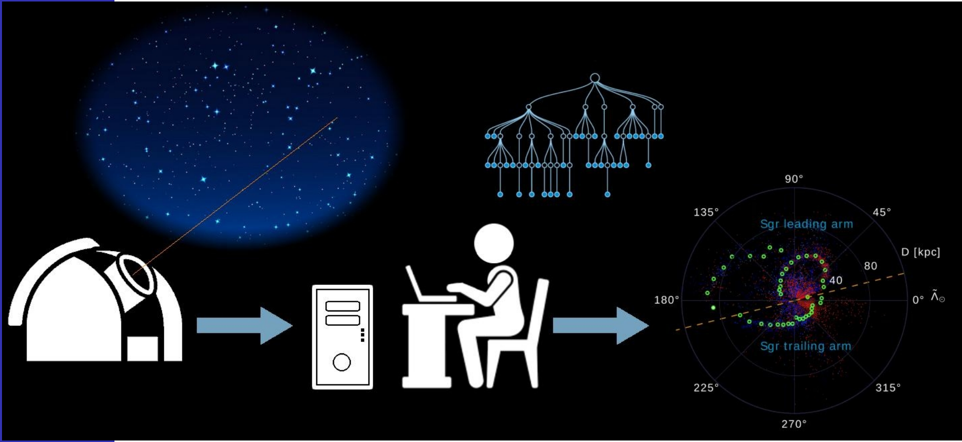
Advanced Machine Learning (Semester 1 2023)

Introduction & Course Logistics

Nina Hernitschek
Centro de Astronomía CITEVA
Universidad de Antofagasta

April 10, 2023

Motivation


90°
45°
135°
Sgr leading arm
D [kpc]
80
40
180°
0°
$\tilde{\Lambda}_{\odot}$
Sgr trailing arm
225°
315°
270°

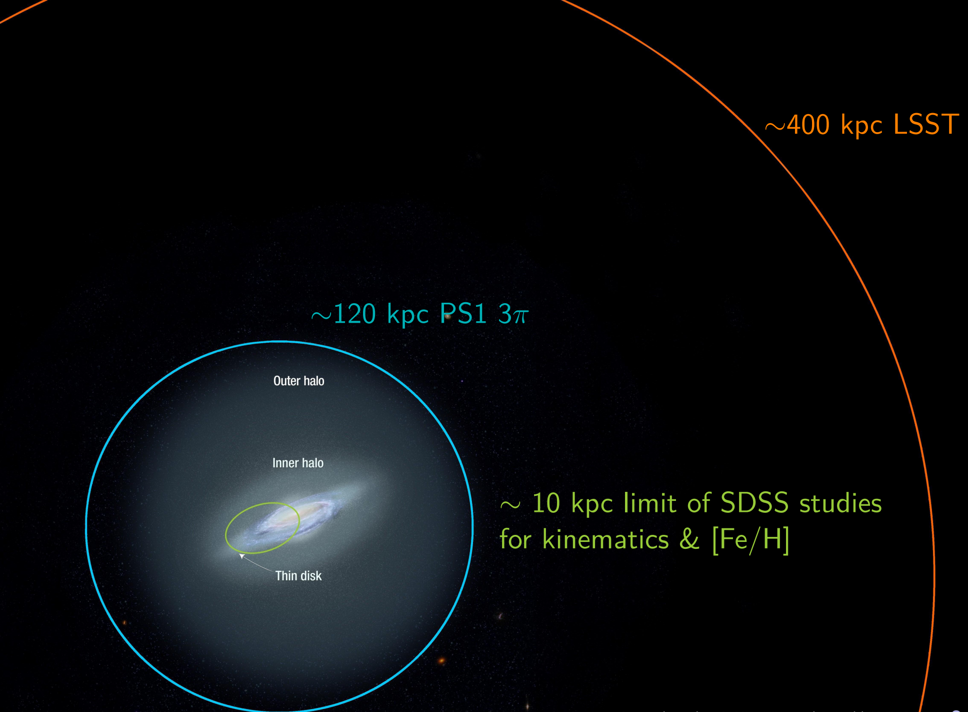
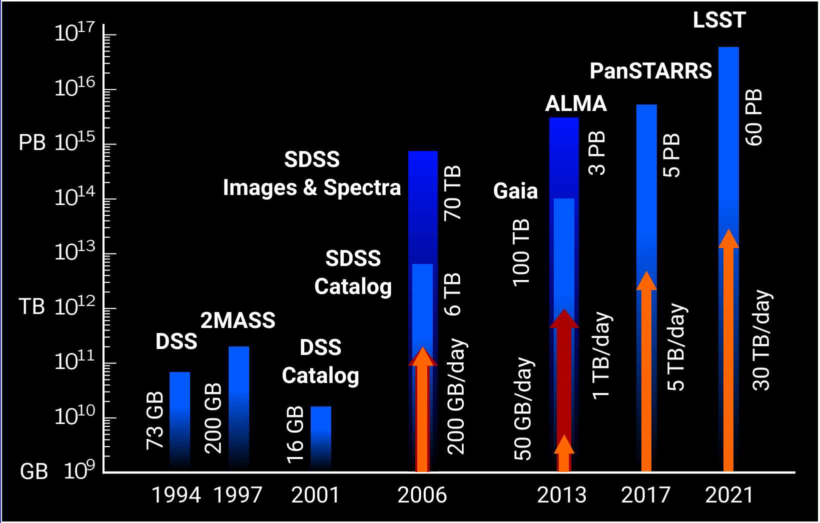


