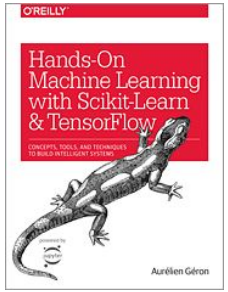
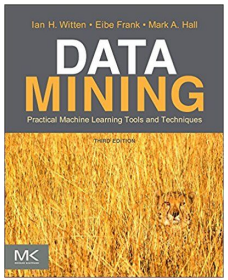
Machine Learning on Galaxy Spectra

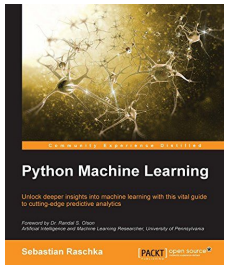
Graziano Ucci
Kapteyn Astronomical Institute, Groningen

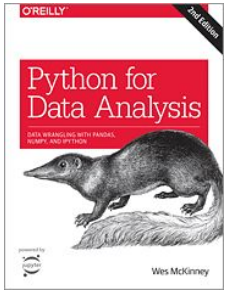
rijksuniversiteit
groningen

Suggested Readings









Aurélien Géron

Hands-On Machine Learning with Scikit-Learn and TensorFlow (Concepts, Tools, and Techniques to Build Intelligent Systems)

O'Reilly Media

Ian H. Witten, Eibe Frank, Mark A. Hall

Data Mining: Practical Machine Learning Tools and Techniques

Morgan Kaufmann Series

Sebastian Raschka

Python Machine Learning

Packt Publishing

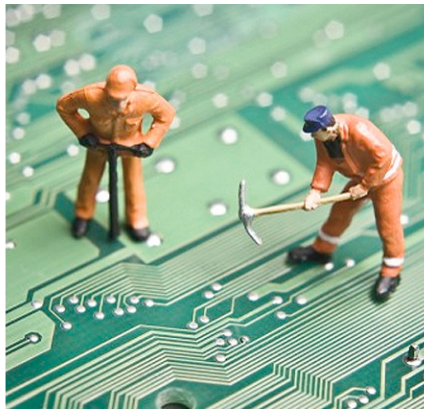
Wes McKinney

Python for Data Analysis

O'Reilly Media

Data Analytics

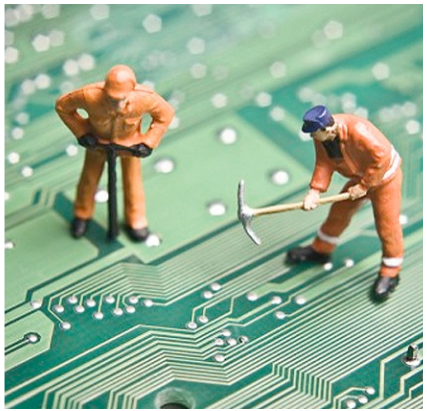
The process of extracting **useful insights from (raw) data.**



Data Analytics

The process of extracting **useful insights from (raw) data.**

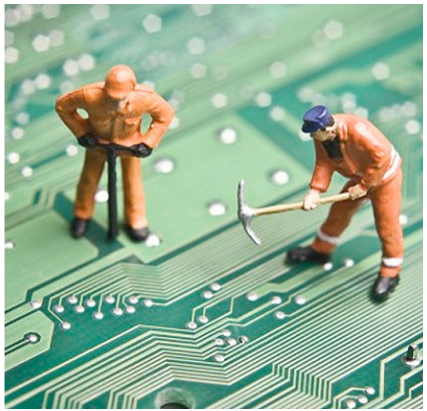
Same as **Data Mining:** the process of discovering valuable information from (large) databases using algorithms able to find hidden patterns in data (Feldman & Dagan 1995).



Data Analytics

The process of extracting **useful insights from (raw) data.**

Same as **Data Mining:** the process of discovering valuable information from (large) databases using algorithms able to find hidden patterns in data (Feldman & Dagan 1995).

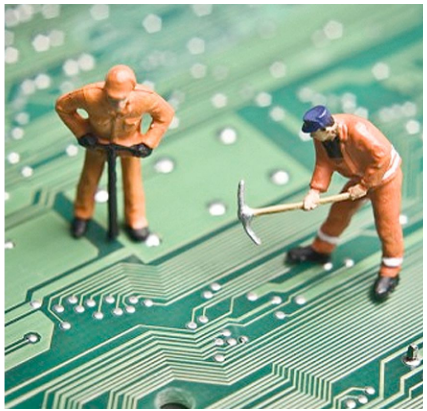


OECD 2015 report: countries could be getting much more out of **Data Analytics** in terms of economic and social gains.

Data Analytics

The process of extracting **useful insights from (raw) data.**

Same as **Data Mining:** the process of discovering valuable information from (large) databases using algorithms able to find hidden patterns in data (Feldman & Dagan 1995).



OECD 2015 report: countries could be getting much more out of **Data Analytics** in terms of economic and social gains.

We need a **“Data-drive Innovation”** where **the usage of Data Analytics improves or foster new products, methods and markets.**

Big Data is a stable presence in science. The volume and rate of data produced in many scientific disciplines in some cases could **exceed our ability to effectively treat and analyse them.**

Big Data is a stable presence in science. The volume and rate of data produced in many scientific disciplines in some cases could **exceed our ability to effectively treat and analyse them.**

- **cosmological simulations;**

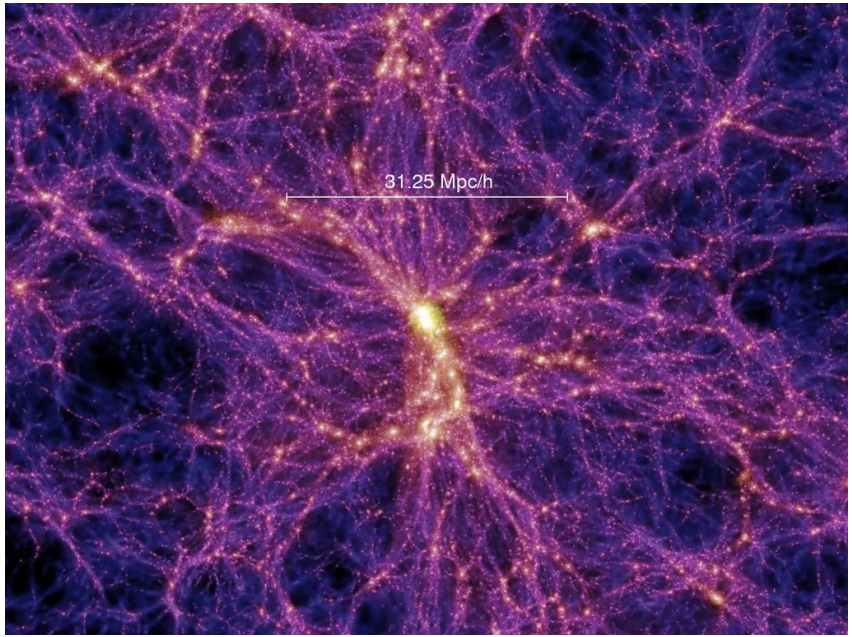


Image credit: Springel et al. (2005)

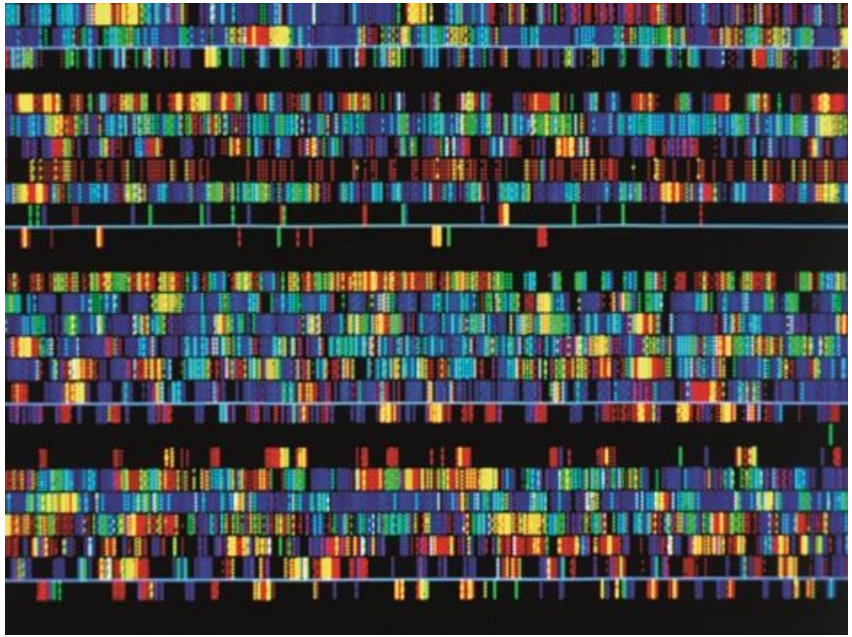
Big Data is a stable presence in science. The volume and rate of data produced in many scientific disciplines in some cases could **exceed our ability to effectively treat and analyse them.**

- cosmological simulations;
- data from **telescopes;**

Image credit: ESO

Big Data is a stable presence in science. The volume and rate of data produced in many scientific disciplines in some cases could **exceed our ability to effectively treat and analyse them.**

- **cosmological simulations;**
- data from **telescopes;**
- **genome sequencing;**



Source: James King-Holmes / Science Photo Library

Big Data is a stable presence in science. The volume and rate of data produced in many scientific disciplines in some cases could **exceed our ability to effectively treat and analyse them.**

- **cosmological simulations;**
- data from **telescopes;**
- **genome sequencing;**
- extreme phenomena in **particle physics.**

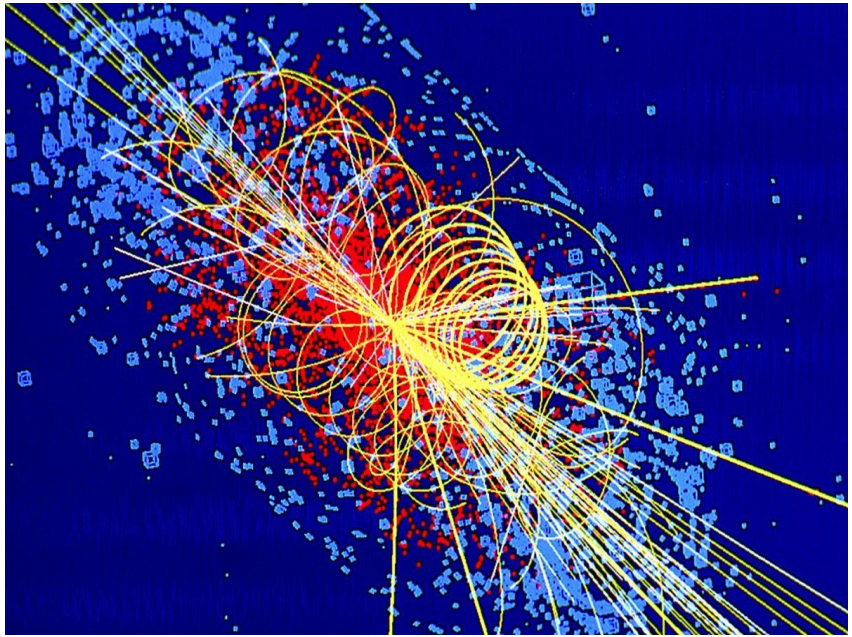


Image credit: Luca Taylor

What **“Big Data”** really is?

The problem arose in the late 1990s within the META Group. For analysts it was becoming evident that their clients were increasingly encumbered by their data assets.

Big Data: a first definition (Gartner Group IT Glossary; Laney 2001)

"Big data is high-**volume**, **-velocity** and **-variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"

Gartner 3 Vs (Laney 2001)

VOLUME: the amount of data that could be generated (**TB up to EB** of data to process, records, transactions, tables, dataframes)

VARIETY: the different types of data we could have in our data sets (structured, semi-structured, unstructured, mixed, multimedia)

VELOCITY: the speed at which new data is generated, collected and analyzed

IBM 4 Vs (2012)

VOLUME: the amount of data that could be generated (**TB up to EB** of data to process, records, transactions, tables, dataframes)

VARIETY: the different types of data we could have in our data sets (structured, semi-structured, unstructured, mixed, multimedia)

VELOCITY: the speed at which new data is generated, collected and analyzed

VERACITY: the messiness or trustworthiness of the data (missing values, authenticity, origin/reputation, uncertainties due to data inconsistency and incompleteness, ambiguities)

5 Vs (2015)

VOLUME: the amount of data that could be generated (**TB up to EB** of data to process, records, transactions, tables, dataframes)

VARIETY: the different types of data we could have in our data sets (structured, semi-structured, unstructured, mixed, multimedia)

VELOCITY: the speed at which new data is generated, collected and analyzed

VERACITY: the messiness or trustworthiness of the data (missing values, authenticity, origin/reputation, uncertainties due to data inconsistency and incompleteness, ambiguities)

VALUE: The ability to turn data into value/money, scientific/business models can be associated to the data (statistics/events, correlations, hypothesis)

Science is basically asking questions: traditionally, scientists **“query the world”** and data is coupled to a specific hypothesis.

Science is basically asking questions: traditionally, scientists **“query the world”** and data is coupled to a specific hypothesis.

The rapid growth in data is leading to a sort of **“eScience”**: we now **download the world** trying to massively acquire in support of many hypotheses.

Science is basically asking questions: traditionally, scientists **“query the world”** and data is coupled to a specific hypothesis.

The rapid growth in data is leading to a sort of **“eScience”**: we now **download the world** trying to massively acquire in support of many hypotheses.

Science in most cases is driven by data more than by computation: because the cost of storing and acquisition has dropped precipitously, dealing with the scientific stored data requires several levels of analysis and various processing/algorithms to obtain complex models.

Science is basically asking questions: traditionally, scientists **“query the world”** and data is coupled to a specific hypothesis.

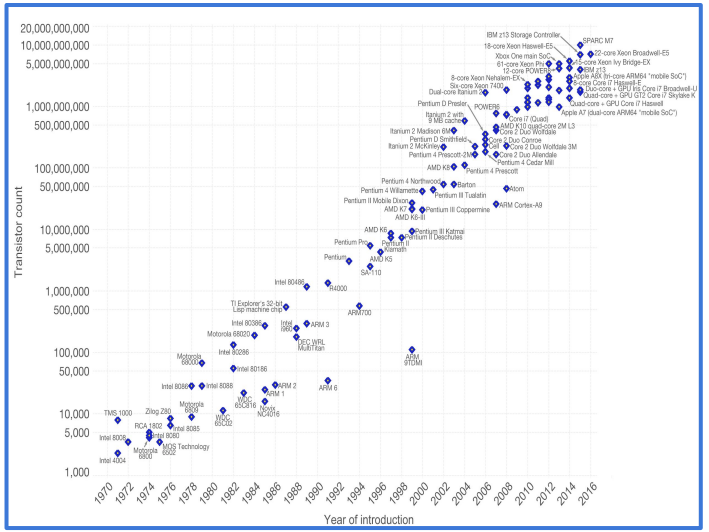
The rapid growth in data is leading to a sort of **“eScience”**: we now **download the world** trying to massively acquire in support of many hypotheses.

Science in most cases is driven by data more than by computation: because the cost of storing and acquisition has dropped precipitously, dealing with the scientific stored data requires several levels of analysis and various processing/algorithms to obtain complex models.

In short words, **we will not face in science** (especially in Astrophysics) **an extreme computational complexity, we are already facing it.**

A new paradigm

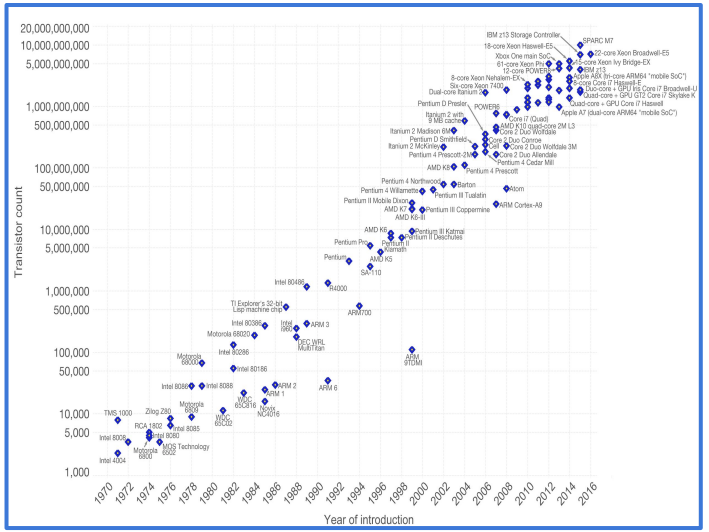
Moore's Law has held for decades (the processing power roughly doubles every two years).



A new paradigm

Moore's Law has held for decades (the processing power roughly doubles every two years).

However, processing speed is no longer the (only) problem: getting the data to the processor is the bottleneck.

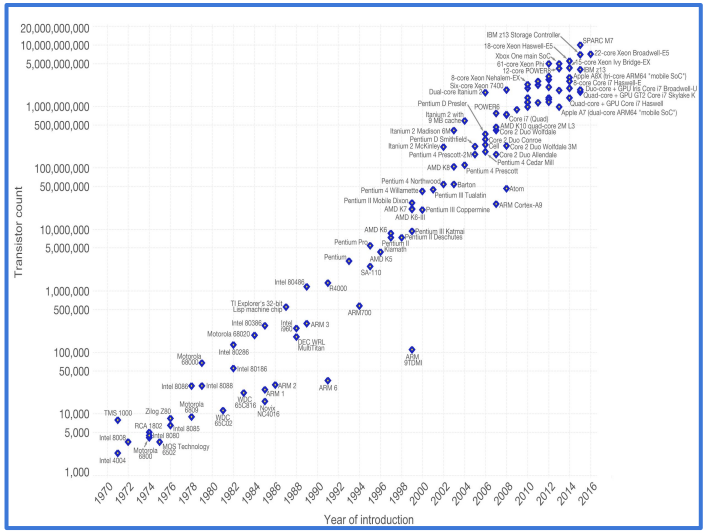


A new paradigm

Moore's Law has held for decades (the processing power roughly doubles every two years).

However, processing speed is no longer the (only) problem: getting the data to the processor is the bottleneck.

Example: at a disk transfer rate of 75 MB/sec, the time taken to transfer 100 GB of data is 22 mins (higher if servers have less than 100 GB of RAM).

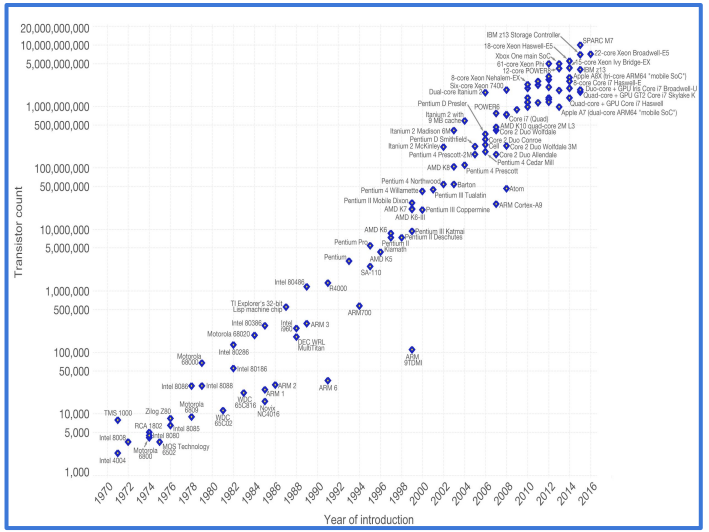


A new paradigm

Moore's Law has held for decades (the processing power roughly doubles every two years).

