Physics Without Frontiers: Chile
School on machine learning in physics

13-17 JANUARY 2025 | VALPARAÍSO, CHILE

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

CCTVal

AI/ML Applications in Astrophysics

Pía Amigo
Departamento de Física, USM
pia.amigo@usm.cl

Physics without Frontiers: Chile | School of Machine Learning, UTFSM, January 13-17 2025

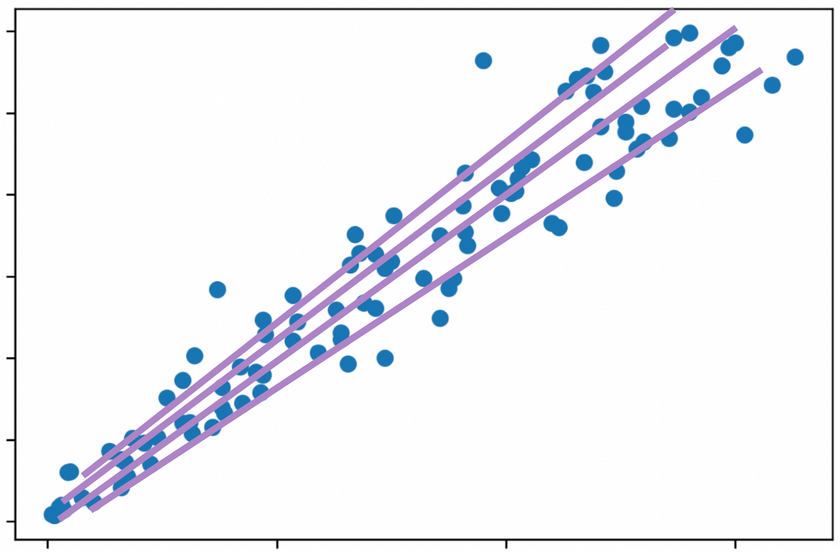
Outline

- Regression problems
- Evaluate and improve your model
- Trees and forests
- Photometric redshift

Regression problems

- In regression problems the prediction is a continuous variable (instead of a class)

- It finds the best-fitting line (plane/hyperplane) that describes the relationship between the variables.



- The linear regression predicts the values
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

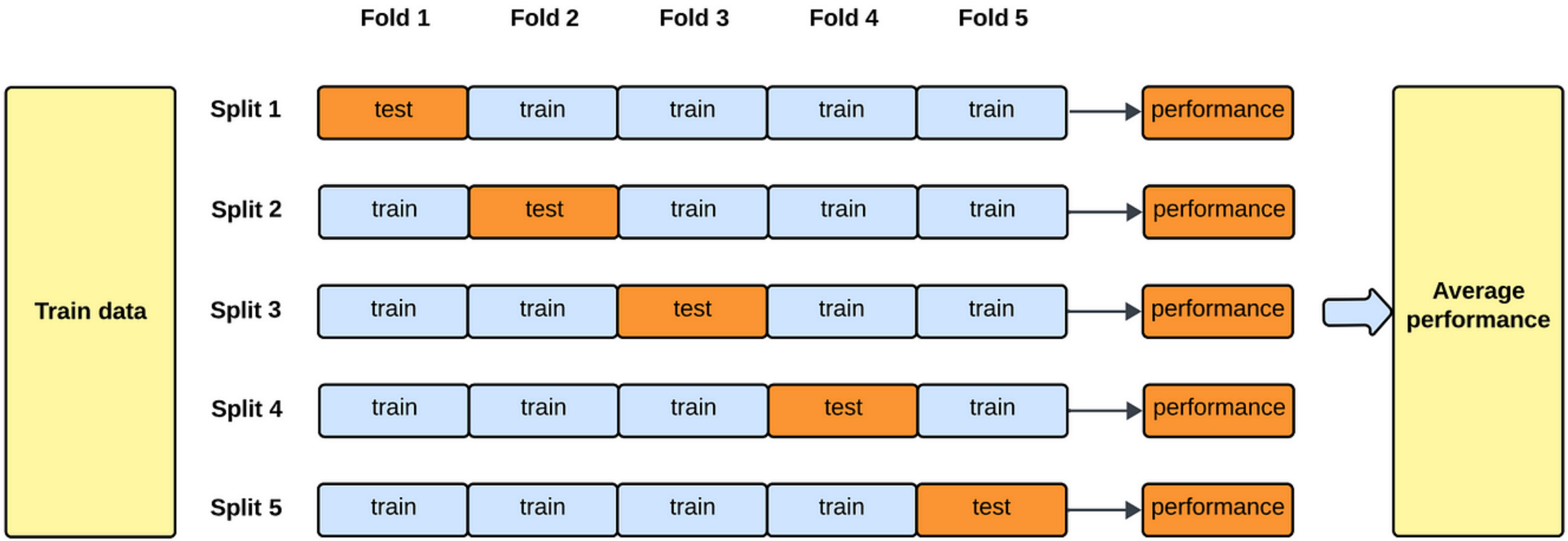
- We need to find the set of parameters β that minimizes the loss function Mean Square Error

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

Cross-validation

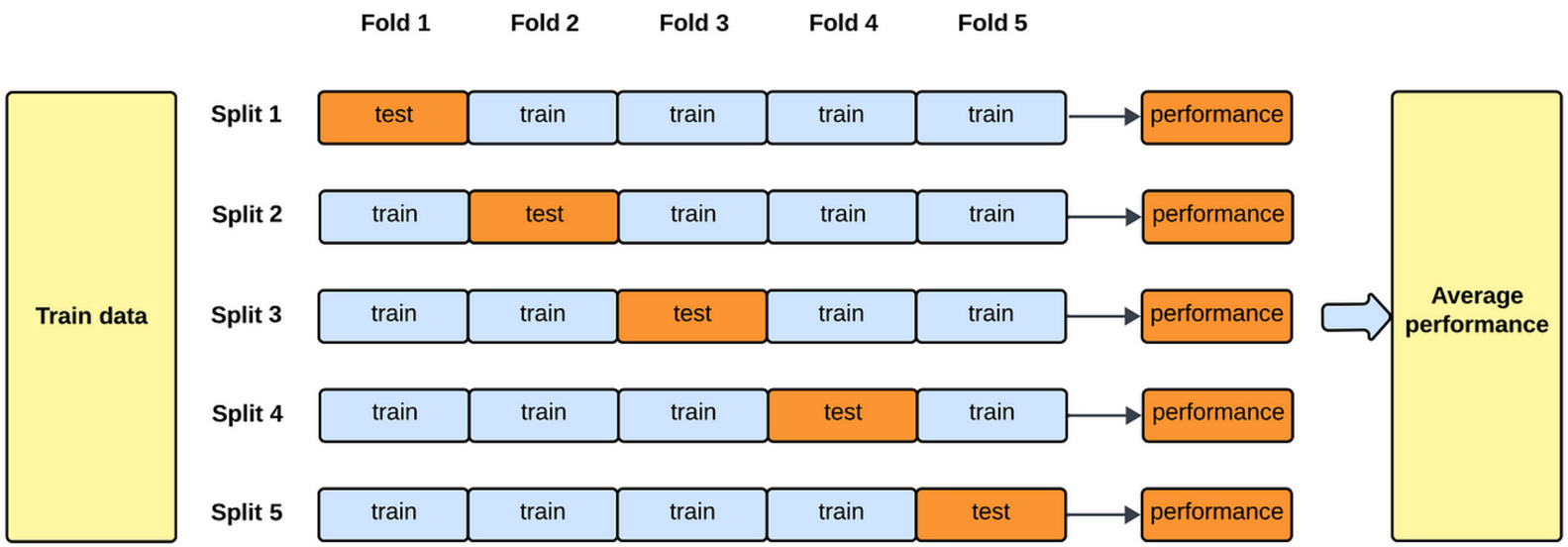
- The choice of training/testing sets can significantly affect model performance
- We want to estimate the uncertainty associated with the statistical variability of the data
- **Cross-validation** is a technique used to evaluate the generalization ability of a machine learning model, i.e., how well it will perform on new or unseen data.
- Its main goal is to prevent the model from overfitting the training set
- The **k-fold** technique is commonly used, but there are other variants, such as **Leave-One-Out** and **Stratified k-fold.**

K-fold Cross-validation (visual explanation)



K-fold Cross-validation (visual explanation)

- We use all training data (all objects are equally represented
- The mean and standard deviation give us an idea of the average performance and uncertainty
- Of course, it takes more time
- How many k? 5-10 is recommended, depending on how long it takes for the model to run.



Diagnosing an ML algorithm

BIAS

The algorithm cannot capture the complexities in the underlying relationships between variables

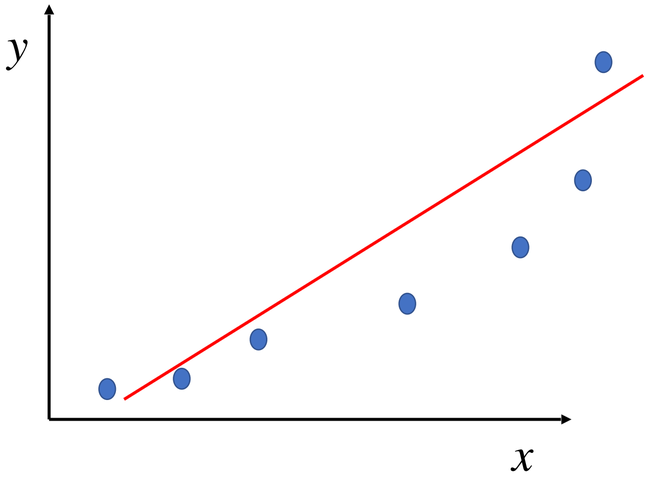
VARIANCE

The algorithm is fitting all the small variations in the training data and cannot generalize

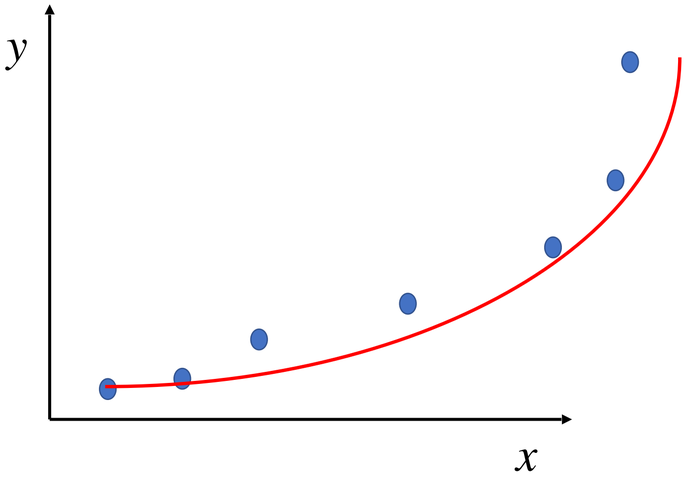
UNDERFITTING

OVERFITTING

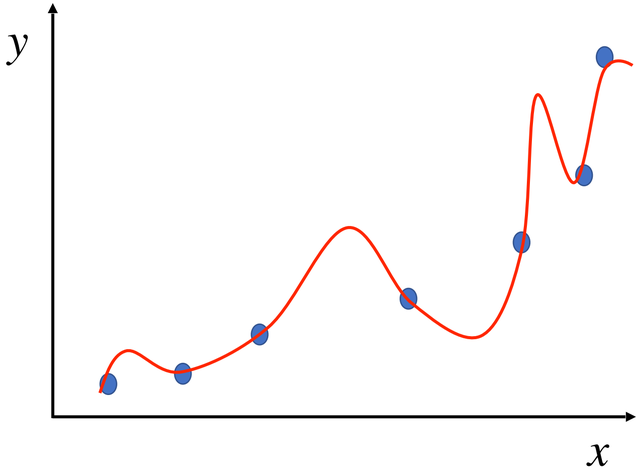
Diagnosing an ML algorithm



Not complex enough (high bias/underfitting)