image based on NASA, ESA, and A. Feild (STScI)

Challenges in Data Handling

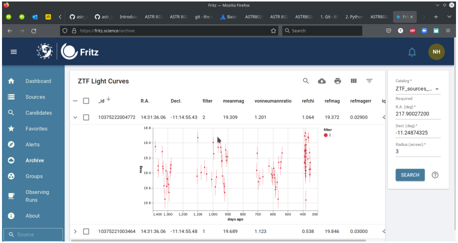
increasing data volume in astronomical surveys



what you will learn in this class

this course will prepare you for "doing science" with current and upcoming large astronomical surveys:

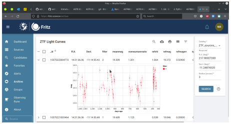
accessing astronomical survey data



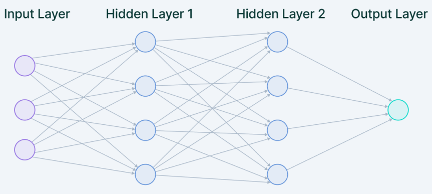
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

accessing astronomical survey data



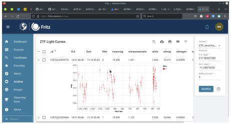
artificial neural networks



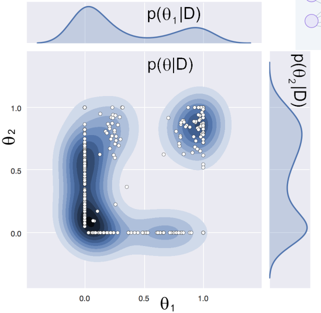
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

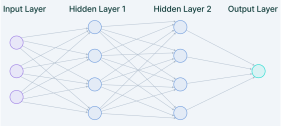
accessing astronomical survey data



statistical methods



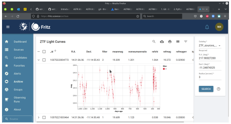
artificial neural networks



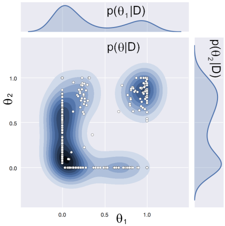
what you will learn in this class

this course will prepare you for “doing science” with current and upcoming large astronomical surveys:

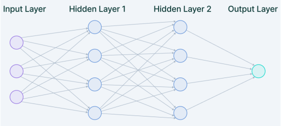
accessing astronomical survey data



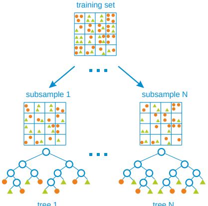
statistical methods



artificial neural networks



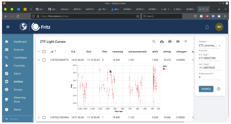
machine learning



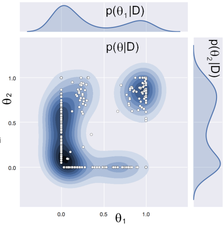
what you will learn in this class

this course will prepare you for "doing science" with current and upcoming large astronomical surveys:

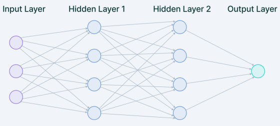
accessing astronomical survey data



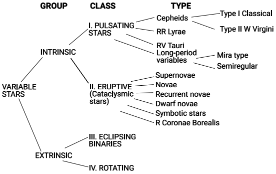
statistical methods



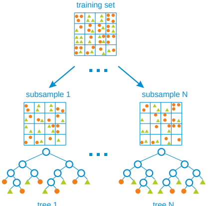
artificial neural networks



classification



machine learning



Course Logistics

course content:

- lecture: Monday 10:00 - 12:00
- practice: Wednesday 10:00 - 11:30
- preparation of a paper presentation (of your choice)
- identification of a problem related to your research to be solved with machine learning, i.e. neural networks: 2 presentations, report

grading:

- participation: 10 %
- paper presentation: 20 %
- project presentations (project idea + project status + final): 35 %
- project report: 35 %

deliverables: your github repository

contact and course material:

- e-mail: `nina.hernitschek@uantof.cl`
- github: `https://github.com/ninahernitschek/advanced_machine_learning_2023_1`

Course Logistics

April 10 Lecture 1: Introduction & Course Logistics, April 12 Practice 1

April 17 Lecture 2: Artificial Neural Networks (I) April 19 Practice 2

April 24 Lecture 3: Artificial Neural Networks (II), April 26 Practice 3

May 8 Lecture 4: Training of Neural Networks, May 10 Practice 4

May 15 *presentation on paper*, May 17 Q&A Session

May 22 Lecture 5: Convolutional Neural Networks, May 24 Practice 5

May 29 Lecture 6: Recurrent Neural Networks, May 31 Practice 6

June 5 *presentation on project idea*, June 7 Q&A Session

June 12 Lecture 7: Autoencoders, June 14 Practice 7

June 19 Lecture 8: Reinforcement Learning

July 3 Lecture 9: Architecture of Machine Learning Projects, July 5 Practice 8

July 10 *presentation on project status*

July 17 optional Q&A Session

July 31 optional Q&A Session

August 4 *final presentation project*

Rules for Coding, Presentations, Report

same as for Advanced Astroinformatics project:

coding: If you have a question when something doesn't work, summarize what you tried - often this will even lead to the solution.

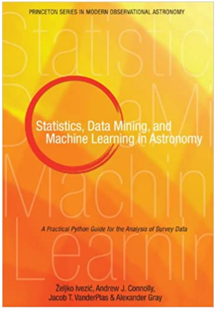
project report and presentation:

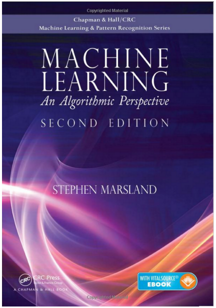
- LaTeX
- figures: all own figures should be in vectorized pdf format
- abstract: concise summary of your project that gives the big picture
- data and aim of project: data description (incl. citation)
- own work: properly cite what is not your own work; discuss how the previous work is similar to or different from your own work
- implementation: medium-level implementation description with libraries/ software frameworks (incl. citation), project milestones ⇒ more details than in a research paper
- related work: include both work aimed at similar problems and work that employs similar solutions to yours
- discussion: reflect your approach (strengths, weaknesses, limitations), lessions learned
- bibliography: bibtex/ref mechanism, ADS/Bibtex information

Textbooks

Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data
Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, A. Gray