However, processing speed is no longer the (only) problem: getting the data to the processor is the bottleneck.

Example: at a disk transfer rate of 75 MB/sec, the time taken to transfer 100 GB of data is 22 mins (higher if servers have less than 100 GB of RAM).



This can be considered as the **end of the Moore's Law as we know it**. Increasing the performances cannot be achieved just through increasing hardware speed. We need a new approach: distributed computation must be exploited.

Machine Learning

One of the fundamental families of algorithms used to extract information from data is represented by **Machine Learning (ML).** ML is a process able to map inputs to the output.

Machine Learning

One of the fundamental families of algorithms used to extract information from data is represented by **Machine Learning (ML).** ML is a process able to map inputs to the output.

Statistics models the process that gave rise to data

ML tries to make an accurate prediction, given the data

Machine Learning

One of the fundamental families of algorithms used to extract information from data is represented by **Machine Learning (ML).** ML is a process able to map inputs to the output.

Statistics models the process that gave rise to data

ML tries to make an accurate prediction, given the data

Traditional techniques rely on relatively small samples combined with heavy assumptions about data and its distributions.

ML automatically discovers regularities in data using computational models that generalize the patterns found into new but similar data.

Machine Learning

One of the fundamental families of algorithms used to extract information from data is represented by **Machine Learning (ML).** ML is a process able to map inputs to the output.

ML tends to make no pre-assumptions. The usual approach is, in most cases, empirical: the accuracy or even the applicability of the model is checked *a posteriori.*

Machine Learning

“Machine Learning is the designing of computational models able to learn from data without being explicitly programmed”

Machine Learning

as subfield of the Artificial Intelligence field, ML is mainly concerned with using computers for learning

"Machine Learning is the designing of **computational models** able to learn from data without being explicitly programmed"

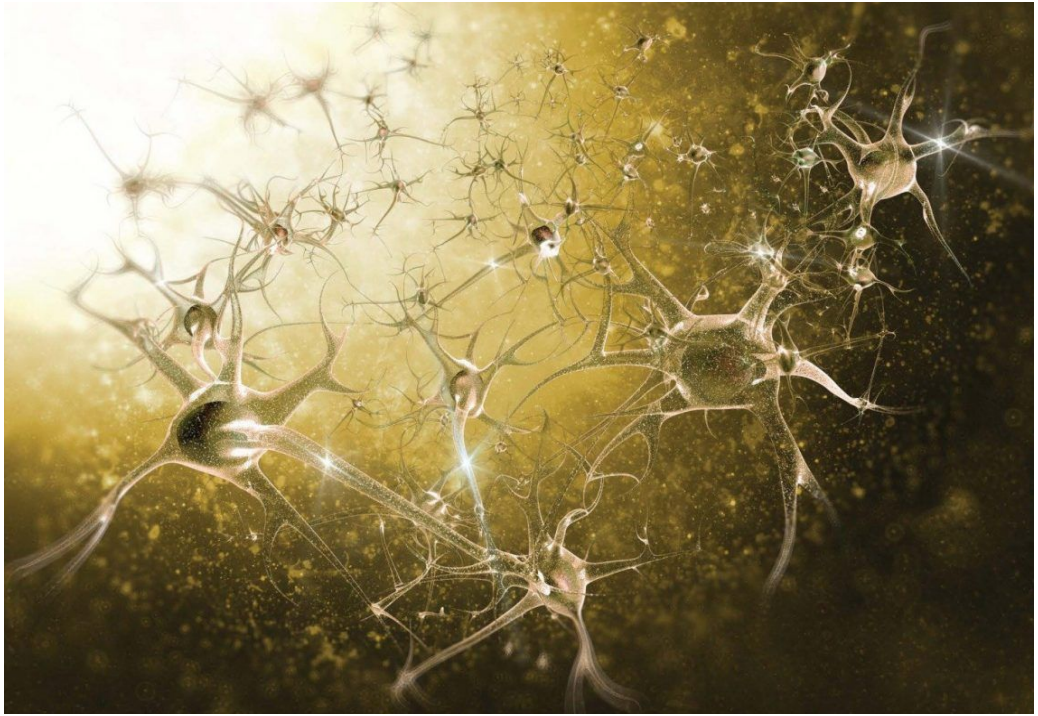
Machine Learning

as subfield of the Artificial Intelligence field, ML is mainly concerned with using computers for learning

"Machine Learning is the designing of **computational models** able to **learn from data without being explicitly programmed**"

with little to no human involvement, learning means any process whereby the system improves its performance based on experience

Machine Learning

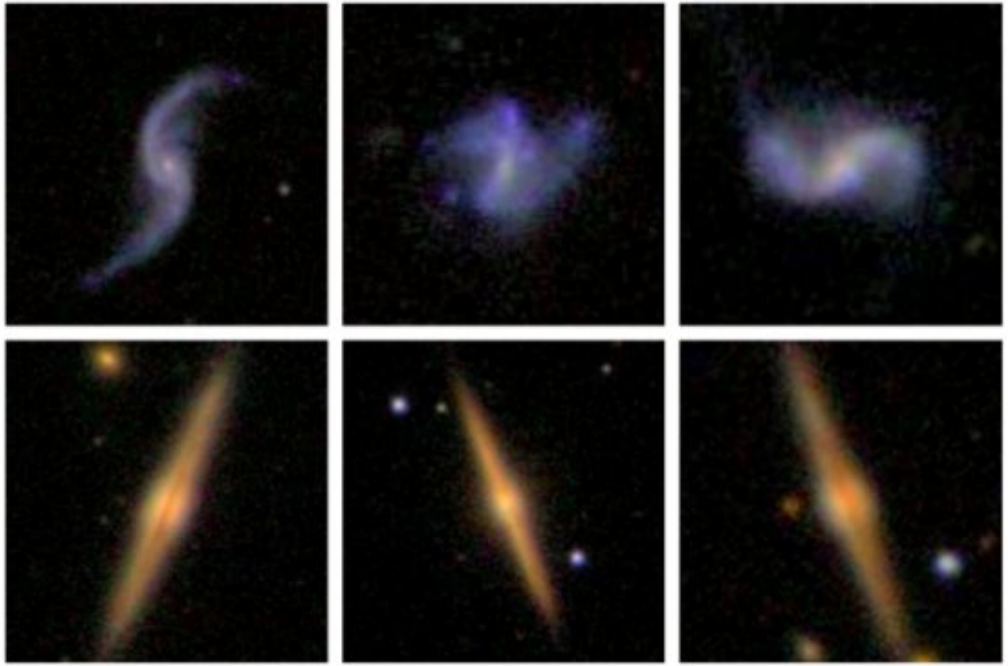


Credit Shantanu N. Zagade for Packt Publishing

In the last decade ML has spread rapidly throughout **computer science** and beyond

web search, spam filters, recommender systems, ads placement, credit scoring, fraud detection, trading, drug design...

Machine Learning



Adapted from Fig. 13 of Dieleman et al. (2015)

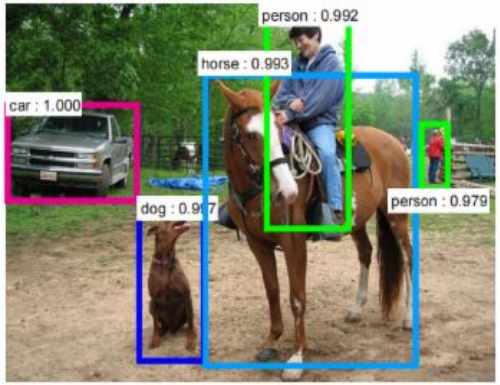
In the last decade ML has spread rapidly throughout computer science and **beyond**

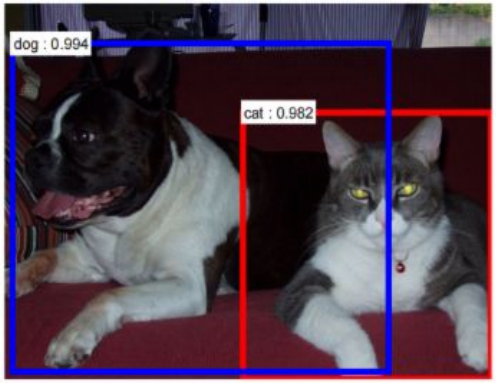
galaxy morphology classification,
photometric redshift determination,
cosmological simulations...

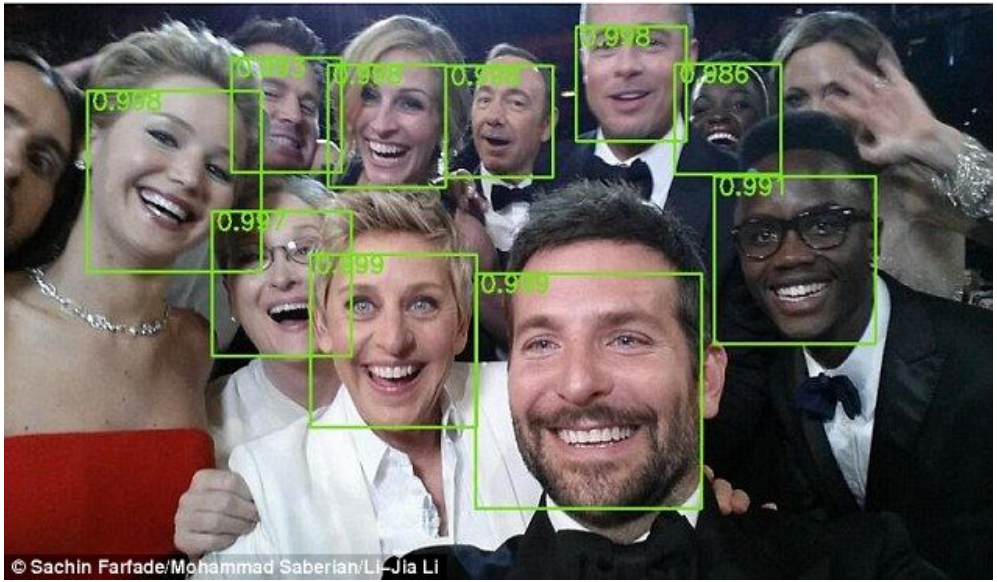
Case study: AlphaGo

October 2015: AlphaGo defeats Fan Hui (“It’s not a human move. I’ve never seen a human play this move [...] So beautiful. So beautiful.”)

March 2016: AlphaGo defeats Lee Sedol

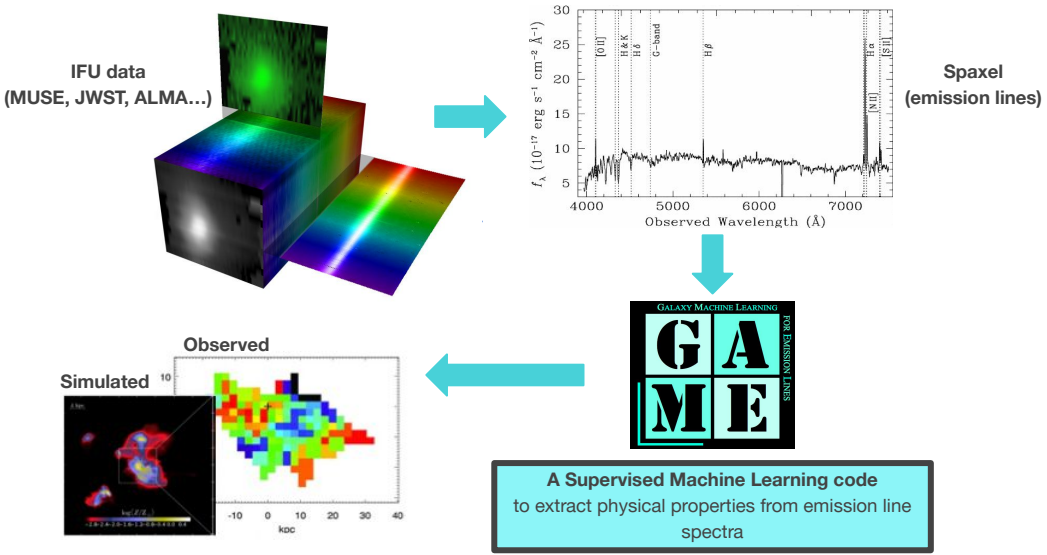






Case study: Image recognition

ML automatically understands the content of images and associate relevant keywords. Deep Convolutional Neural Networks spot faces even if they are partially hidden/upside down.



Case study: GAME

"GAlaxy Machine learning for Emission lines": a code used for astrophysical applications able to infer the physical properties of galaxies from spectra (Ucci et al. 2017, 2018, 2019).

online at: game.sns.it

Case study: Amazon Go

Amazon Go is a store with no checkout required. An experience made possible by computer vision and sensor fusion, combined with deep learning technologies.

The “just walk out technology” automatically detects when products are taken from or returned to the shelves and keeps track of them in a virtual cart.

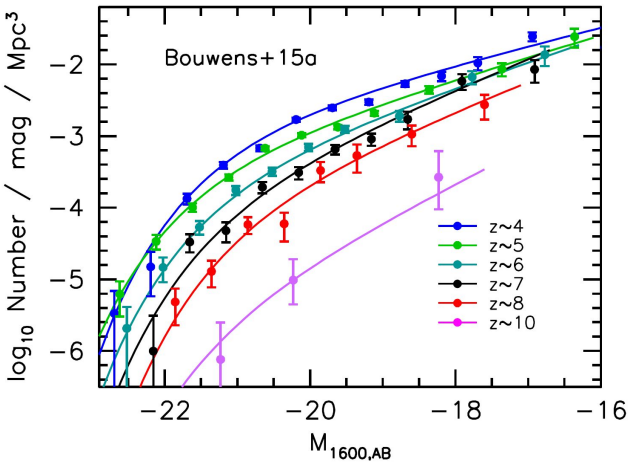
Why Machine Learning?

Why Machine Learning?

ML is useful when we **do not have a simple and clear algorithmic and/or an analytic representation of our problem**

Why Machine Learning?

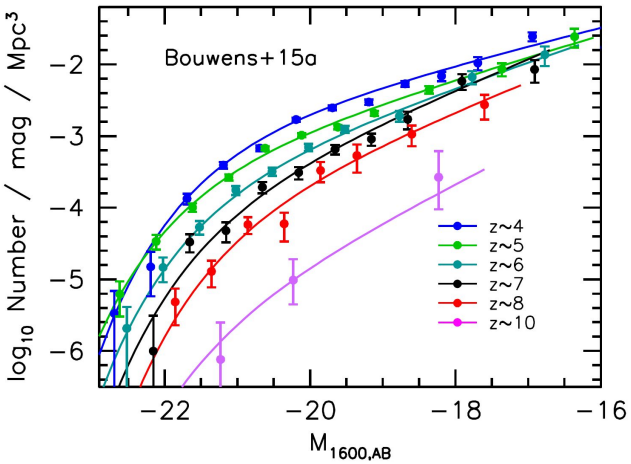
ML is useful when we **do not have a simple and clear algorithmic and/or an analytic representation of our problem**



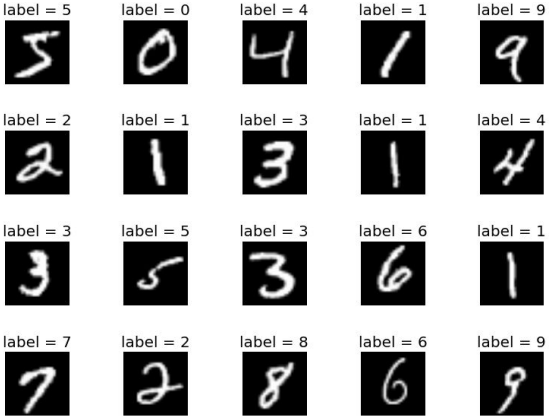
luminosity function
representation: Schechter function

Why Machine Learning?

ML is useful when we **do not have a simple and clear algorithmic and/or an analytic representation of our problem**



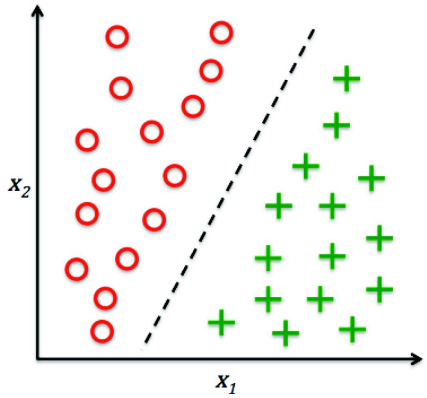
luminosity function
representation: Schechter function



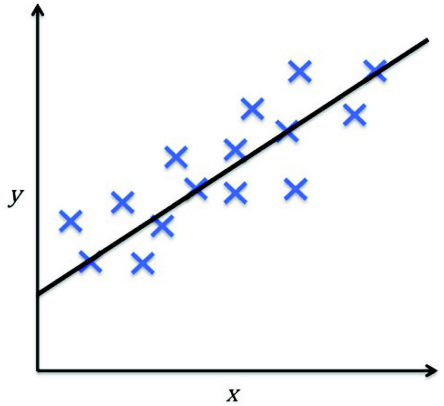
handwritten digits
representation: ???

Supervised Learning

ML algorithms can figure out how to perform tasks by generalizing from examples. The main goal in Supervised Learning is to learn a model from **labeled data** that allows us to make predictions about unseen or future data.

