ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

# Data Mining & Machine Learning: Intro to Scikit-Learn

Dr. Nina Hernitschek
April 4, 2022

# Data Mining



**Big Data** needs different approaches

- parallelism & data-side processing
- need of ways to formally specify computations

# Data Mining



huge, uniform, multivariate databases are emerging from **large-scale surveys**:

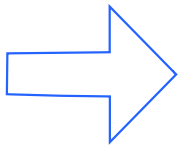
- $10^9$-object photometric catalogs from 2MASS, SDSS, VISTA
- $10^9$-object $\times 10^2$ epochs photometric lightcurve catalogs from ZTF, VVV, PS1 $3\pi$, LSST...
- $10^{6-8}$-galaxy redshift catalogs from SDSS, LAMOST...
- $10^{6-7}$-source radio/IR/X-ray catalogs from eROSITA, WISE

# Data Mining



huge, uniform, multivariate databases are emerging from **large-scale surveys**:

- $10^9$-object photometric catalogs from 2MASS, SDSS, VISTA
- $10^9$-object $\times 10^2$ epochs photometric lightcurve catalogs from ZTF, VVV, PS1 $3\pi$, LSST...
- $10^{6-8}$-galaxy redshift catalogs from SDSS, LAMOST...
- $10^{6-7}$-source radio/IR/X-ray catalogs from eROSITA, WISE



powerful astrostatistical & machine-learning tools are needed to **derive scientific insights**

# Data Mining



**shifting use cases:**

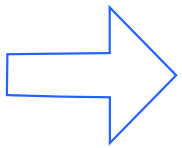
As data become more plentiful, finding the *right* data has become more important.

# Data Mining



**shifting use cases:**

As data become more plentiful, finding the *right* data has become more important.



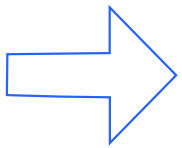
**Data Mining**

# Data Mining



**shifting use cases:**

As data become more plentiful, finding the *right* data has become more important.



**Data Mining**

Individual measurements giving way to **statistics, clustering, patterns** in the data.
Data processing needs to be **highly automatized**.
Analysis growing more exploratory rather than pre-defined/ scripted.

# Data Mining



**Examples:**

Finding and classifying variable stars in PS1 $3\pi$ required processing of $10^9$ sparse light curves $\Rightarrow$ 44,000 RRab stars.

Transient science (gravitational wave follow-up, GRBs, unknowns from LSST) requires rapid access to data sets of what is already known, anywhere on sky.

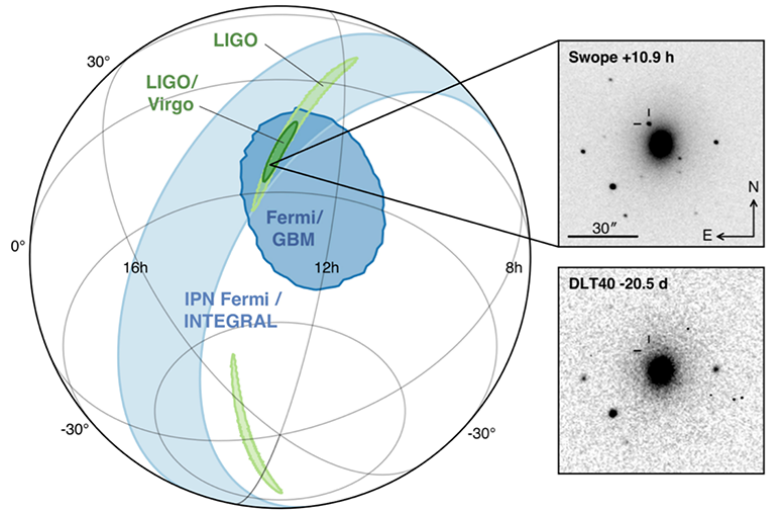



image credit: LIGO

# Data Mining



## Big Data: What is hard?

- Scalability can be challenging: Learn techniques from distributed systems, parallel databases - do not reinvent the wheel!
- Leverage new architectures: clouds, GPUs, multiple cores
- Lots of diverse and dirty data: Use and combine different cleaning and integration techniques

# Data Mining



(Jupyter/IPython) Notebook-driven science/analysis

very flexible tool to create readable analyses: keep code, images, comments, formula and plots together

IP[y]: Notebook spectrogram Last Checkpoint: a few seconds ago (autosaved)

IPython (Python 3)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

## Simple spectral analysis

An illustration of the Discrete Fourier Transform using windowing, to reveal the frequency content of a sound signal.