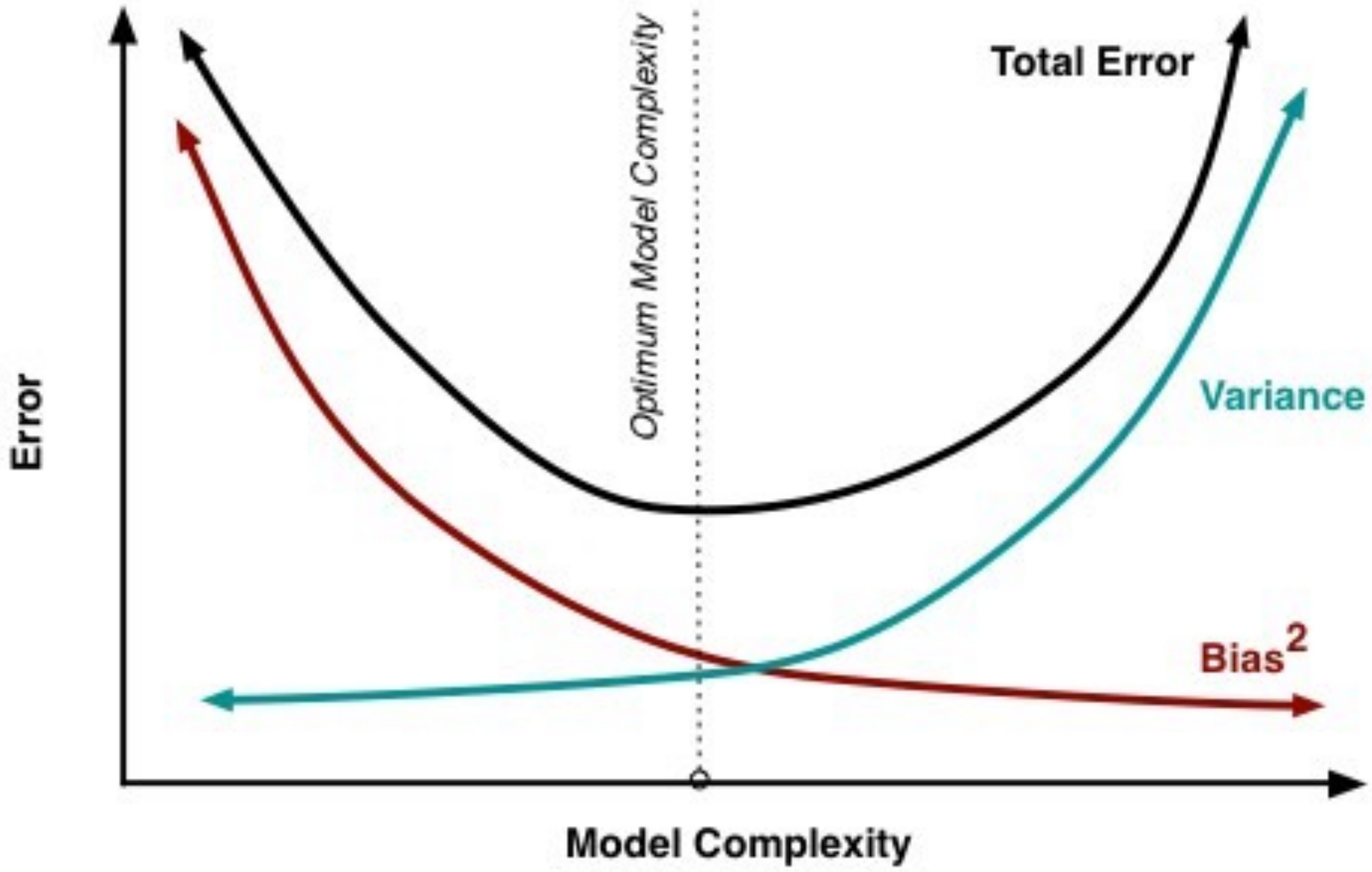


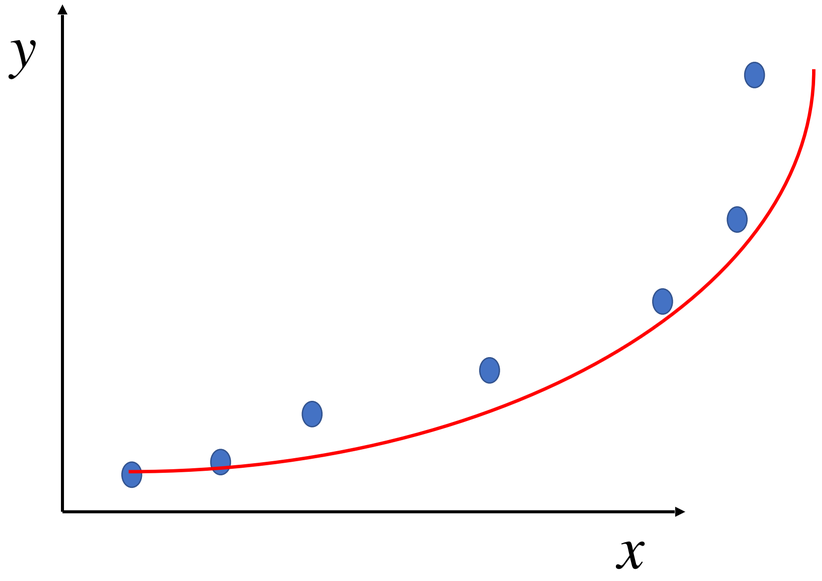
Good trade-off between complexity and performance



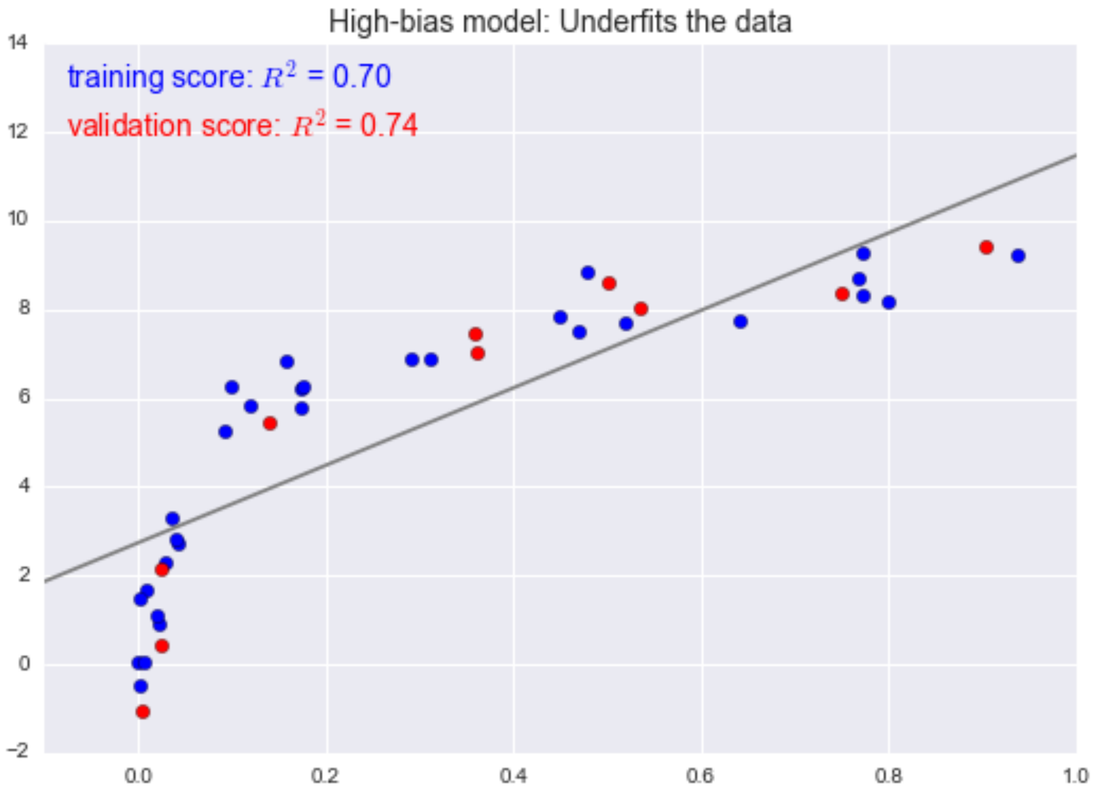
Too complex (high variance/overfitting)

Diagnosing an ML algorithm

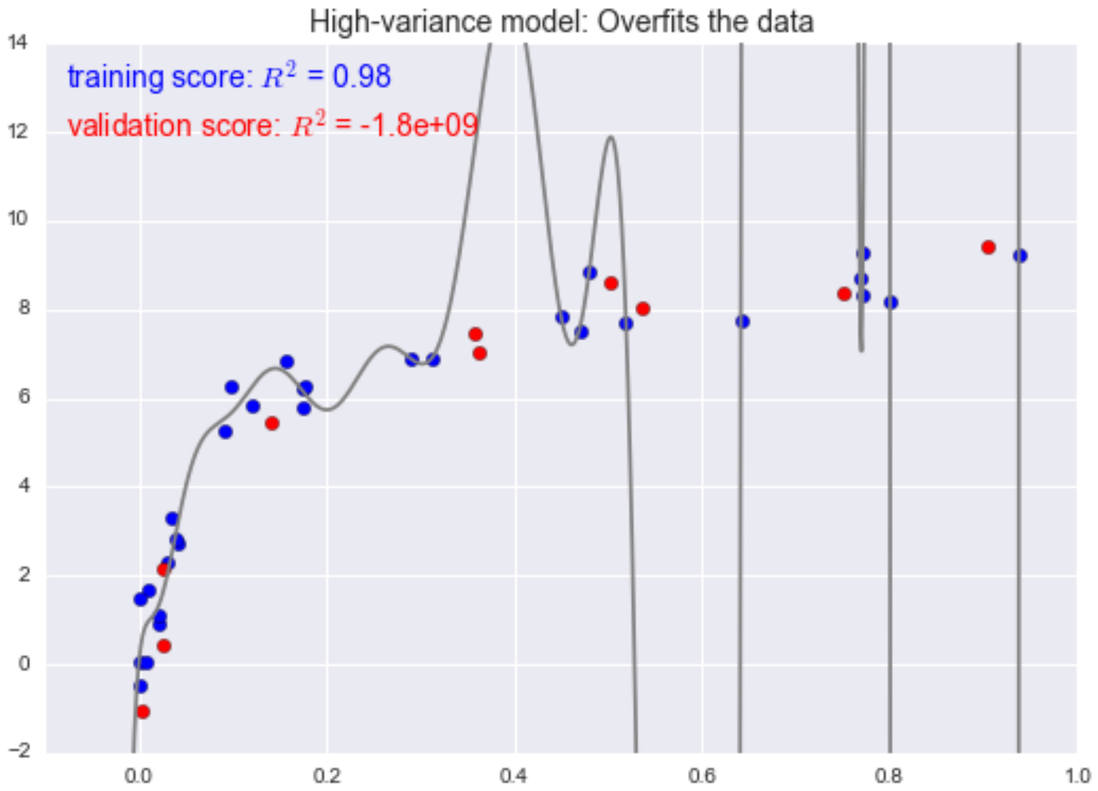




High bias vs high variance



High bias: train and test errors are similar



High variance: there is a gap between test and train error because the algorithm does not generalize well

https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html

How can we improve the model?