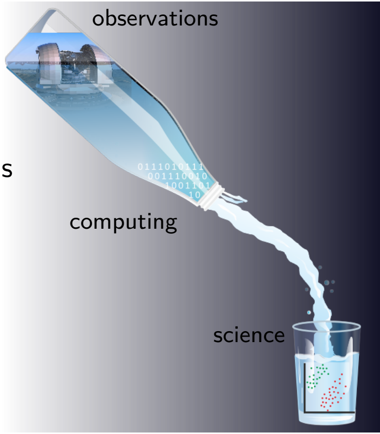
Machine Learning - An Algorithmic Perspective
Stephen Marsland





Challenges in Data Handling

astronomy is largely determined by computational capacity



⇒ telescopes & instruments as front-ends for data processing systems

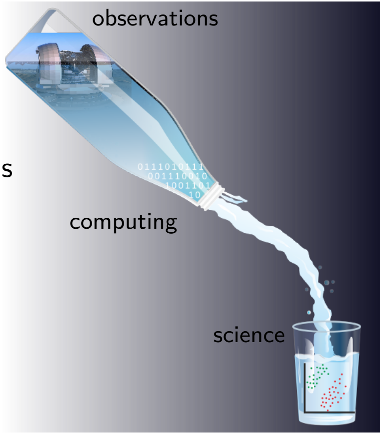
⇒ **challenge and chance:**
understanding complex phenomena requires complex data

Challenges in Data Handling

astronomy is largely determined by computational capacity

$\Rightarrow$ telescopes & instruments as front-ends for data processing systems

$\Rightarrow$ **challenge and chance:** understanding complex phenomena requires complex data



Big Data is transforming how and which discoveries are made

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

shifting use cases:
As data becomes big data, finding the *right* data has become more important.

Big Data

Laney et al. 2001: data growth challenge is **three-dimensional**

Big Data is data with at least one big dimension:

- volume
- velocity: bandwidth, response speed
- variety: number and size of individual assets

shifting use cases:
As data becomes big data, finding the *right* data has become more important.

$\Rightarrow$ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

Big Data

shifting use cases:

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

$\Rightarrow$ powerful astrostatistical & machine-learning tools are needed to derive scientific insights

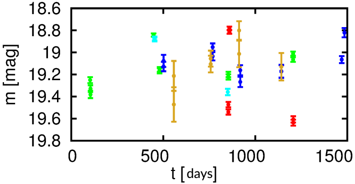
Individual measurements giving way to **statistics, clustering, patterns** in the data.

Data processing needs to be **highly automatized**.
Analysis growing more exploratory rather than pre-defined/scripted.

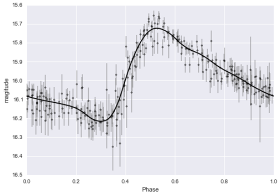
Big Data

one **example** for finding the *right* data :

Pan-STARRS1 3π survey with about 10^9 light curves like that:



goal: finding RR Lyrae* stars whose light curves look like that (if better sampled):



*less than 1 % of the light curves are expected to be from that type

Statistical Data Analysis

Data-driven methods like statistical methods can reliable **quantify information** embedded in scientific data **without the biases of physical models**.

Requirements:

- find the right method(s): modern statistics is vast in its scope and methodology
- scientific inferences should not depend on arbitrary choices in methodology and variable scale
- correct interpretation of the meaning of a statistical result w.r.t. the scientific goal: (astro-)statistics and machine learning are only tools!

(Astro-)Statistics

a lot is possible:

galaxy clustering	spatial point processes, clustering
galaxy morphology	regression, mixture models
weak lensing morphology	geostatistics, density estimation
strong lensing morphology	shape statistics
faint source detection	false discovery rate
variable source preclassification	structure functions + classifier