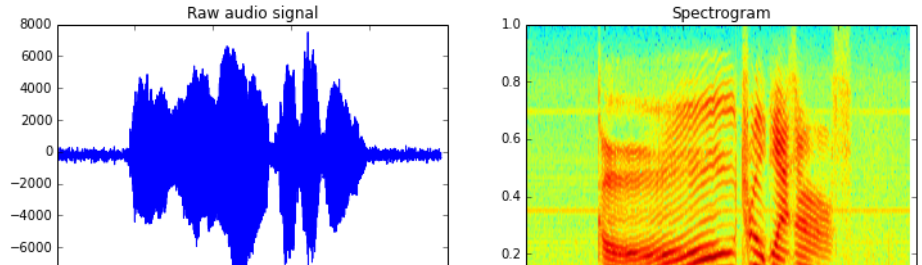
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

We begin by loading a datafile using SciPy's audio file support:

In [1]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

In [2]:
```
%matplotlib inline
from matplotlib import pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



# Data Mining



common technologies, many implementers for **data exploration & analysis**

# Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

# Machine Learning



... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

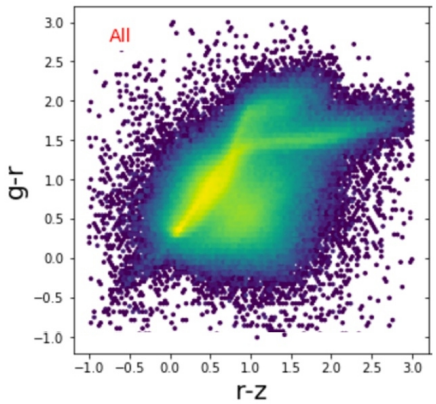
# Machine Learning



... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

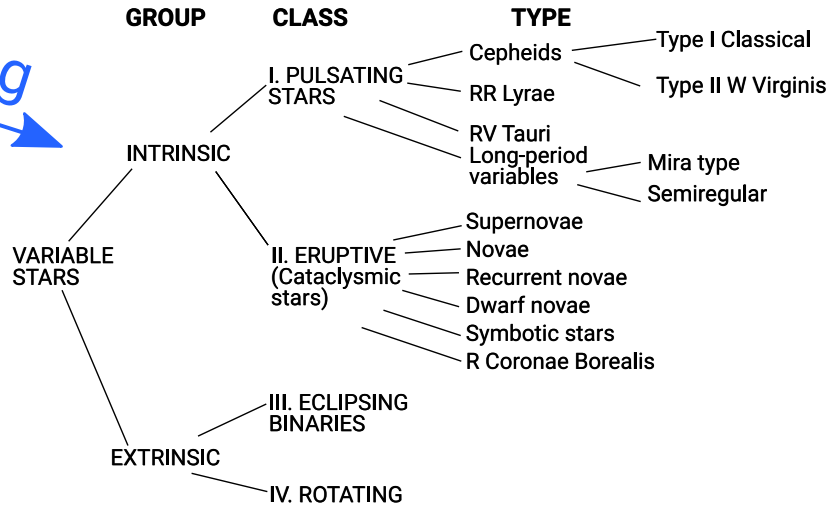
$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

**parameter space of measurements**




machine learning

**parameter space of astrophysical objects**



# Machine Learning



... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