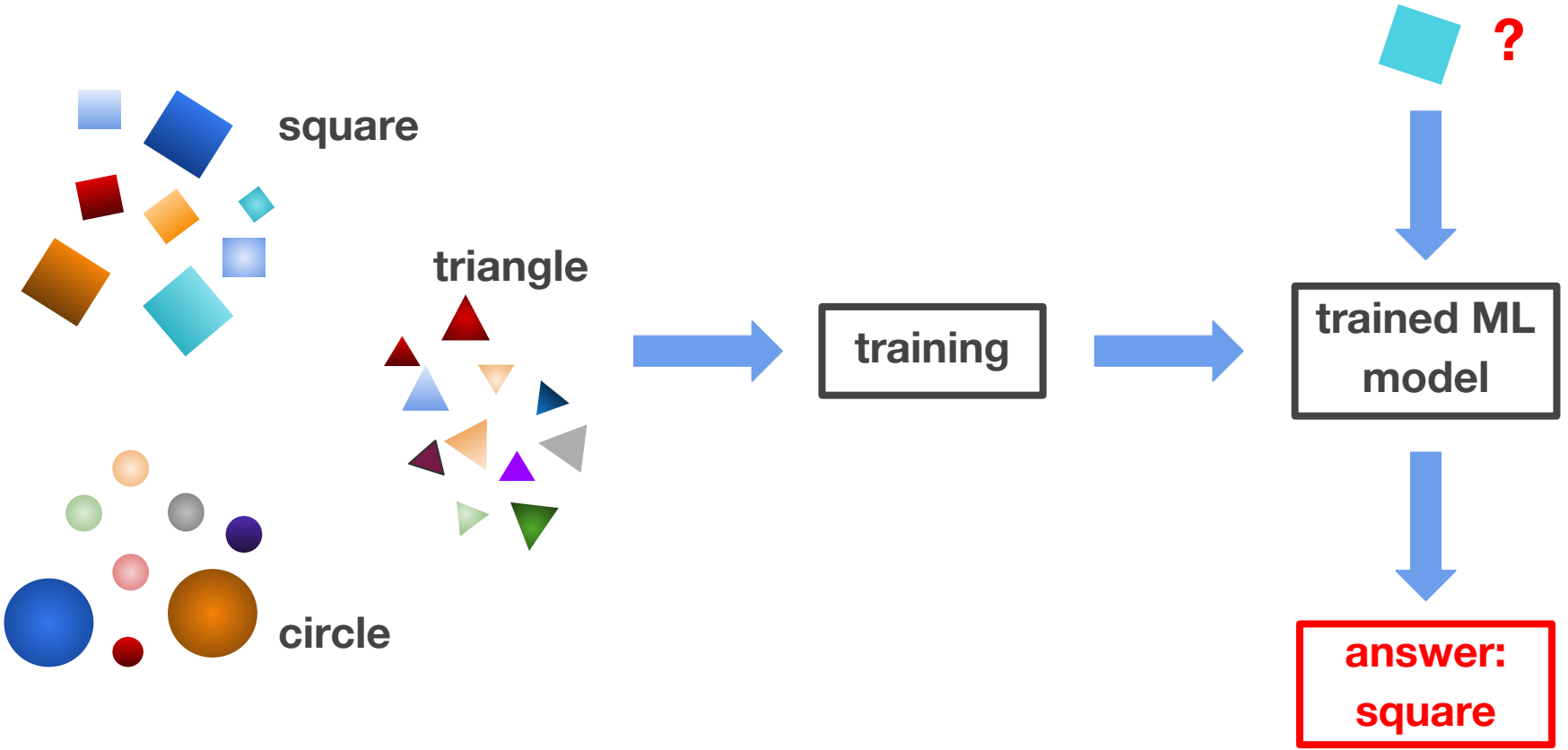


classification for predicting class labels



regression for predicting continuous outcomes

Supervised Learning

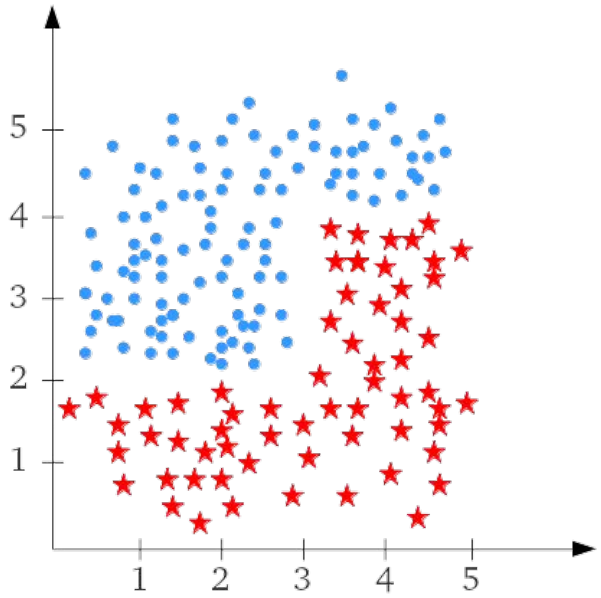


Supervised Learning: Decision Trees

Decision Trees recursively partition the input feature space into an increasing number of “leaves”. Each leaf represents a value for the target variable.

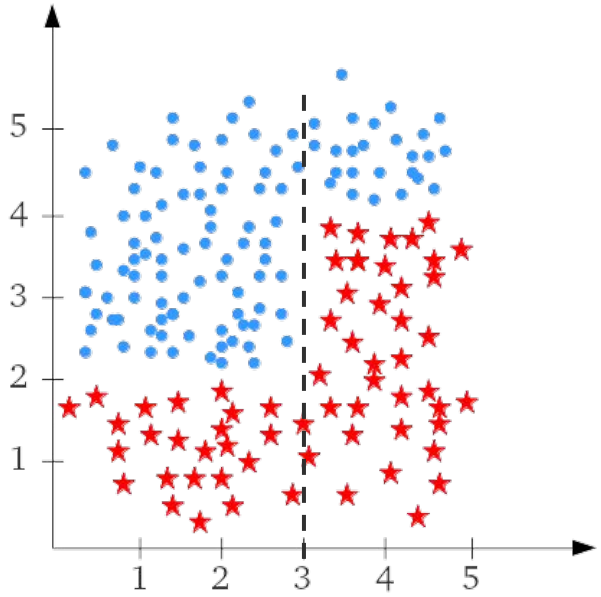
Supervised Learning: Decision Trees

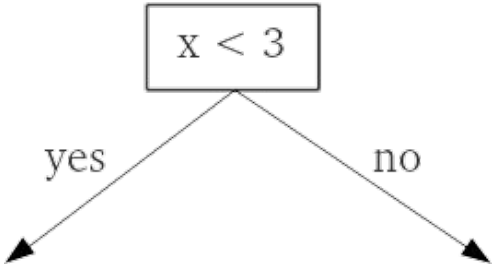
Decision Trees recursively partition the input feature space into an increasing number of “leaves”. Each leaf represents a value for the target variable.



Supervised Learning: Decision Trees

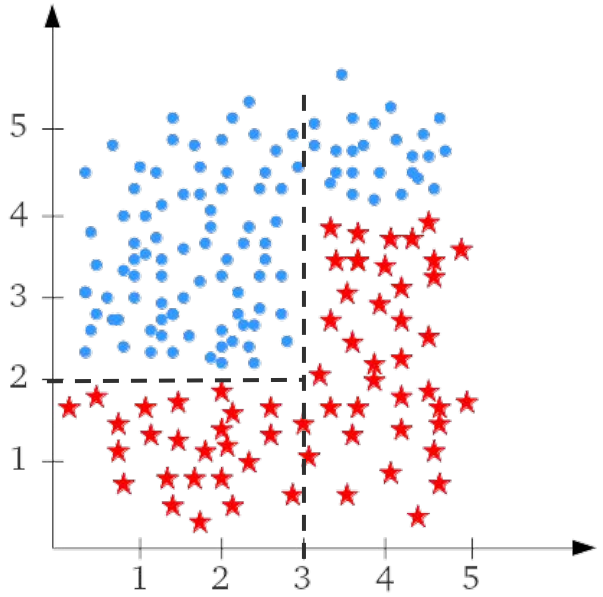
Decision Trees recursively partition the input feature space into an increasing number of “leaves”. Each leaf represents a value for the target variable.

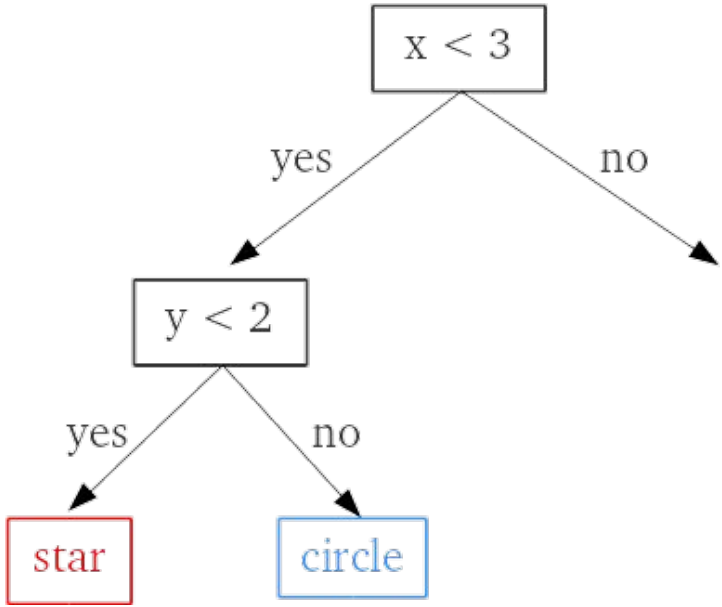




Supervised Learning: Decision Trees

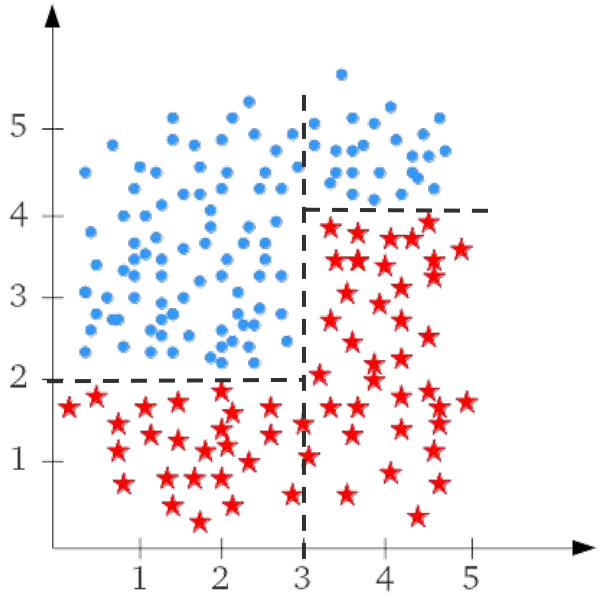
Decision Trees recursively partition the input feature space into an increasing number of “leaves”. Each leaf represents a value for the target variable.

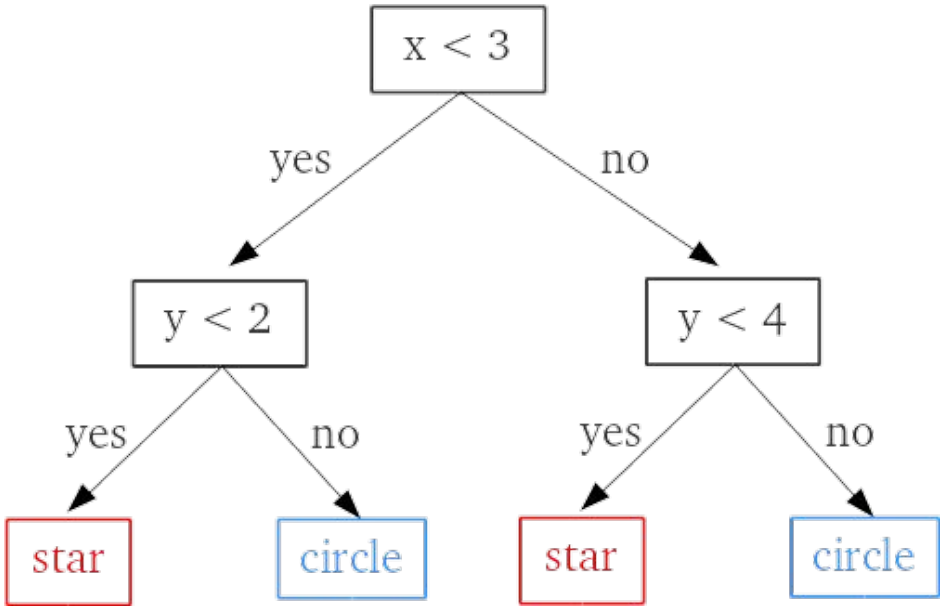




Supervised Learning: Decision Trees

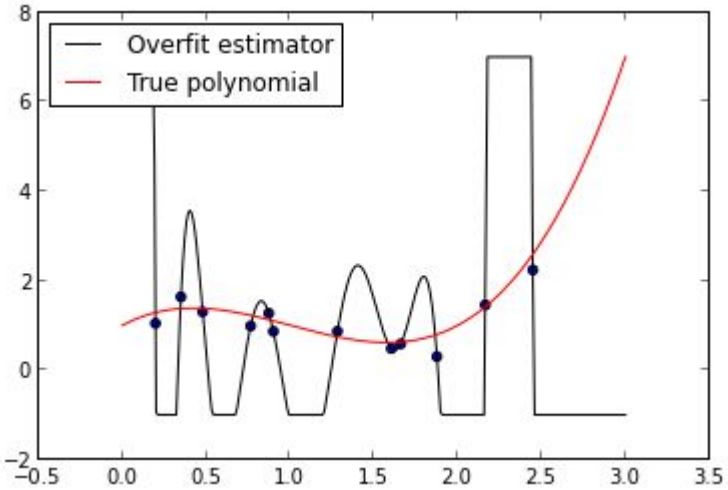
Decision Trees recursively partition the input feature space into an increasing number of “leaves”. Each leaf represents a value for the target variable.





Supervised Learning: Decision Trees

Decision Trees find final decision boundaries automatically based on the data.



PROs

valid for a **large range of applications**

extremely **fast and simple**

easy implementation

CONs

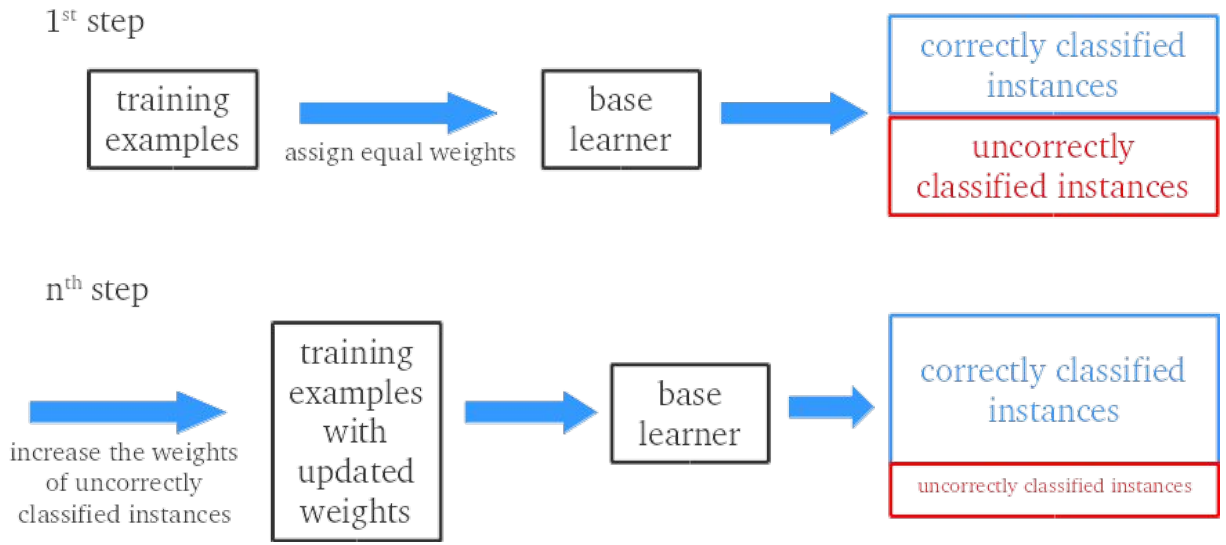
they are prone to **overfitting**

Ensemble methods

Many base learning algorithms **(Trees)** can be combined into an **“ensemble learner” (Forest)** which can achieve better results

Ensemble method: AdaBoost

AdaBoost method adds decision trees sequentially to generate a forest



In each iteration we apply the base learners on the training set with updated weights

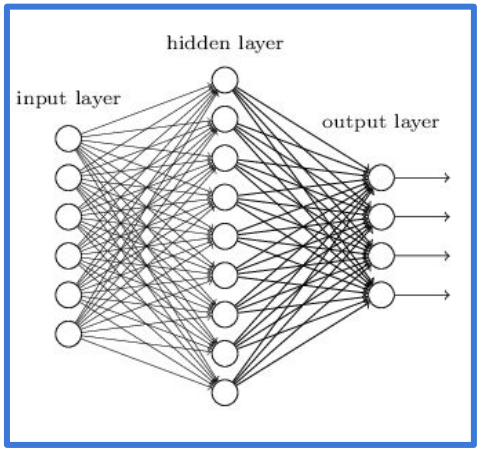
The **final model** is the weighted sum of the n learners.


Deep Learning

Deep Learning

Deep learning could be defined as a set of algorithms that were developed to train in the most efficient way artificial neural networks with many layers.

Neural network

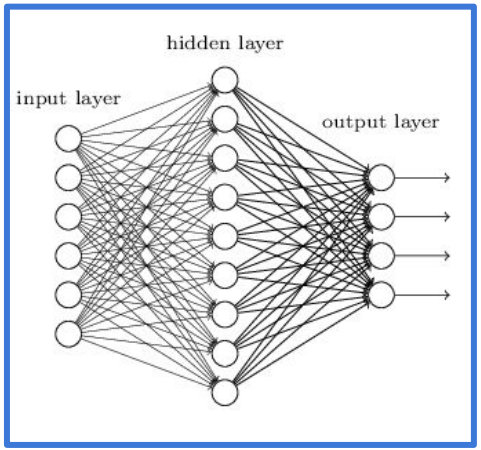


Adapted from: http://neuralnetworksanddeeplearning.com/chap5.html

Deep Learning

“Simple” neural networks (i.e. with one hidden layer) can easily **classify handwritten digits with an accuracy better than 98%.**

Neural network

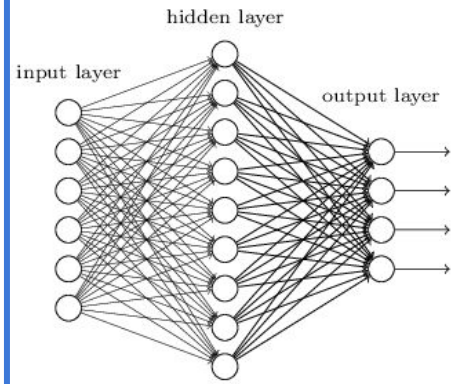


Adapted from: http://neuralnetworksanddeeplearning.com/chap5.html

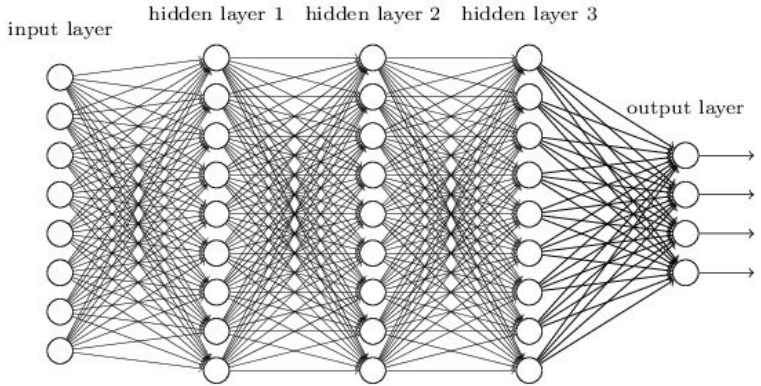
Deep Learning

Nonetheless, networks with **more hidden layers** can be even more powerful.