HIGH BIAS

- Using different features
- Creating new features
- Trying new parameters or more complex algorithms

HIGH VARIANCE

- Reducing the number of features
- Trying less complex algorithms

We can also check if we need more data

Learning curves

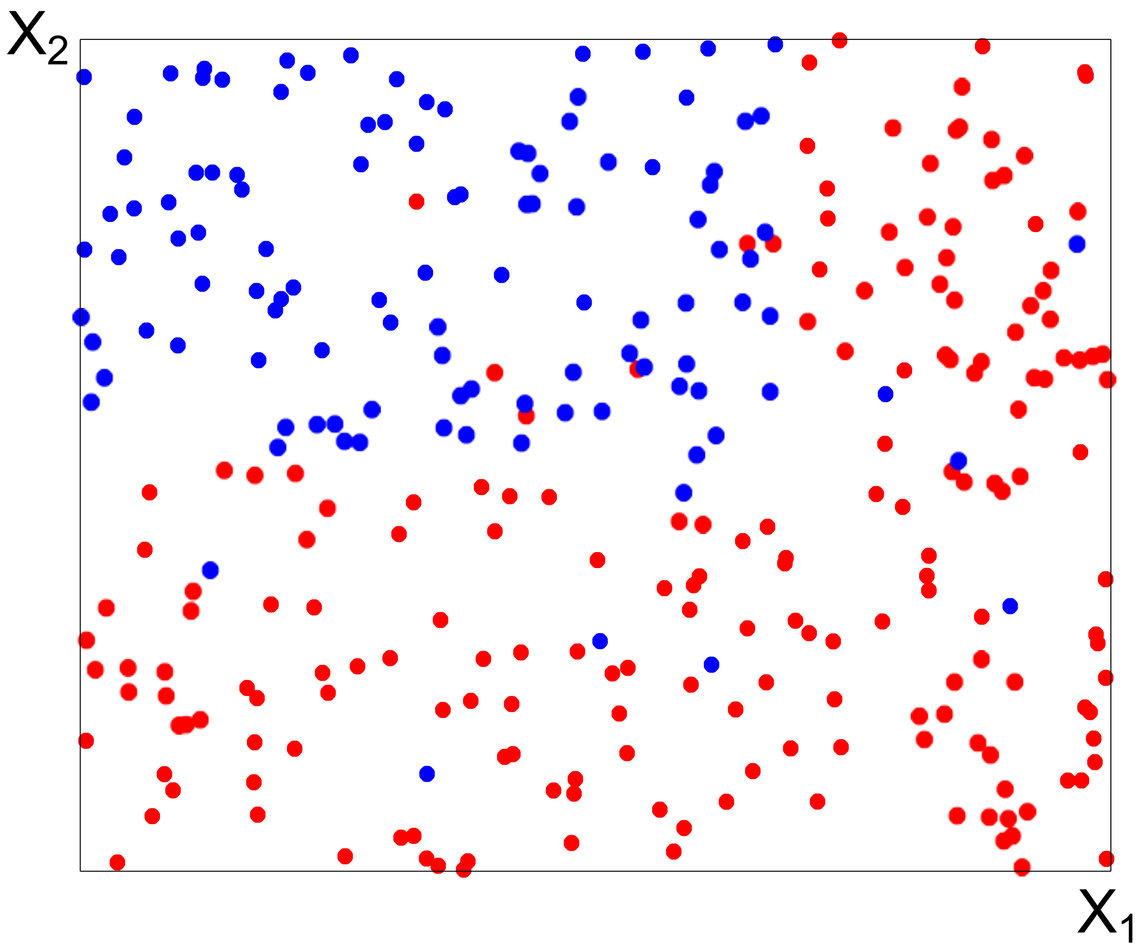
Algorithm performance for training and testing sets as a function of training set size

Decision trees

- It works by splitting the data based on different values of the variables or features.
- If the variables are categorical, the split is based on yes/no.
- If the variable is numerical, the split is based on a certain value
- They are easy to interpret and visualize.

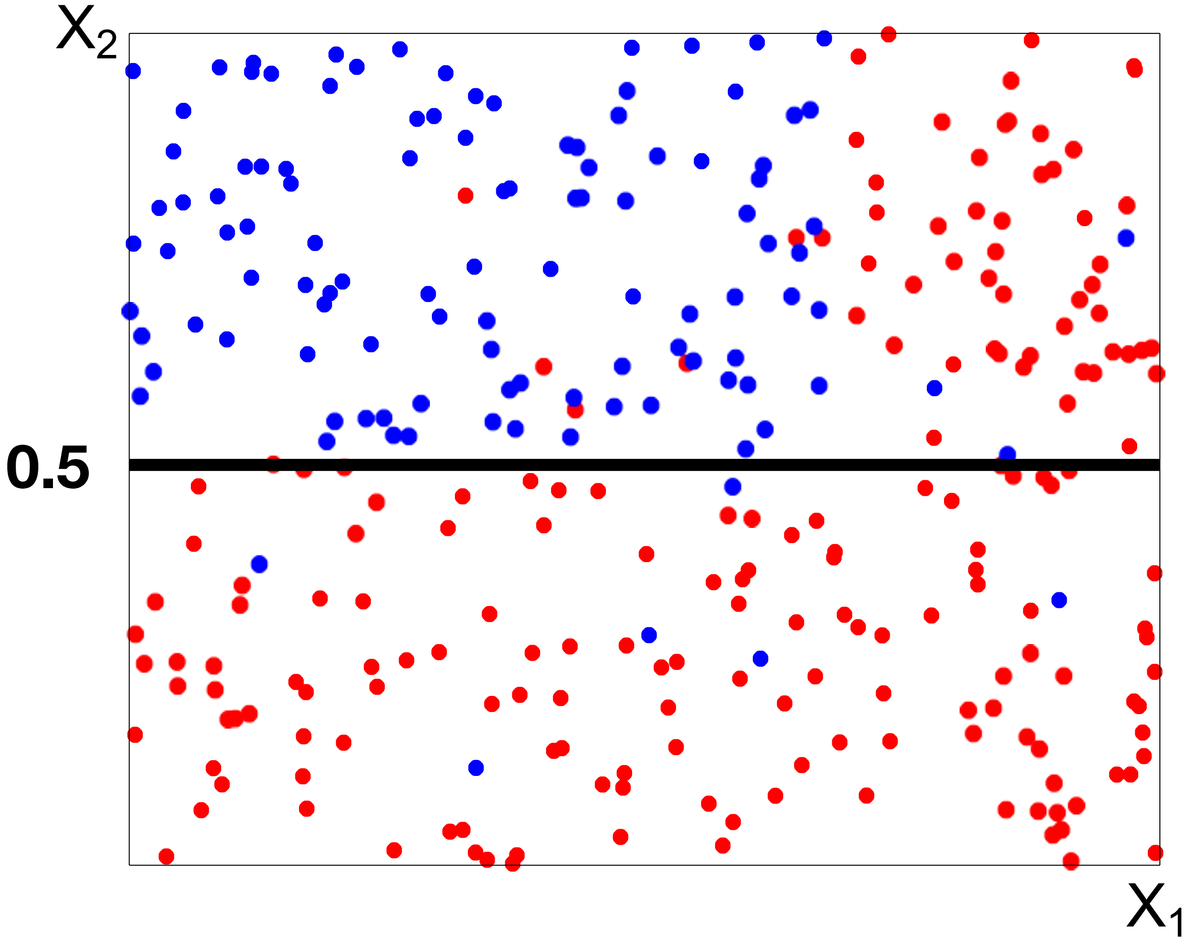
Example

How do we separate the classes in this dataset with two features?



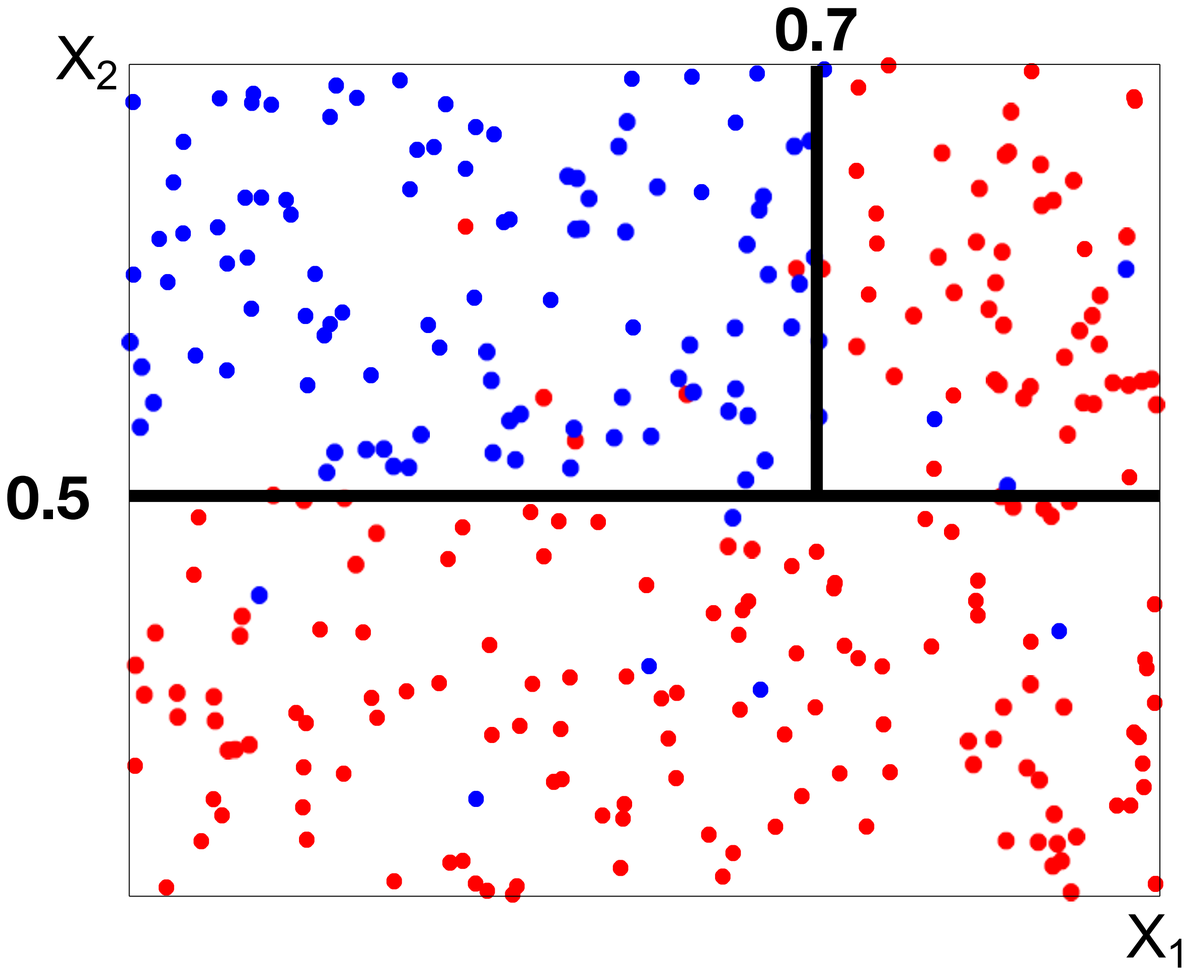
Example

How do we separate the classes in this dataset with two features?