$\Rightarrow$ allows to **uncover hidden correlation patterns** through iterative learning by sample data

$\Rightarrow$ allows **to model a survey**:

- describing data quality $\rightarrow$ outlier
- describing light curve characteristics $\rightarrow$ "features"
- classifying sources $\rightarrow$ catalogs
- finding substructure $\rightarrow$ clumps, overdensities, ...

# Unsupervised vs. Supervised Learning



**unsupervised learning** or "learning without labels"

***Clustering:***
Find subtypes or groups that are not defined a priori based on measurements
$\Rightarrow$ members of the same cluster are "close" in some sense

vs.

**supervised learning** or "learning witho labels"

***Classification:***
Use a priori group labels in analysis to assign new observations to a particular group or class

***Regression:***
instead of having training data with discrete labels, the "truth" is a continuous property and we are trying to predict the values of that property for the test data

# Unsupervised vs. Supervised Learning



**supervised learning** or "learning witho labels"

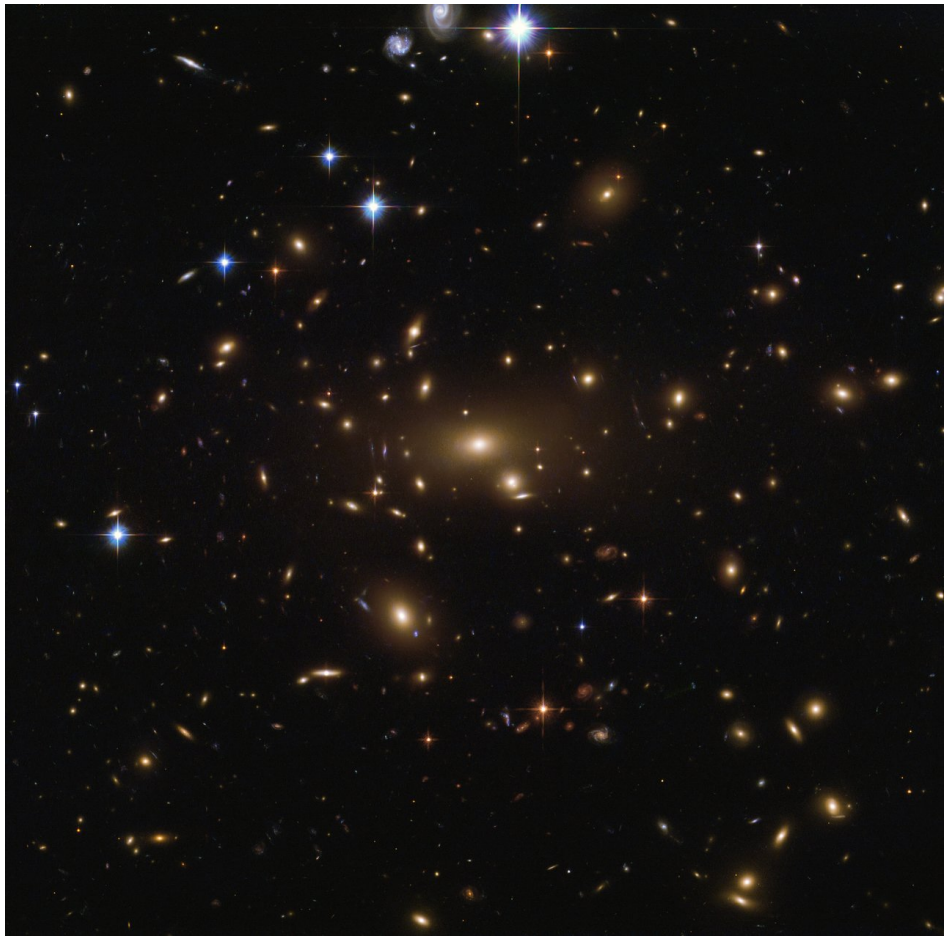
***Classification:***
Use a priori group labels in analysis to assign new observations to a particular group or class

***Regression:***
instead of having training data with discrete labels, the "truth" is a continuous property and we are trying to predict the values of that property for the test data

**example:**
The task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. On the other hand, we might wish to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

# Clustering Methods

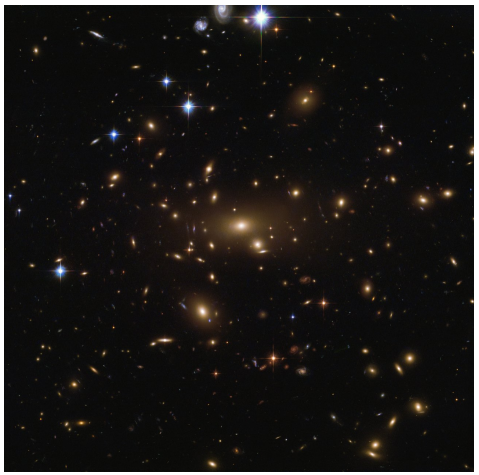
- Abell clustering richness class (Abell 1958)



# Clustering Methods



- Abell clustering richness class (Abell 1958)



- Gamma Ray Bursts: use properties of GRBs (e.g. location on the sky, arrival time, duration) to find classes of events

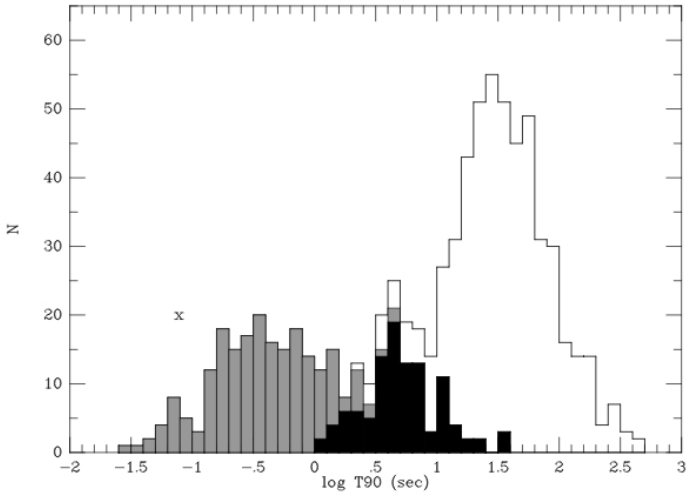



image credit: (Mukherjee+1998)