Neural network



“Deep” neural network



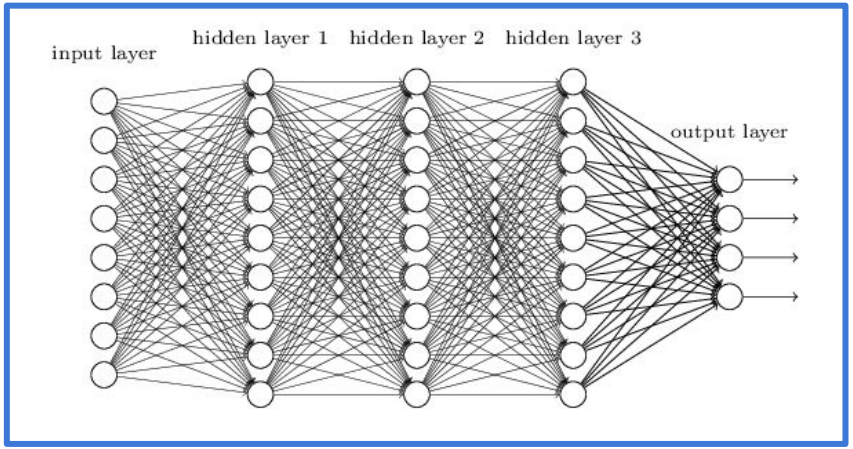
Adapted from: http://neuralnetworksanddeeplearning.com/chap5.html

Deep Learning

Deep networks use the intermediate layers to **build up multiple levels of abstraction.**

For example, in visual pattern recognition, the first layer might learn to recognize edges, the neurons in the second layer could learn to recognize more complex shapes, the third layer would then recognize still more complex shapes, and so on.

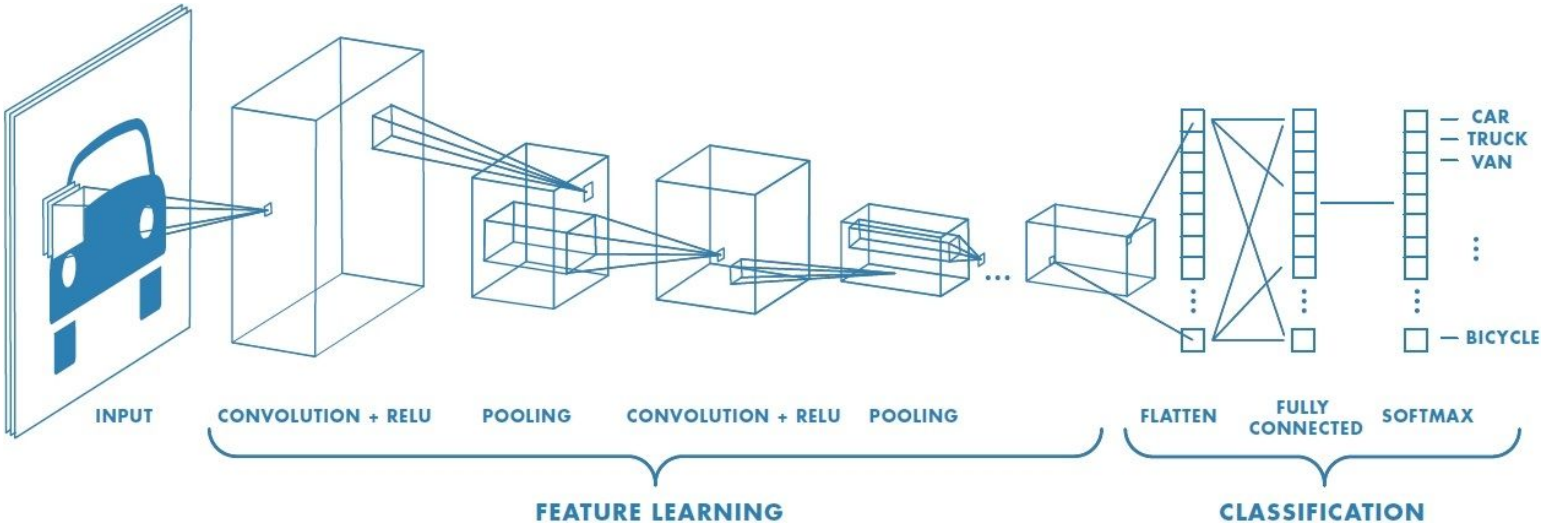
Deep neural network



Adapted from: http://neuralnetworksanddeeplearning.com/chap5.html

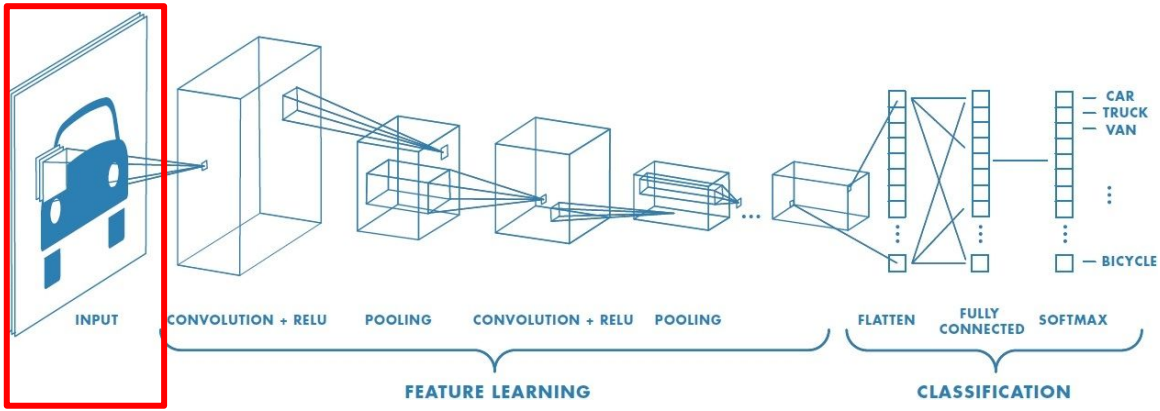
A simple Deep Learning Net for image recognition

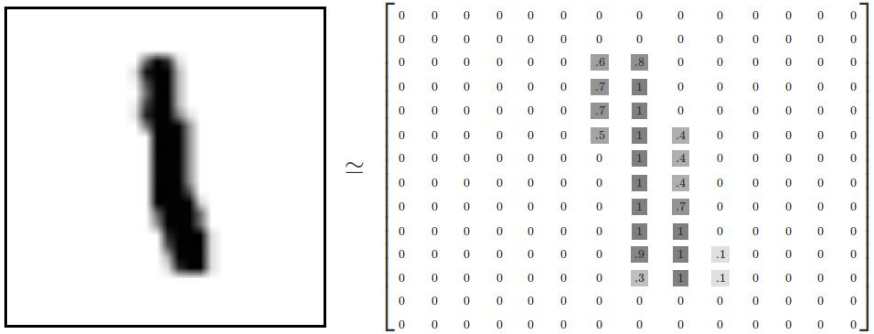
Convolutional Neural Networks

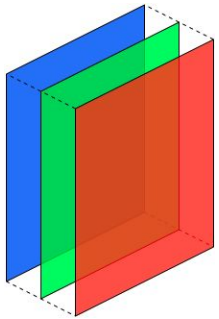


Adapted from: https://www.mathworks.com/

Convolutional Neural Networks

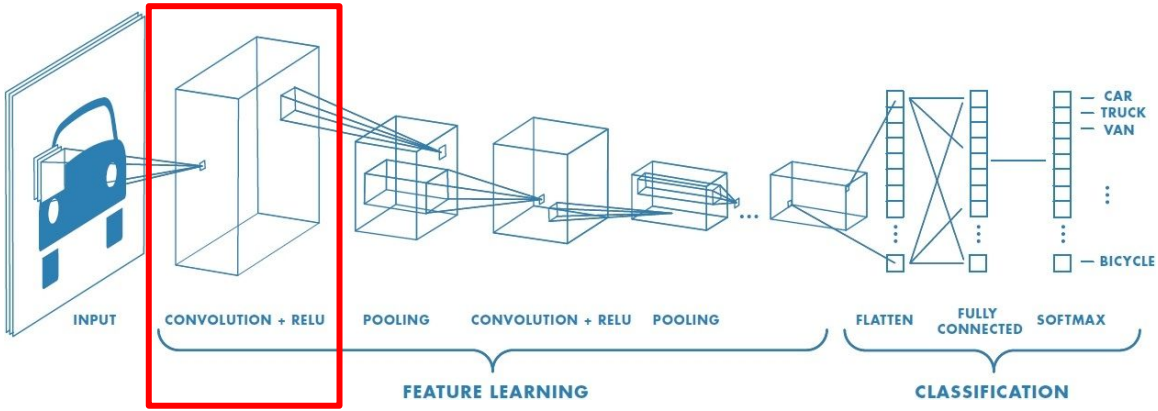






Images are matrices of pixel values, eventually one for each channel (i.e. RGB)

Convolutional Neural Networks



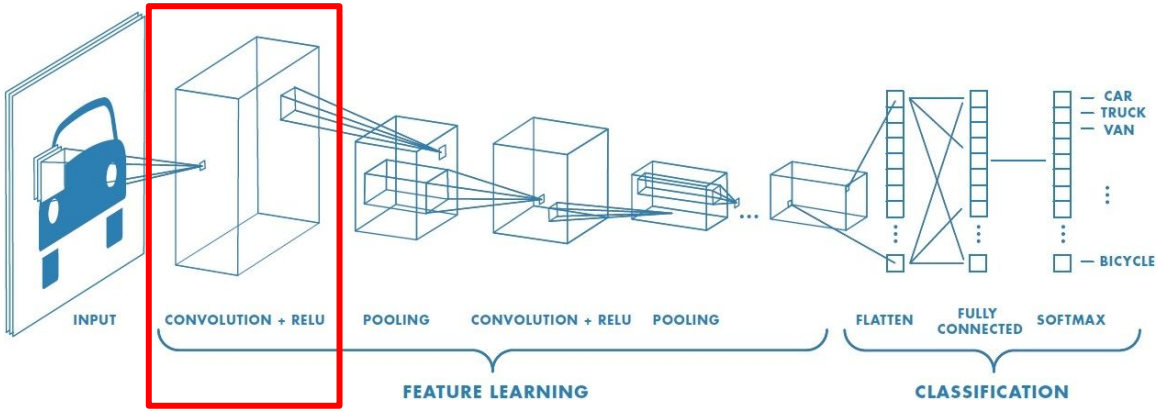
$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

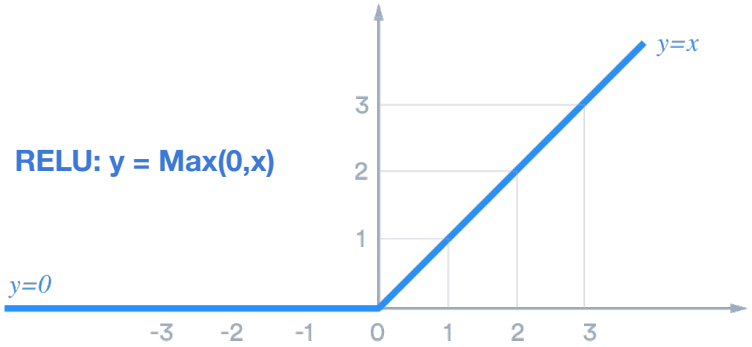
Image

4		

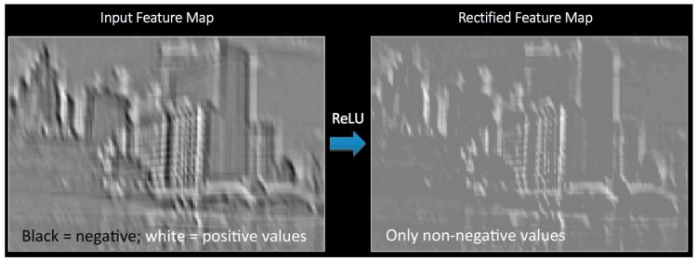
Convolved Feature

Convolutional Neural Networks

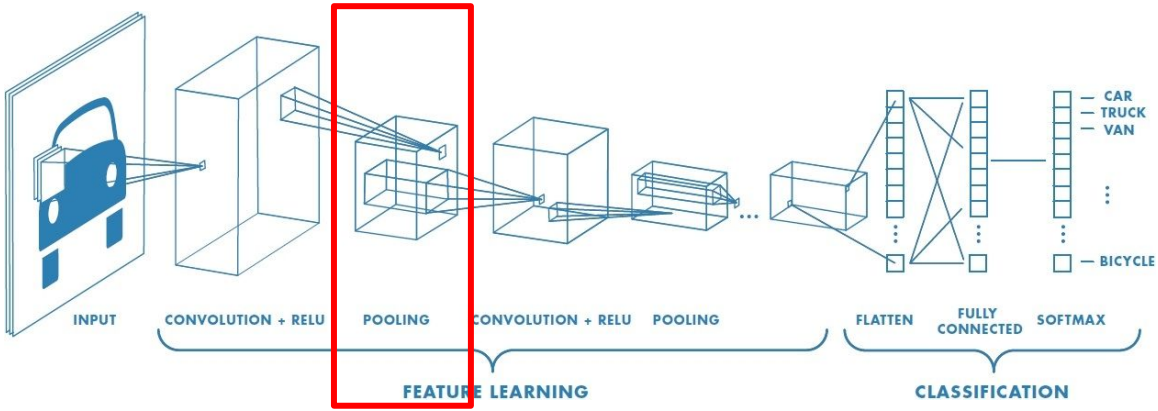




The purpose of RELU is to introduce non-linearity (given that convolution is a linear operation, i.e. matrix multiplication)



Convolutional Neural Networks



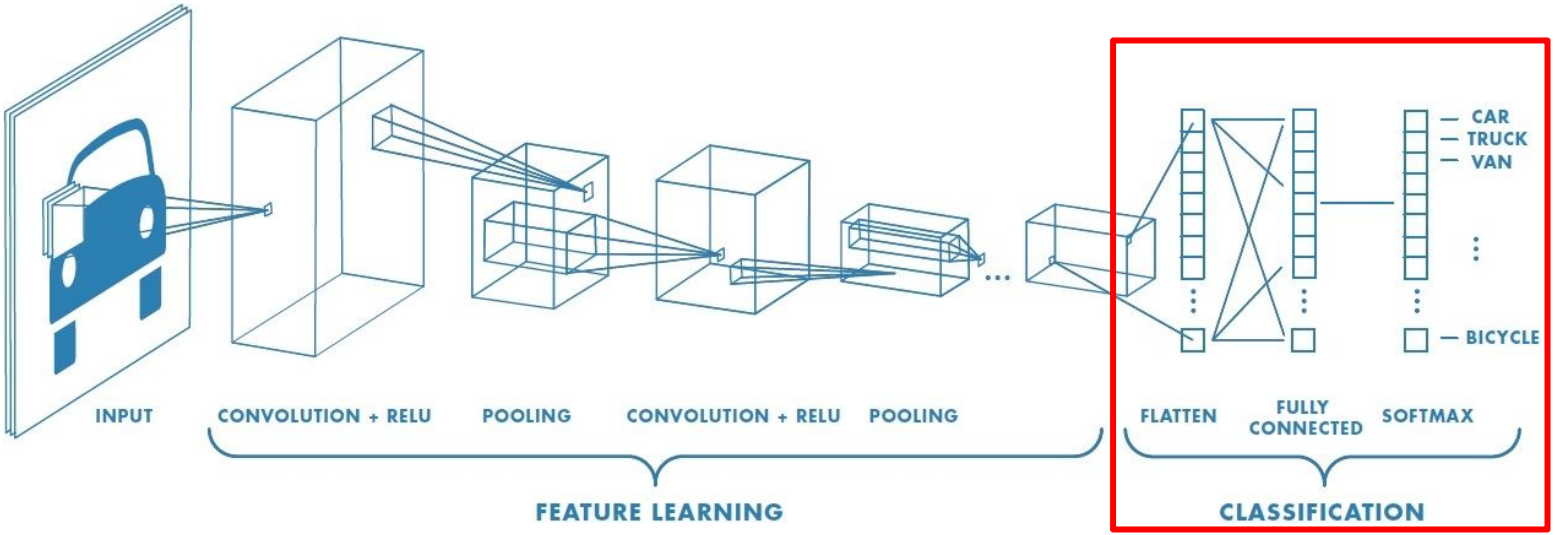
12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

2×2 Max-Pool →

20	30
112	37

Pooling (also called subsampling or downsampling) downsample, i.e. reduces the dimensionality of each feature map.

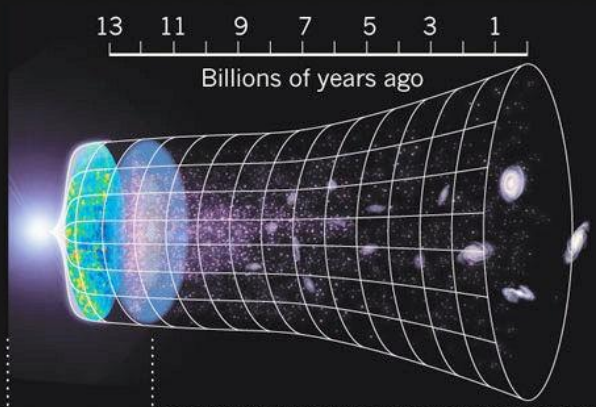
Convolutional Neural Networks



Putting all together with a fully connected neural net at the end.


