An Introduction to Machine Learning in Astronomy

Dalya Baron
Tel Aviv University

LIneA Webinar, June 2021

The Data Revolution

THE SPECTRA OF NARROW-LINE SEYFERT 1 GALAXIES[1]

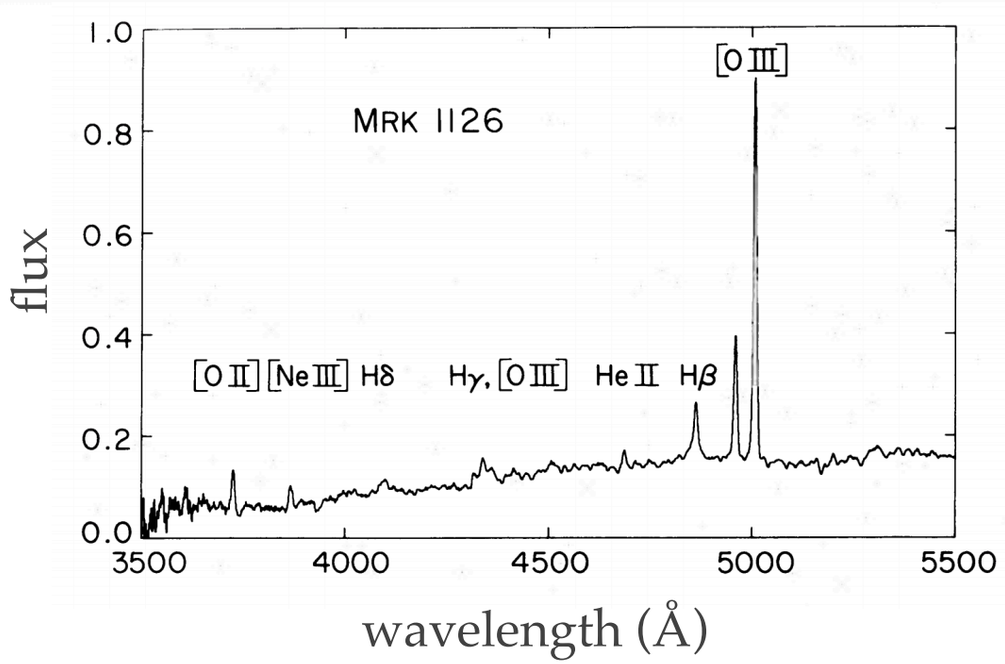
DONALD E. OSTERBROCK AND RICHARD W. POGGE
Lick Observatory, Board of Studies in Astronomy and Astrophysics, University of California, Santa Cruz
Received 1985 January 2; accepted 1985 April 9

ABSTRACT

Measurements are presented of a group of active galactic nuclei with all the properties of Seyfert 1 or 1.5 galaxies, but with unusually narrow H I lines. They include Mrk 42, 359, and 1239 (previously studied by other authors) as well as Mrk 493, 766, 783, and 1126. One other somewhat similar object, Mrk 1388, is also included in the discussion; measurements of its spectrum have been published elsewhere. For these objects, narrow-line widths, relative intensities of the emission lines, etc., are all similar to those in other Seyfert 1 galaxies. Some, in particular Mrk 493 and Mrk 42, have relatively strong Fe II emission; in others, especially Mrk 359, 783, and 1126, it is quite weak.

As a group, these narrow-line Seyfert 1 galaxies have approximately normal luminosities. Their Hβ emission-line equivalent widths are, on the average, somewhat smaller than in typical Seyfert 1's. Overall, these narrow-line Seyfert 1 galaxies show a wide variety of deviations from the properties of typical Seyfert 1 objects. They clearly demonstrate that the Seyfert phenomenon is not a simple one-parameter effect.

Subject headings: galaxies: nuclei — galaxies: Seyfert



The Data Revolution

THE SPECTRA OF NARROW-LINE SEYFERT 1 GALAXIES[1]

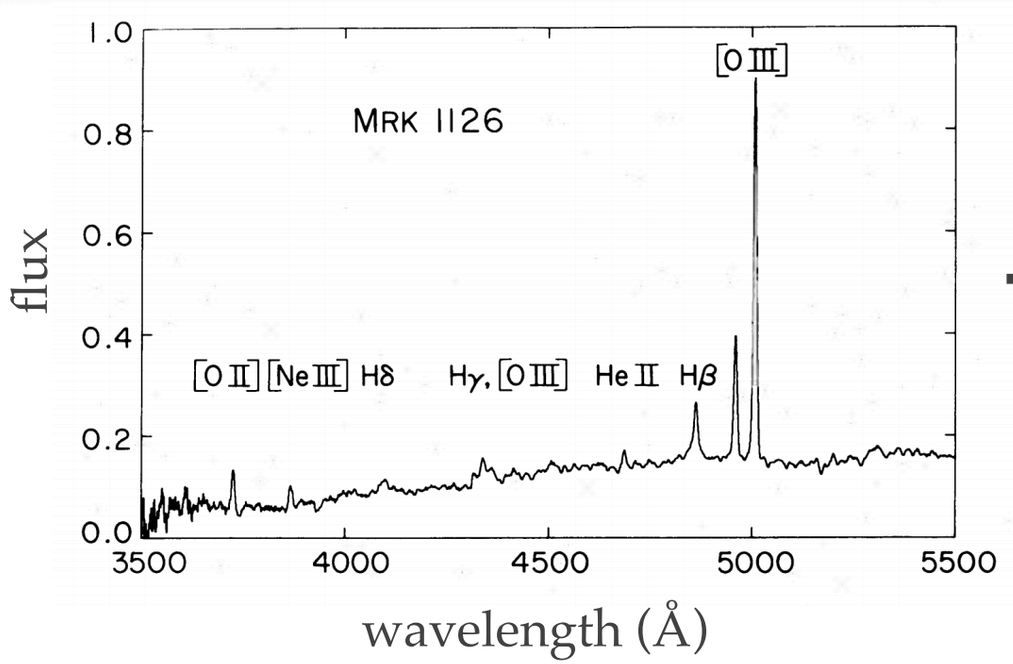
DONALD E. OSTERBROCK AND RICHARD W. POGGE
Lick Observatory, Board of Studies in Astronomy and Astrophysics, University of California, Santa Cruz
Received 1985 January 2; accepted 1985 April 9

ABSTRACT

Measurements are presented of a group of active galactic nuclei with all the properties of Seyfert 1 or 1.5 galaxies, but with unusually narrow H I lines. They include Mrk 42, 359, and 1239 (previously studied by other authors) as well as Mrk 493, 766, 783, and 1126. One other somewhat similar object, Mrk 1388, is also included in the discussion; measurements of its spectrum have been published elsewhere. For these objects, narrow-line widths, relative intensities of the emission lines, etc., are all similar to those in other Seyfert 1 galaxies. Some, in particular Mrk 493 and Mrk 42, have relatively strong Fe II emission; in others, especially Mrk 359, 783, and 1126, it is quite weak.

As a group, these narrow-line Seyfert 1 galaxies have approximately normal luminosities. Their Hβ emission-line equivalent widths are, on the average, somewhat smaller than in typical Seyfert 1's. Overall, these narrow-line Seyfert 1 galaxies show a wide variety of deviations from the properties of typical Seyfert 1 objects. They clearly demonstrate that the Seyfert phenomenon is not a simple one-parameter effect.

Subject headings: galaxies: nuclei — galaxies: Seyfert



- **Redshift.**
- **Stellar population:** velocity and age.
- **Gas properties:** velocity, dust content, metal content, density, temperature, ionization state, and ionizing source.

The Data Revolution

The volume and rate of information grows exponentially: sky surveys now generate ~PBs of data + derived products.

- **SDSS:** multi-band images of ~1/3 of the sky, and spectra of millions of galaxies and stars.
- **PAN-STARRS and ZTF:** multi-band time series data of numerous variable stars, supernovae, AGN, and more.
- **Gaia:** positions and velocities of ~1 billion stars in the Milky Way.

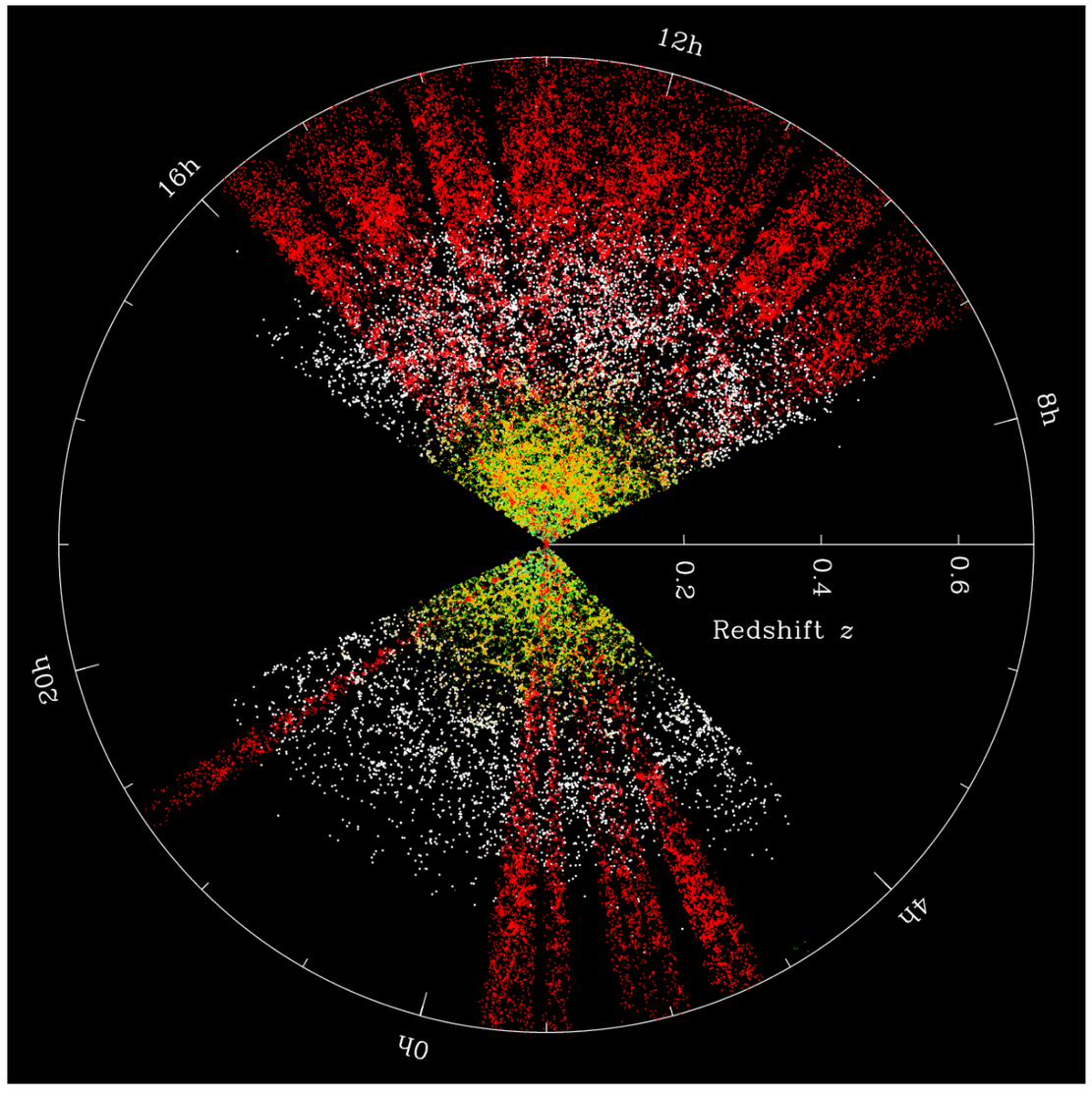


image credit: Michael Blanton and the SDSS-III Collaboration

The Data Revolution

There is a great increase in data dimensionality and complexity.

- Different telescopes and instruments map astronomical objects throughout the entire electromagnetic spectrum.
- Time domain surveys now provide multi-band time series of astronomical objects.
- Integral Field Units (IFU) provide spatially-resolved spectroscopic information.