# Clustering Methods

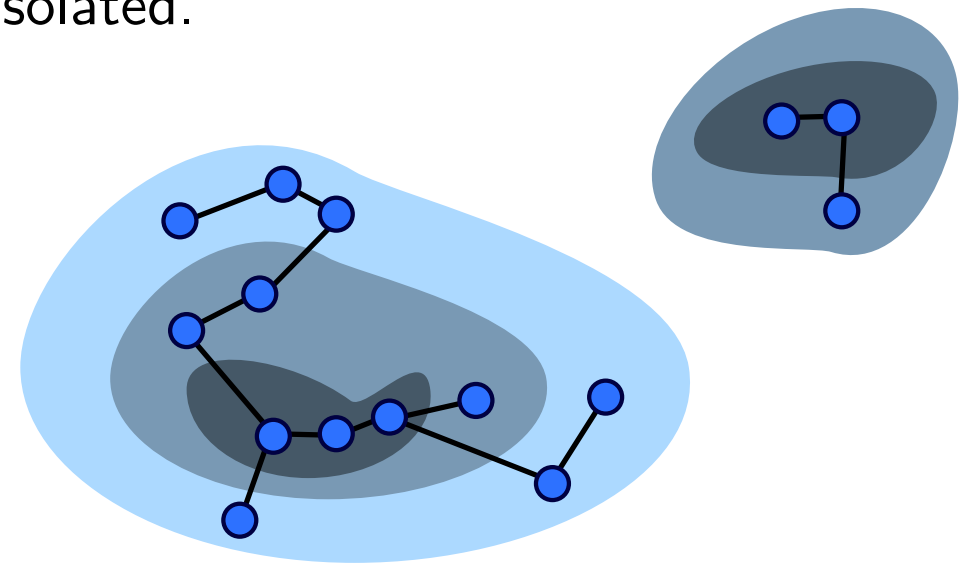


**Percolation or 'Friends of Friends (FoF)' algorithm**

1. Plot data points in a 2-dimensional diagram (or: calculate distances using a metric).

2. Find the closest pair, and call the merged object a *cluster*.

3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated.

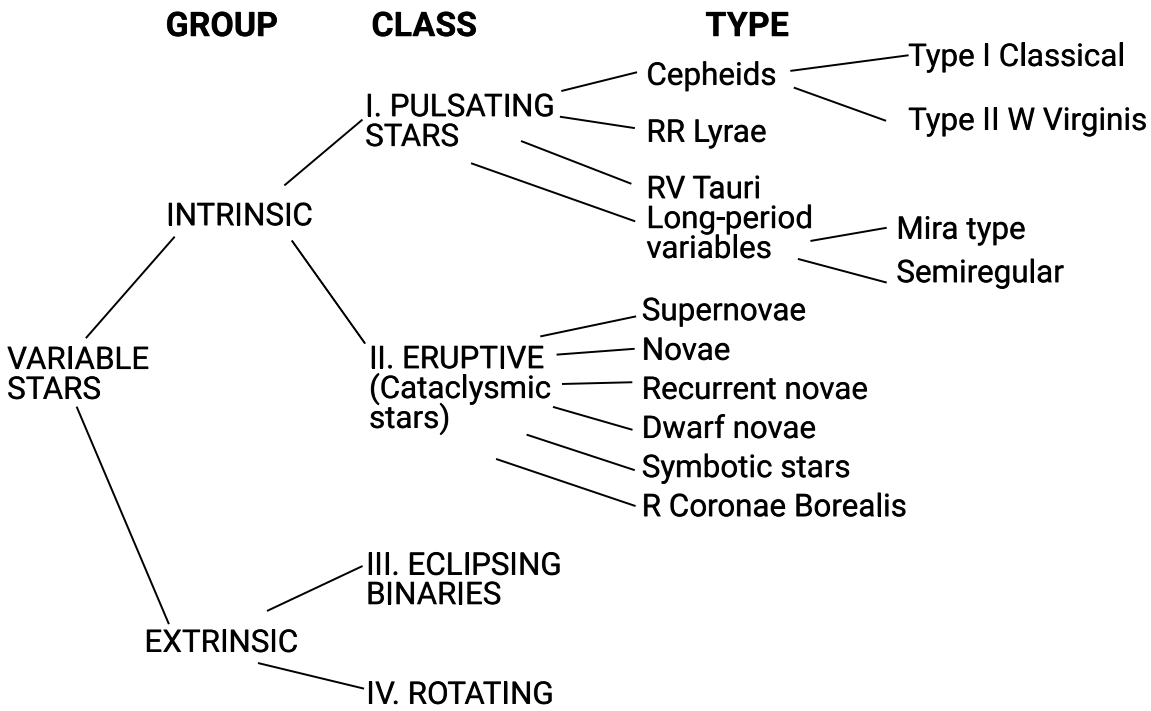


# Classification Methods

## Classification

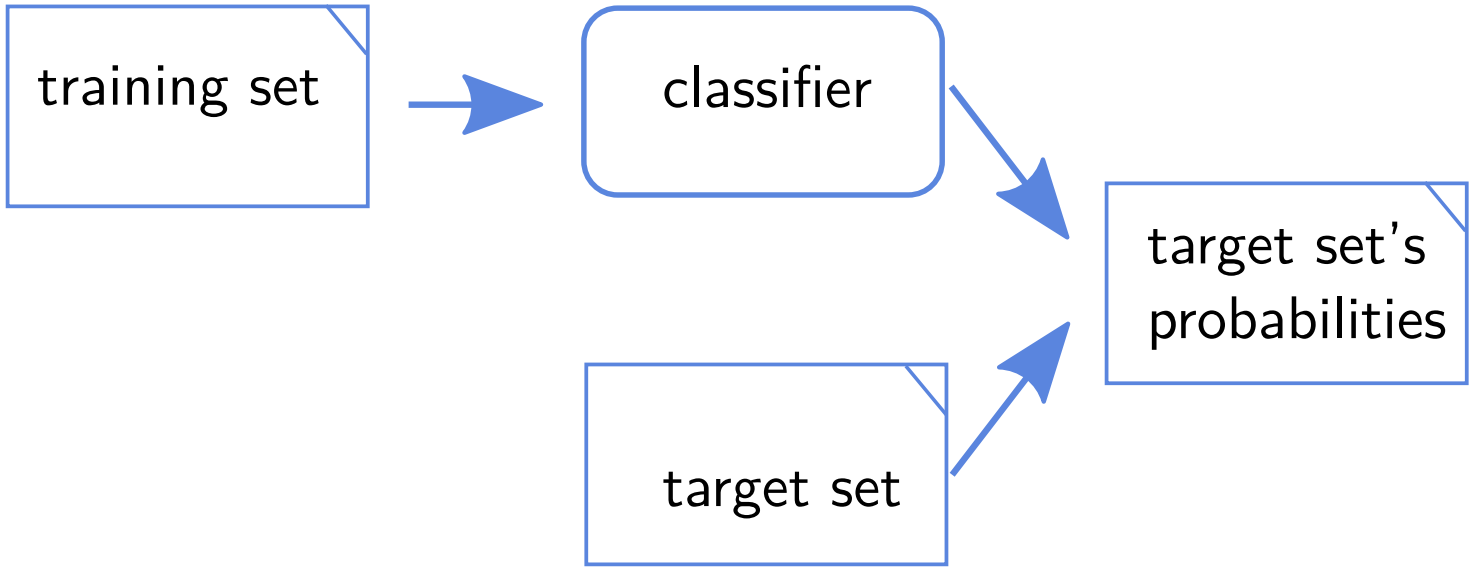
Use a priori group labels in analysis to assign new observations to a particular known group or class.

$\Rightarrow$ *supervised learning* or "learning with labels".



# Concepts of Supervised Classification



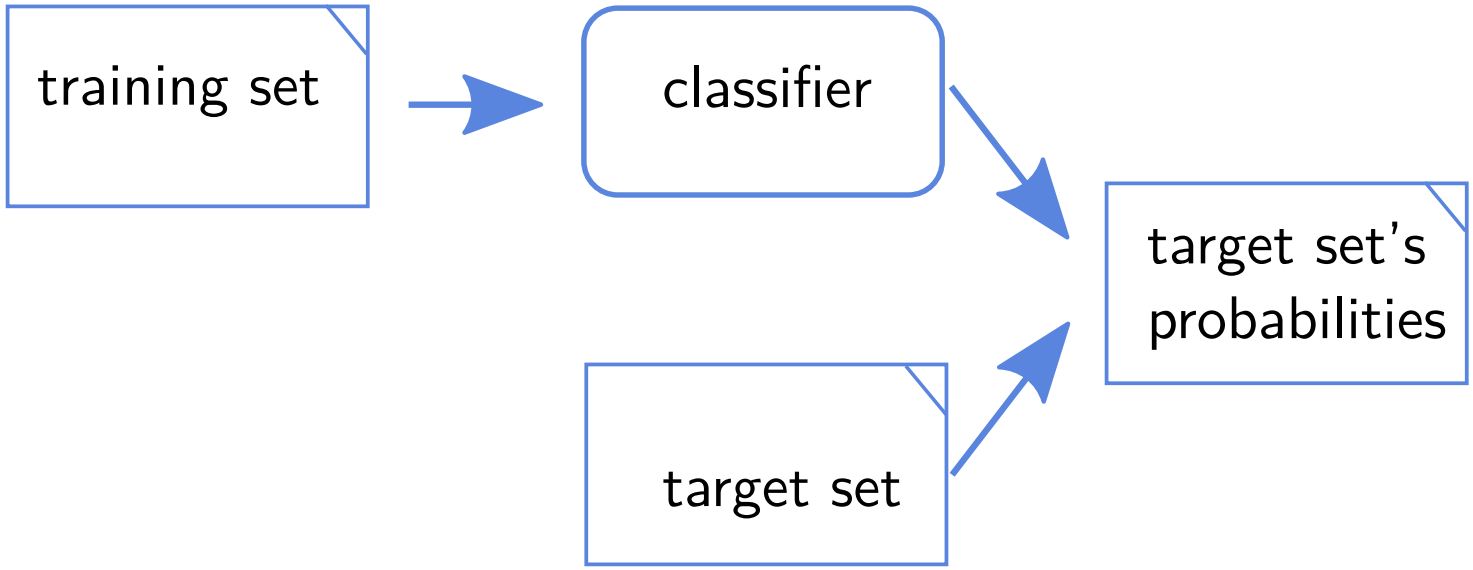



training set:

- set of sources inside/outside the category we are looking for
- same data quality as found in target set

# Concepts of Supervised Classification






training set:

- set of sources inside/outside the category we are looking for