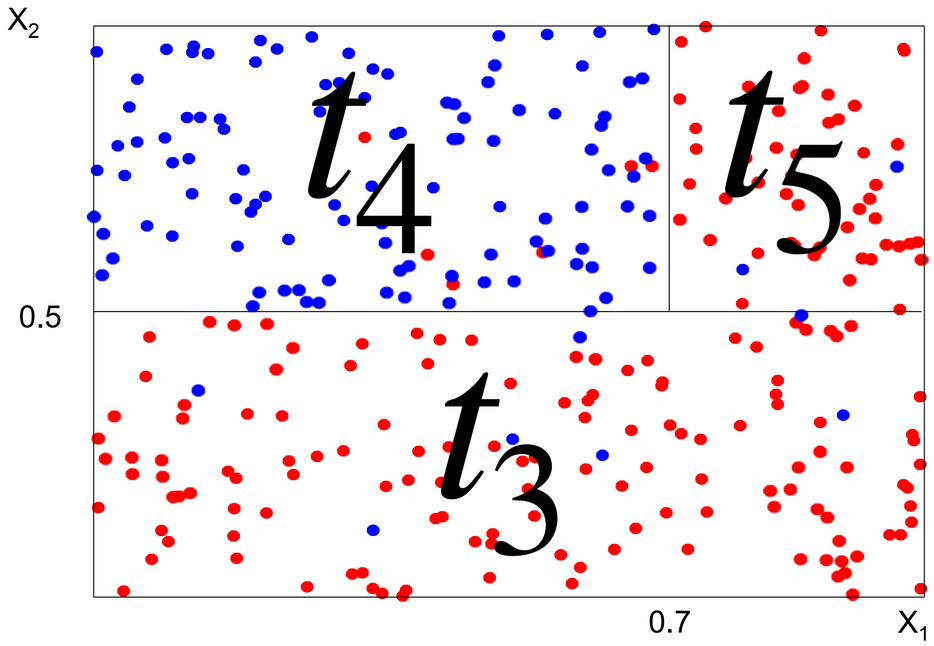


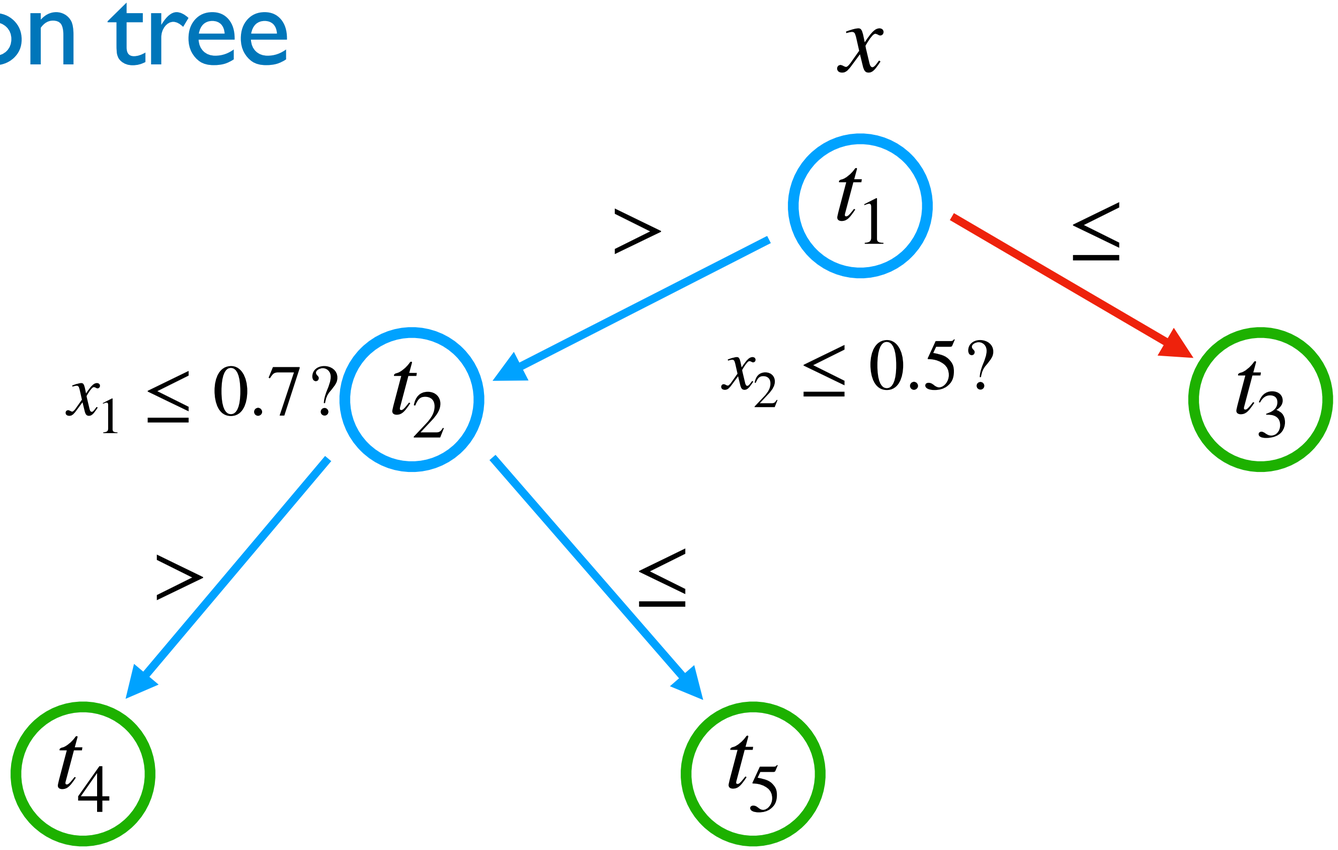
Example

How do we separate the classes in this dataset with two features?



Nodes define the decision tree



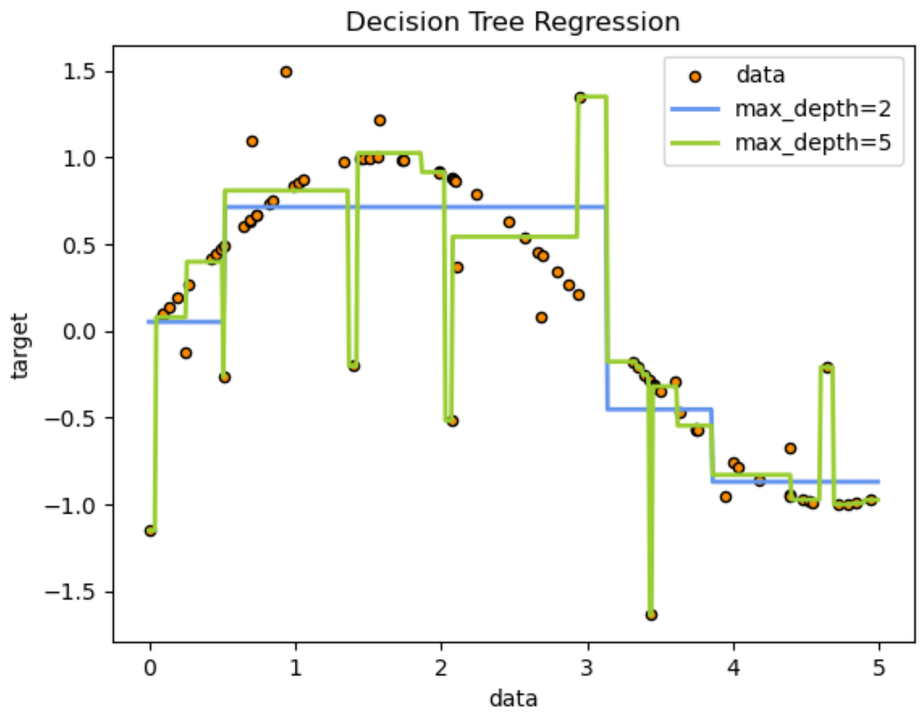


Split node

Terminal node

At a terminal node (leaf), the model has completed the classification (and all objects in that leaf belong to the same class)

Decision trees for regression



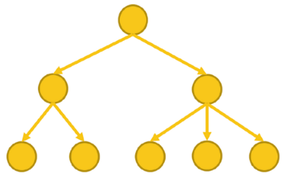
- At each node, the feature and threshold that minimize the MSE in the resulting groups are chosen.
  The split is performed recursively until a stopping criterion is met (e.g., maximum depth, minimum observations per node).

  For each possible split:

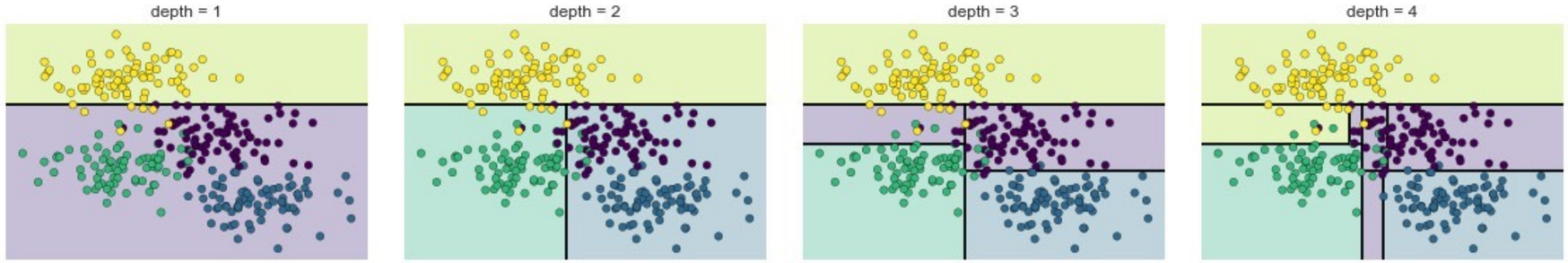
  - The mean of the target variable is calculated for each group.
  - The MSE is measured as the average of the squared differences between the values and the group mean.
  - The split that minimizes the total MSE (weighted average) is selected.

  The prediction for a new observation is the mean value of the target variable in the leaf where the observation falls.

