders and SOM

Outlier detection

Advanced topics and summary

Tutorials

Questions?