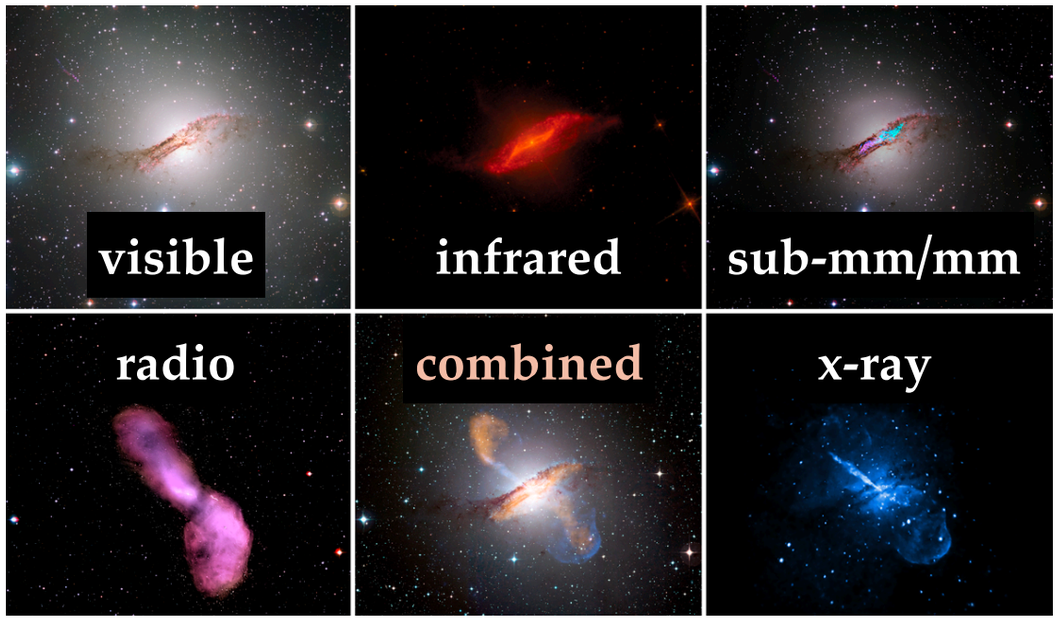


image from: www.almaobservatory.org

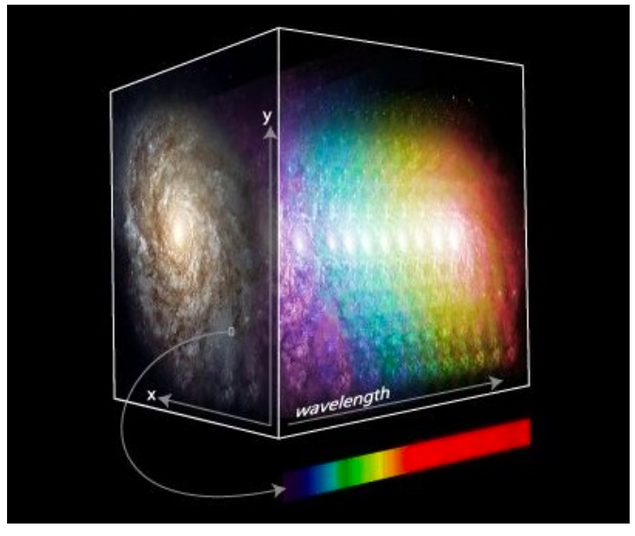
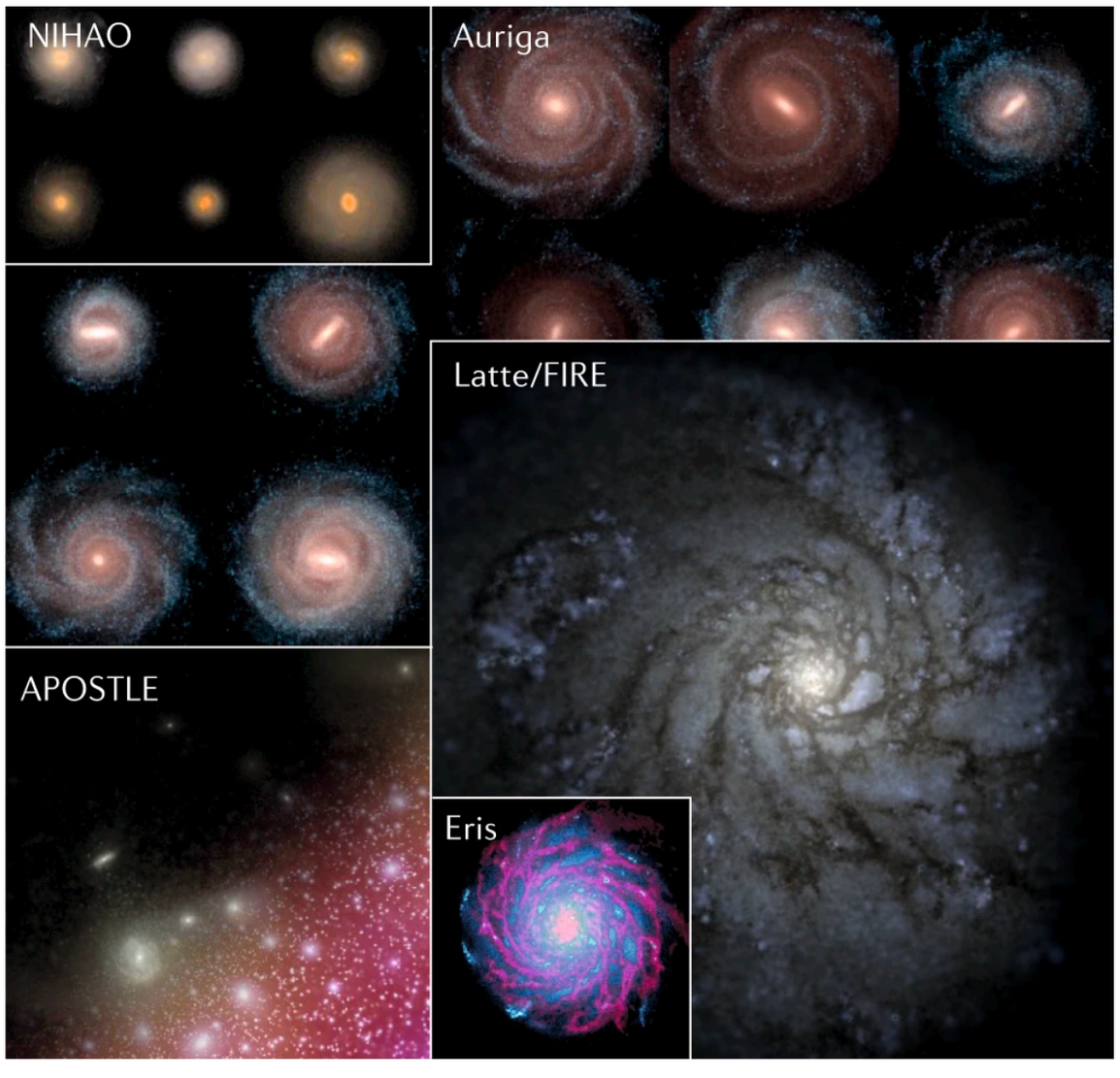


image credit: Marc White (RSAA-ANU)

The Data Revolution

We no longer collect information only from our Universe: simulations now produce rich and complex datasets and provide us with mock observations.



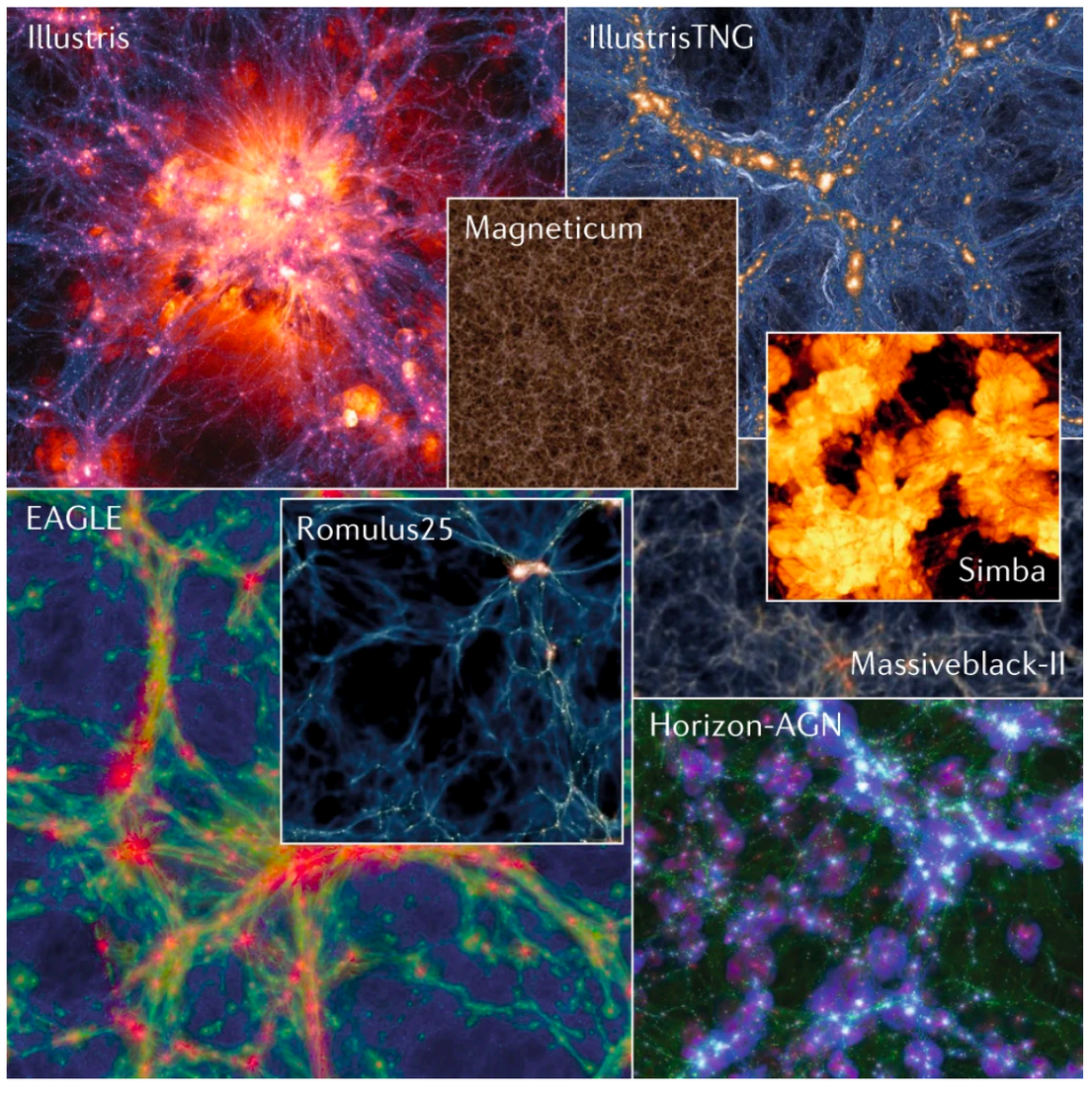
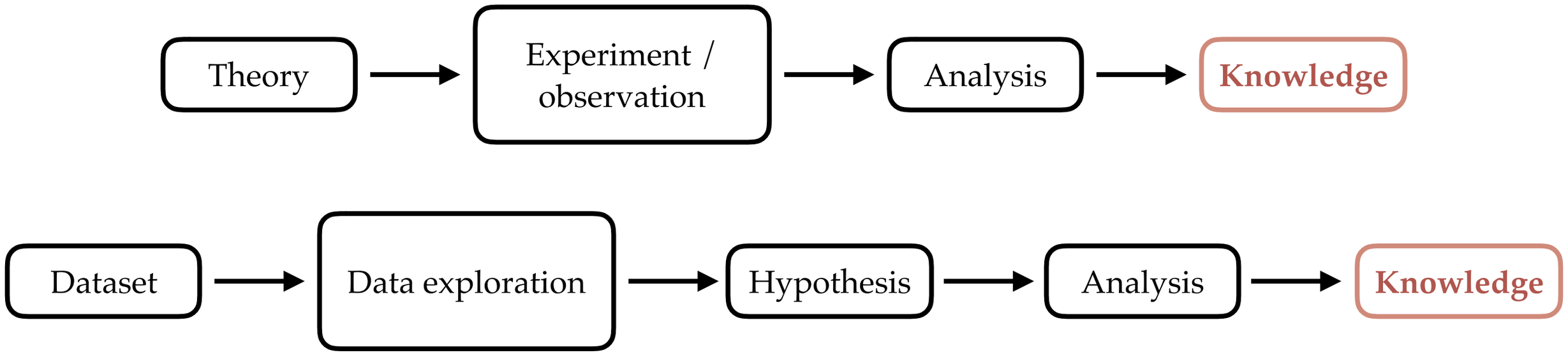


image credit: Vogelsberger+ (2020)

The Data Revolution

What is fundamentally different now?

- We can no longer inspect all of the data by eye.
- Patterns and correlations can no longer be visualized.
- The increase in the data information content gives rise to a new scientific methodology:

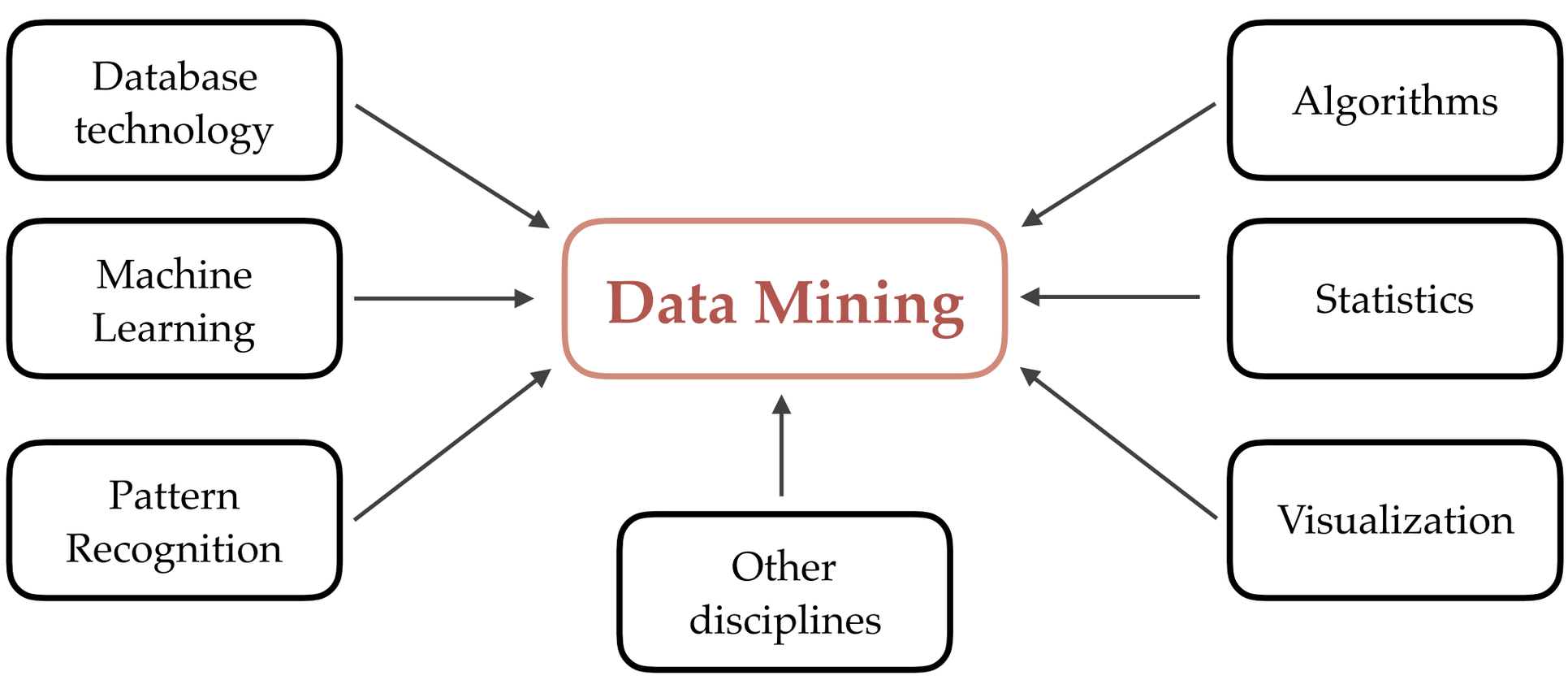


- Large-scale or high-resolution simulations are very expensive to re-run.

Challenges in the Big Data Era

Challenges in the Big Data Era:

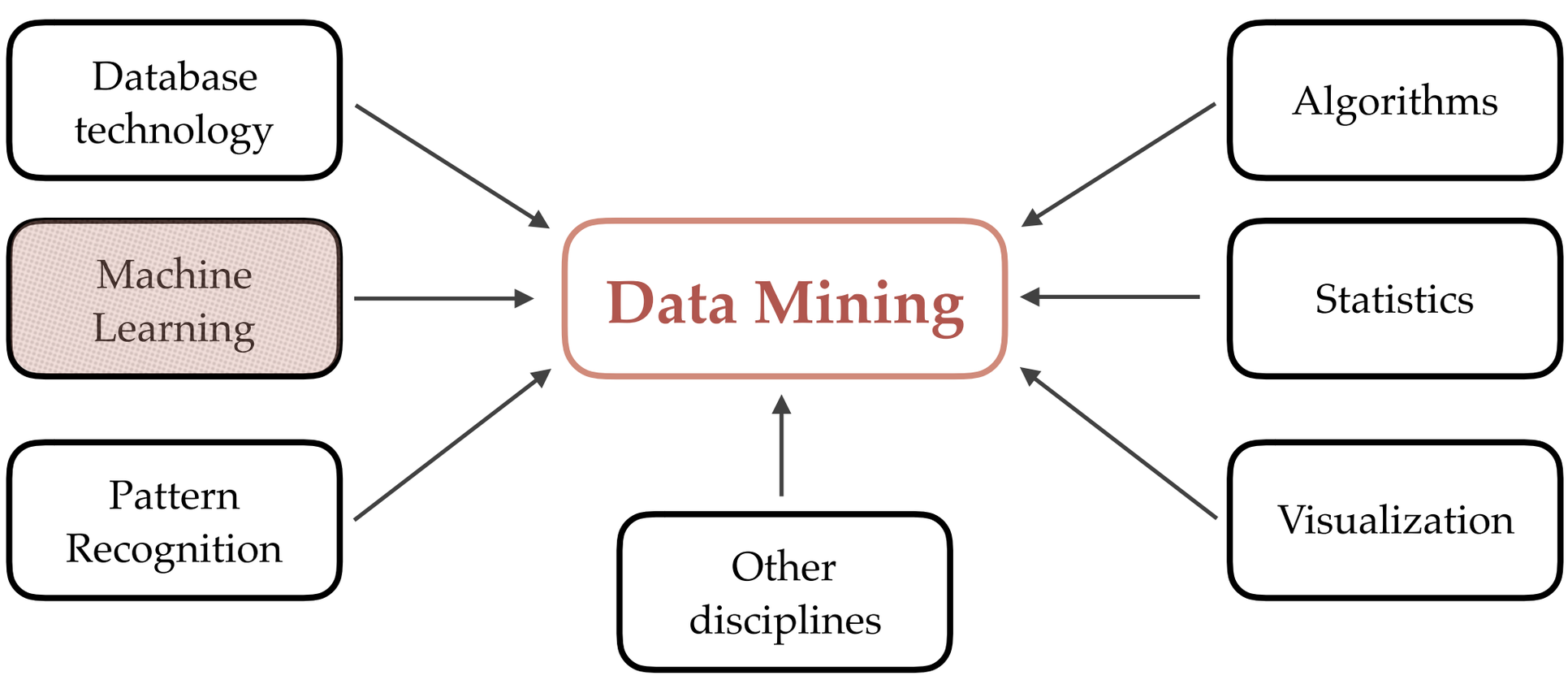
- Cleaning, sorting, and storing the data.
- Classification and regression.
- Generation of new data.
- Discoveries of new phenomena.



Machine Learning as a tool in the astronomers toolkit

Challenges in the Big Data Era:

- Cleaning, sorting, and storing the data.
- Classification and regression.
- Generation of new data.
- Discoveries of new phenomena.



Machine Learning as a tool in the astronomers toolkit

Machine Learning in Astronomy: a practical overview

Dalya Baron

Astronomy is experiencing a rapid growth in data size and complexity. This change fosters the development of data-driven science as a useful companion to the common model-driven data analysis paradigm, where astronomers develop automatic tools to mine datasets and extract novel information from them. In recent years, machine learning algorithms have become increasingly popular among astronomers, and are now used for a wide variety of tasks. In light of these developments, and the promise and challenges associated with them, the IAC Winter School 2018 focused on big data in Astronomy, with a particular emphasis on machine learning and deep learning techniques. This document summarizes the topics of supervised and unsupervised learning algorithms presented during the school, and provides practical information on the application of such tools to astronomical datasets. In this document I cover basic topics in supervised machine learning, including selection and preprocessing of the input dataset, evaluation methods, and three popular supervised learning algorithms, Support Vector Machines, Random Forests, and shallow Artificial Neural Networks. My main focus is on unsupervised machine learning algorithms, that are used to perform cluster analysis, dimensionality reducti
Unsupervised learning algorithms are of particular importance to scientific research, since they can be used to extract new know
facilitate new discoveries.