- same data quality as found in target set

**What's happening internally?**

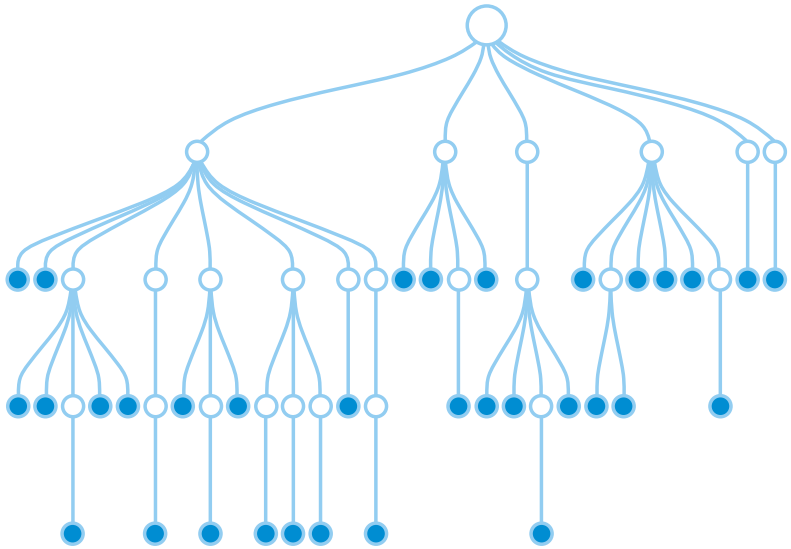
# Concepts of Supervised Classification



The learning process ("training"):

To build a decision tree, the set is divided into smaller and smaller subsets by **splitting** w.r.t. a single **feature** at a time.

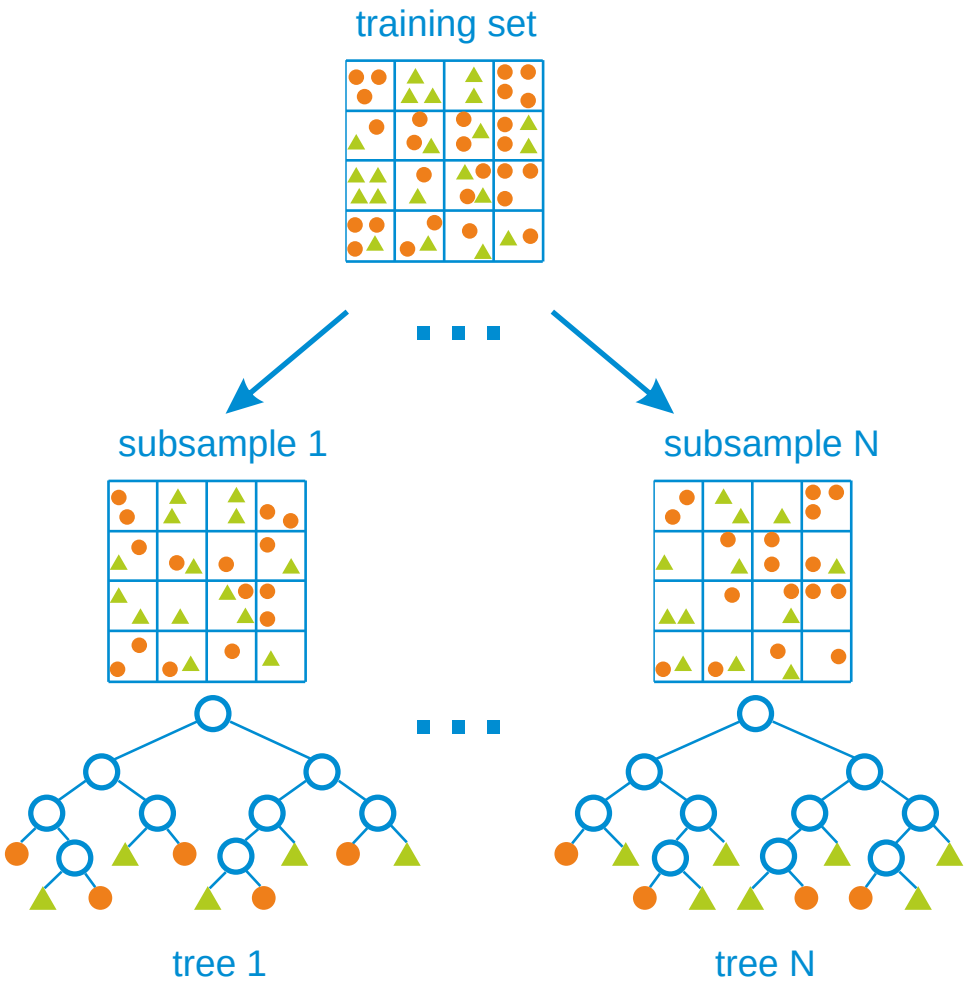
Split criteria: select feature and split point to produce the smallest impurity in the two resultant nodes based on the **training set**.



# Supervised Classification - Ensemble Methods



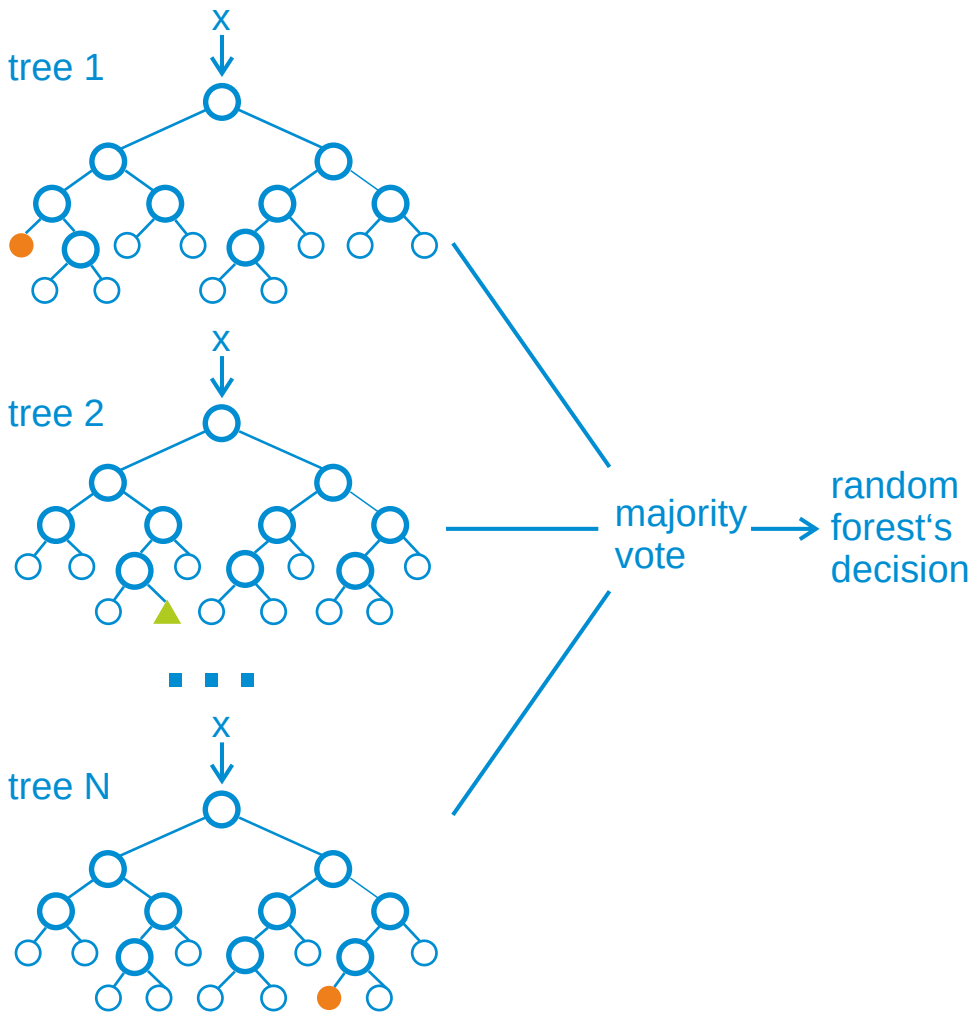
**Random Forest Classifier** as ensemble method: many trees are grown from subsets of the training set



# Supervised Classification - Ensemble Methods

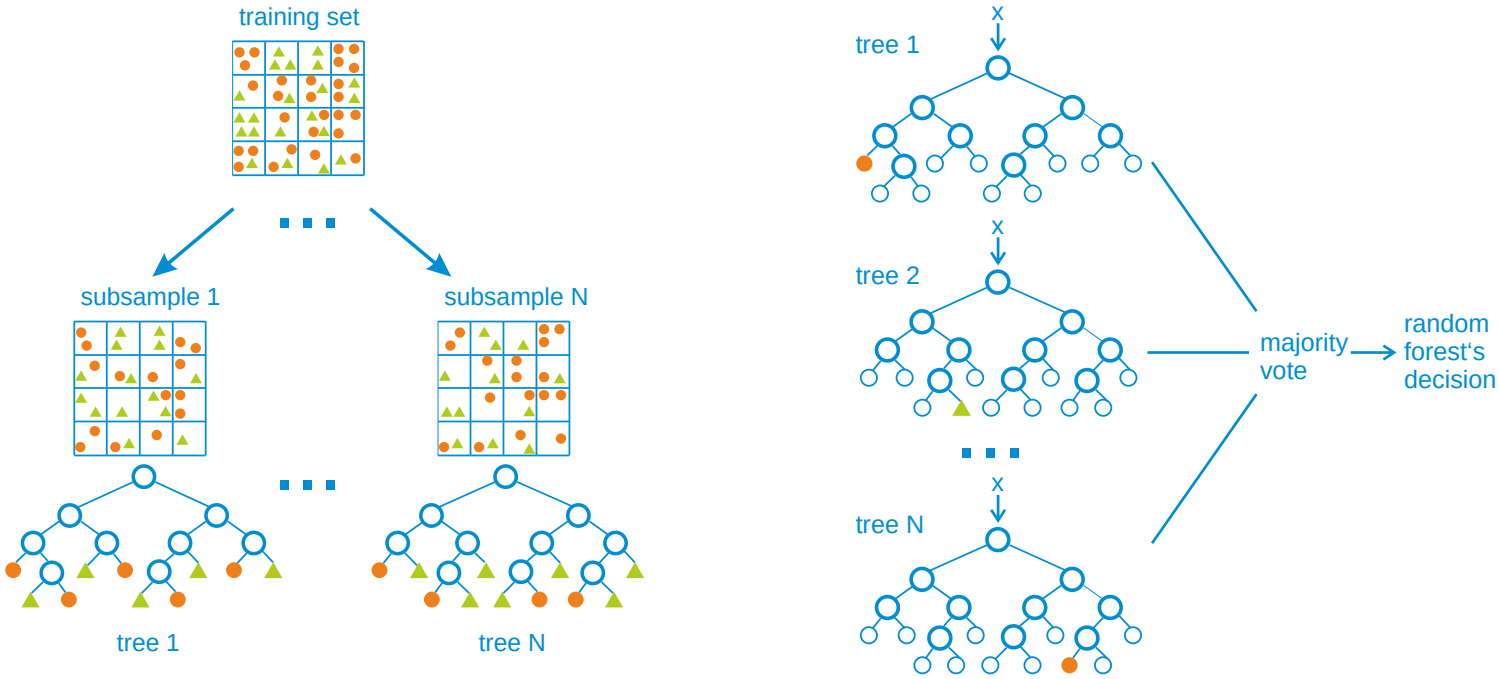


**Random Forest Classifier** as ensemble method: ... and are "voting" for classification



# Supervised Classification - Ensemble Methods






**divide-and-conquer approach** improves classification performance

- less sensitive to training set variances