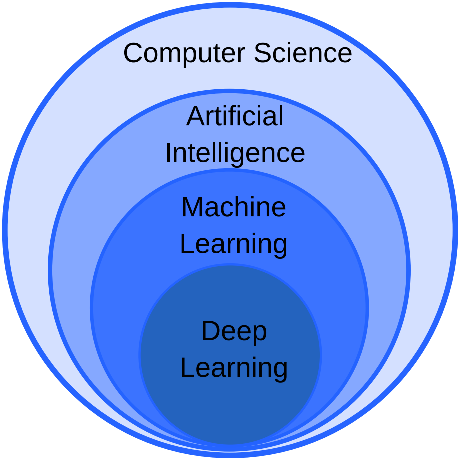
(Astro-)Statistics

a lot is possible:

galaxy clustering	spatial point processes, clustering
galaxy morphology	regression, mixture models
weak lensing morphology	geostatistics, density estimation
strong lensing morphology	shape statistics
faint source detection	false discovery rate
variable source preclassification	structure functions + classifier

$\Rightarrow$ **fitting models**

Machine Learning - Terminology



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

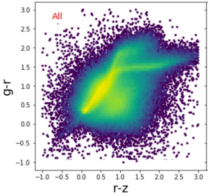
Machine Learning

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

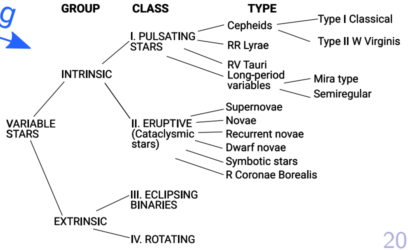
⇒ allows to **uncover hidden correlation patterns** through iterative learning by sample data

⇒ **in astronomy**:

parameter space of measurements



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

$\Rightarrow$ **in astronomy**: 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, ...

Functions of Machine Learning Systems

Descriptive the system uses the data to explain data properties; tools: simple statistical tools such as averages, percentages

Functions of Machine Learning Systems

Descriptive	the system uses the data to explain data properties; tools: simple statistical tools such as averages, precentages
Predictive	focuses on predicting and understanding future behavior

Functions of Machine Learning Systems

Descriptive	the system uses the data to explain data properties; tools: simple statistical tools such as averages, precentages
Predictive	focuses on predicting and understanding future behavior
Prescriptive	the system uses data to make suggestions about actions to take based on the insights gained

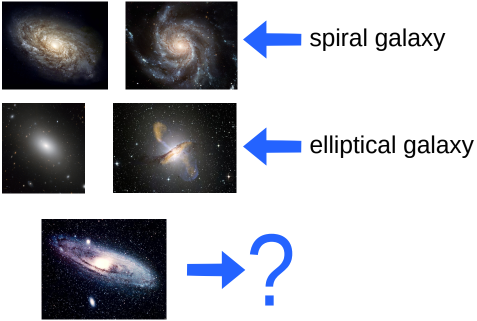
Types of Machine Learning Algorithms

there are different types of machine learning algorithms that differ mostly by **how they use data**

Supervised Machine Learning

labeled data (training data, training set) enable the supervised machine learning algorithm to understand the connection between **features** and **labels**

new observations (target data) are assigned to a group or class



applications: classification problems, regression problems

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

Supervised Machine Learning

The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

During **training**, the algorithm will search for patterns in the data that correlate with the desired outputs.

Supervised Machine Learning

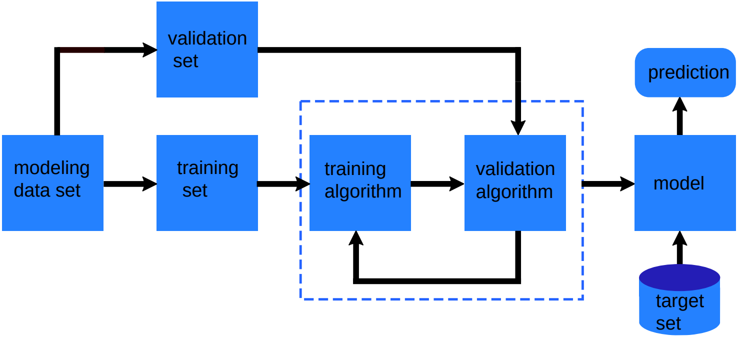
The objective of a supervised learning model is to predict the correct label for newly presented input data.

When training a supervised learning algorithm, the **training set** will consist of inputs paired with the correct outputs. Inputs in the training set should represent the **target set** which we have to classify: composition of the data, data quality.