Arxiv: 1904.07248

Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

Christopher J. Fluke, Colin Jacobs

Machine learning (automated processes that learn by example in order to classify, predict, discover or generate new data) and artificial intelligence (methods by which a computer makes decisions or discoveries that would usually require human intelligence) are now firmly established in astronomy. Every week, new applications of machine learning and artificial intelligence are added to a growing corpus of work. Random forests, support vector machines, and neural networks (artificial, deep, and convolutional) are now having a genuine impact for applications as diverse as discovering extrasolar planets, transient objects, quasars, and gravitationally-lensed systems, forecasting solar activity, and distinguishing between signals and instrumental effects in gravitational wave astronomy. This review surveys contemporary, published literature on machine learning and artificial intelligence in astronomy and astrophysics. Applications span seven main categories of activity: classification, regression, clustering, forecasting, generation, discovery, and the development of new scientific insight. These categories form the b
use of machine learning and artificial intelligence emerges, progresses or becomes established.

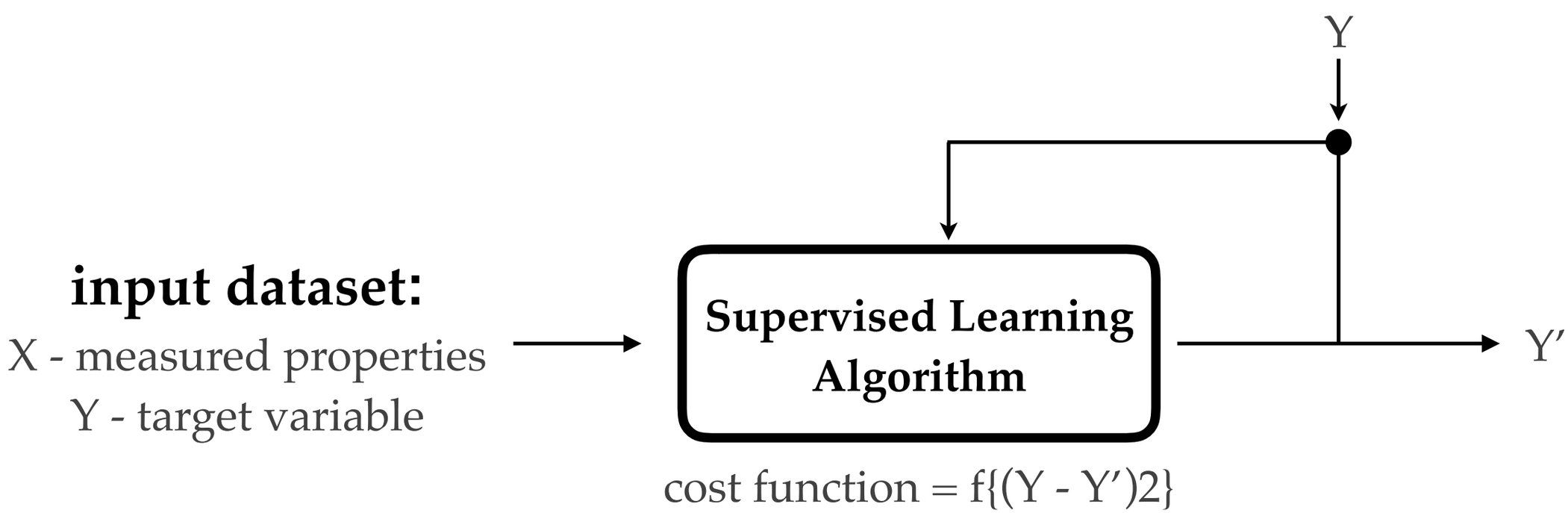
Arxiv: 1912.02934

The two flavors of Machine Learning

Supervised Machine Learning: given a list of objects with measured properties and a target variable, train an algorithm to predict the target variable of previously-unseen examples.

Common tasks: classification and regression.

Popular algorithms: Support Vector Machine (SVM), Decision Trees (DT) and Random Forest (RF), and artificial neural networks (shallow and deep).

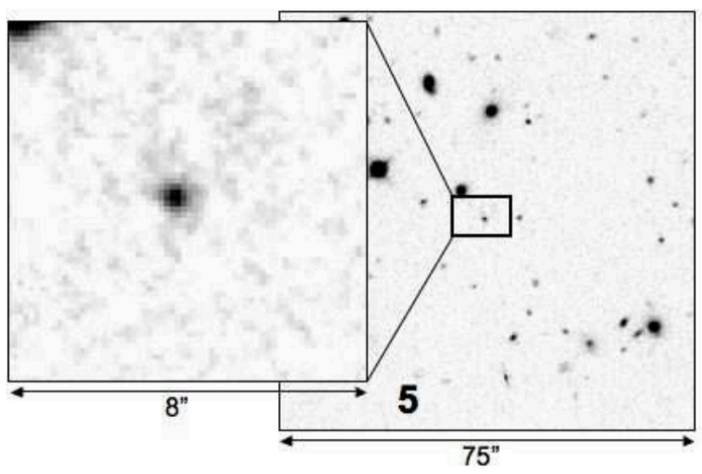


Supervised Learning: examples

Common tasks: **classification** and regression.

Popular algorithms: **Support Vector Machine (SVM)**, Decision Trees (DT) and Random Forest (RF), and artificial neural networks (shallow and deep).

A set of distant galaxies images



A set of measured features:
concentration, asymmetry, smoothness, , moment of Light, Gini Coefficient, etc…

SVM

morphological classification:
late type or early type galaxy

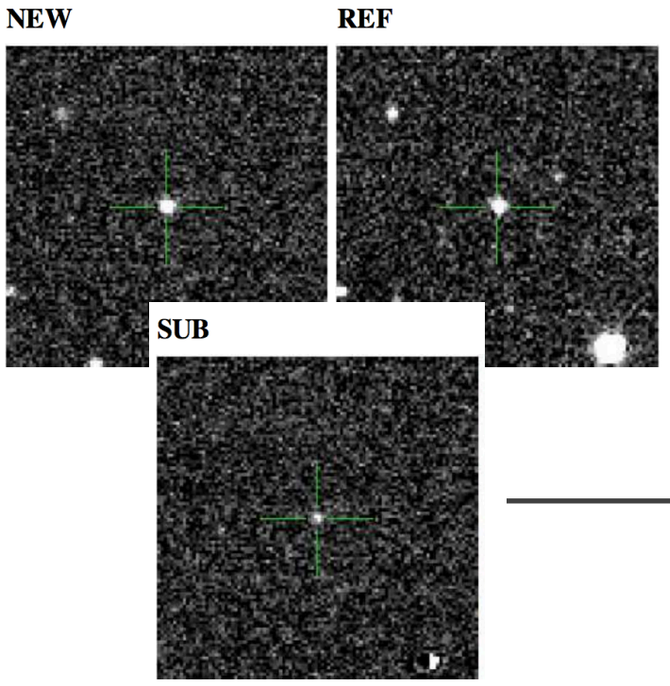
Morphological classification of galaxies by Huertas-Company+ (2008).

Supervised Learning: examples

Common tasks: **classification** and regression.

Popular algorithms: Support Vector Machine (SVM), Decision Trees (DT) and **Random Forest (RF)**, and artificial neural networks (shallow and deep).

A set of images taken at different nights:



A set of features measured from the difference image

RF

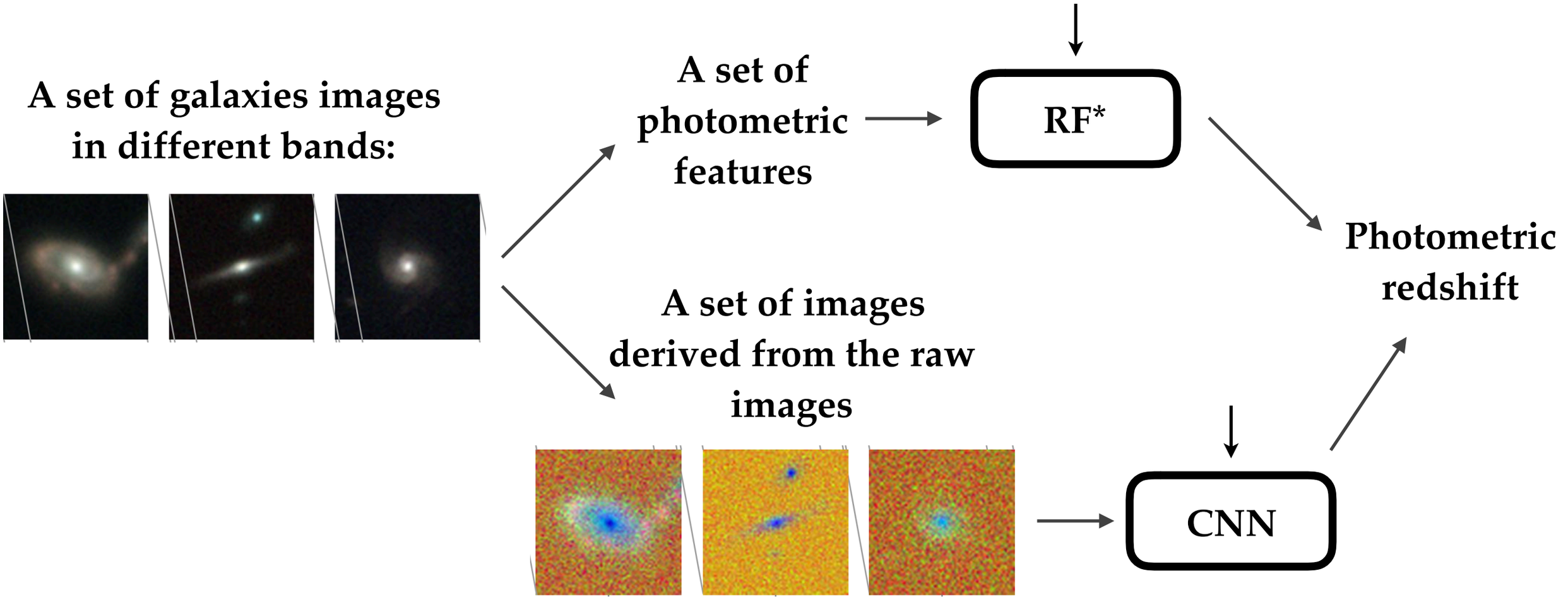
Candidate classification: real variable source or an artifact

Real/bogus classification in the Palomar Transient Factory by Bloom et al. (2012).

Supervised Learning: examples

Common tasks: classification and **regression**.

Popular algorithms: Support Vector Machine (SVM), Decision Trees (DT) and **Random Forest (RF)**, and **artificial neural networks** (shallow and deep).



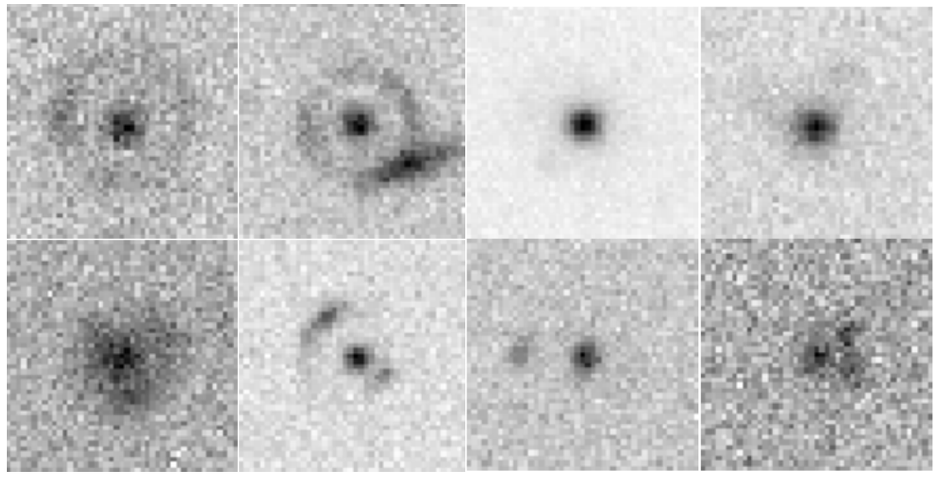
Measuring photometric redshifts by Hoyle (2016).

Supervised Learning: examples

Common tasks: **classification** and regression.

Popular algorithms: Support Vector Machine (SVM), Decision Trees (DT) and Random Forest (RF), and **artificial neural networks** (shallow and deep).

A set of simulated galaxy images



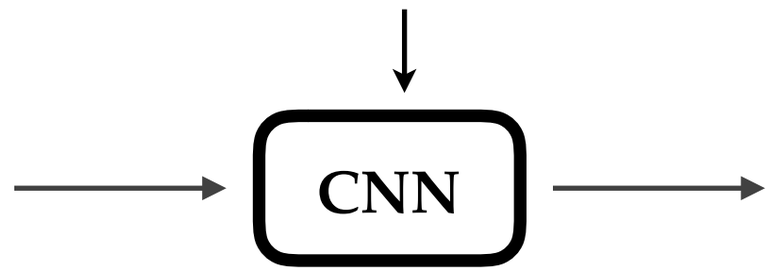


Image classification: contains strong lensing signal or not.

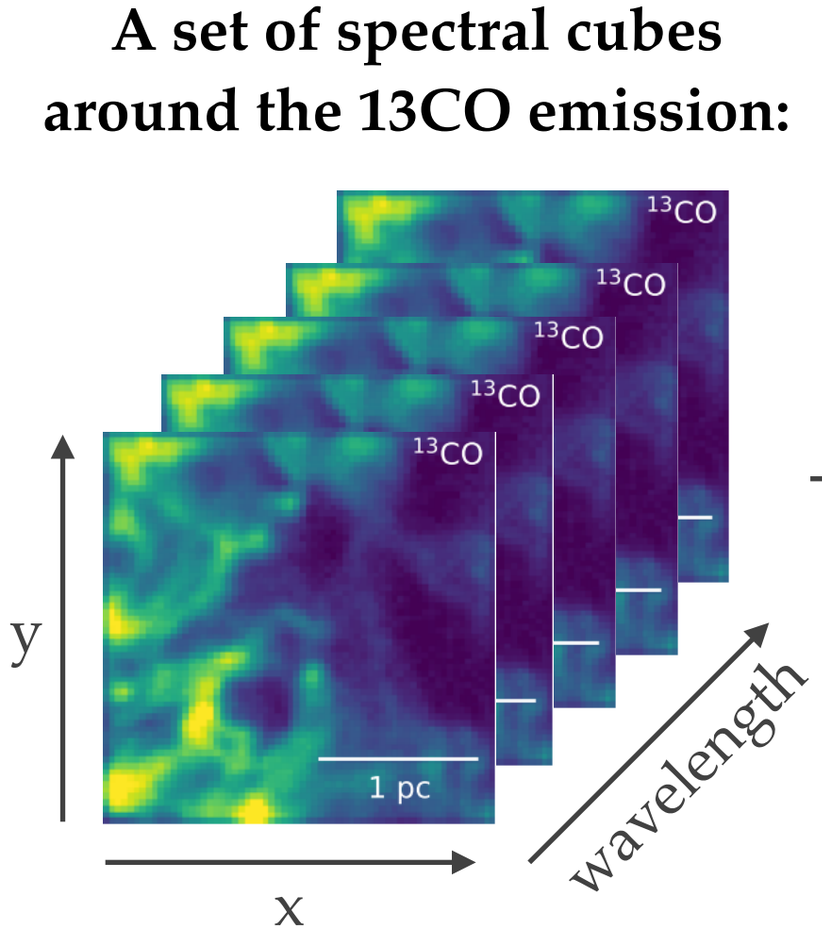
Detection of galaxy-galaxy strong gravitational lenses by Lanusse et al. 2017.