A “not-so-simple” Deep Learning Net for
Cosmological applications

My research focuses on the Interstellar Medium of galaxies. The aim is to understand the **“high redshift Universe”** (z > 6, age < 1 Gyr)



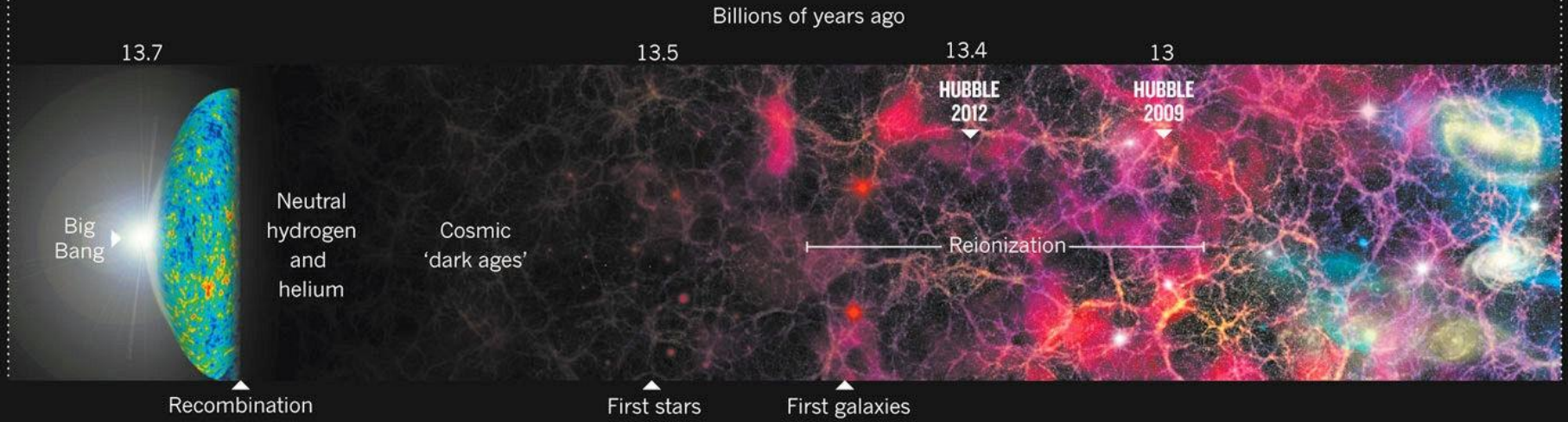
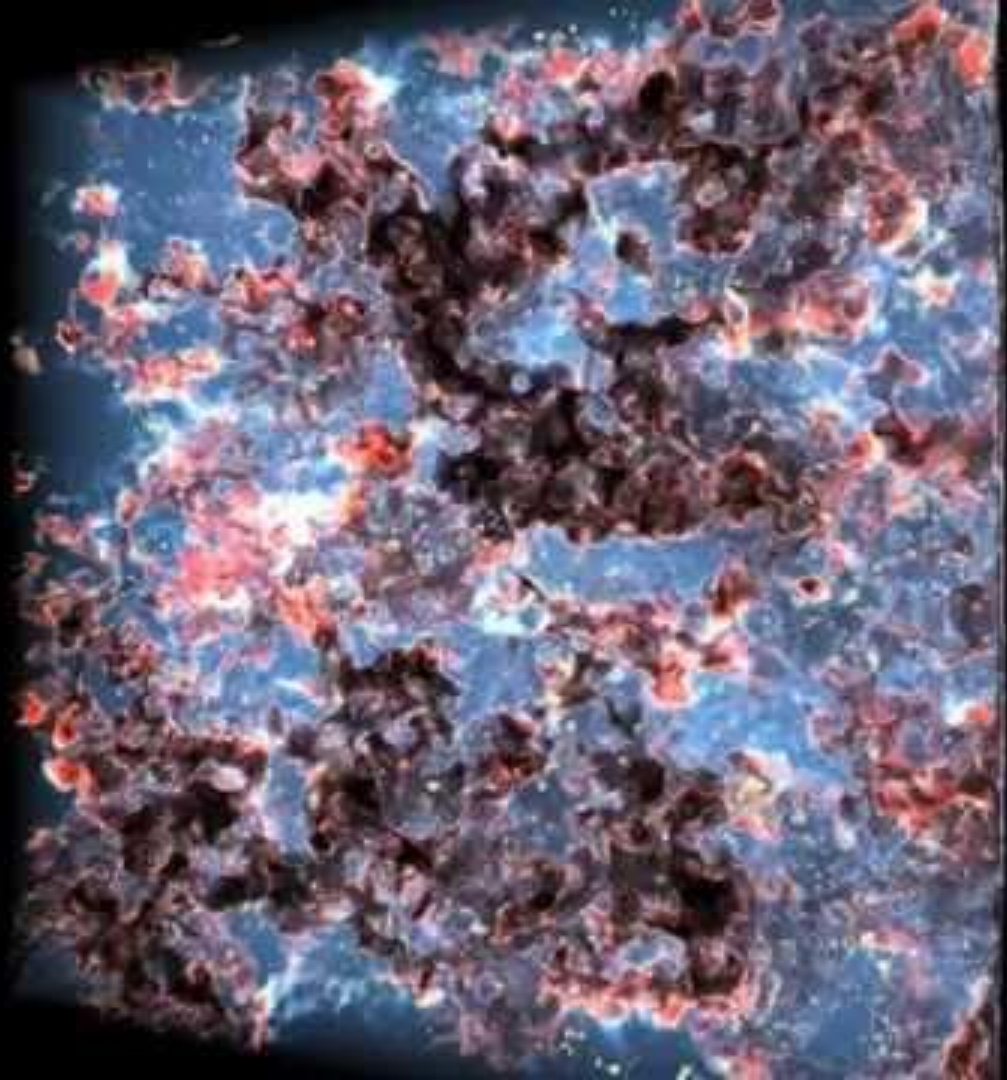
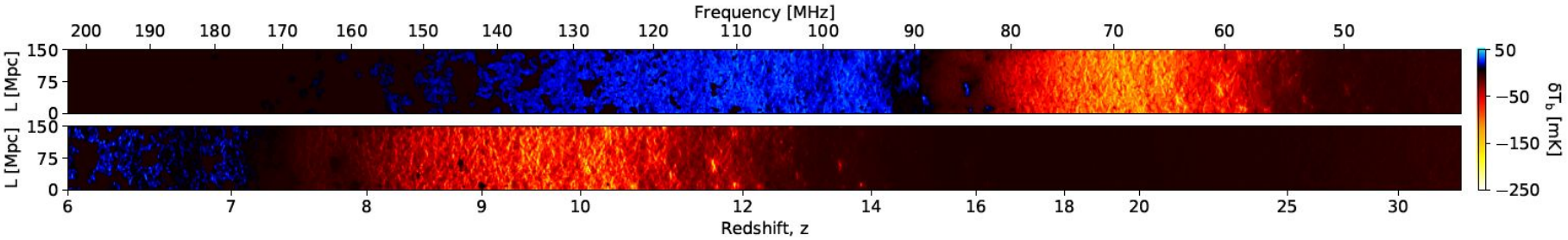
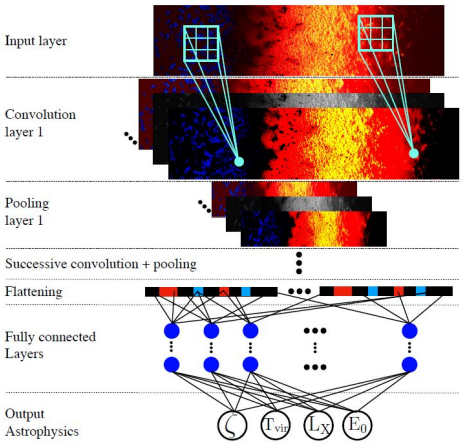


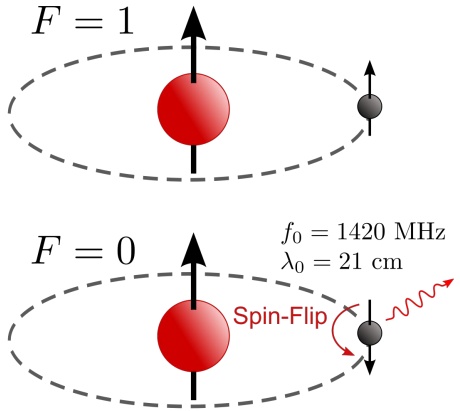
Illustration: Nik Spencer; sources: NASA/WMAP Science Team; R. Ellis (Caltech)



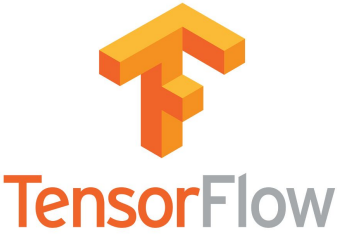
Deep learning 21-cm images of the Cosmic Dawn (Gillet et al. 2019)



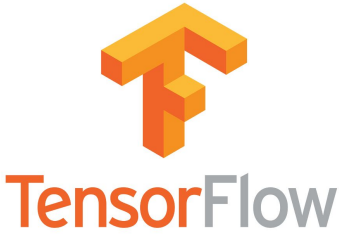




Most common libraries, programming languages



Most common libraries, programming languages




GAME
“GAlaxy Machine learning for Emission lines”

Interstellar Medium (ISM)

Main (baryonic) ingredients of galaxies:

- stars

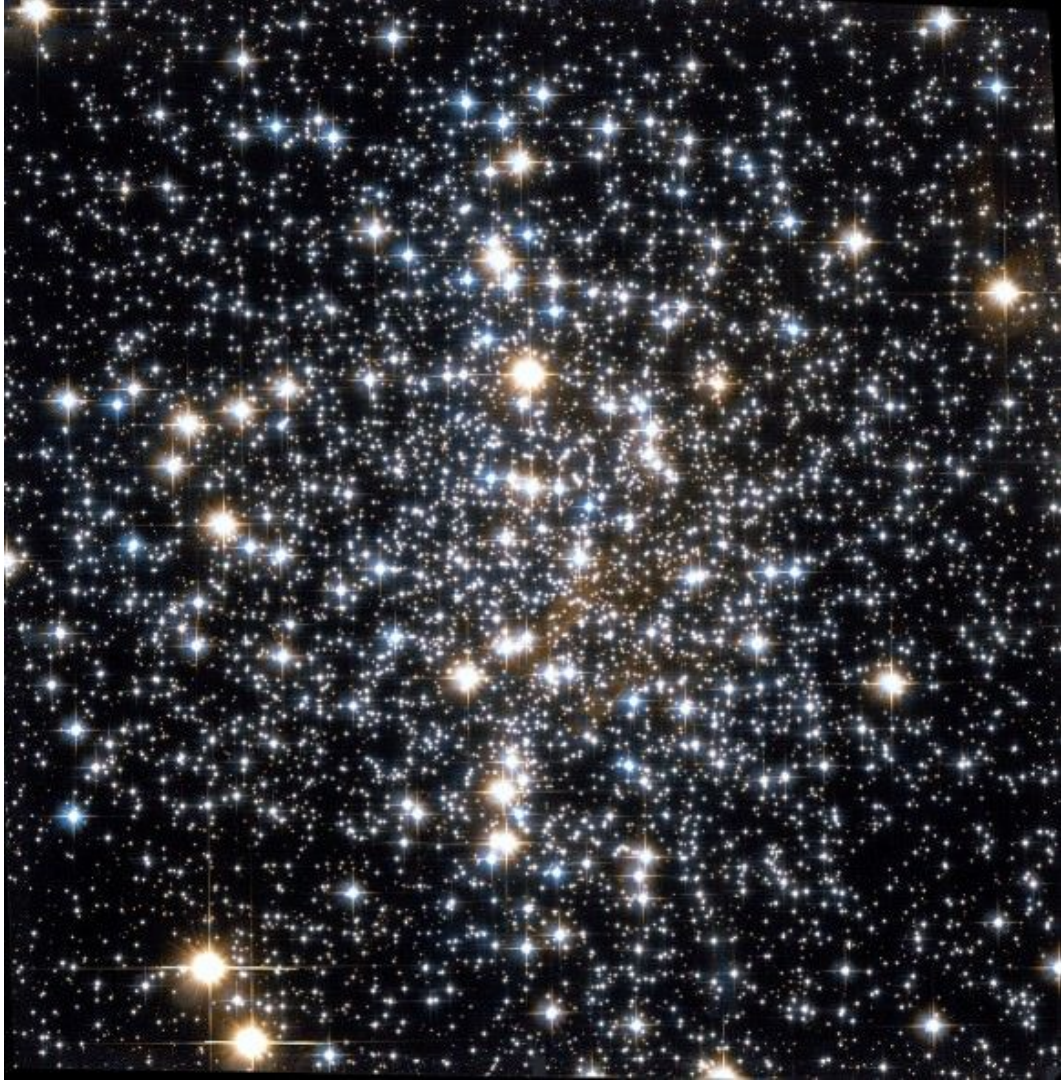


Illustration: Messier 4 by HST, Credit NASA/STScI/WikiSky

Interstellar Medium (ISM)

Main (baryonic) ingredients of galaxies:

- stars
- gas

Illustration: Eagle Nebula (Messier 16), Credit NASA

Interstellar Medium (ISM)

Main (baryonic) ingredients of galaxies:

- stars
- gas
- dust

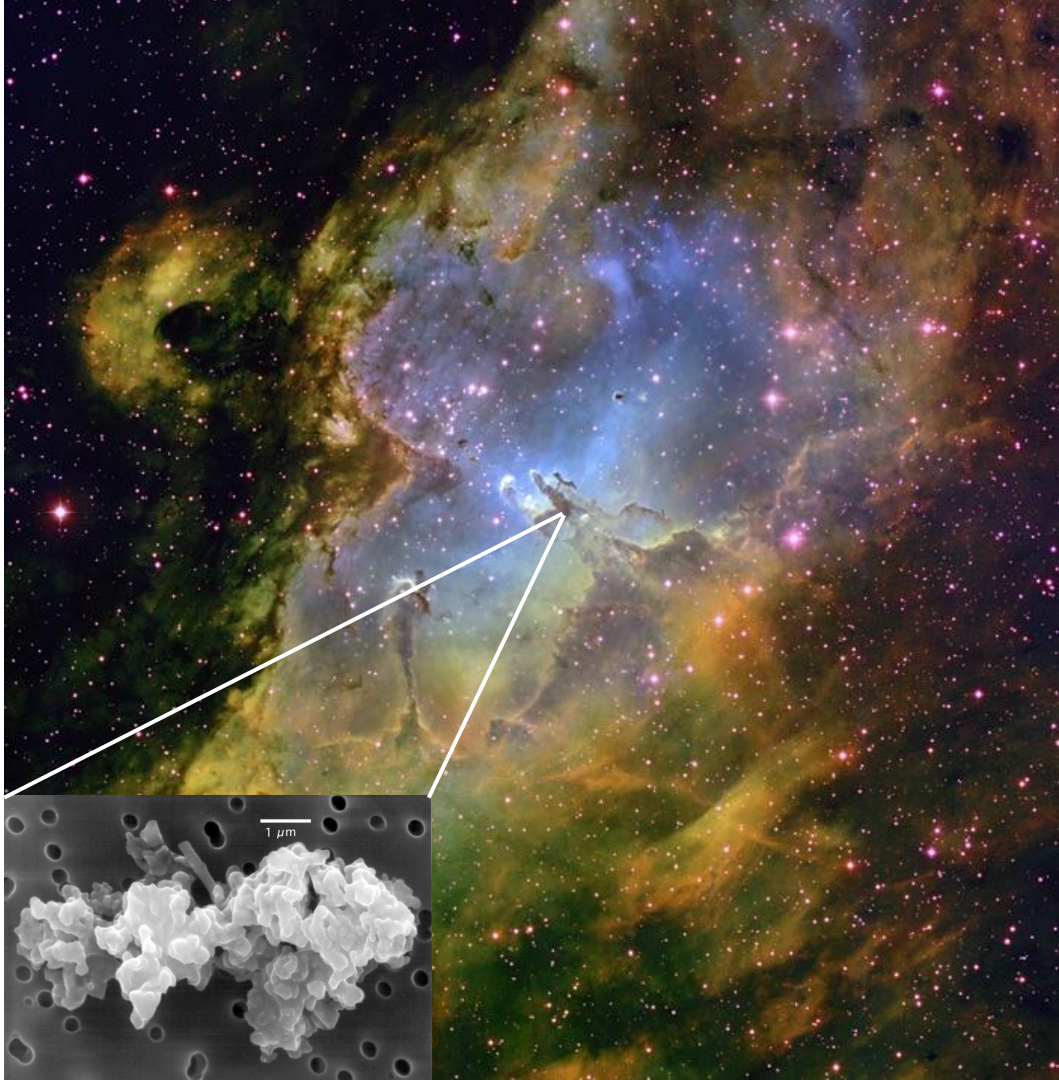


Illustration: Eagle Nebula (Messier 16), Credit NASA

Interstellar Medium (ISM)

Main (baryonic) ingredients of galaxies:

- stars
- gas
- dust
- cosmic rays
- e.m. radiation
- magnetic field

ISM

Illustration: Eagle Nebula (Messier 16), Credit NASA

Typical physical conditions

density of an HII region: 100 cm^{-3}

size: 10 pc ~ 3 x 10^{19} cm

Illustration: Eagle Nebula (Messier 16), Credit NASA

Typical physical conditions

density of an HII region: 100 cm^{-3}

density of the air (sea level): 10^{19} cm^{-3}

size: 10 pc ~ 3 x 10^{19} cm

radius of our Solar System: 4.5 x 10^{12} cm

Illustration: Eagle Nebula (Messier 16), Credit NASA

Typical physical conditions

density of an HII region: 100 cm^{-3}

density of the air (sea level): 10^{19} cm^{-3}

size: 10 pc ~ 3 x 10^{19} cm

radius of our Solar System: 4.5 x 10^{12} cm

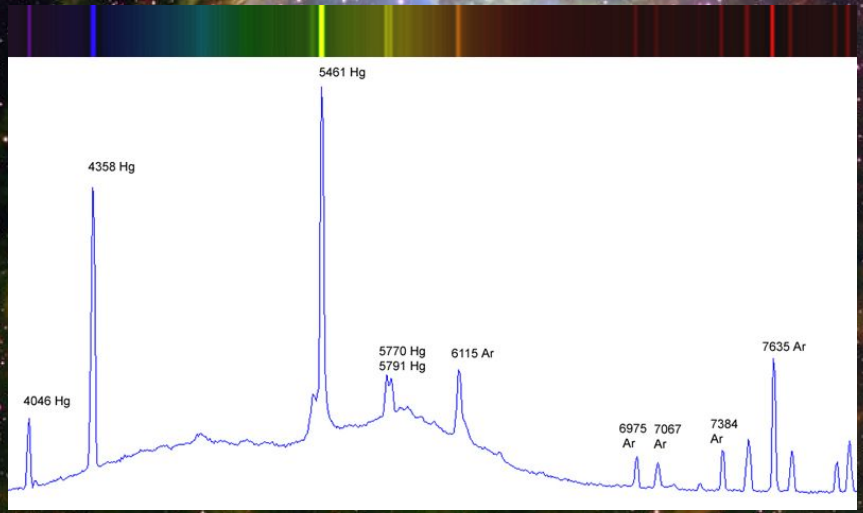


Illustration: Eagle Nebula (Messier 16), Credit NASA

Typical physical conditions

density of an HII region: 100 cm^{-3}

density of the air (sea level): 10^{19} cm^{-3}

size: 10 pc ~ 3 x 10^{19} cm

radius of our Solar System: 4.5 x 10^{12} cm

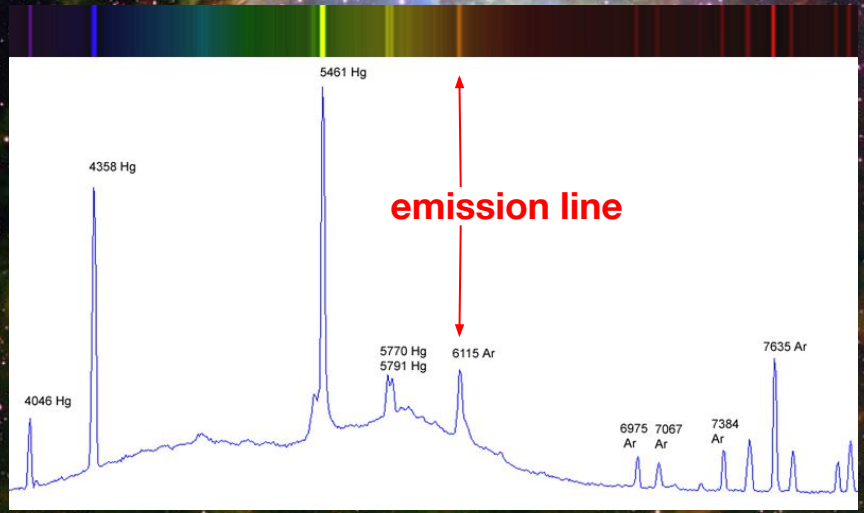


Illustration: Eagle Nebula (Messier 16), Credit NASA

Typical physical conditions

density of an HII region: 100 cm^{-3}

density of the air (sea level): 10^{19} cm^{-3}

size: 10 pc ~ 3 x 10^{19} cm

radius of our Solar System: 4.5 x 10^{12} cm

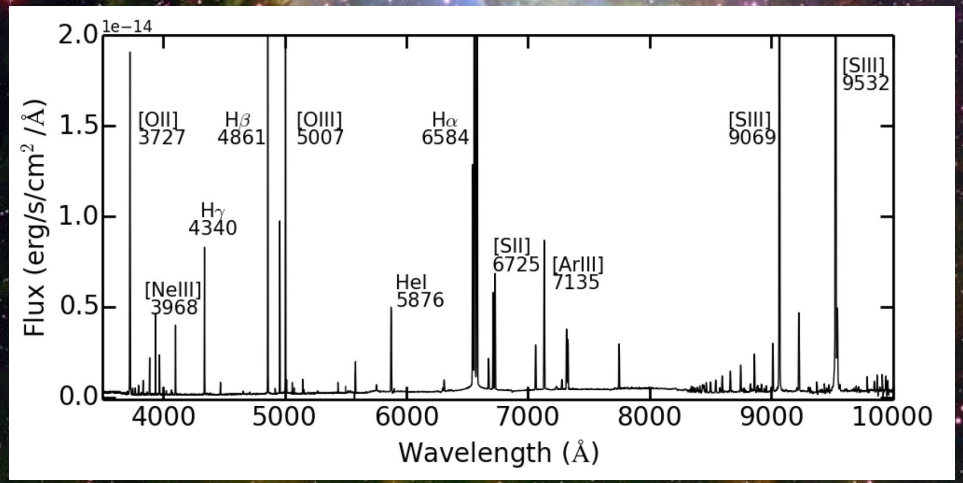
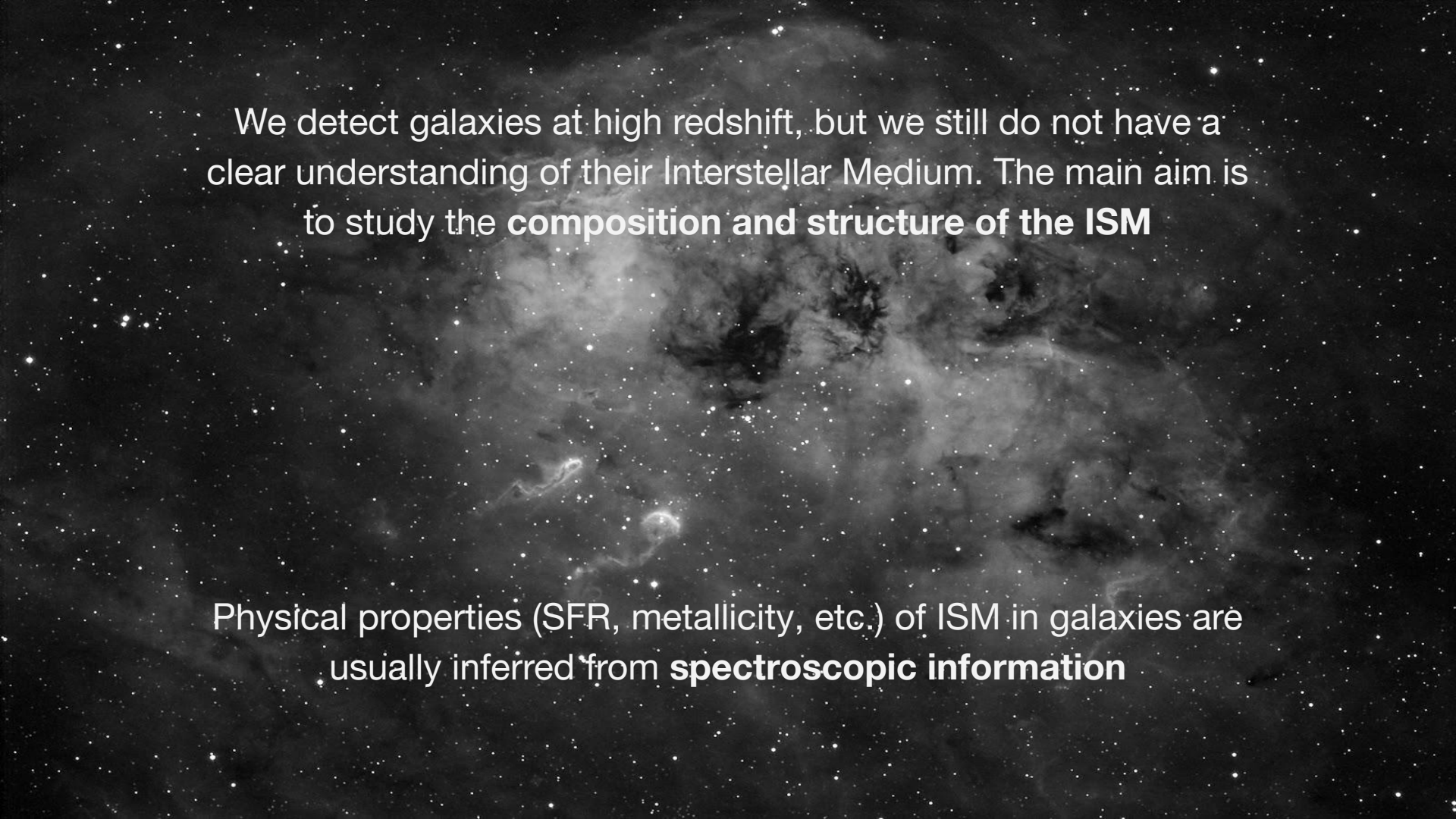


Illustration: Eagle Nebula (Messier 16), Credit NASA



We detect galaxies at high redshift, but we still do not have a clear understanding of their Interstellar Medium. The main aim is to study the **composition and structure of the ISM**

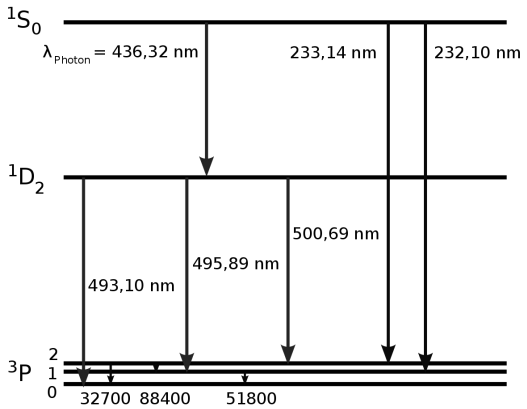
Physical properties (SFR, metallicity, etc.) of ISM in galaxies are usually inferred from **spectroscopic information**

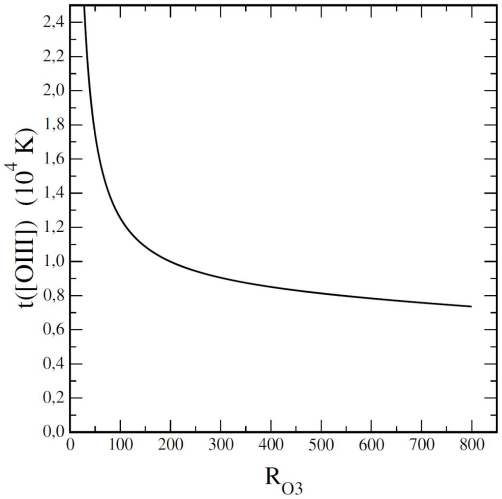
Direct method (electron temperature)

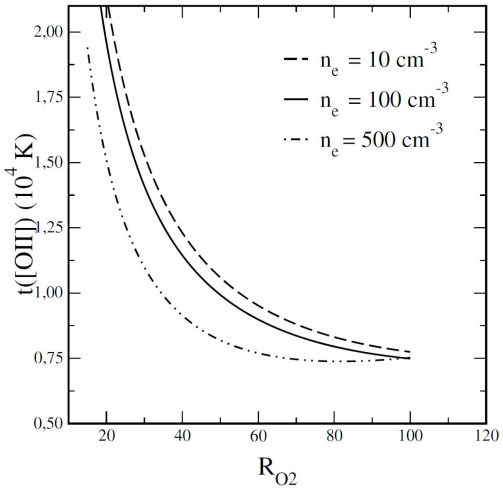
(auroral to nebular line ratios, Pérez-Montero 2017)

R_{O3} = ([OIII] λ4959 + [OIII] λ5007) / [OIII] λ4363

R_{O2} = ([OII] λ3726 + [OII] λ3729) / ([OII] λ7319 + [OII] λ7330)







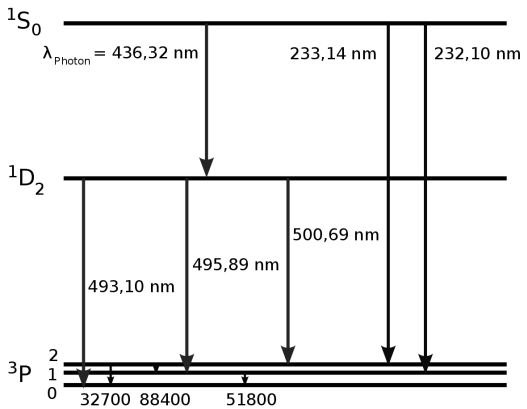
Direct method (electron temperature)

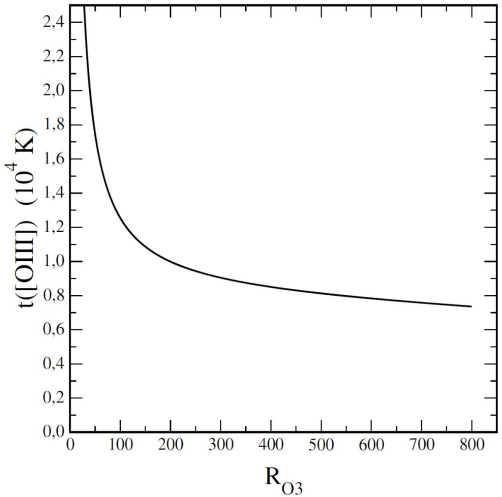
(auroral to nebular line ratios, Pérez-Montero 2017)

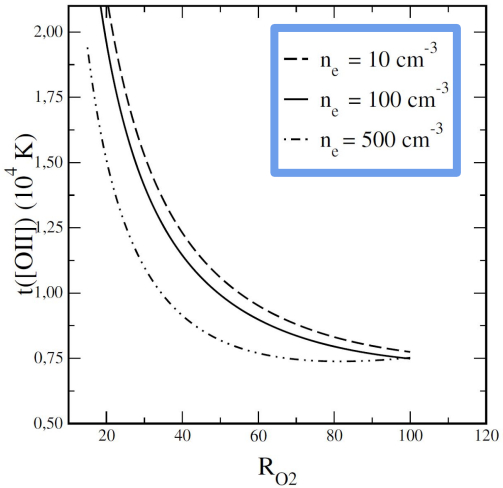
R_{O3} = ([OIII] λ4959 + [OIII] λ5007) / [OIII] λ4363

R_{O2} = ([OII] λ3726 + [OII] λ3729) / ([OII] λ7319 + [OII] λ7330)

relations between the nebular-to-auroral line ratios and the electron temperature as a function of the electron density







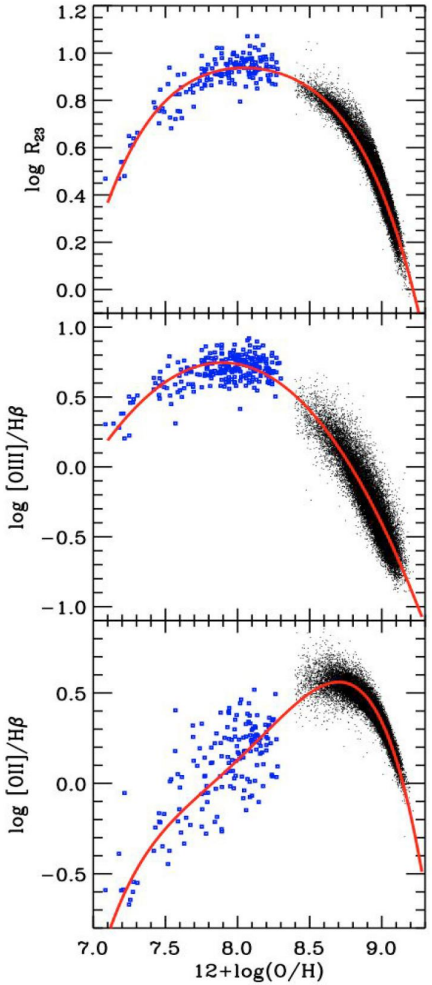
Empirical calibrations

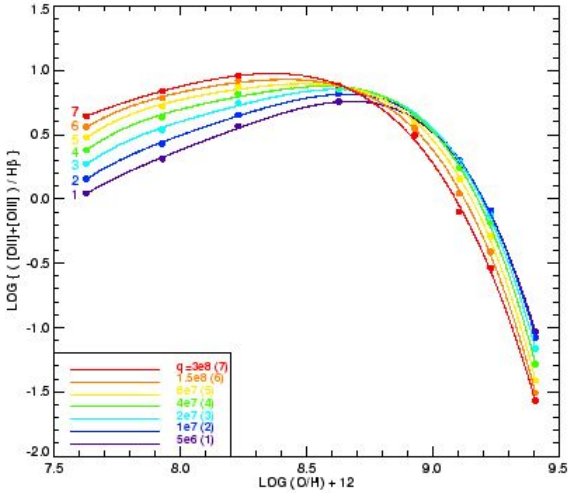
(Pagel et al. 1979, Vilchez & Esteban 1996, Pettini & Pagel 2004,
Maiolino et al. 2008, Nagao et al. 2011, Marino et al. 2013, Curti et al. 2017)

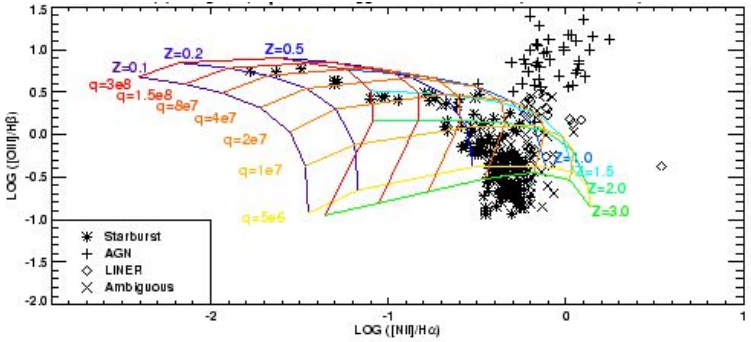
R_{23} = ([OII] λ3727 + [OIII] λ4959 + [OIII] λ5007) / Hβ

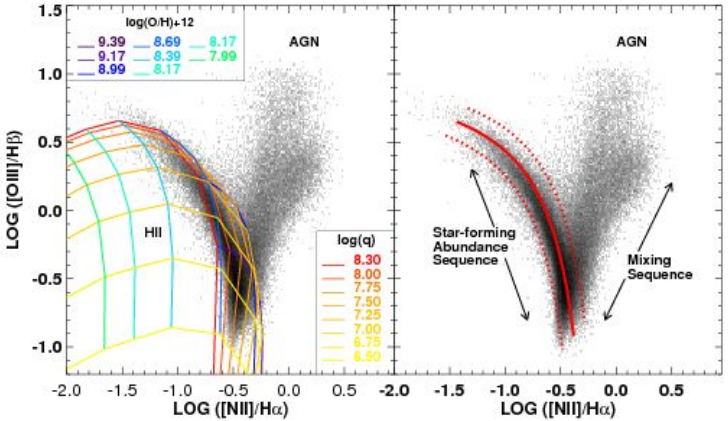
N_2 = [NII] λ6583 / Hα

R = ([OIII] λ51.80 μm + [OIII] λ88.33 μm) / [NIII] λ57.21 μm









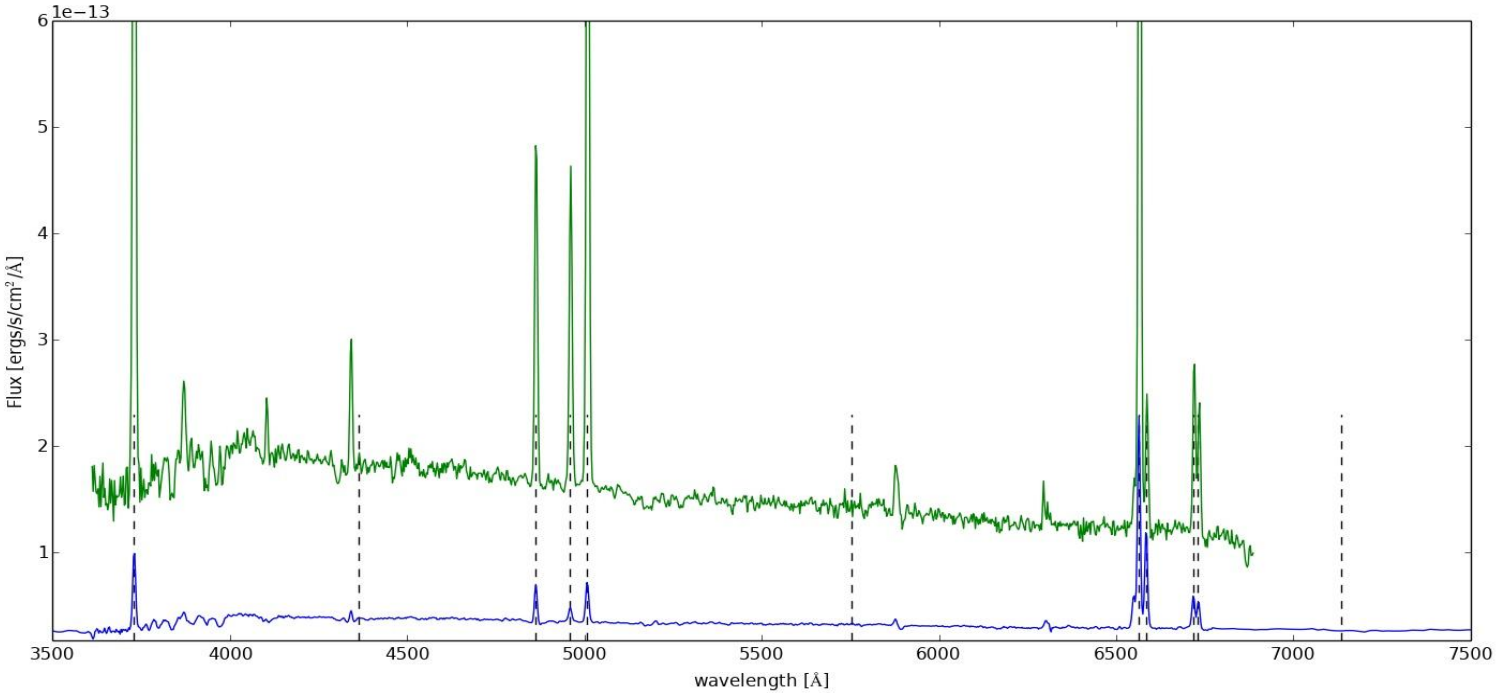
Comparison of theoretical spectra from a grid of photoionization models

(McGaugh 1991, Zaritsky et al. 1994, Kewley & Dopita 2002, Kobulnicky & Kewley 2004, Tremonti et al. 2004, Kewley & Ellison 2008, Dopita et al. 2016)