During **training**, the algorithm will search for patterns in the data that correlate with the desired outputs.

After training, a supervised learning algorithm will take in new unseen inputs and will determine which label the new inputs will be classified based on prior training data.

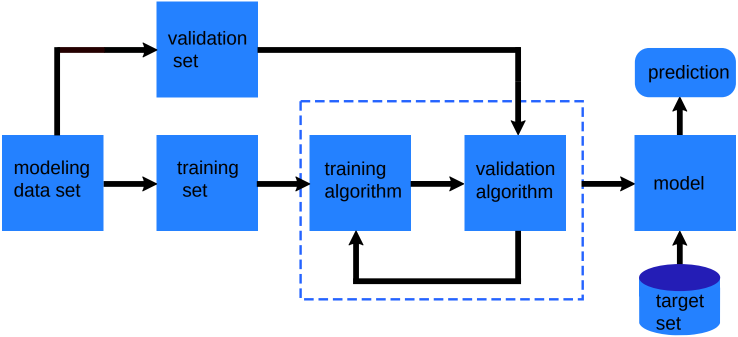
Supervised Machine Learning



split modeling set into training set and validation set:
the validation set (also: test set) is used for the testing the model after the model has been trained on the training set - it is extremely important to test the model on data not being part of the training set

A **fundamental assumption** of supervised machine learning is that the distribution of training examples is identical to the distribution of validation examples and future unseen examples (the target set).

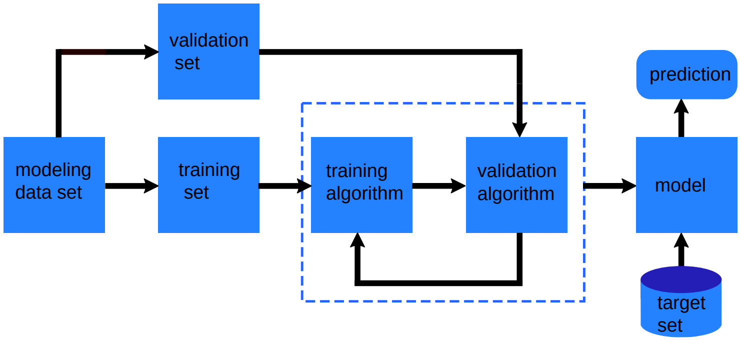
Supervised Machine Learning



Training:
given a training set of labeled examples $\{(x_1, y_1), ..., (x_n, y_n)\}$, estimate the prediction function f and parameters θ which minimizes the prediction error on the training set

Validation:
apply f to validation set x, output predicted value $y = f(x)$
from this we generate performance measures, also called accuracy measures

Supervised Machine Learning

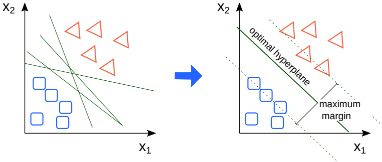


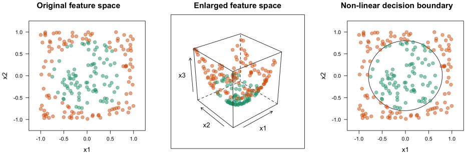
Application:
Run the model on the target set.

Supervised Machine Learning

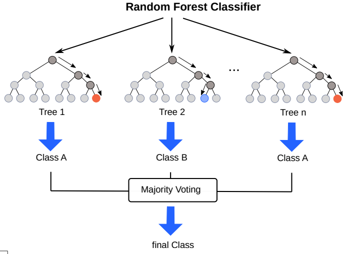
state-of-the-art (before Deep Learning):

Support Vector Machines
binary classification





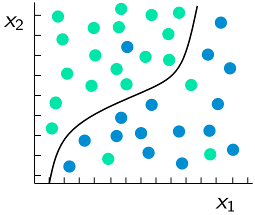
Random Forest Classifiers
multiclass classification