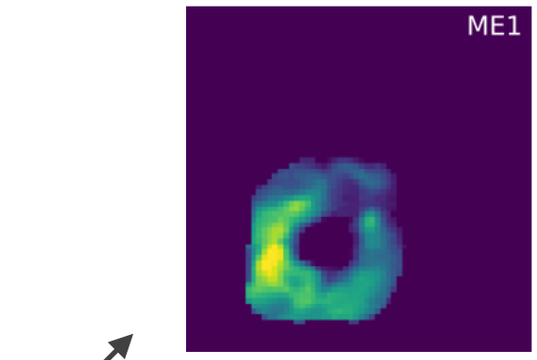
Supervised Learning: examples

Common tasks: **classification** and **regression**.

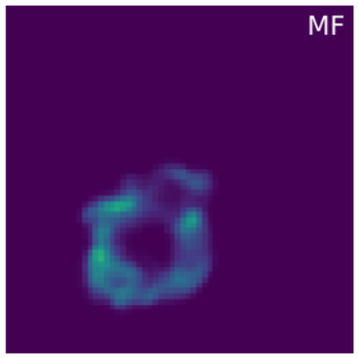
Popular algorithms: Support Vector Machine (SVM), Decision Trees (DT) and Random Forest (RF), and **artificial neural networks** (shallow and deep).

A set of spectral cubes around the 13CO emission:





Pixel classification: position of stellar feedback features



Pixel regression: fraction of the mass coming from feedback

Identification of stellar feedback bubbles in CO emission by Xu et al. (2020).

Which algorithm should we use?

	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000

Which algorithm should we use?

	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000
model complexity	low	mediocare	mediocare	complex

Which algorithm should we use?

	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000
model complexity	low	mediocare	mediocare	complex
number of free parameters	~4	~10	~4	>10

Which algorithm should we use?

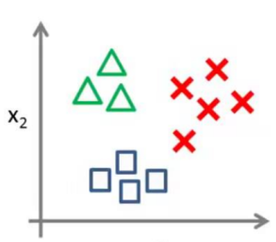
	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000
model complexity	low	mediocare	mediocare	complex
number of free parameters	~4	~10	~4	>10
overfitting?	no	maybe	maybe	more likely

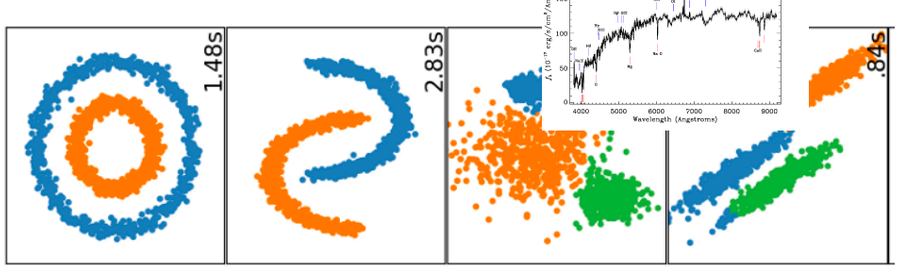
Which algorithm should we use?

	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000
model complexity	low	mediocare	mediocare	complex
number of free parameters	~4	~10	~4	>10
overfitting?	no	maybe	maybe	more likely
easy to interpret?	yes	sometimes	sometimes	more challenging

Which algorithm should we use?

	SVM	ANN	RF	Deep Learning
input dataset	small number of features	larger number of features	larger number of features	raw data
size of training data	can be small	~10s	~10s	>1000
model complexity	low	mediocare	mediocare	complex
number of free parameters	~4	~10	~4	>10
overfitting?	no	maybe	maybe	more likely
easy to interpret?	yes	sometimes	sometimes	more challenging

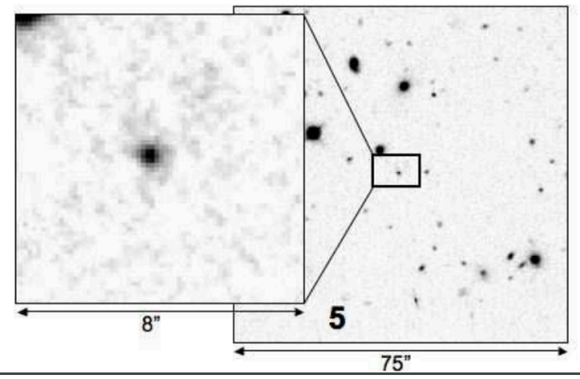




Notes on input data

As our input data, we can either use **measured features** or the **raw data**.

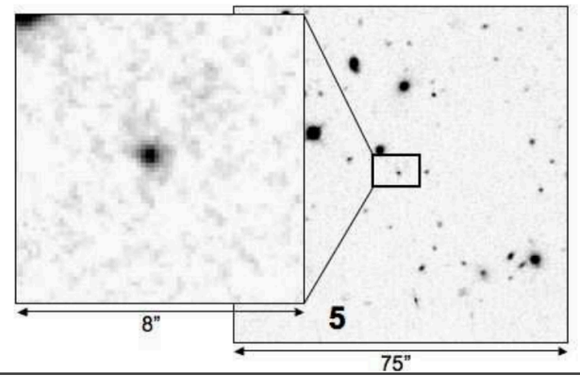
A set of measured features:
concentration, asymmetry, smoothness, ,
moment of Light, Gini Coefficient, etc…



Notes on input data

As our input data, we can either use **measured features** or the **raw data**.

A set of measured features:
concentration, asymmetry, smoothness, ,
moment of Light, Gini Coefficient, etc…

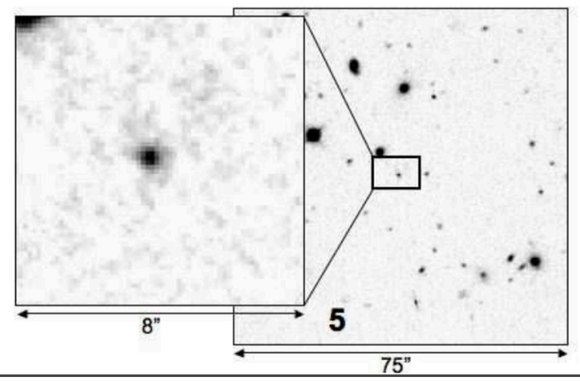


less-complex models: SVM & RF

more-complex models: deep nets

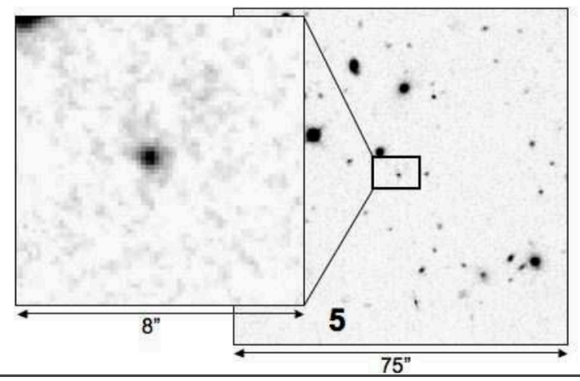
Notes on input data

As our input data, we can either use **measured features** or the **raw data**.

A set of measured features: concentration, asymmetry, smoothness, , moment of Light, Gini Coefficient, etc…	
less-complex models: SVM & RF	more-complex models: deep nets
allows/requires us to incorporate our scientific knowledge into the task	allows us to build a more general tool, but requires us to incorporate our scientific knowledge in a different way

Notes on input data

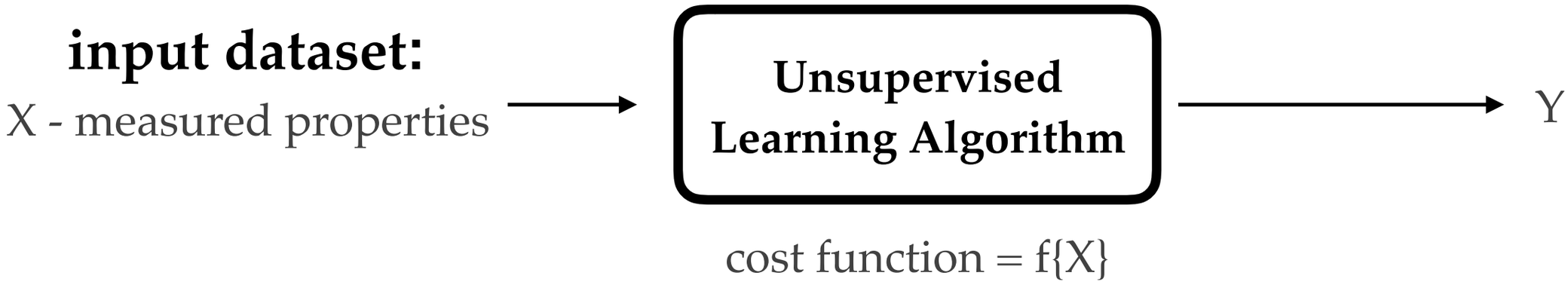
As our input data, we can either use **measured features** or the **raw data**.

A set of measured features: concentration, asymmetry, smoothness, , moment of Light, Gini Coefficient, etc…	
less-complex models: SVM & RF	more-complex models: deep nets
allows/requires us to incorporate our scientific knowledge into the task	allows us to build a more general tool, but requires us to incorporate our scientific knowledge in a different way
information loss	too much information?

The two flavors of Machine Learning

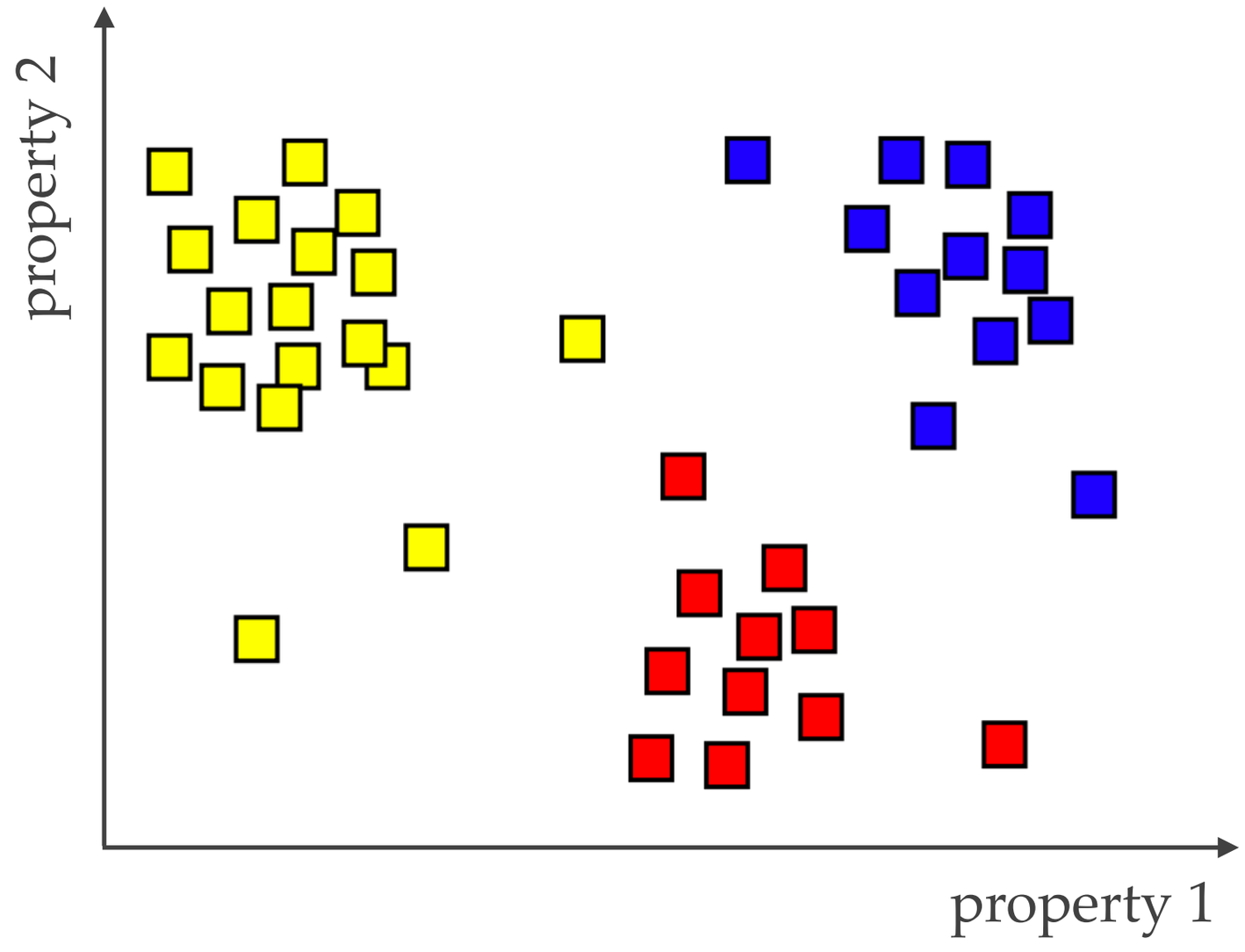
Unsupervised Machine Learning: given a list of objects with measured properties (but no target variable), find patterns in the data.

Common tasks: clustering, dimensionality reduction, and outlier detection.



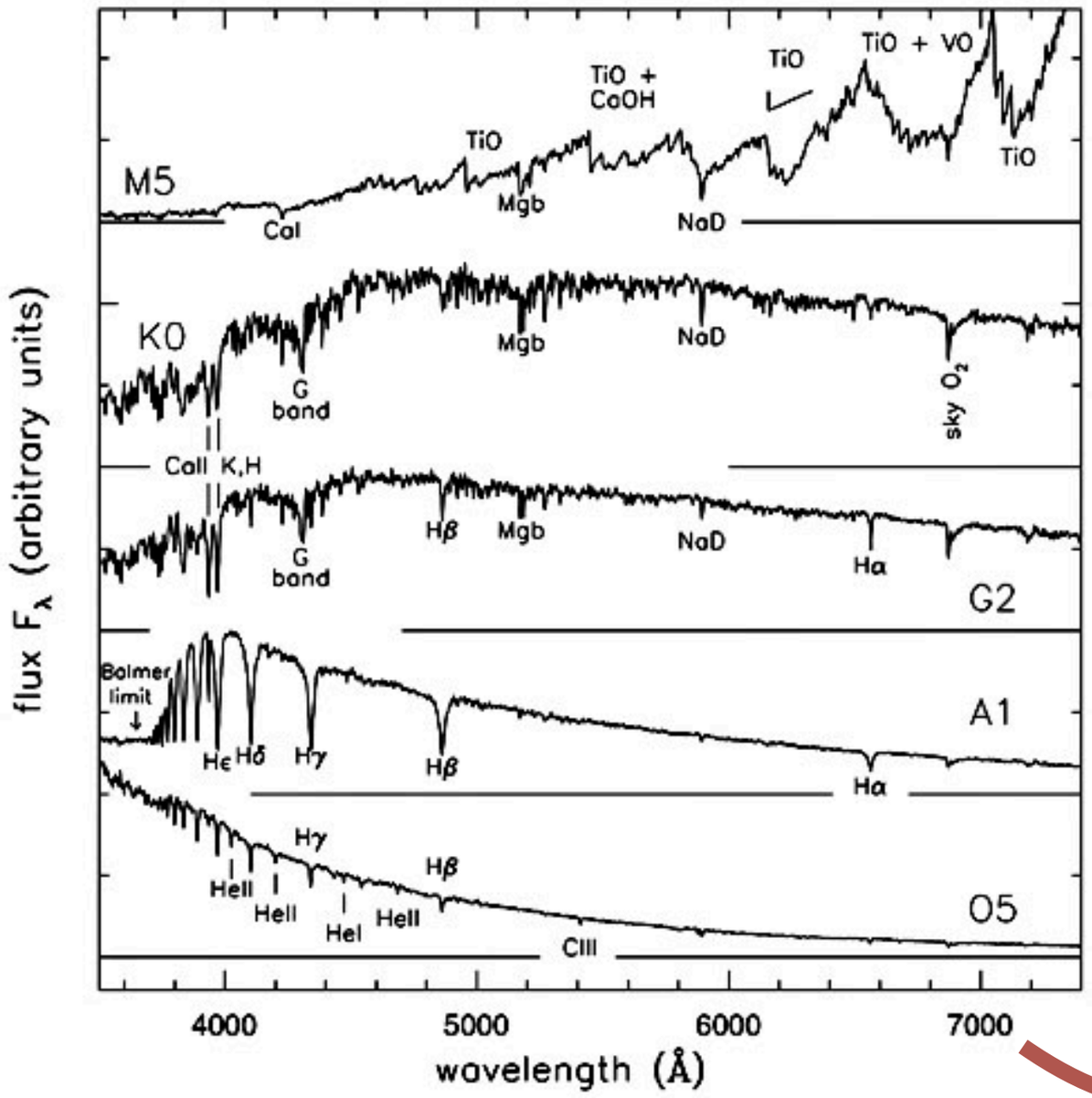
Why should we consider Unsupervised Learning?