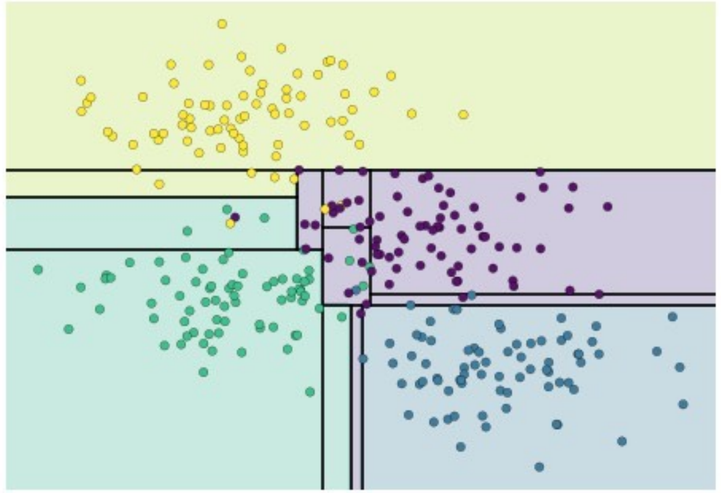
Decision trees for regression



- Fast
- Interpretable
- Low bias
- They usually tend to overfit (high variance)



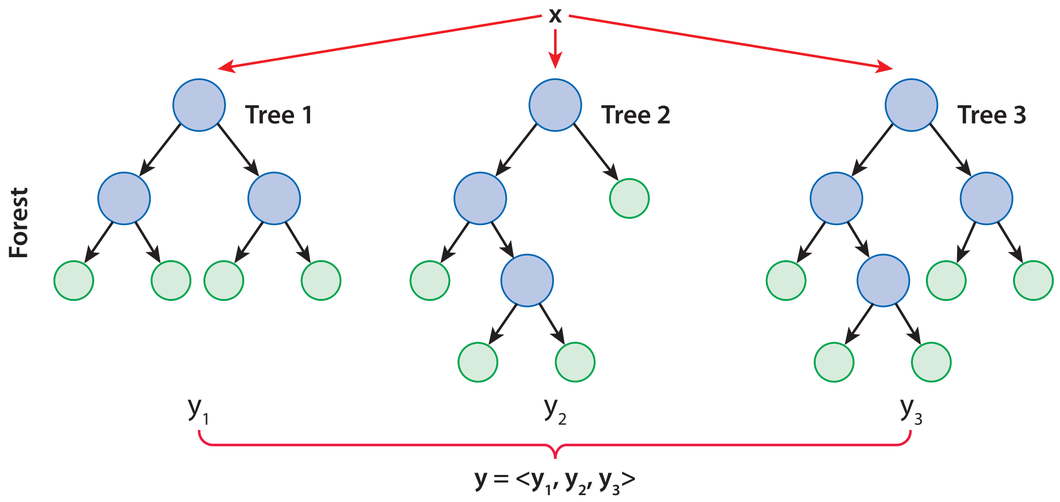
depth=5



We need to "prune" the tree to remove the smallest leaves.
`min_impurity_decrease`, `min_samples_split`, `min_samples_leaf`, `max_depth`

https://jakevdp.github.io/PythonDataScienceHandbook/06.00-figure-code.html#Decision-Tree-Levels

Ensemble methods: Random Forests



- The original dataset is replicated M times using bootstrap sampling with replacement. A decision tree will be built on each of the M training sets.
- When creating the M decision trees, the features participating in the selection of the optimal split are a random subset of all available features.
- The M decision trees are constructed independently, and each tree provides a prediction for each element in the training set.
- **The final prediction is simply the average of all predictions (for regression problems) or the majority vote (for classification problems).**

Hyperparameters optimization

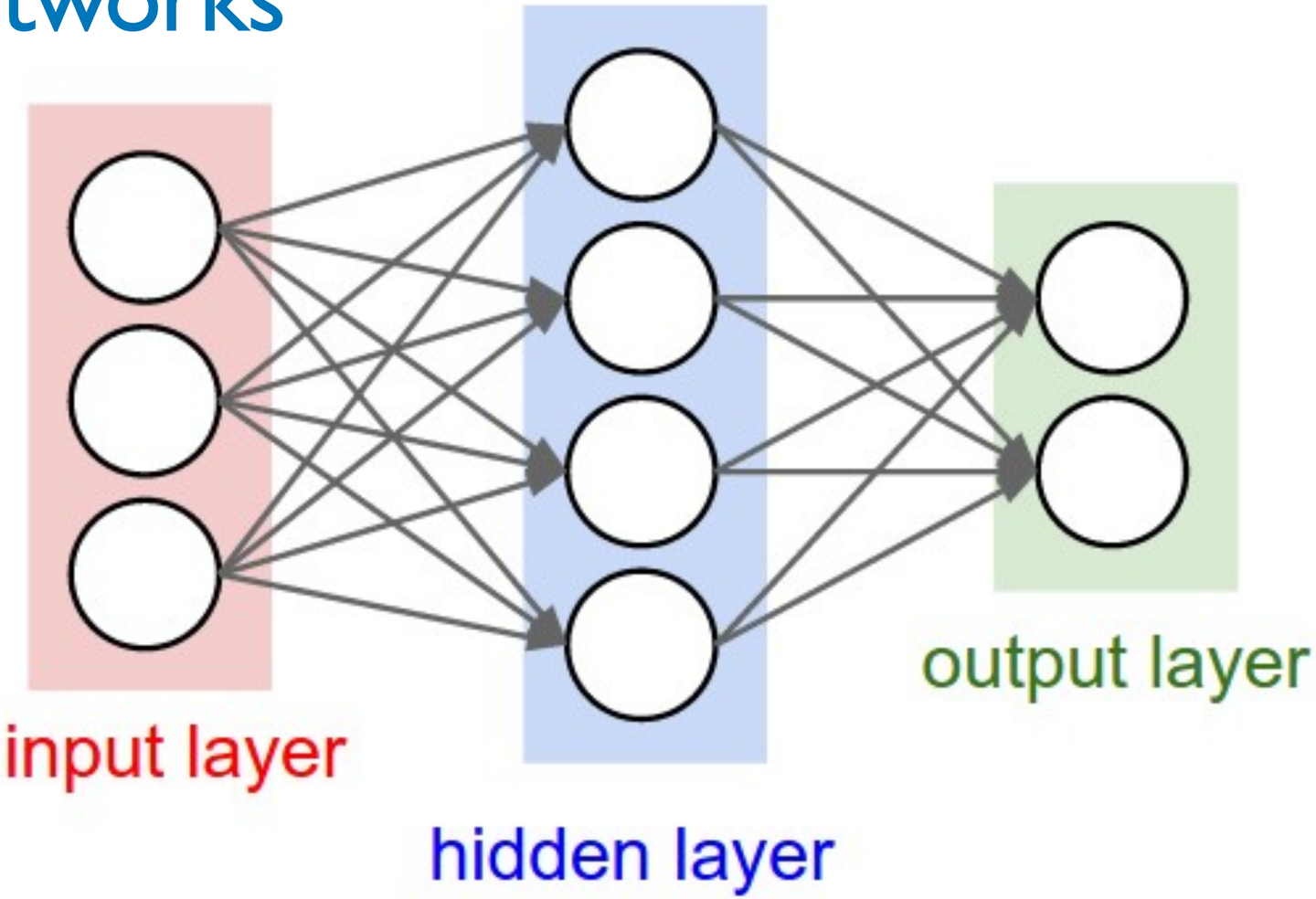
Decision Trees (DT):

- max_depth: Maximum depth of the tree (controls over/underfitting).
- min_samples_split: Minimum samples to split a node.
- min_samples_leaf: Minimum samples in a leaf node (prevents small leaves).
- max_features: Number of features to consider for the best split.
- criterion: Metric for split quality (mse, gini, entropy).

Random Forests (RF):

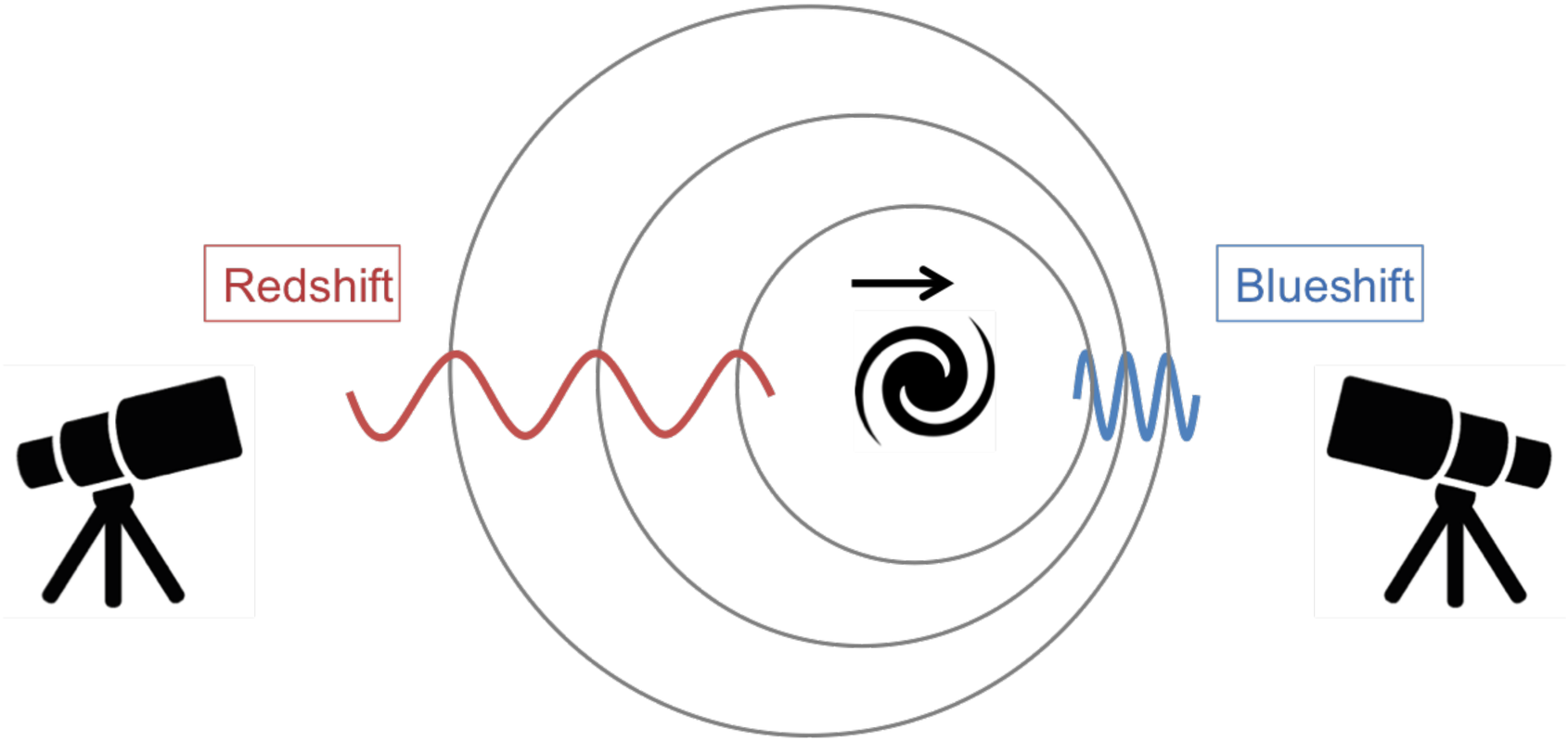
- Inherits DT Hyperparameters.
- n_estimators: Number of trees in the forest.
- bootstrap: Use bootstrap sampling (default: True).
- max_samples: Number of samples per tree (if bootstrap=True).
- max_features: Features considered for splits (sqrt, log2, None).