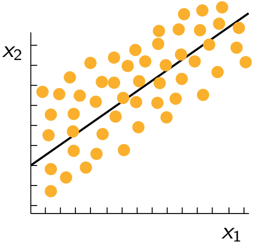
Supervised Machine Learning

two main areas of supervised machine learning:
classification problems and **regression** problems

mapping input value(s) to a discrete value, the class
example: predicting whether an object is a star or a galaxy



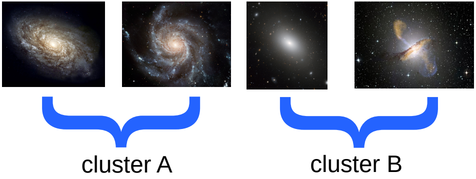
mapping input value(s) to continuous data
example: predicting the surface temperature of a star



Unsupervised Machine Learning

unlabeled data enable the unsupervised machine learning algorithm to **understand the data** and **find patterns in data themselves**

data is clustered, new data is assigned to clusters



applications: data exploration, data clustering, anomaly

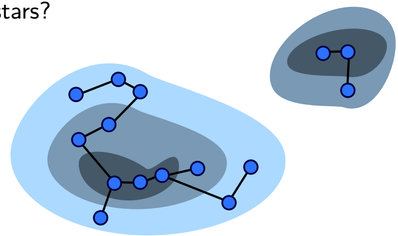
Unsupervised Machine Learning

three main areas of unsupervised machine learning:
clustering, **association** and **dimensionality reduction**

find hidden patterns in the data based on similarities or differences
example: are there subtypes within a given type of stars?

find the probability of co-occurrence of items in a collection
example: which stars likely host exoplanets?

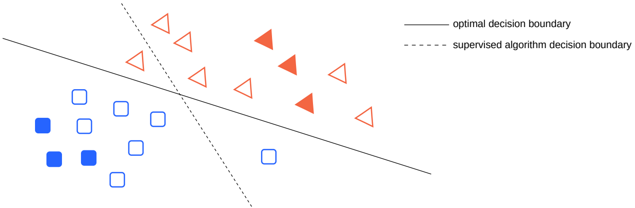
reduce the dimensions of the data
example: feature rextraction to reduce the number of random variables



Semi-Supervised Learning

takes the **main advantages from both** supervised and unsupervised learning

it uses a **smaller labeled data set** to guide classification and performs unsupervised feature extraction from a **larger, unlabeled data set**

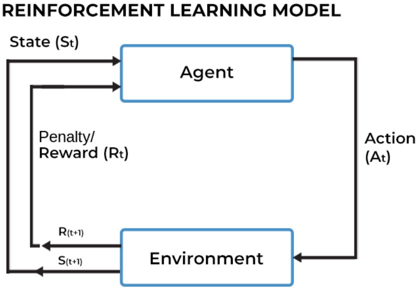


applications: solve problems when there is not enough labeled data present to train a model

Reinforcement Learning

an **agent** learns to behave in an environment by performing actions and adjusting its further course to feedback

data is not labeled, agent learns by experience: good actions are rewarded, bad actions result in a penalty



applications: playing chess; optimizing astronomical survey

Types of Machine Learning - Overview

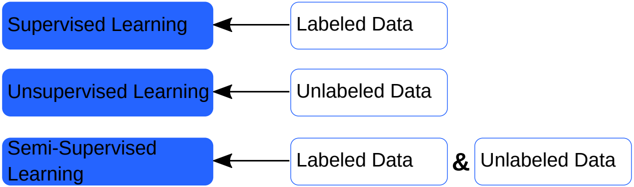
Machine Learning

- Supervised Learning
  - Classification: table, time series classifiction, image classification
  - Regression: predictions, continuous variables
- Unsupervised Learning
  - Clustering: anomaly detection, data exploration
  - Association: co-occurrence
  - Dimensionality Reduction: data visualization, feature extraction, structure discovery
- Semi-supervised Learning: using small labeled data set and large unlabeled
- Reinforcement Learning: robot navigation, survey strategy planning

The Role of Data in Training Process

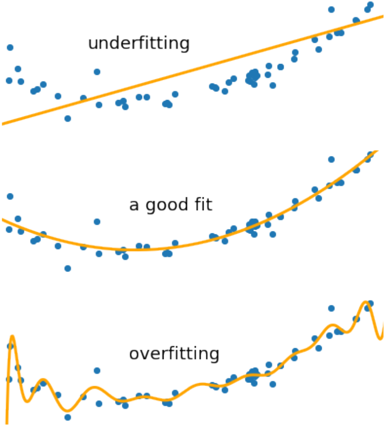
Supervised Learning learns from labeled training set data by iteratively making predictions on the data and adjusting for the correct answer. This makes supervised Learning models **more accurate** than unsupervised learning models.