Numerical codes

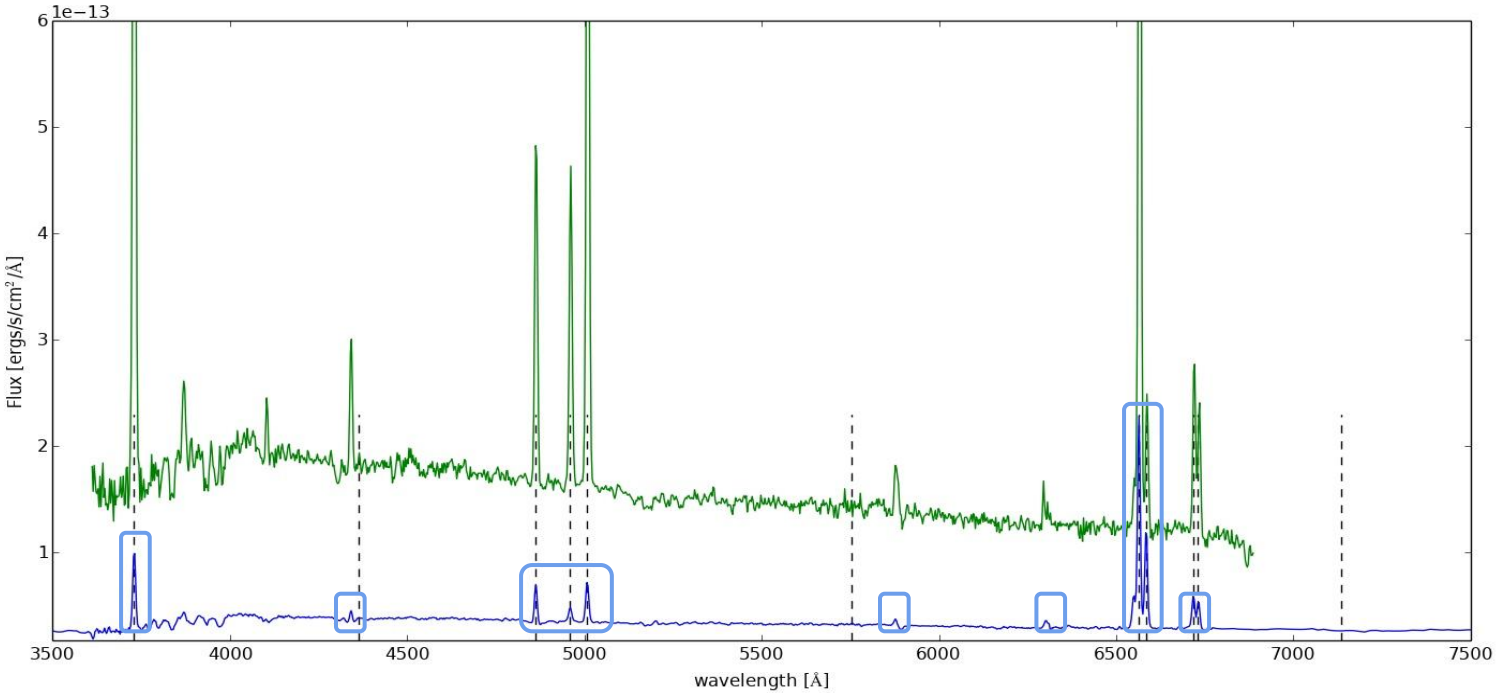
(*IZI* Blanc et al. 2015, *pyqz* Dopita et al. 2013,
HII-CHI-mistry Pérez-Montero 2014,
BOND Vale Asari et al. 2016)

GAME (**GA**laxy **M**achine learning for **E**mission lines, Ucci et al. 2017, 2018) is a new, fast code able to reconstruct the physical properties of the ISM in (distant) galaxies



GAME (**GA**laxy **M**achine learning for **E**mission lines, Ucci et al. 2017, 2018) is a new, fast code able to reconstruct the physical properties of the ISM in (distant) galaxies

by using all the available information encrypted in spectra



Inspiration...

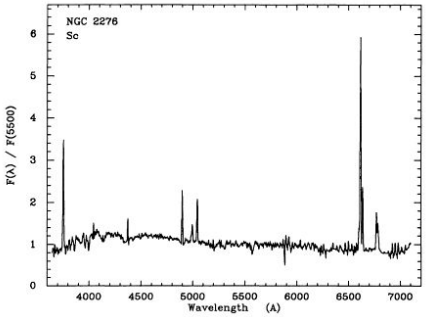
Inspiration... Shazam!

you are in a shop and you like the music you are hearing: **tap the button**

it there is a **match** you are then given the **metadata of the audio** (title, lyrics, video, artist…)

shazam analyzes the captured sound by creating the fingerprint of the audio, **it starts the search for matches in the database** (Wang 2003)

A “Shazam for galaxies”

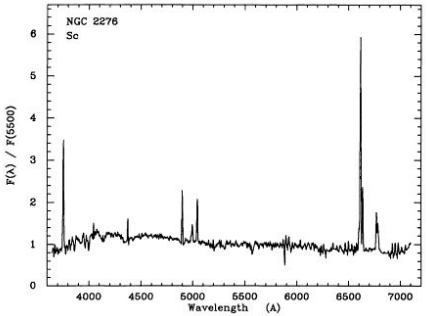


you have an observational/synthetic **spectrum of a galaxy**

you are then given the **“metadata” of the galaxy** (gas volume density, ionization state, metallicity…)

the algorithm works by **analyzing the emission line intensities**, it works on the models in the database

A “Shazam for galaxies”



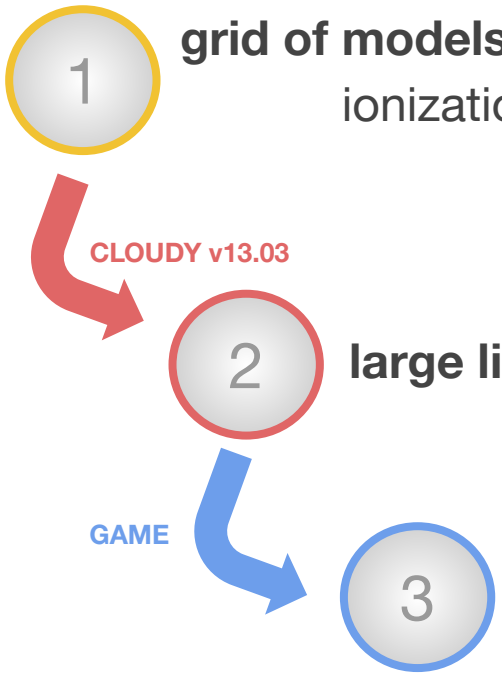
you have an observational/synthetic **spectrum of a galaxy**

you are then given the **“metadata” of the galaxy** (gas volume density, ionization state, metallicity…)

the algorithm works by **analyzing the emission line intensities**, it works on the models in the database

but...

Method (Ucci et al. 2017, 2018)



grid of models: metallicity, column density, ionization parameter, density

large library of **emission line intensities** (50,000 models)

Supervised Machine Learning (SML): physical properties are inferred from an input spectrum using emission line intensities as input features

1

grid of models: metallicity, column density, ionization parameter, density

Parameter	minimum	maximum
$\log(Z/Z_{\odot})$	-3.0	0.5
$\log(n/\mathrm{cm}^{-3})$	-3.0	5.0
$\log(U)$	-4.0	3.0
$\log(N_H/\mathrm{cm}^{-2})$	17.0	23.0

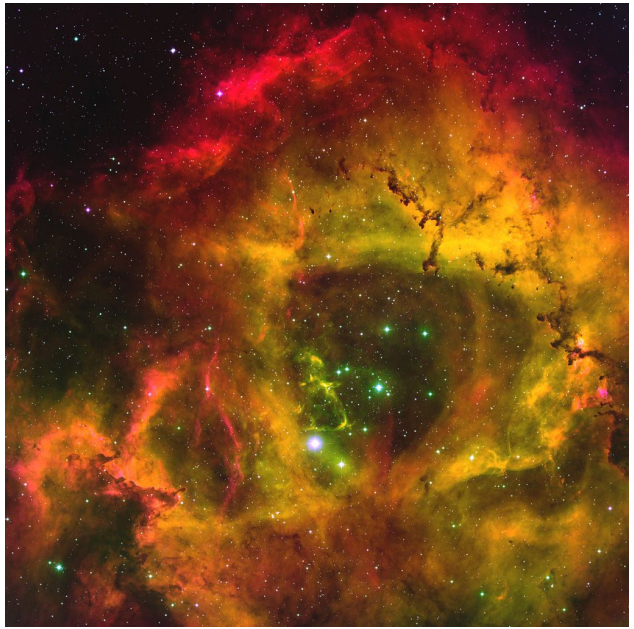
1 Photoionization models

- metallicity ($Z / Z_{\odot}$)
- ionization state (ionization parameter U or FUV flux in the Habing band 6 - 13.6 eV)

$$U = \frac{1}{4\pi R_S^2 nc} \int_{\nu_e}^{\infty} \frac{L_\nu}{h\nu} d\nu = \frac{F(H)}{4\pi R_S^2 nc}$$

- density (n / cm^{-3})
- column density (N_H / cm^{-2})

$$N_H = \int_0^R n ds$$



The Rosette nebula (NGC 2237)
(Credit WIYN and NOAO/AURA/NSF)

1

grid of models: metallicity, column density, ionization parameter, density

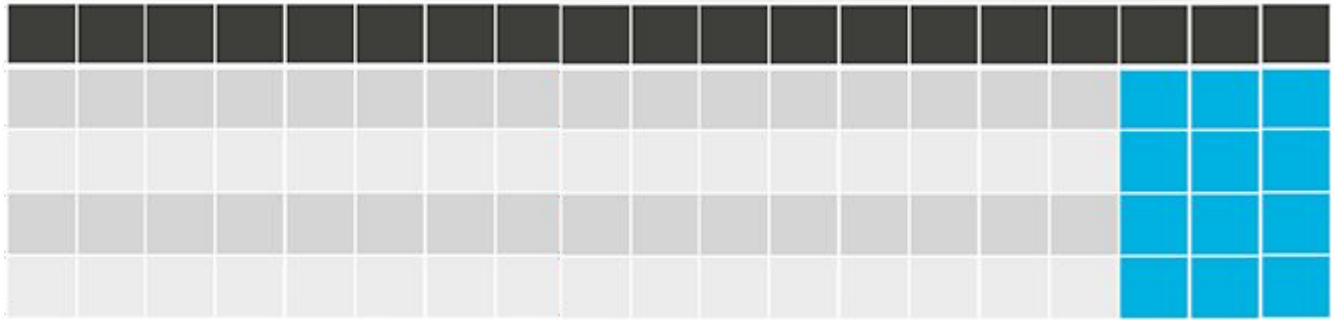
Parameter	minimum	maximum
$\log(Z/Z_{\odot})$	-3.0	0.5
$\log(n/\mathrm{cm}^{-3})$	-3.0	5.0
$\log(U)$	-4.0	3.0
$\log(N_H/\mathrm{cm}^{-2})$	17.0	23.0

CLOUDY v13.03

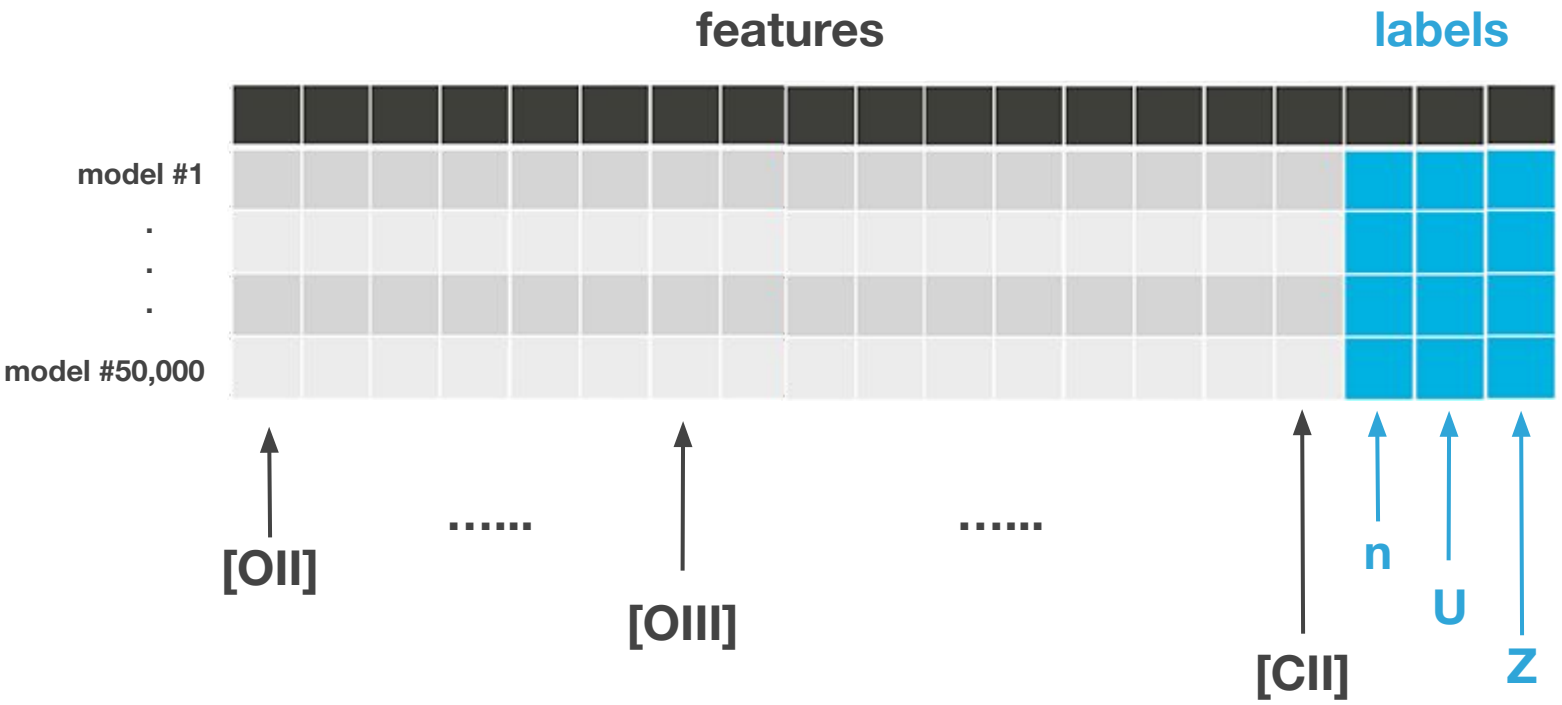
2

large library of **emission line intensities** (50,000 models)

2 Library of emission line intensities



2 Library of emission line intensities



1

grid of models: metallicity, column density, ionization parameter, density

Parameter	minimum	maximum
$\log(Z/Z_{\odot})$	-3.0	0.5
$\log(n/\mathrm{cm}^{-3})$	-3.0	5.0
$\log(U)$	-4.0	3.0
$\log(N_H/\mathrm{cm}^{-2})$	17.0	23.0

CLOUDY v13.03

2

large library of **emission line intensities** (50,000 models!)

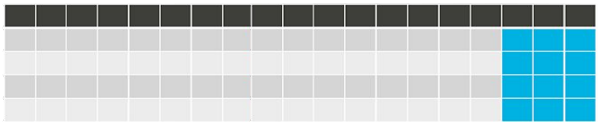
GAME

3

Supervised Machine Learning (SML): physical properties are inferred from an input spectrum using emission line intensities as input features

3 Supervised Machine Learning (SML)

library of emission line intensities

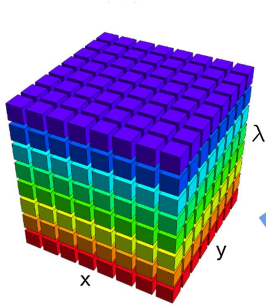


training

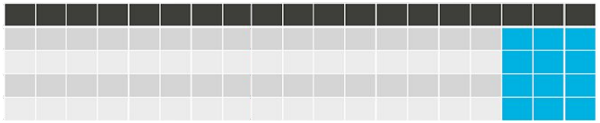
GALAXY MACHINE LEARNING FOR EMISSION LINES
GAME

3 Supervised Machine Learning (SML)

data cube

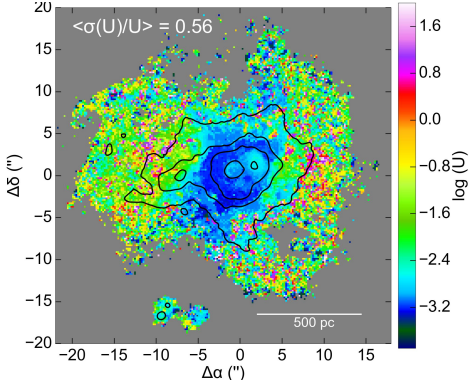


library of emission line intensities

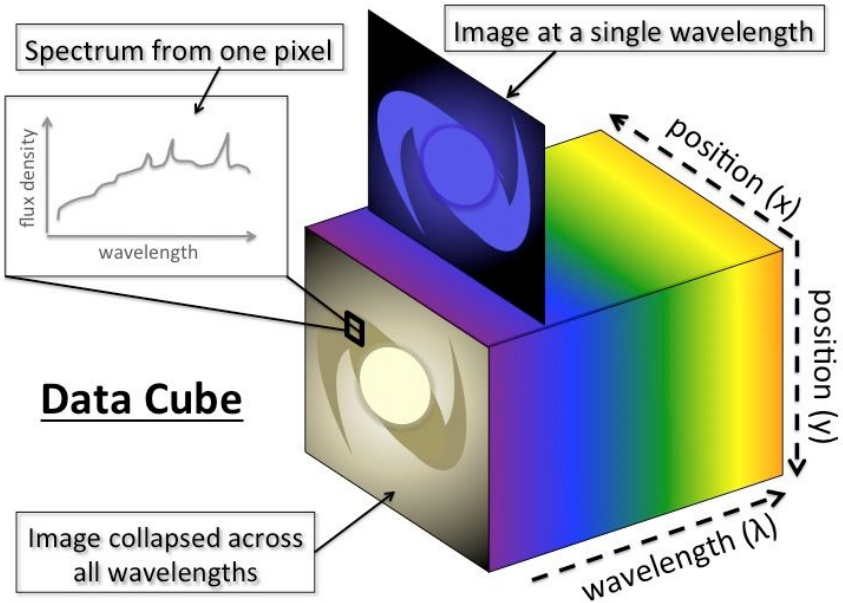


training

GALAXY MACHINE LEARNING FOR EMISSION LINES
GAME

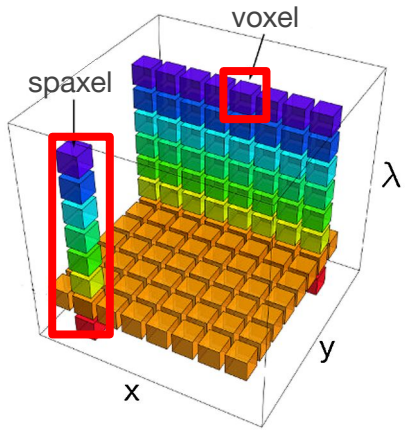


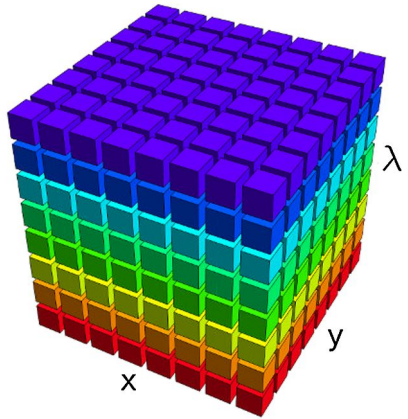
Integral Field Spectroscopy (IFS)



combines imaging and spectroscopy: “3D spectroscopy”