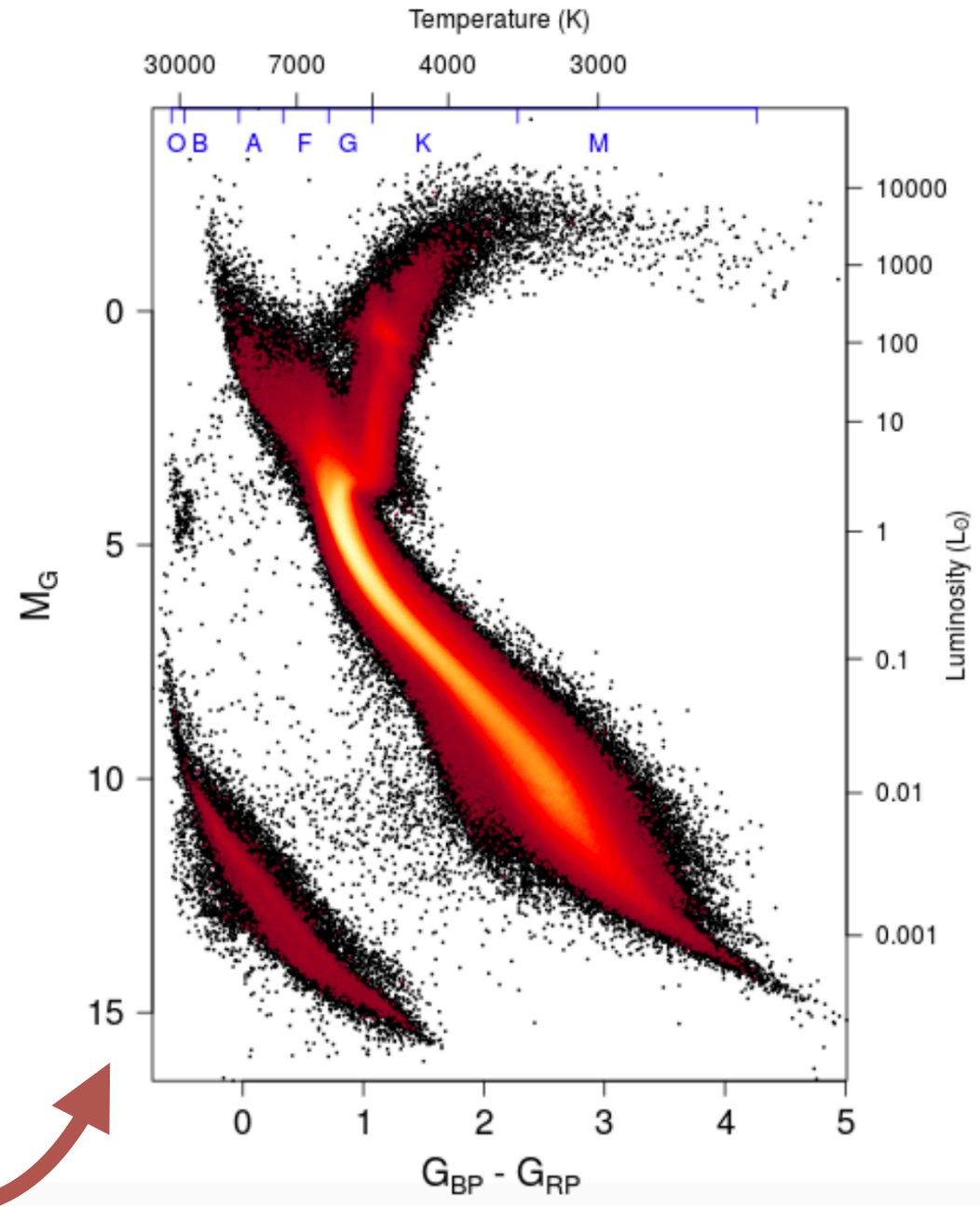
1. Clustering



Why should we consider Unsupervised Learning?

2. Dimensionality Reduction

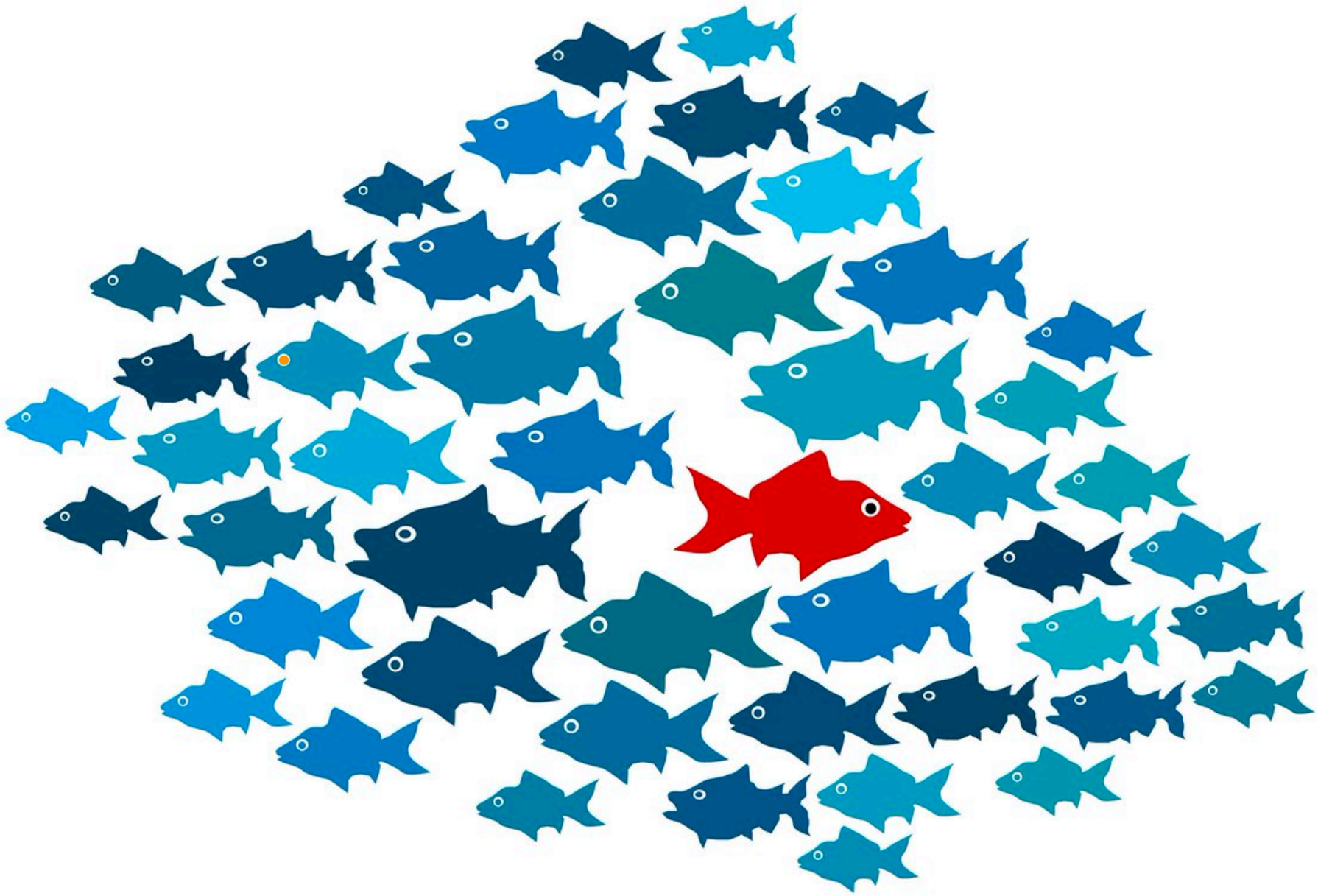




From: *Gaia* Collaboration et al. 2018

Why should we consider Unsupervised Learning?

3. Outlier Detection



Machine Learning as a tool in the astronomers toolkit

Machine Learning in Astronomy: a practical overview

Dalya Baron

Astronomy is experiencing a rapid growth in data size and complexity. This change fosters the development of data-driven science as a useful companion to the common model-driven data analysis paradigm, where astronomers develop automatic tools to mine datasets and extract novel information from them. In recent years, machine learning algorithms have become increasingly popular among astronomers, and are now used for a wide variety of tasks. In light of these developments, and the promise and challenges associated with them, the IAC Winter School 2018 focused on big data in Astronomy, with a particular emphasis on machine learning and deep learning techniques. This document summarizes the topics of supervised and unsupervised learning algorithms presented during the school, and provides practical information on the application of such tools to astronomical datasets. In this document I cover basic topics in supervised machine learning, including selection and preprocessing of the input dataset, evaluation methods, and three popular supervised learning algorithms, Support Vector Machines, Random Forests, and shallow Artificial Neural Networks. My main focus is on unsupervised machine learning algorithms, that are used to perform cluster analysis, dimensionality reducti
Unsupervised learning algorithms are of particular importance to scientific research, since they can be used to extract new know
facilitate new discoveries.

Arxiv: 1904.07248

Surveying the reach and maturity of machine learning and artificial intelligence in astronomy

Christopher J. Fluke, Colin Jacobs

Machine learning (automated processes that learn by example in order to classify, predict, discover or generate new data) and artificial intelligence (methods by which a computer makes decisions or discoveries that would usually require human intelligence) are now firmly established in astronomy. Every week, new applications of machine learning and artificial intelligence are added to a growing corpus of work. Random forests, support vector machines, and neural networks (artificial, deep, and convolutional) are now having a genuine impact for applications as diverse as discovering extrasolar planets, transient objects, quasars, and gravitationally-lensed systems, forecasting solar activity, and distinguishing between signals and instrumental effects in gravitational wave astronomy. This review surveys contemporary, published literature on machine learning and artificial intelligence in astronomy and astrophysics. Applications span seven main categories of activity: classification, regression, clustering, forecasting, generation, discovery, and the development of new scientific insight. These categories form the b
use of machine learning and artificial intelligence emerges, progresses or becomes established.

Arxiv: 1912.02934

(1) How can we discover new types of objects?

(2) How can we discover new trends and correlations?

The Search for the Unknown Unknowns

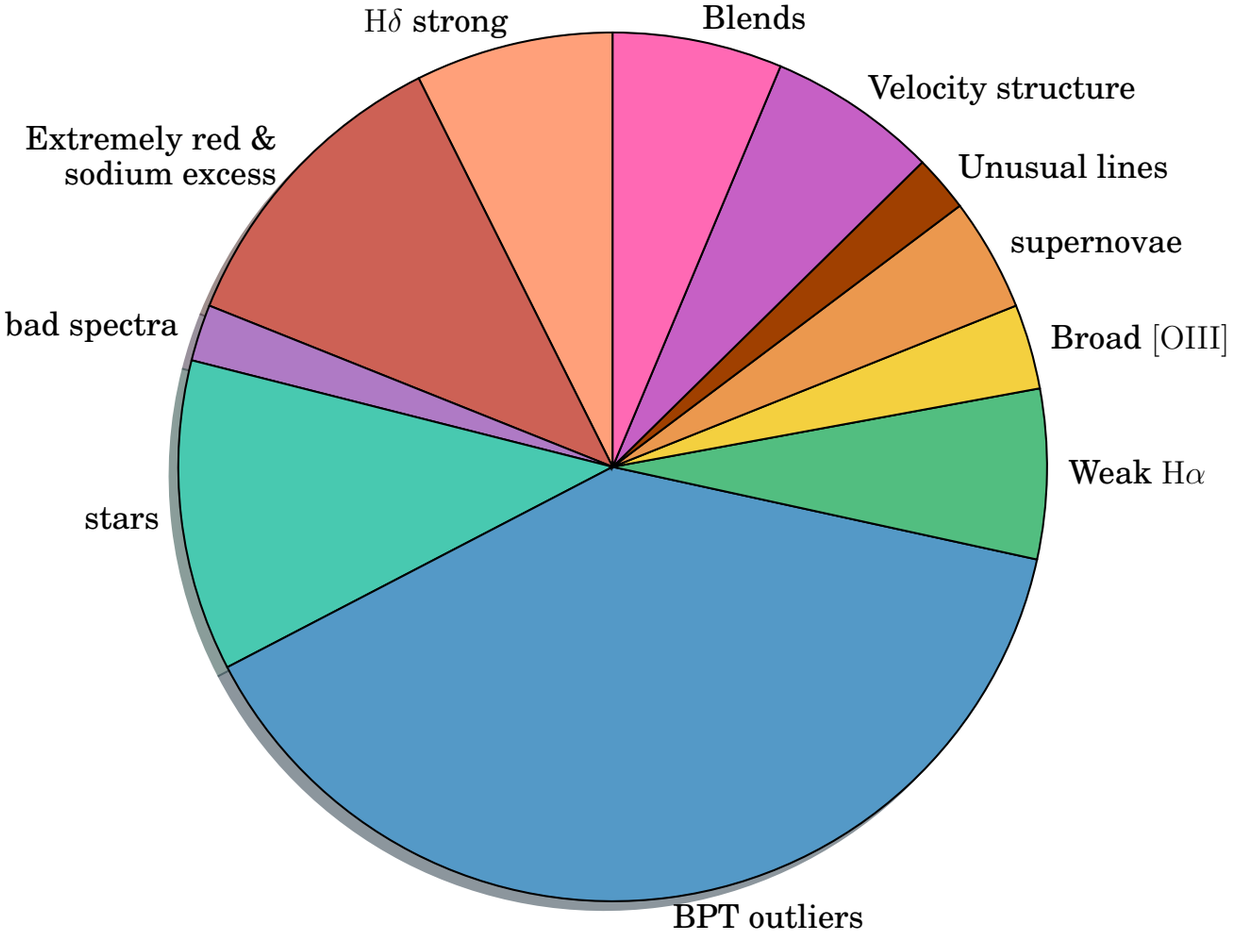
What are outliers?

- **"Bad object"**: problem with the instrument, faulty observation, pipeline error, etc.
- **Misclassified object**: a star in a catalog of galaxies.
- **Tail of a distribution:** the most massive black hole, the most luminous supernova.
- **Unknown unknowns:** completely new objects we did not know we should be looking for.

In astronomy: processes which happen on shorter time scales.

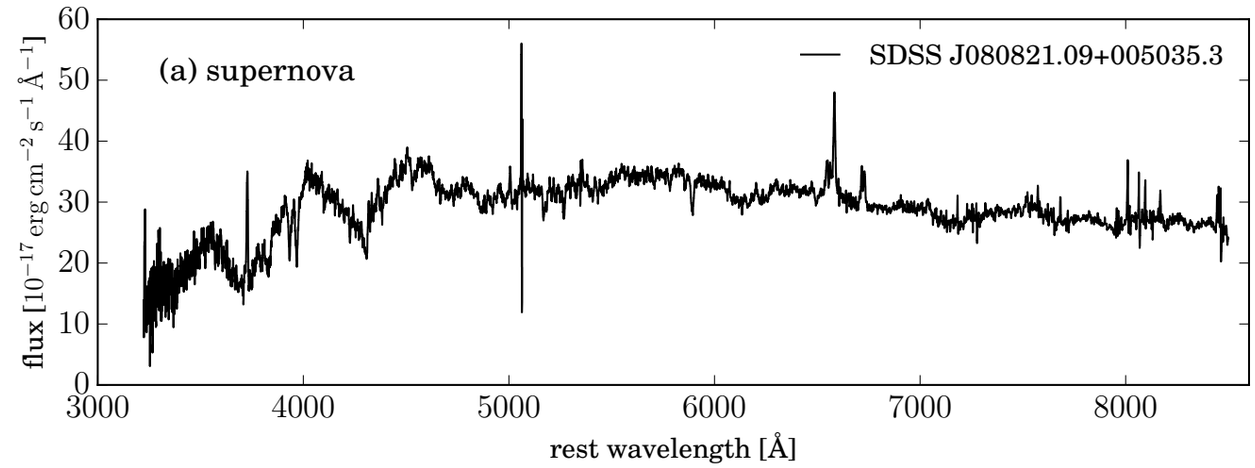
The Search for the Unknown Unknowns

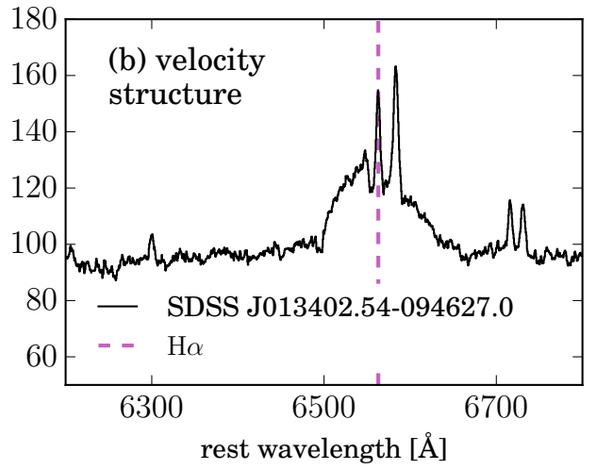
- We used an Unsupervised Random Forest algorithm to assign distances between ~2M galaxy spectra from the SDSS.
- The algorithm learns the data structure through the correlations between different features.
- The galaxies with the largest distances from the rest were defined as outliers.
- A manual inspection of the 400 weirdest galaxies revealed a diverse group of outliers.