Unsupervised Learning models work on their own to discover the inherent structure of unlabeled data. The unsupervised learning algorithm works with unlabeled data, in which the output is based solely on the collection of perceptions. This makes unsupervised methods **more flexible** to deal with new data.



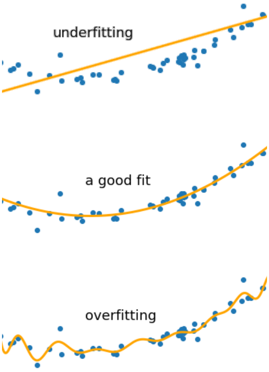
Challenges and Limitations

In most scenarios, the cause of the poor performance of any machine learning algorithm is due to **underfitting or overfitting**.



Challenges and Limitations

In most scenarios, the cause of the poor performance of any machine learning algorithm is due to **underfitting or overfitting**.



Underfitting is a scenario where the machine learning model can neither learn the relationship between variables in the data nor predict a new data point correctly. In other words, the machine learning system hasn't found a correlation between data.

Overfitting occurs when the machine learning model learns from the training data a little too much, attempting to fit every point on the curve and, as a result, memorizes the data patterns. In other words, it narrowed its focus too much on the examples given, making it unable to see the bigger picture and fails to predict new data points.

Challenges and Limitations

Underfitting can occur when:

- The model was trained using the wrong features.
- The model is too simple and can't remember enough features.
- The target data is too varied or complex - the training set doesn't represent the target data's distribution realistically.

Challenges and Limitations

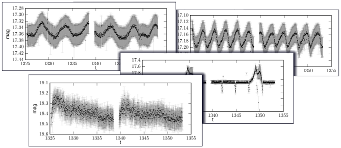
Underfitting can occur when:

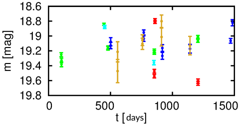
- The model was trained using the wrong features.
- The model is too simple and can't remember enough features.
- The target data is too varied or complex - the training set doesn't represent the target data's distribution realistically.

Overfitting can occur when:

- The model was trained using the wrong parameters and over-observed the training data.
- The model complexity is too high for the presented data variability.
- The training data's labels are too restrictive.

example: Training on 'nice' (high cadence, long baseline, good S/N) light curves - applied to worse. Don't do that!





Key Takeaways: Machine Learning Basics

- Machine learning is a concept that allows computers to learn and improve from experience without being explicitly programmed.
- Machine learning works by the approach of *find the pattern, apply the pattern*.
- Machine Learning consists of Supervised, Unsupervised, Reinforcement, and Semi-Supervised Learning.
- Supervised learning is useful when dealing with purely labeled datasets for training and knowing how the output should look like.
- Unsupervised Learning is useful for finding the hidden patterns.
- A machine learning model is underfitted when it fails to capture the relationship between the input and output.
- If a machine learning model performs better on the training set than on the test set, then it is likely overfitting: it memorizes the data it was trained on without being able to generalize.
- Machine learning is part of many nowadays everyday applications such as Google Maps, Alexa, Youtube...
- It is increasingly important for astronomy for such as source classification, anomaly detection, survey strategy planning...

An Outlook: Neural Networks & Deep Learning

three **ingredients**:

- discriminative neural network models (supervised learning)
- large labeled datasets
- lots of computer power

in particular **useful** for:

working with data sets for which computers typically don't perform well and specific algorithms are hard to write

- images
- videos
- speech/ time series data

⇨ ideal for **mining large astronomical survey datasets**