ML01 — Demystifying Machine Learning: introduction

Romain Gautron

CIAT

March 5, 2019

Introduction to ML

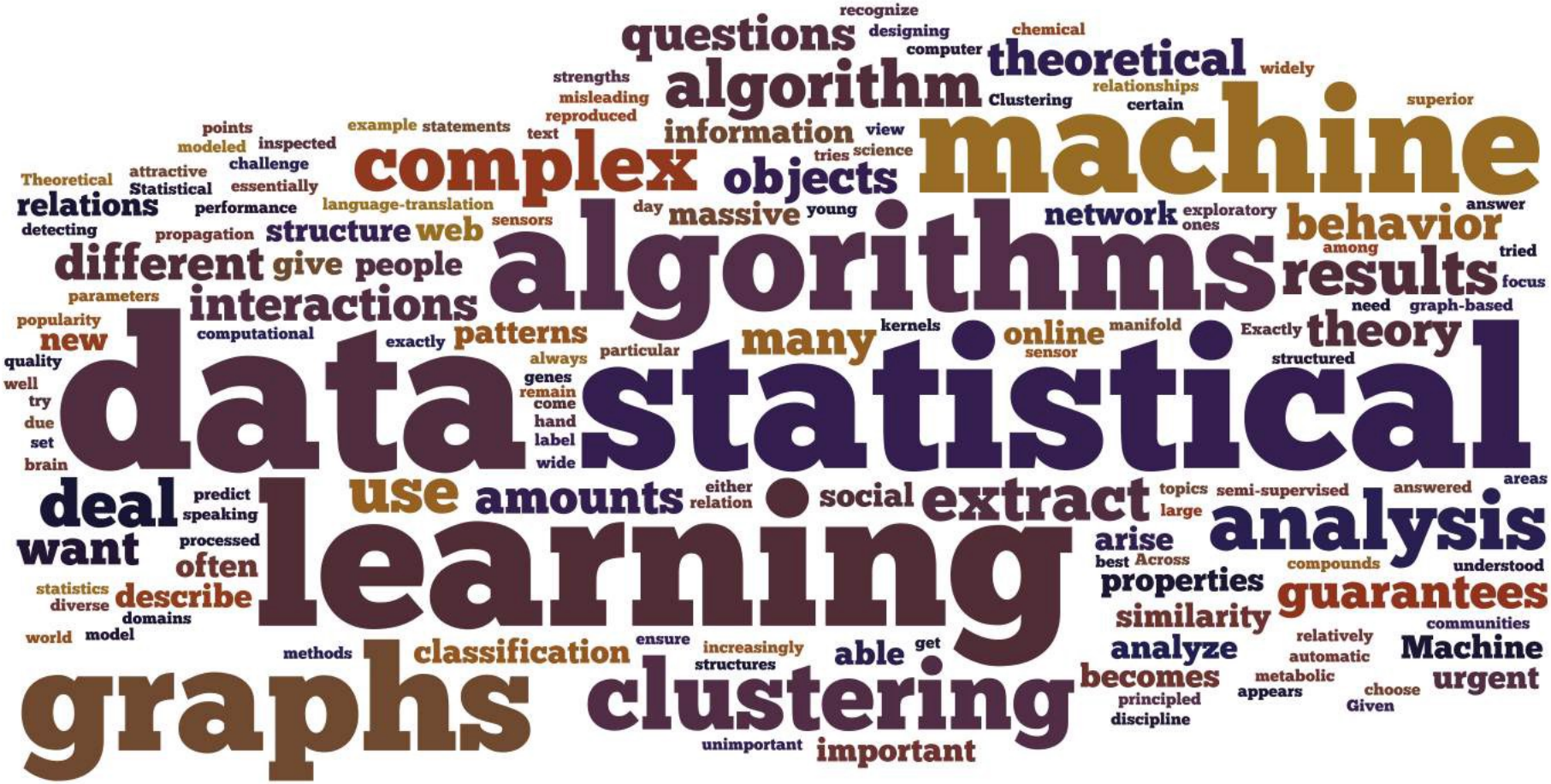
→ What is Machine Learning?