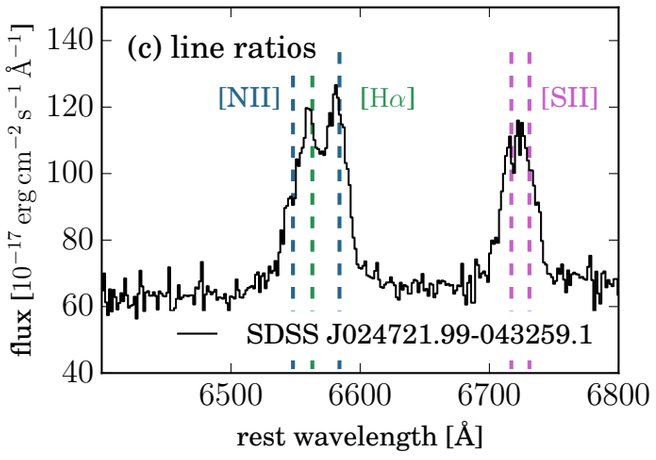


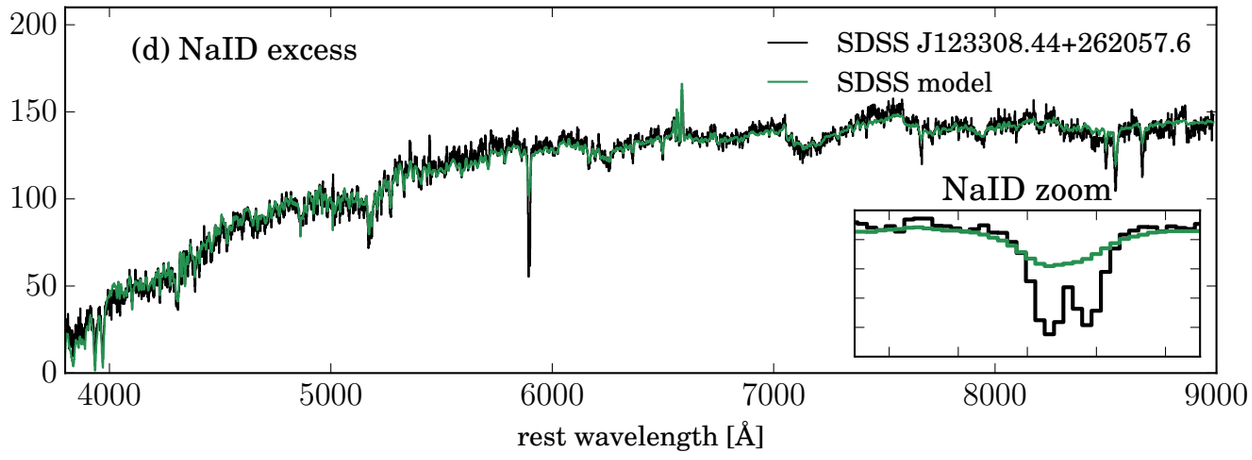
Baron & Poznanski (2017), Reis+ (2018)

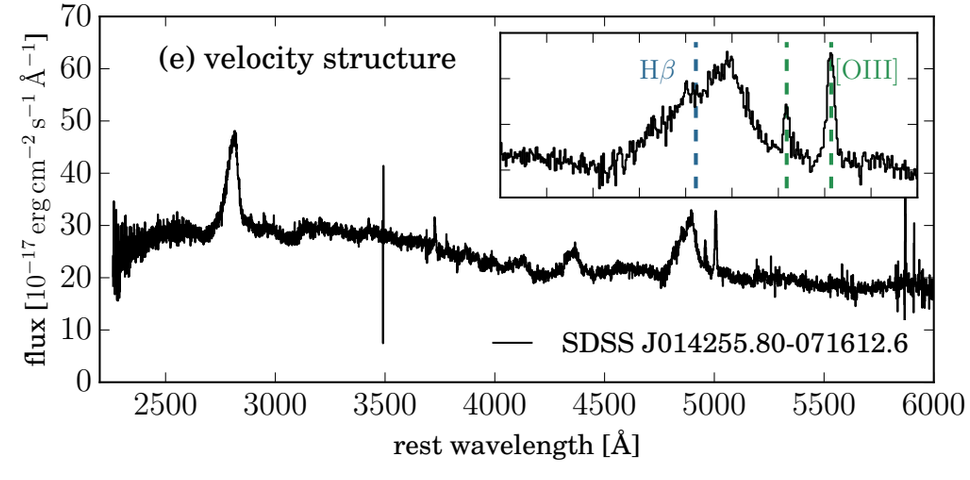
The Search for the Unknown Unknowns

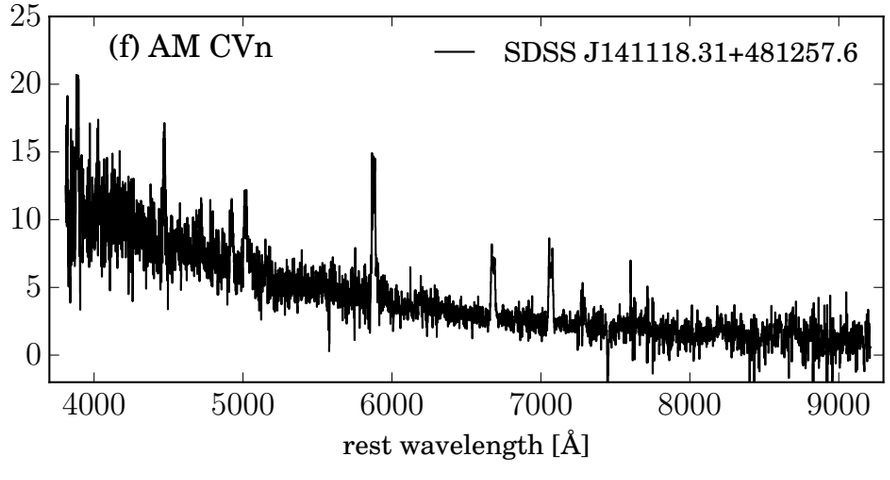






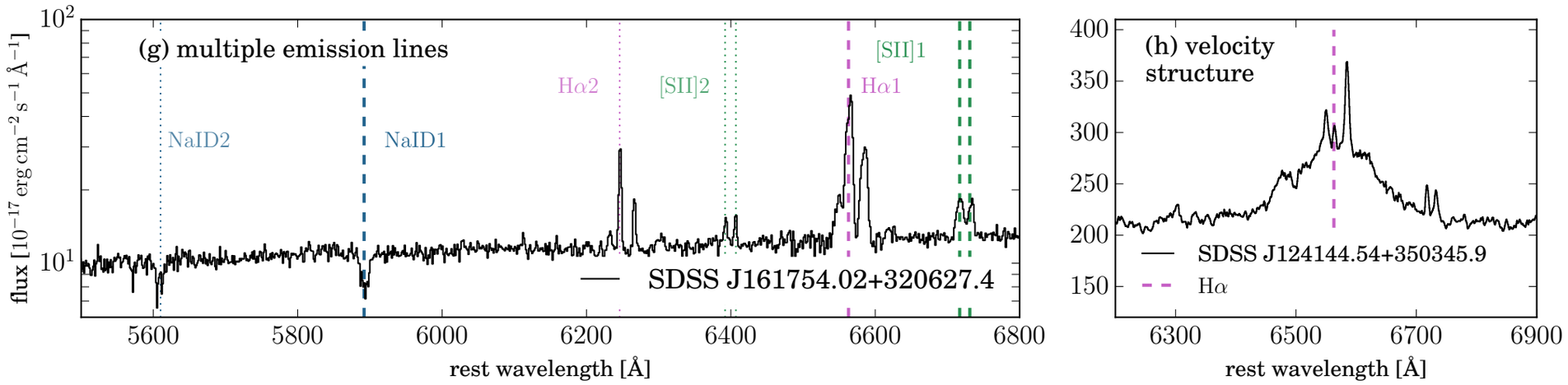






Baron & Poznanski (2017), Reis+ (2018)

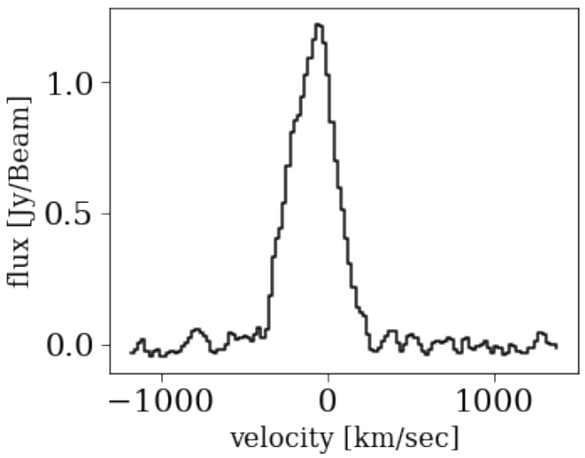
The Search for the Unknown Unknowns



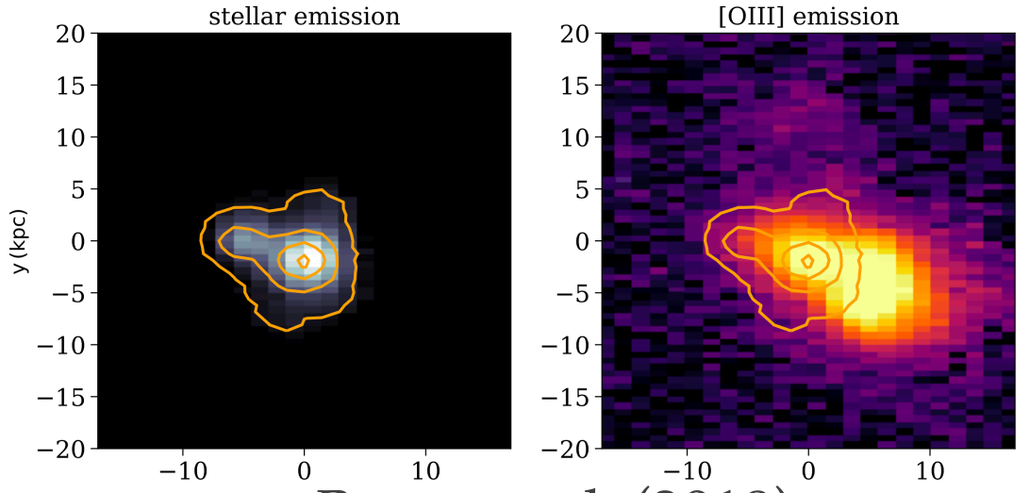
Baron & Poznanski (2017), Reis+ (2018)

Post starburst E+A galaxies & AGN

ALMA

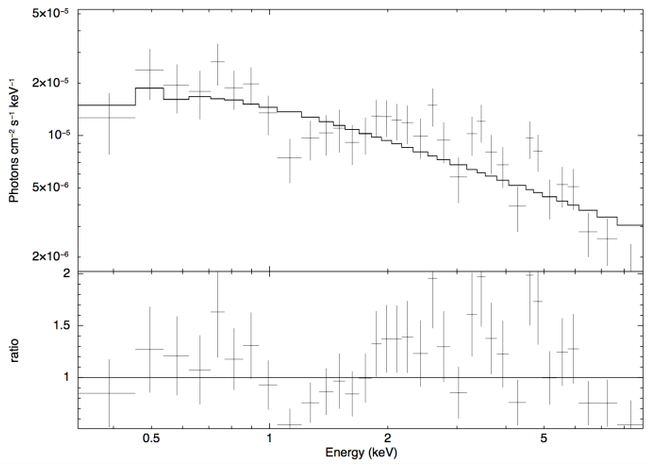


KCWI@Keck

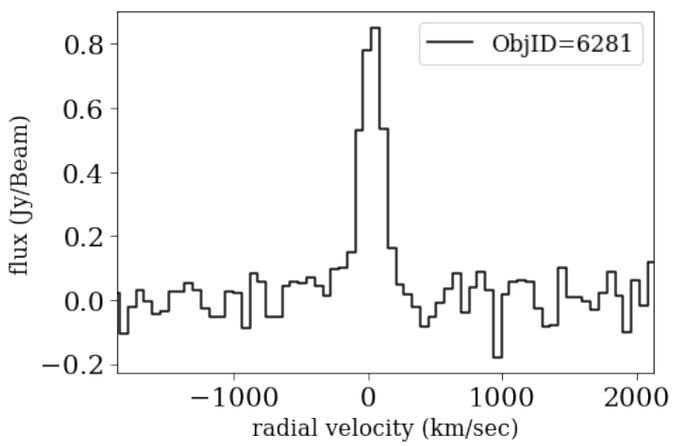


Baron et al. (2018)