Neural networks



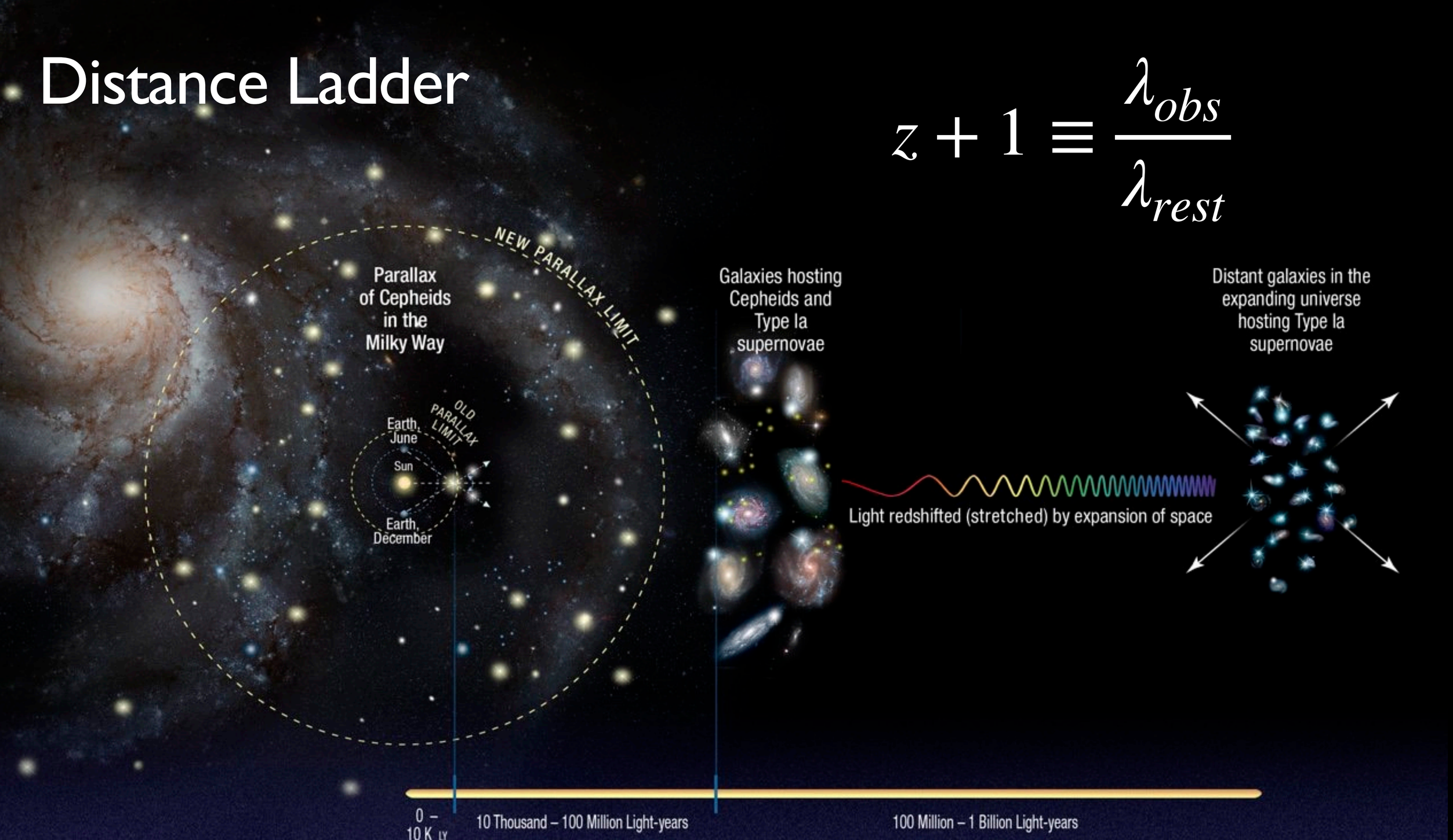
PLAYGROUND.TENSORFLOW.ORG

A 3D map of the universe

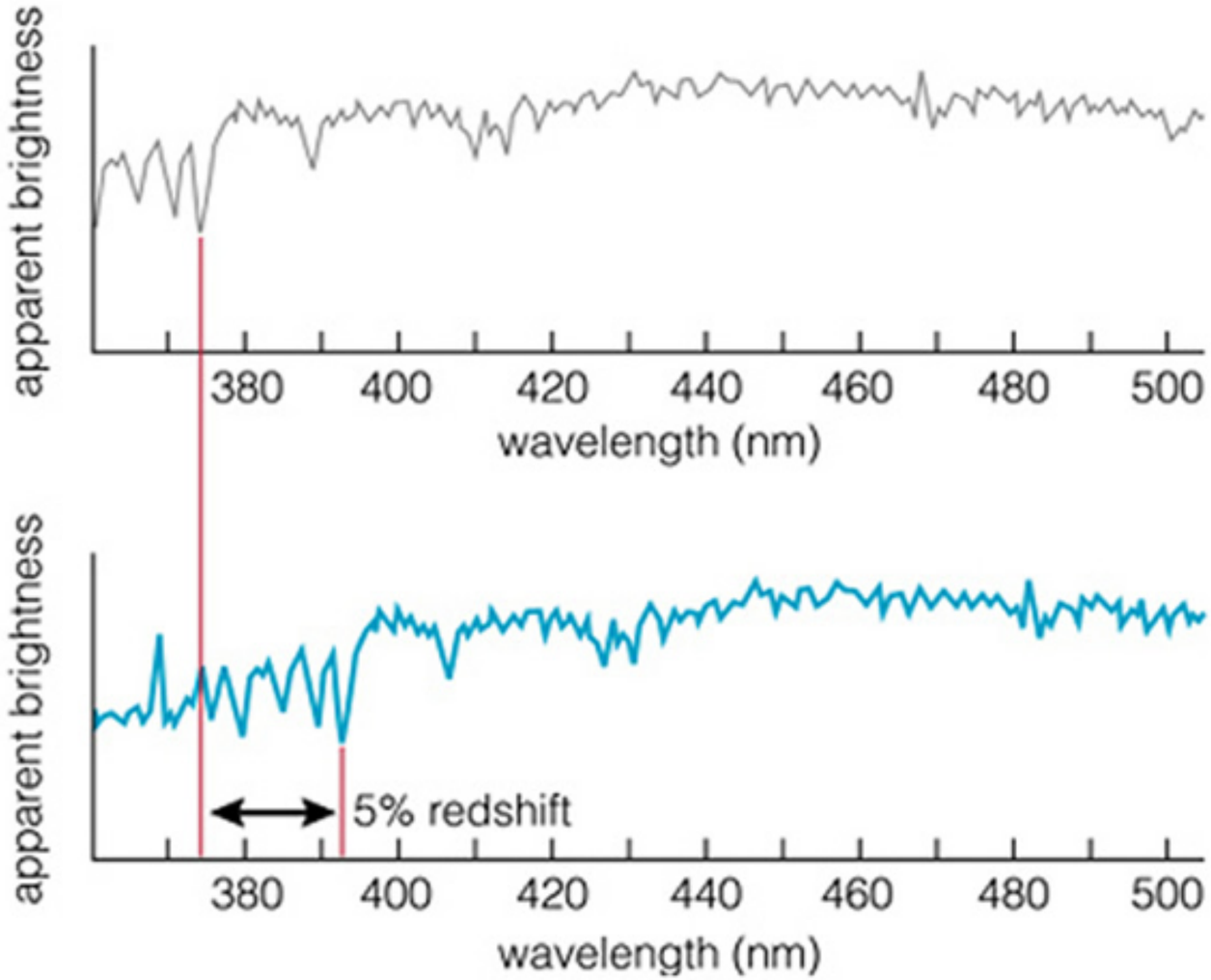


Distance Ladder

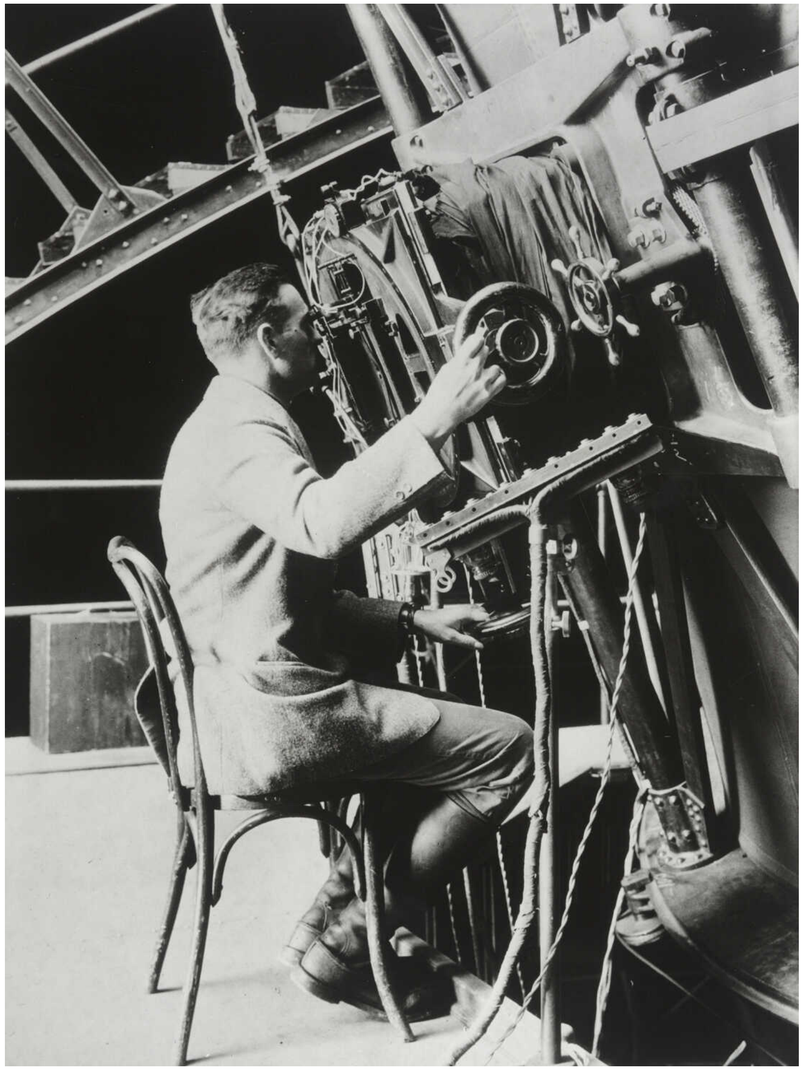
$$z + 1 \equiv \frac{\lambda_{obs}}{\lambda_{rest}}$$