- robust to outliers
- training and classification can be parallelized

# Supervised Classification - Verification



don't apply a classifier as a "black box"!

several concepts for **verification**

# Supervised Classification - Verification



don't apply a classifier as a "black box"!

several concepts for **verification**

make usage of the training set $\Rightarrow$ **10-fold cross-validation**
10 % held out $\Rightarrow$ train on 90%, apply to 10% in turn

# Supervised Classification - Verification

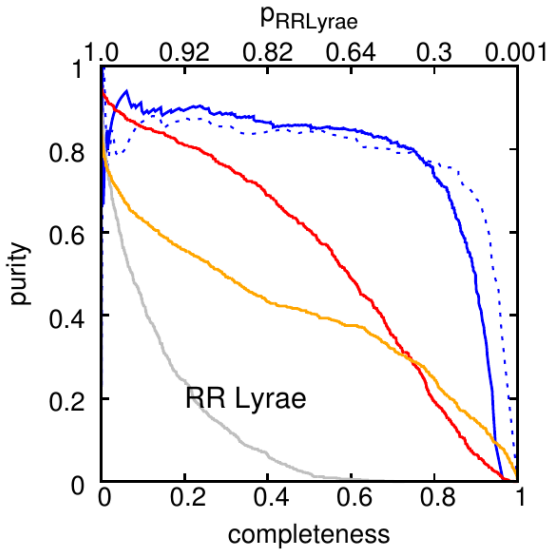


don't apply a classifier as a "black box"!

several concepts for **verification**

make usage of the training set $\Rightarrow$ **10-fold cross-validation**
10 % held out $\Rightarrow$ train on 90%, apply to 10% in turn

**purity-completeness (or precision-recall) curves**




completeness:
# selected true RR Lyrae / # true RR Lyrae

purity:
# selected true RR Lyrae / # all selected sources

# scikit-learn




scikit-learn is a popular Python package containing a collection of tools for **machine learning**

it includes algorithms used for classification, regression and clustering

it comes with an extensive **online documentation**: http://scikit-learn.org/stable/tutorial/basic/tutorial.html

scikit-learn is built upon Python's NumPy (Numerical Python) and SciPy (Scientific Python) libraries, which enable efficient in-core numerical and scientific computation within Python.

# scikit-learn



scikit-learn uses 3 steps for **developing, applying and testing** machine learning algorithms:

- Train the model using an existing data set describing the phenomena you need the model to predict.
- Test the model on another existing data set to ensure it performs well.
- Use the model to predict phenomena as needed for your project.

# Break & Questions



afterwards we continue with **`lecture_10.ipynb`** from the `github` repository