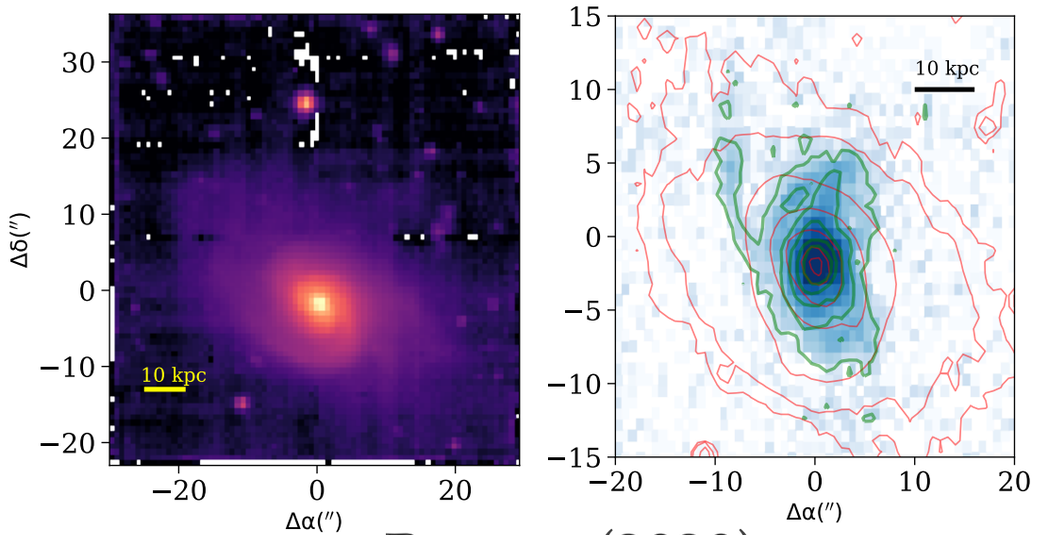
XMM-Newton



NOEMA



MUSE@VLT

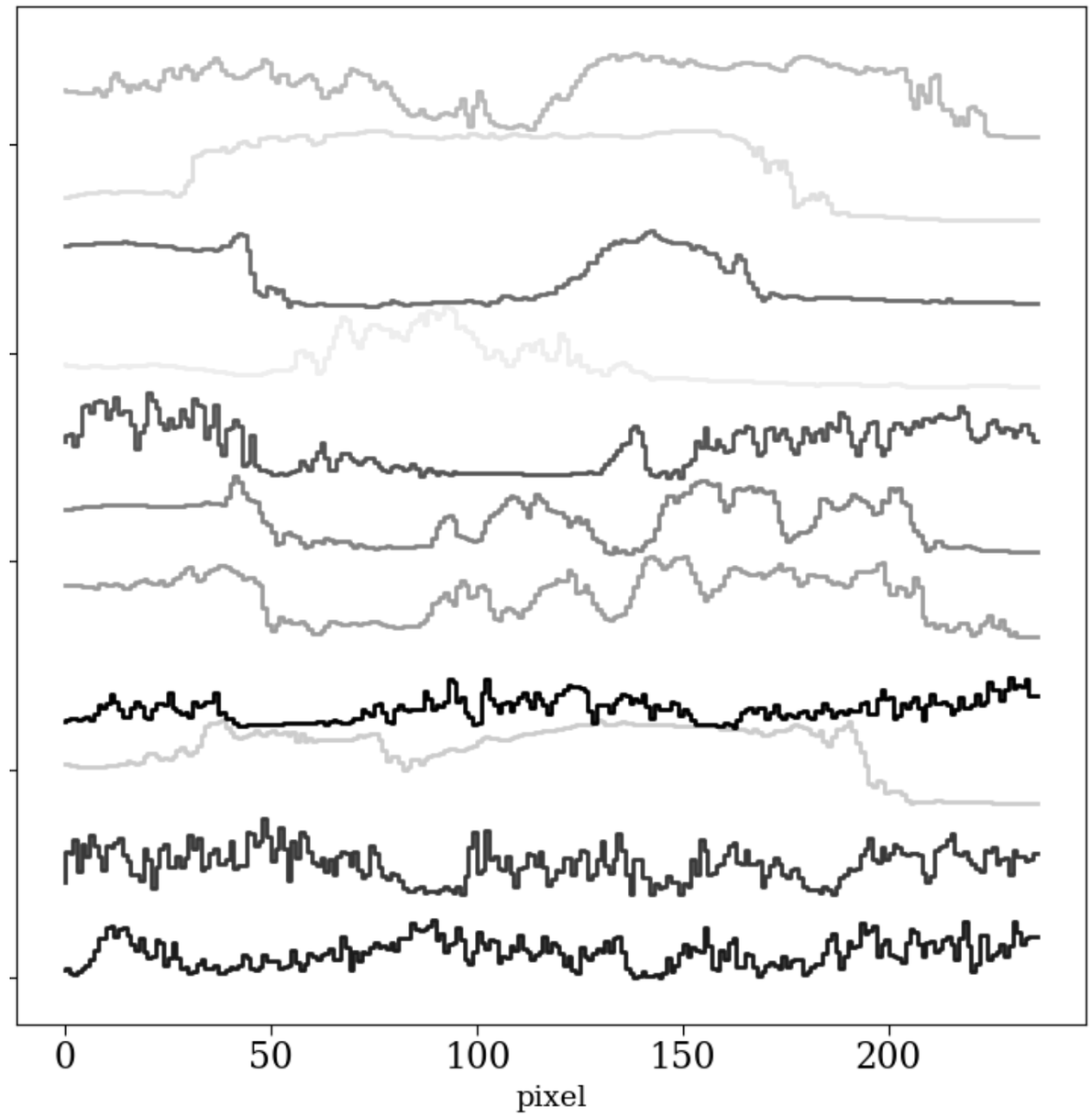


Baron+ (2020)

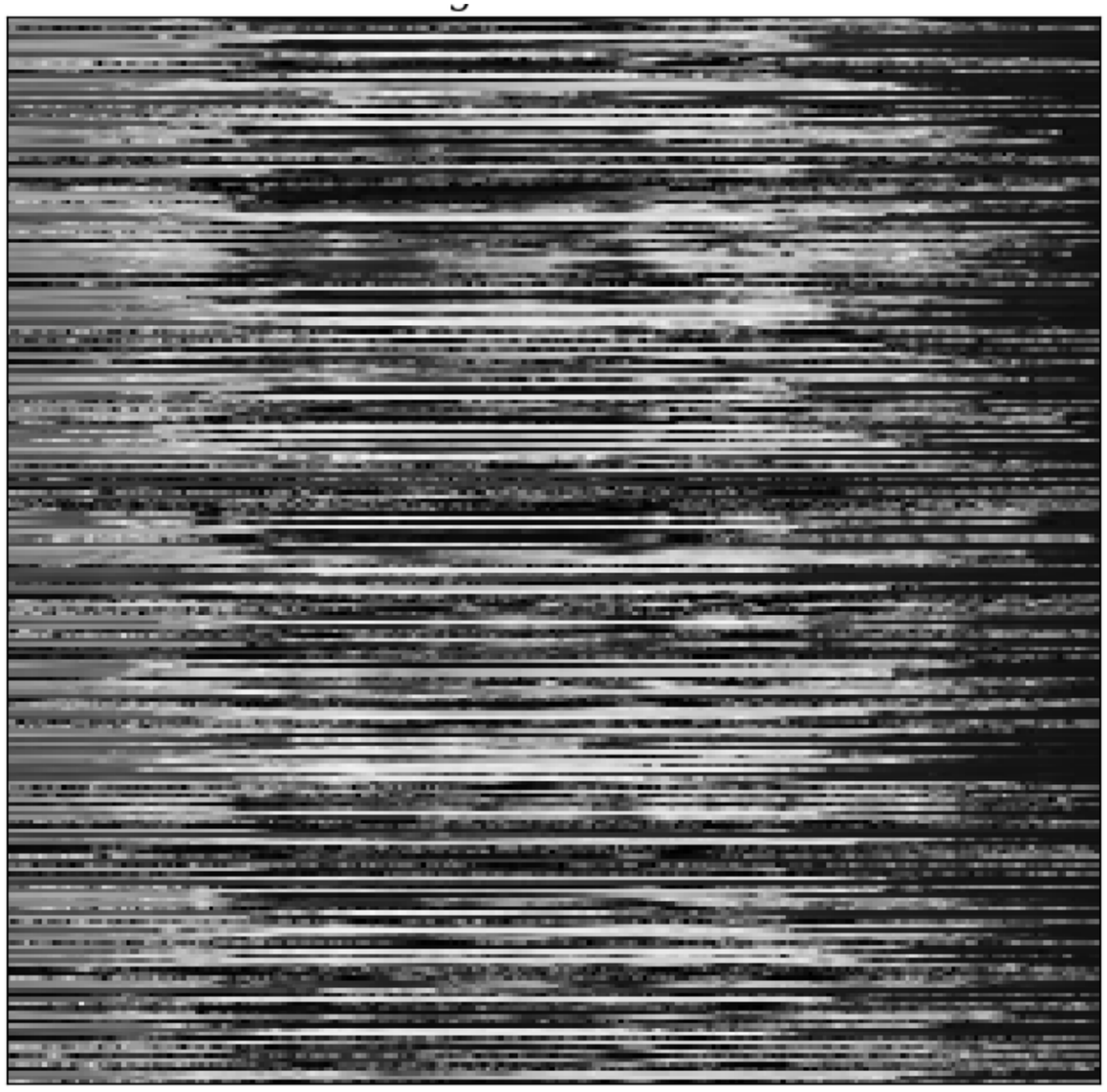
sub-mm → optical → X-ray (Energy)

What is there in my dataset?

sample of the dataset

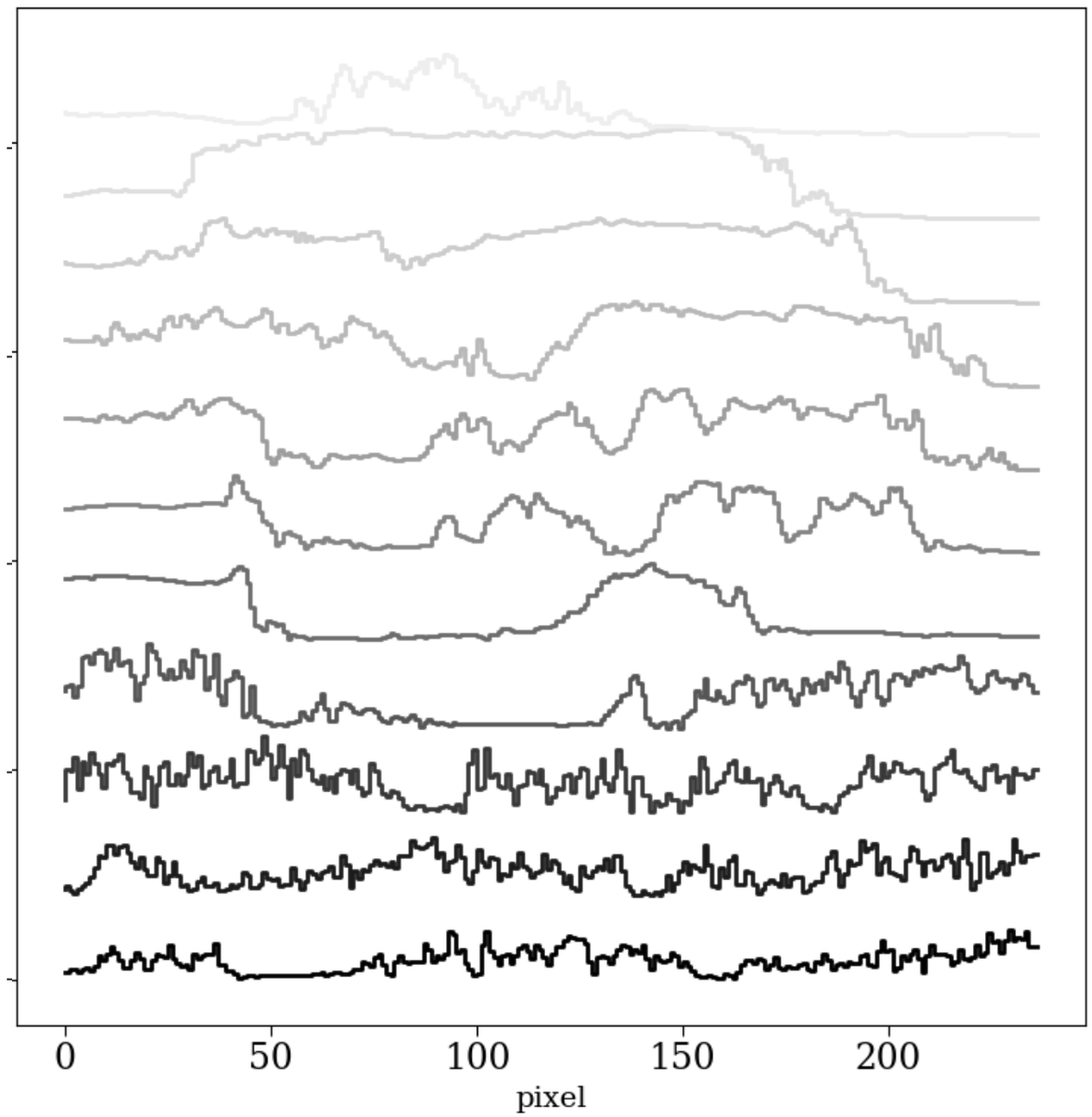


entire dataset

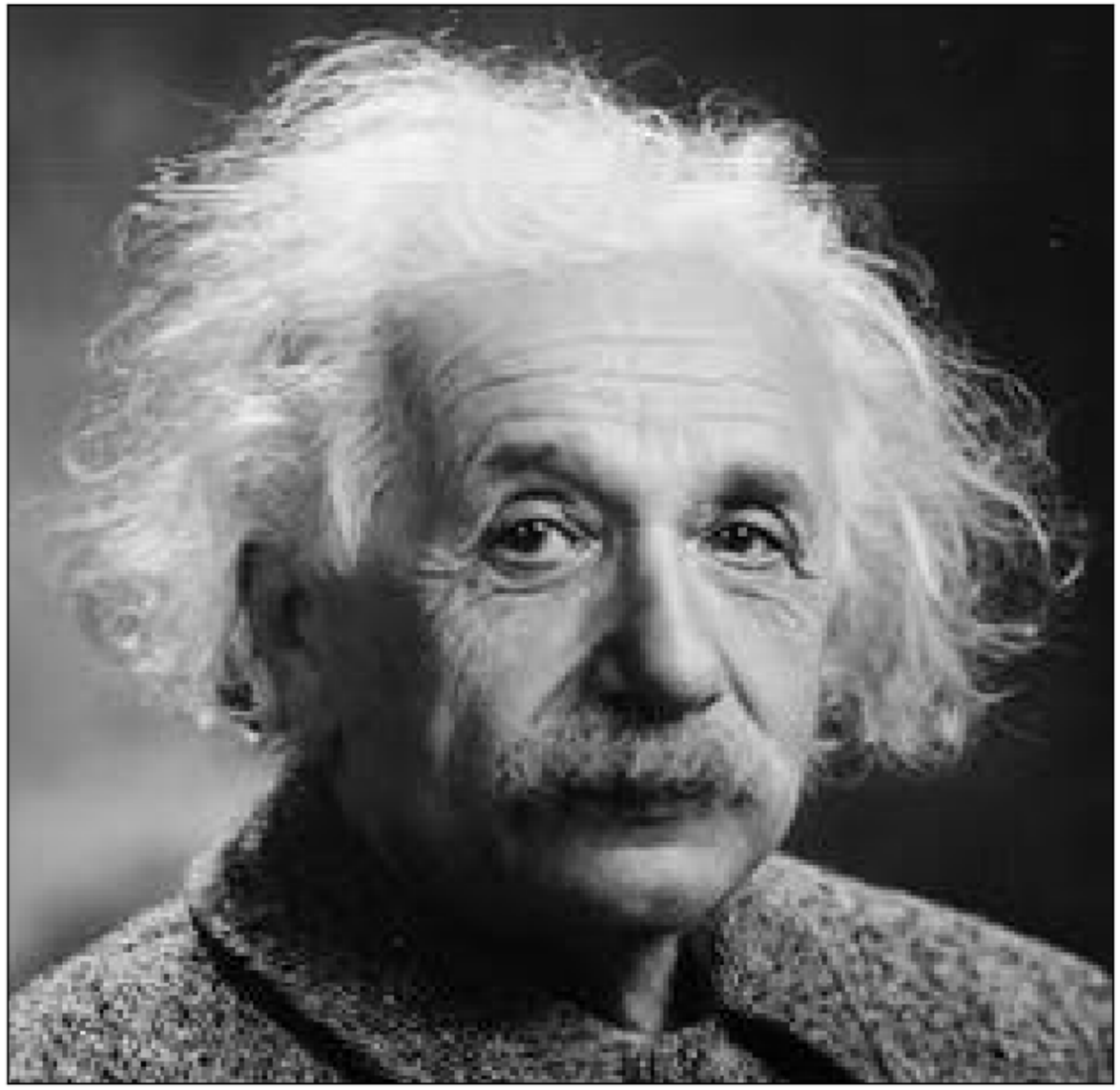


What is there in my dataset?