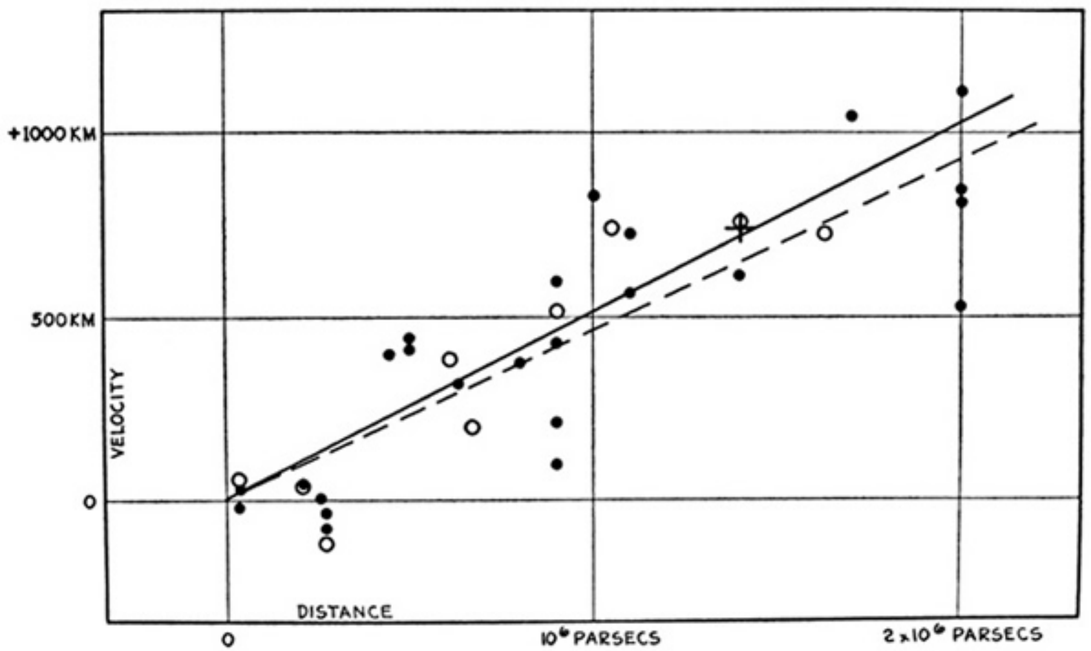


Spectroscopic redshift



Expansion of the universe

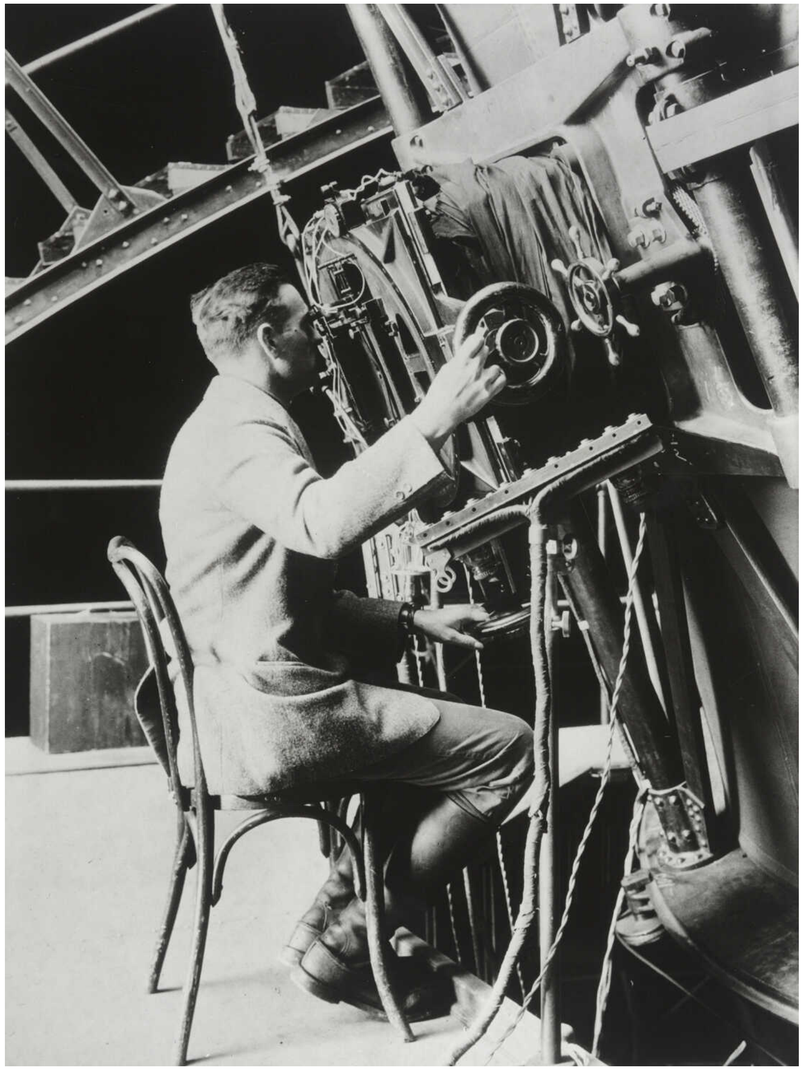


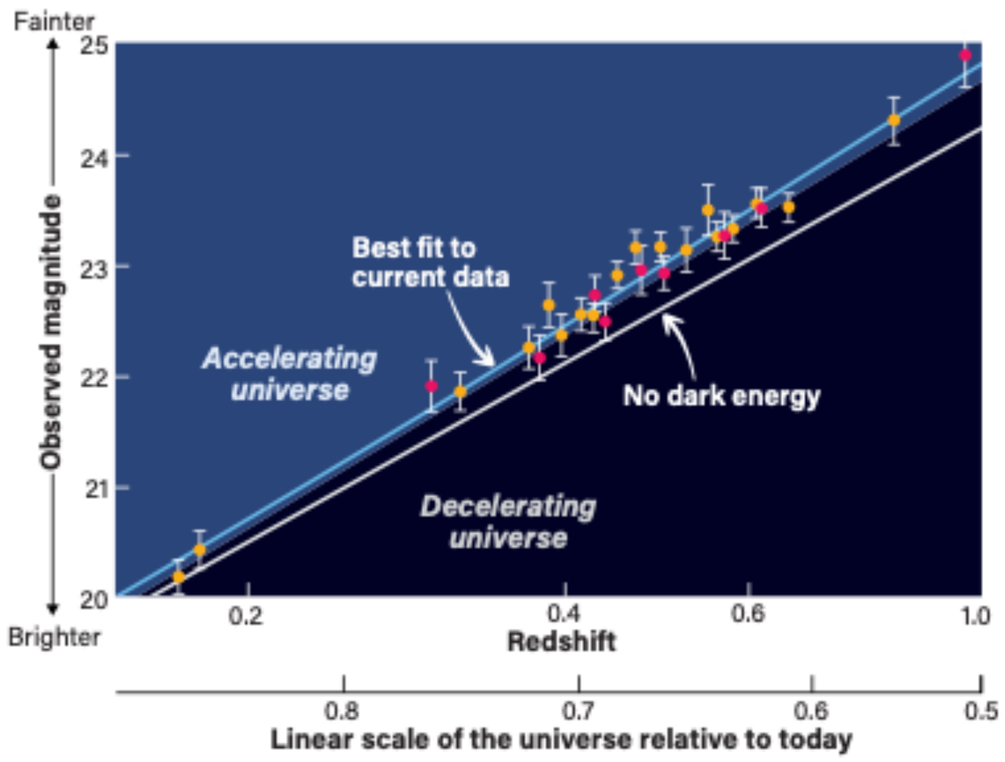


$$V = H_o d$$

The universe is expanding!

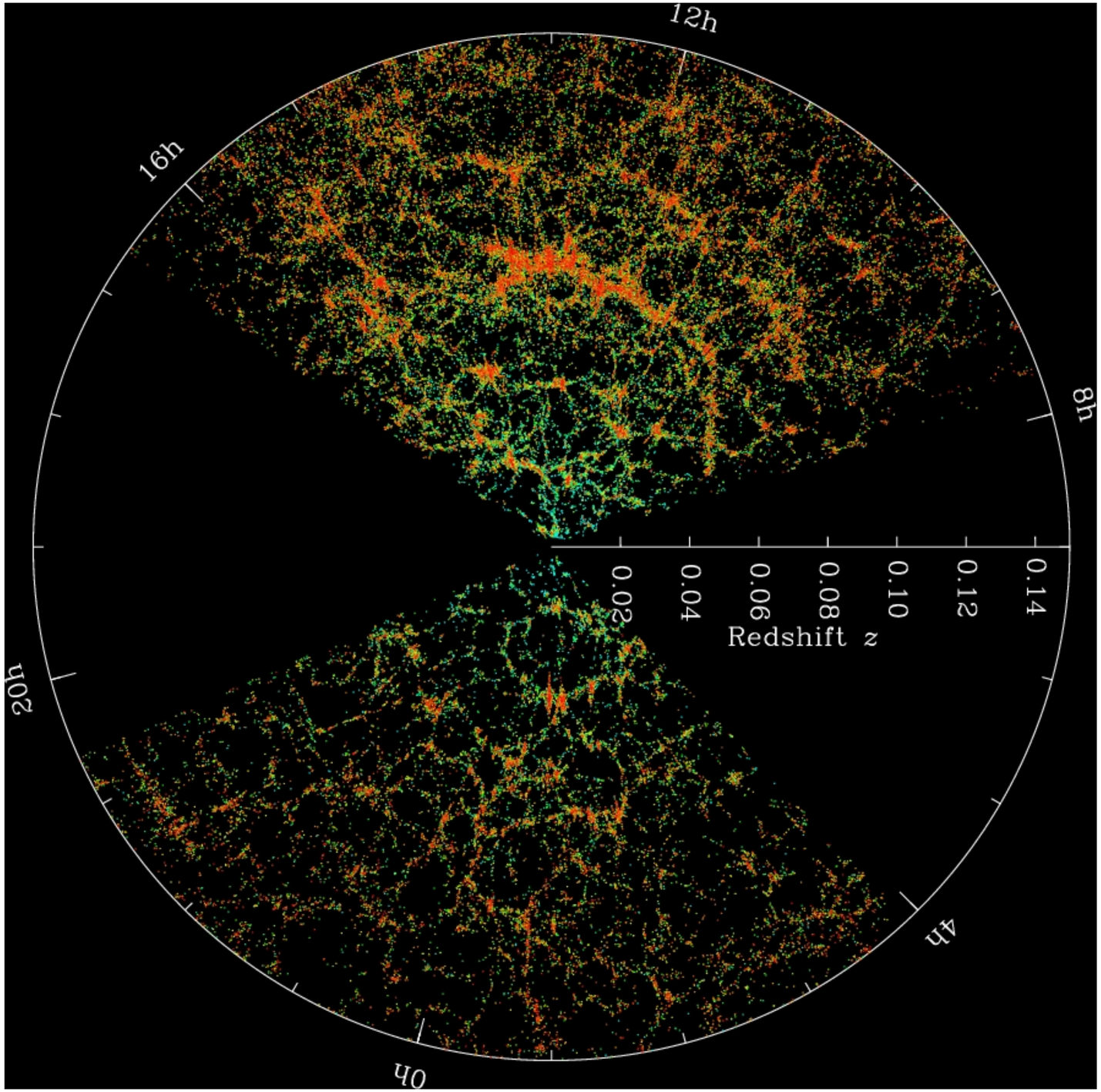
Expansion of the universe





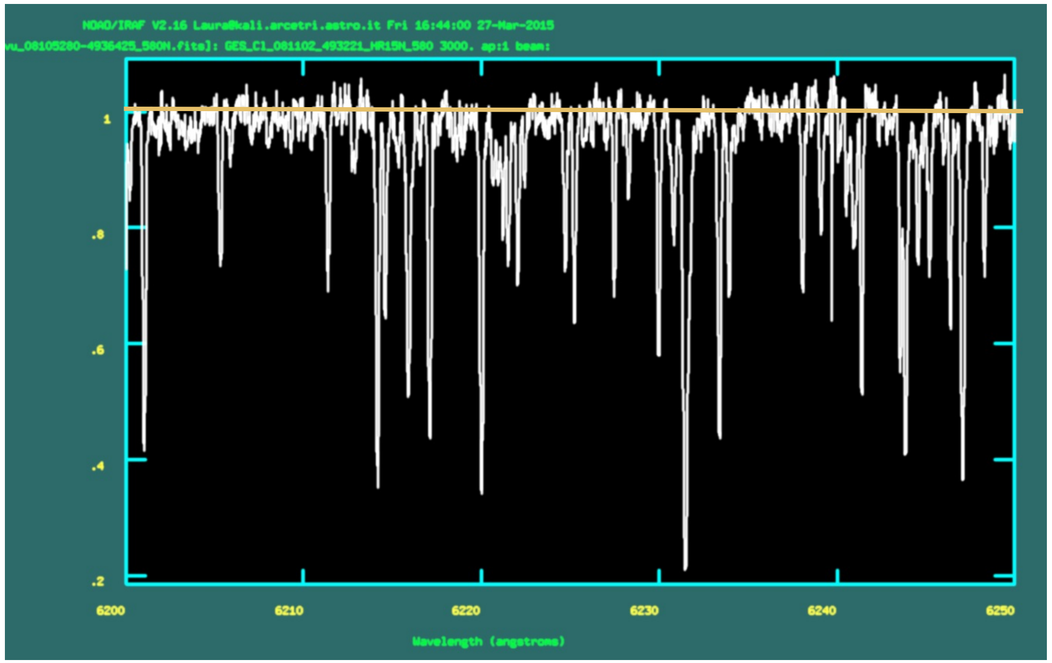
Accelerated expansion!