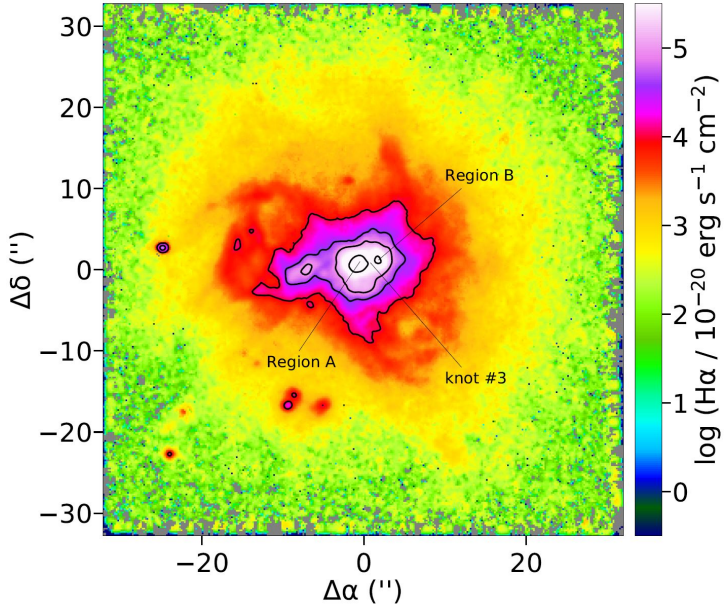
output: data cube





pixel: a data point on a CCD
spaxel: a spectrum in a data cube
voxel: a data point in a data cube

Results: application to IFU observations (Ucci et al. 2018, 2019)



MUSE observations of **Henize 2-10** (Cresci et al. 2017)

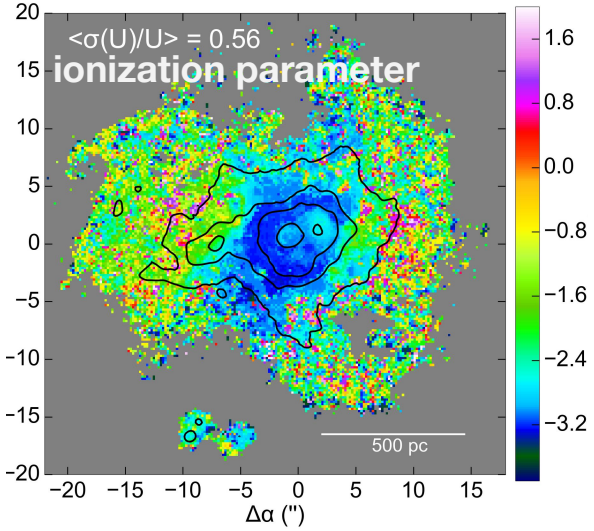
prototype of starburst HII galaxies (Allen et al. 1976)

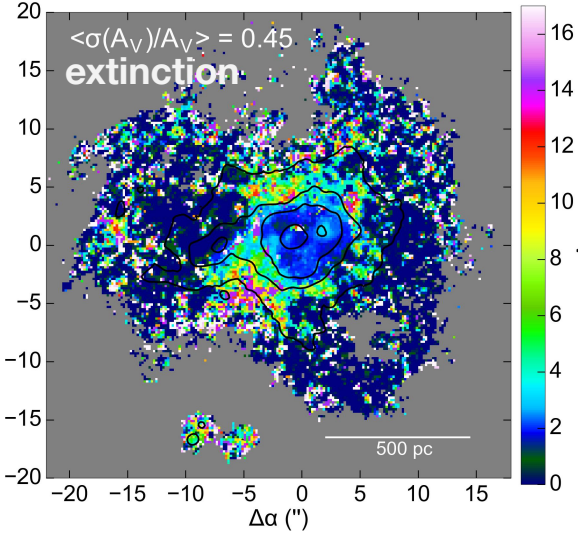
D = 8.23 Mpc (1 arcsec = 40 pc)

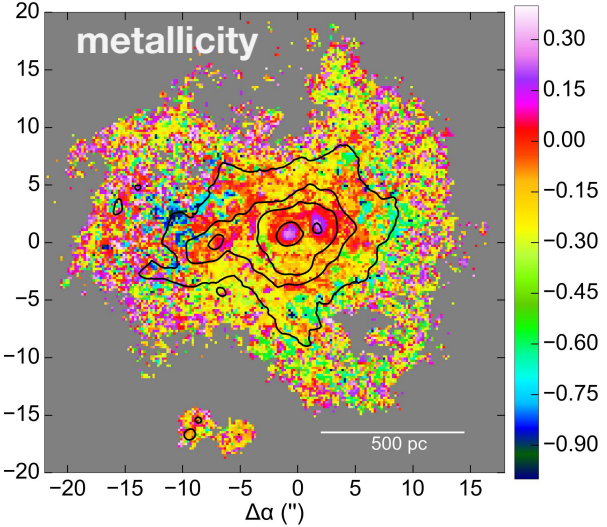
$M_{\star} \sim 3.7 \times 10^{9} M_{\odot}$

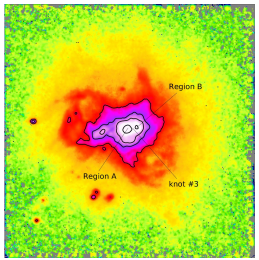
SFR $\sim 1.9 M_{\odot}\ yr^{-1}$ (Reines et al. 2011)

Results: application to IFU observations (Ucci et al. 2018, 2019)



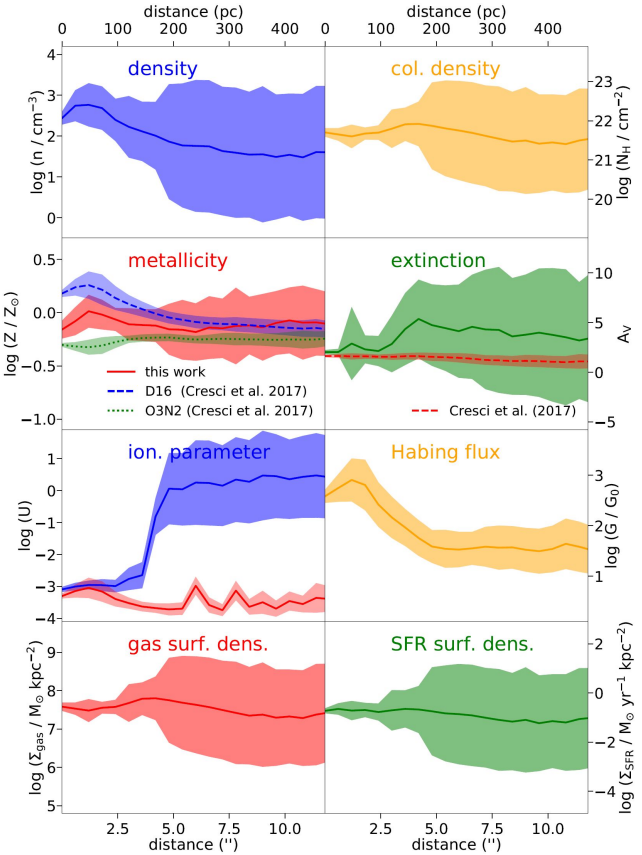


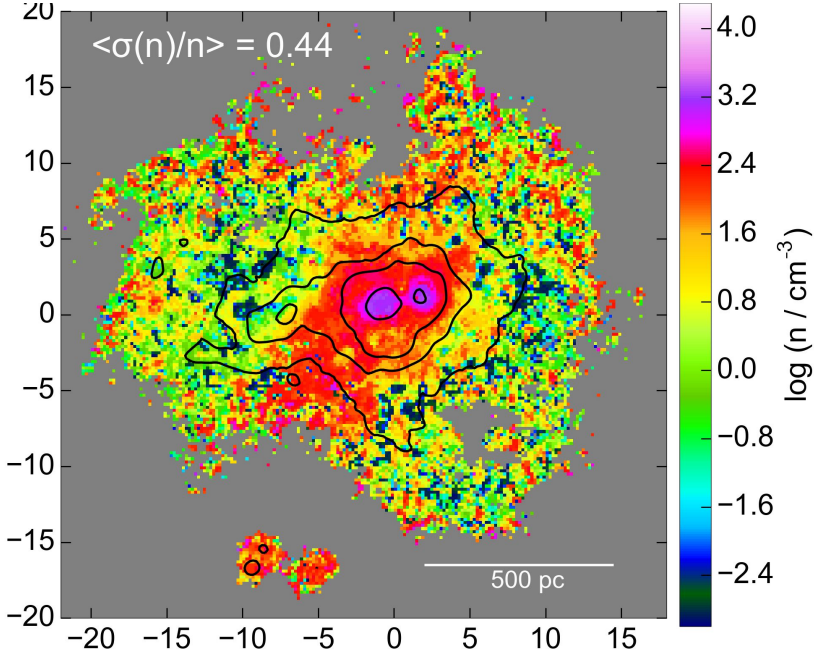




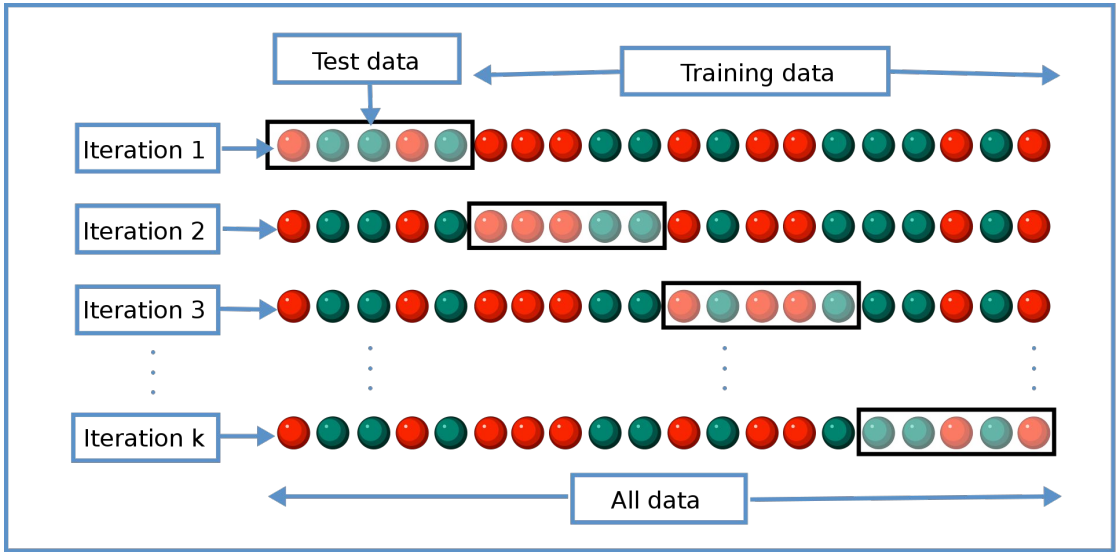
More than **75,000 spaxels** processed on a desktop PC in **less than 4 hr**

Results: radially-averaged profiles (Ucci et al. 2018, 2019)





GAME Predictive performances (k-fold cross validation)

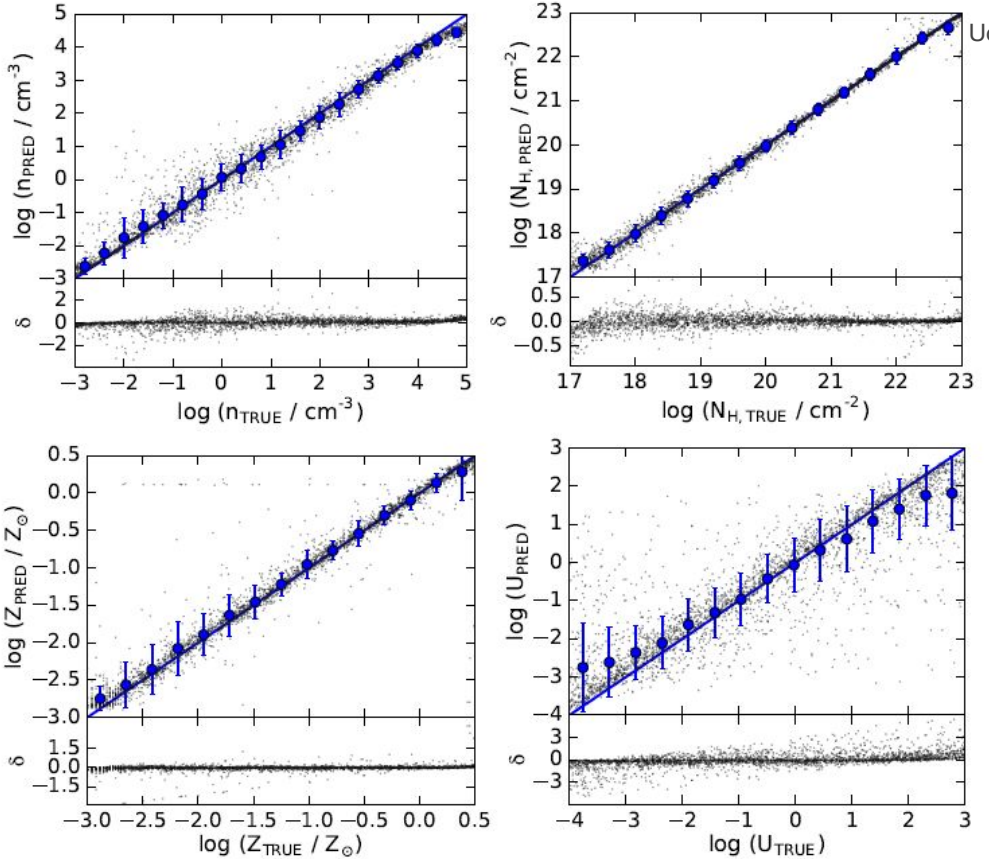


the library is split into **k folds**

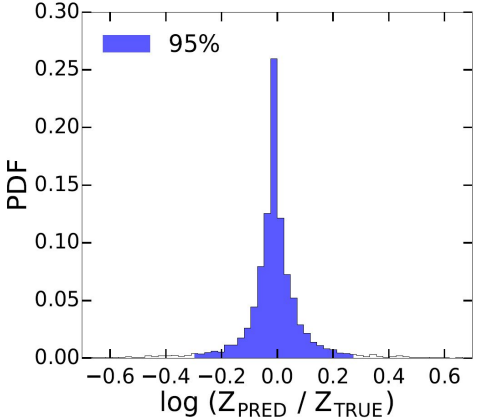
the code **trains the ML algorithm on k-1** and then tests on the left-out

GAME computes the score k consecutive times, and gives back the mean of these scores for each physical property

GAME Predictive performances (results on 1 fold)

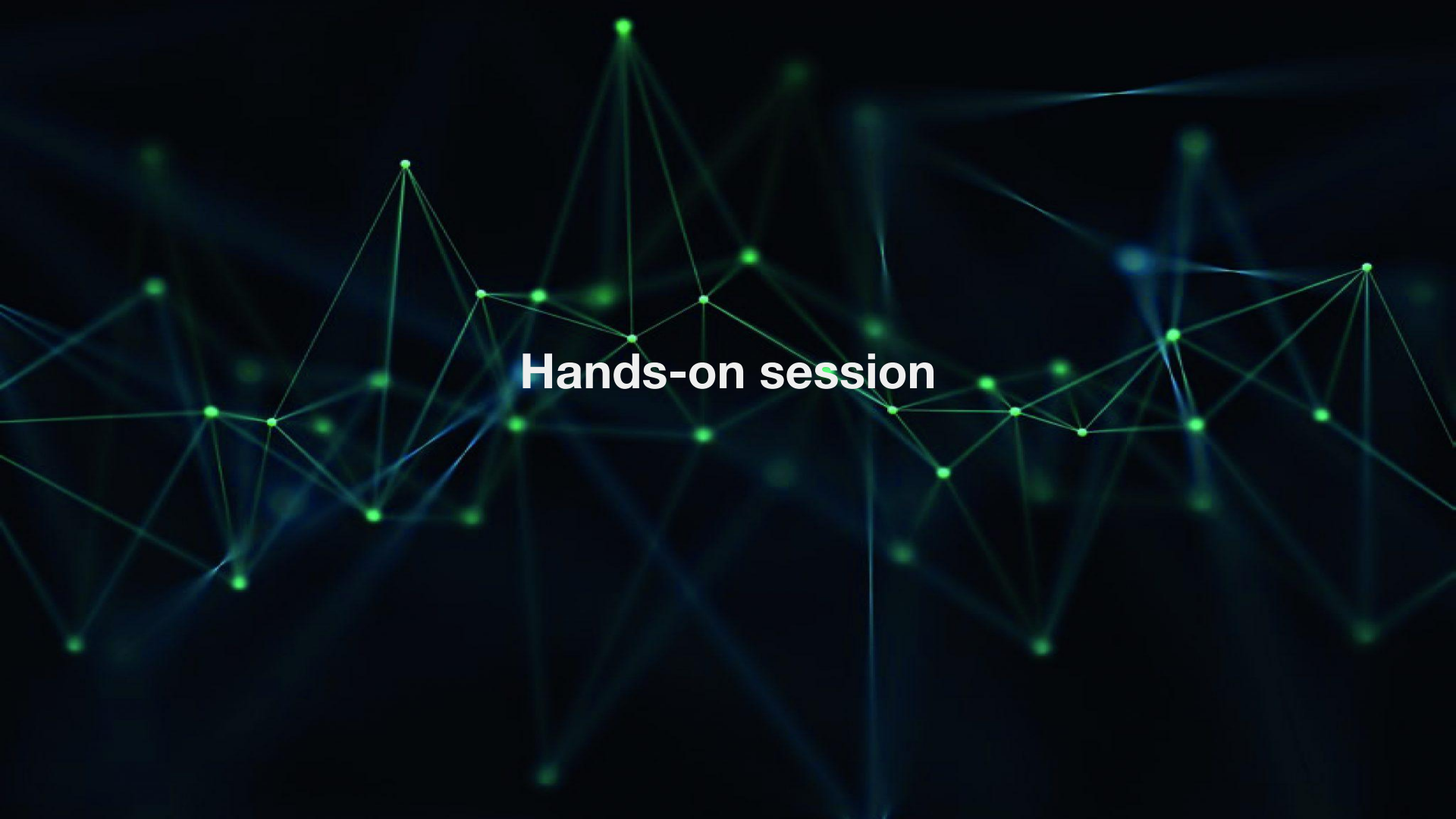


Ucci et al. (2017)



line	wavelength [Å]
Hβ	4861
[O III]	5007
He I	5876
[O I]	6300
Hα	6563
[N II]	6584
He I	6678
[S II]	6717
[S II]	6731

Ucci et al. (2018)


Hands-on session

```
> cd /media/...

> cd astrophysics_classes/

> cd ucci_machine_learning_on_galaxy_spectra/

> module load python/3.7.2

> jupyter-notebook &
```