sample of the dataset

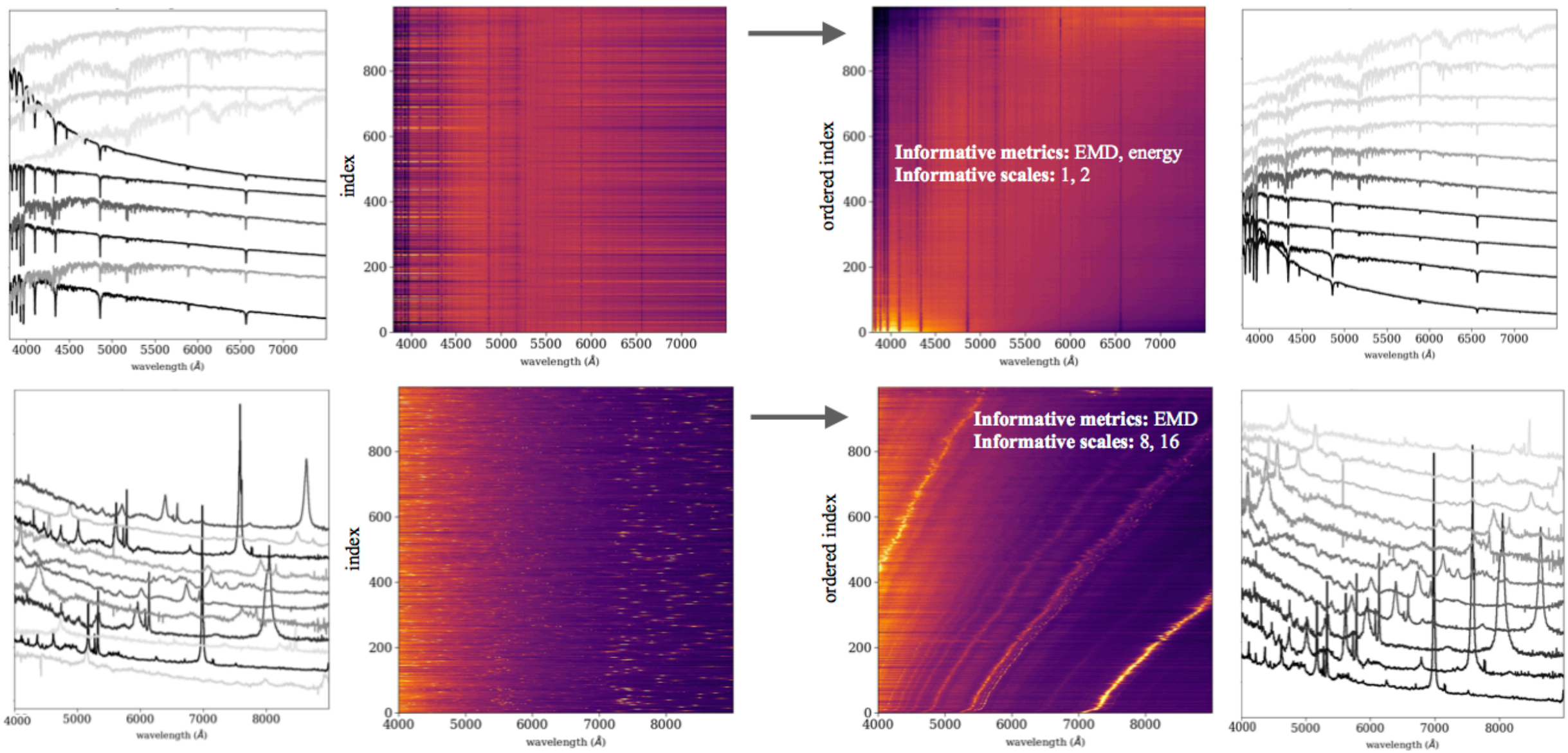


entire dataset



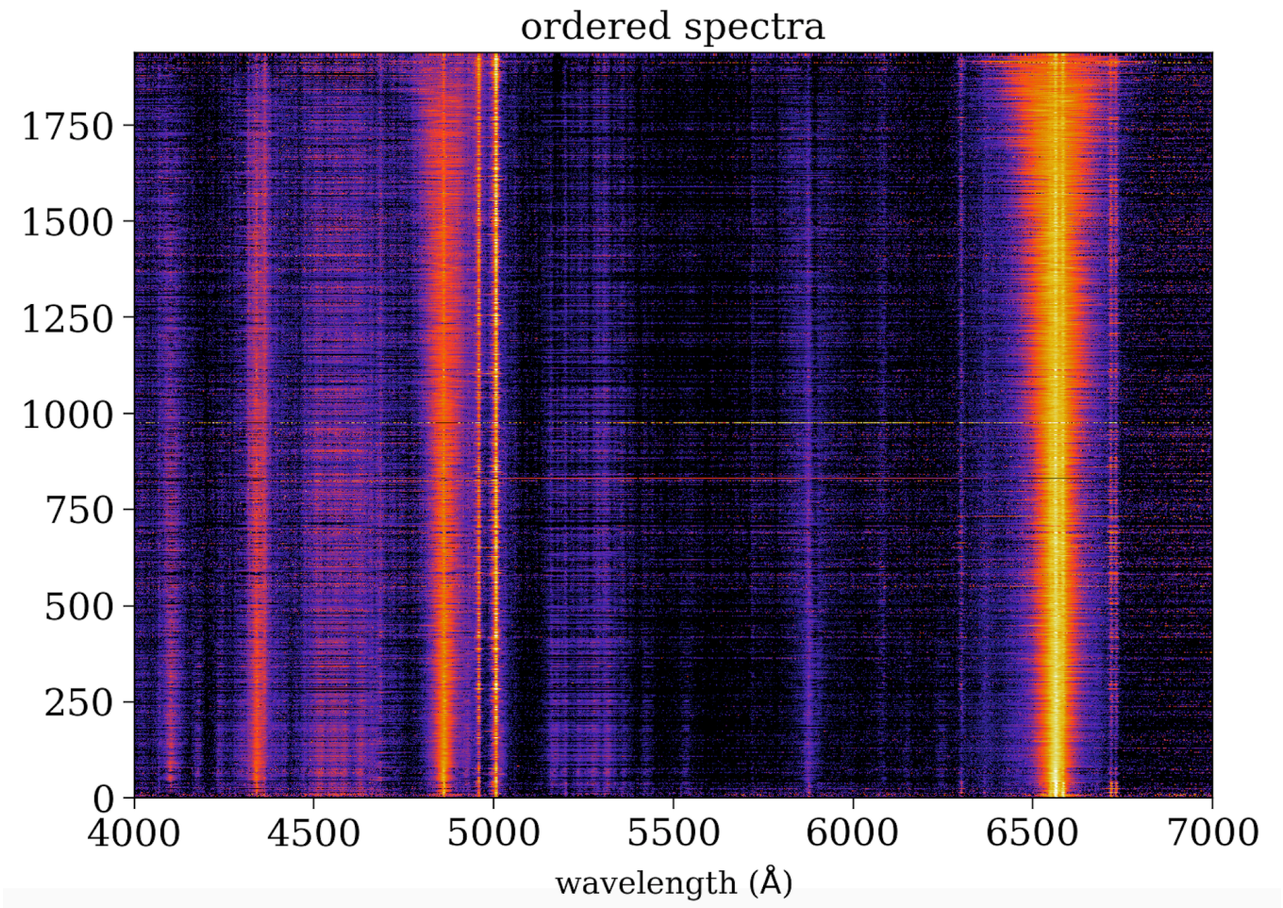
In Baron & Ménard (2020), we presented the Sequencer algorithm.
The algorithm searches for a sequence in the data. If such sequence is found, it reorders the objects according to the detected sequence.

Examples from Astronomy



Sequencer-Assisted Discovery

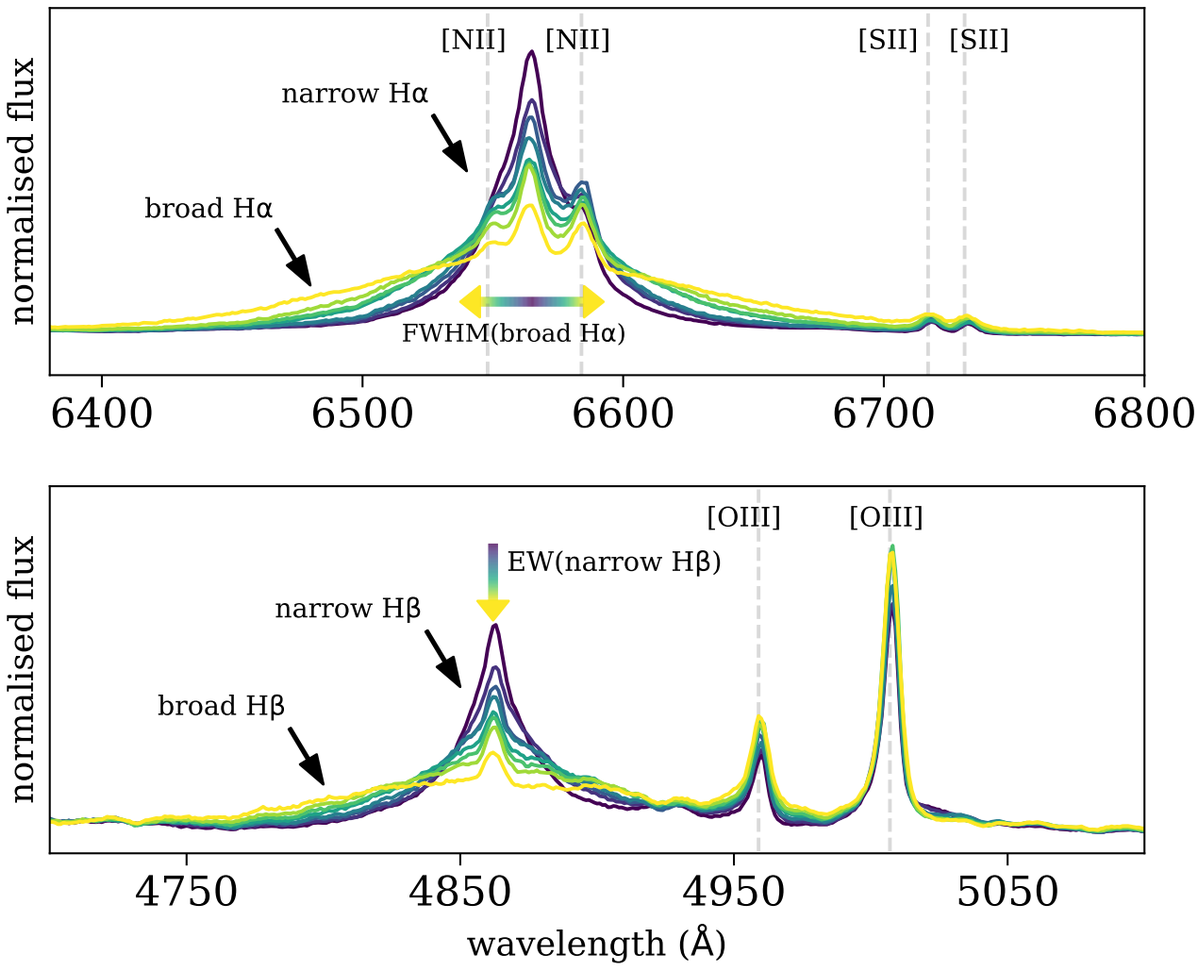
We applied the Sequencer to spectra of type I AGN. The ordered dataset reveals a new correlation between the narrow and broad emission lines!

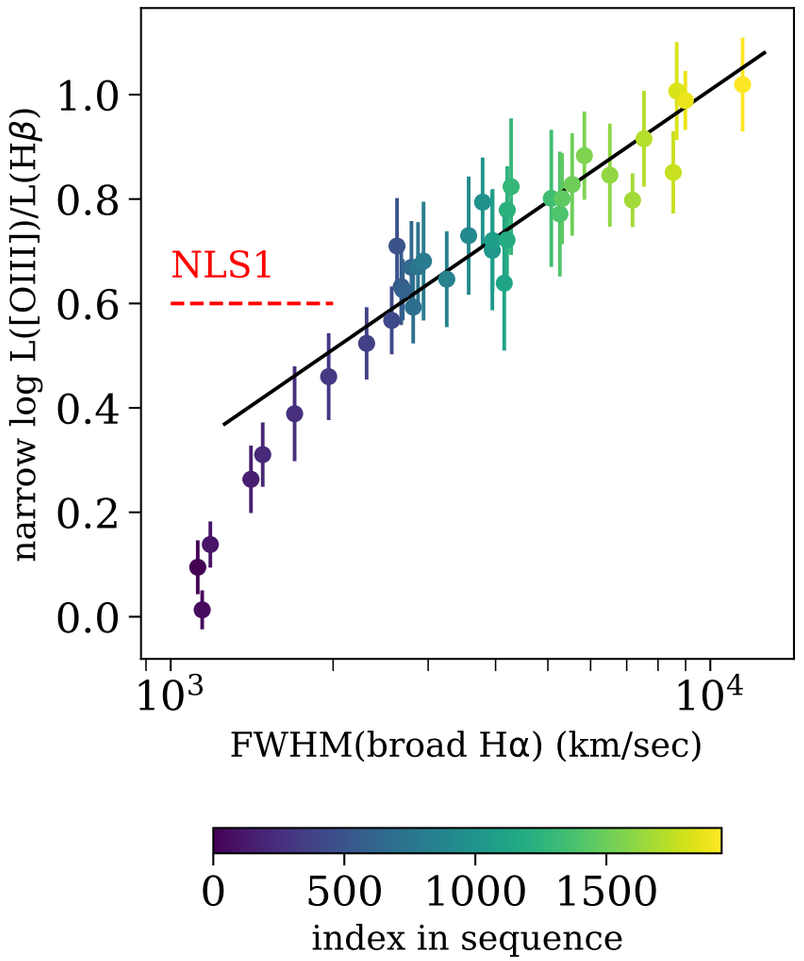


Baron & Ménard (2019)

Sequencer-Assisted Discovery

We applied the Sequencer to spectra of type I AGN. The ordered dataset reveals a new correlation between the narrow and broad emission lines!

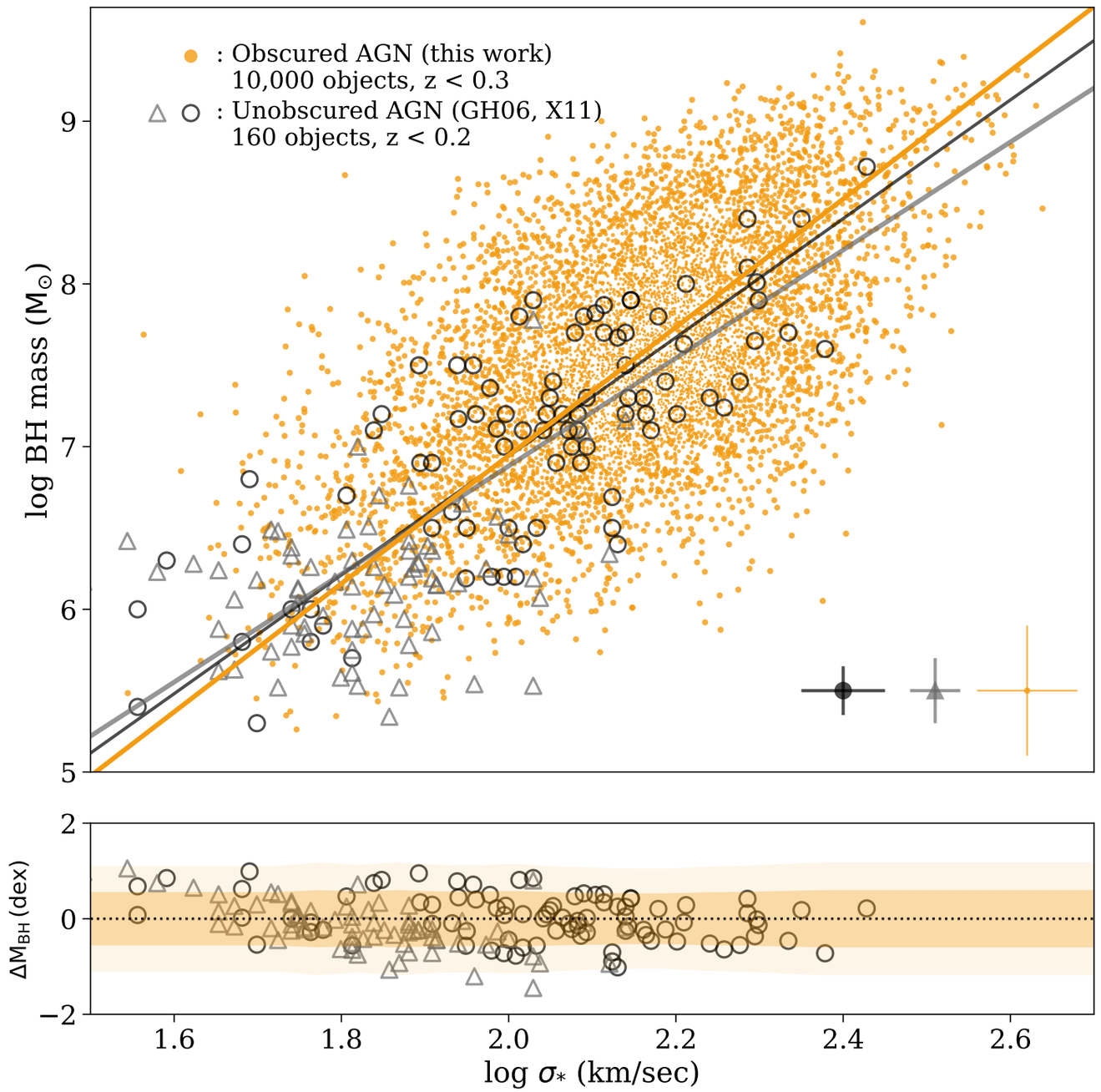




Baron & Ménard (2019)

Sequencer-Assisted Discovery

Using this newly-discovered correlation, we proposed a method to estimate the black hole mass in obscured type II AGN.



Baron & Ménard (2019)

Final remarks

- The outlier detection algorithm and the Sequencer algorithm are available online at: https://github.com/dalya/.

- Supervised Learning algorithms have allowed us to automate various tasks, providing speed and computational ease.
- In recent years, the use of Supervised Learning algorithms have allowed us to do things we could not do before (e.g., detection of strong lenses, detection of stellar feedback features, etc).
- Unsupervised Learning algorithms can be used to extract new knowledge from existing datasets, and can thus facilitate new discoveries.

Thanks! :)

Dalya Baron (TAU)
dalyabaron@gmail.com