Large-scale structures in the Universe

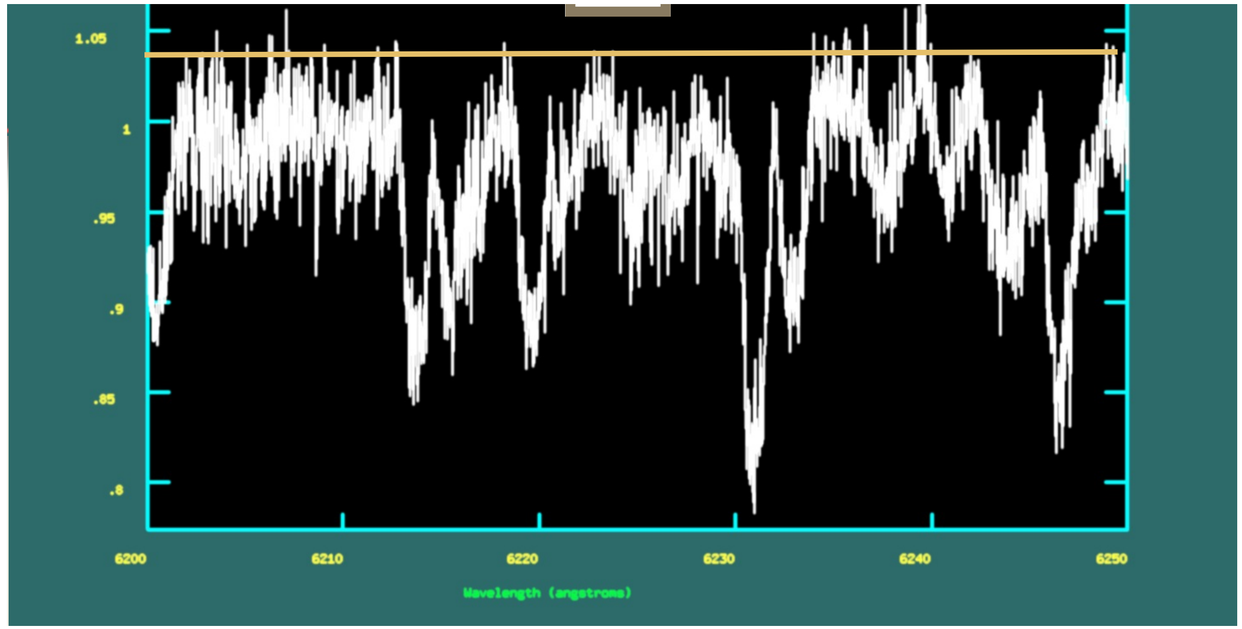


The problem….

- Spectroscopy is expensive, and you need enough resolution and good quality data to measure the redshift



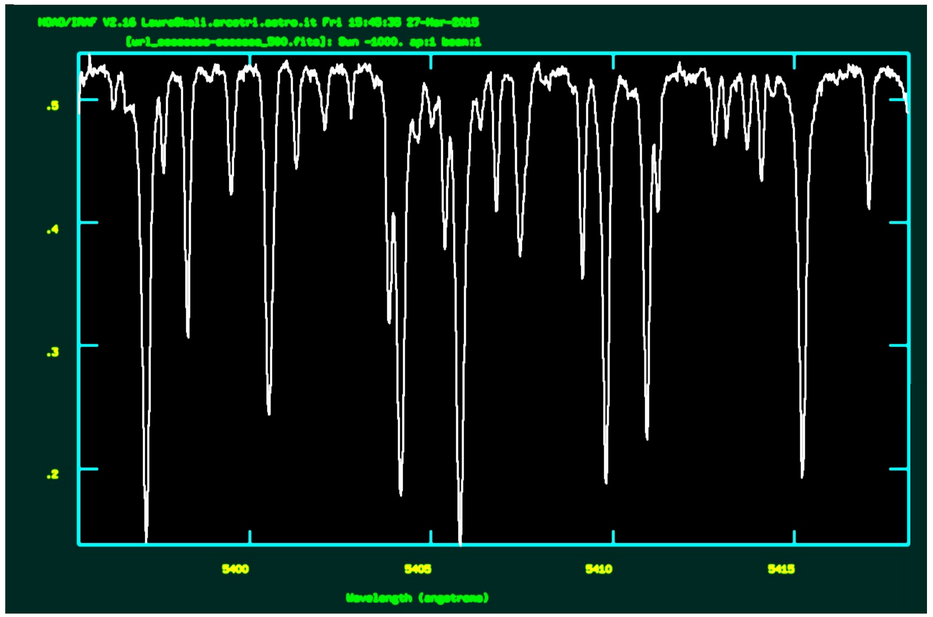
High signal-to-noise ratio to define the continuum and to measure faint lines



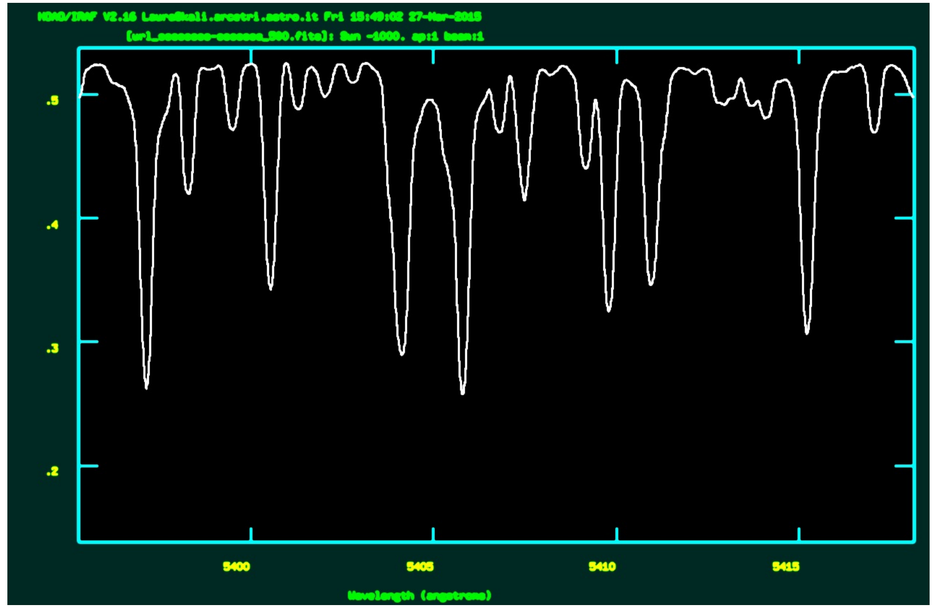
Low signal-to-noise ratio: impossible to define the continuum and to measure lines

The problem….

- Spectroscopy is expensive, and you need enough resolution and good quality data to measure the redshift



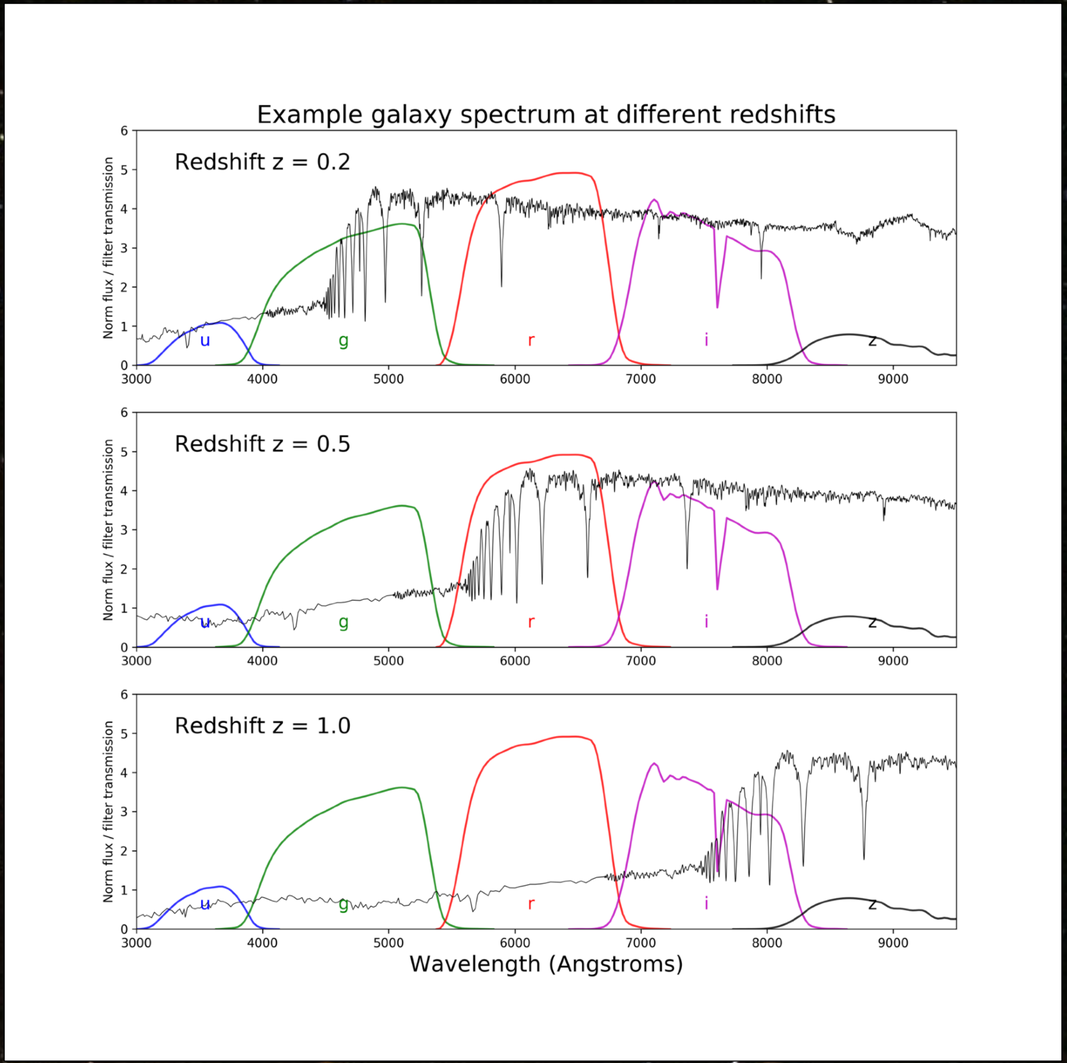
High resolution



Low resolution

Easier way…photometry!

- Photometry is more available for much larger samples
- Can we predict redshift based on photometric information?



Find out this afternoon in the hands-on session!