→ What can I do with ML?

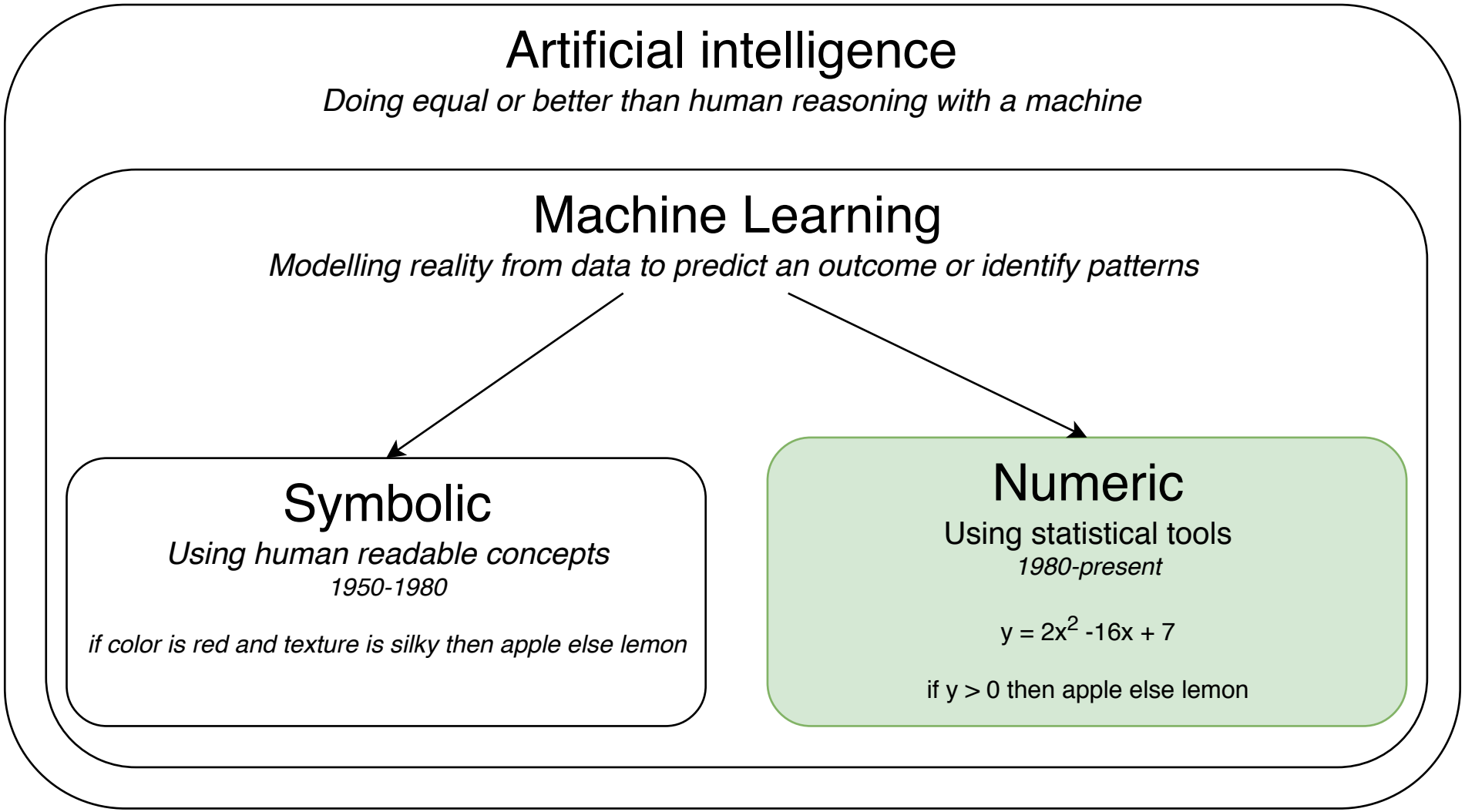
→ What is the fundamental problem we want to solve?

→ The best algorithm for problem solving

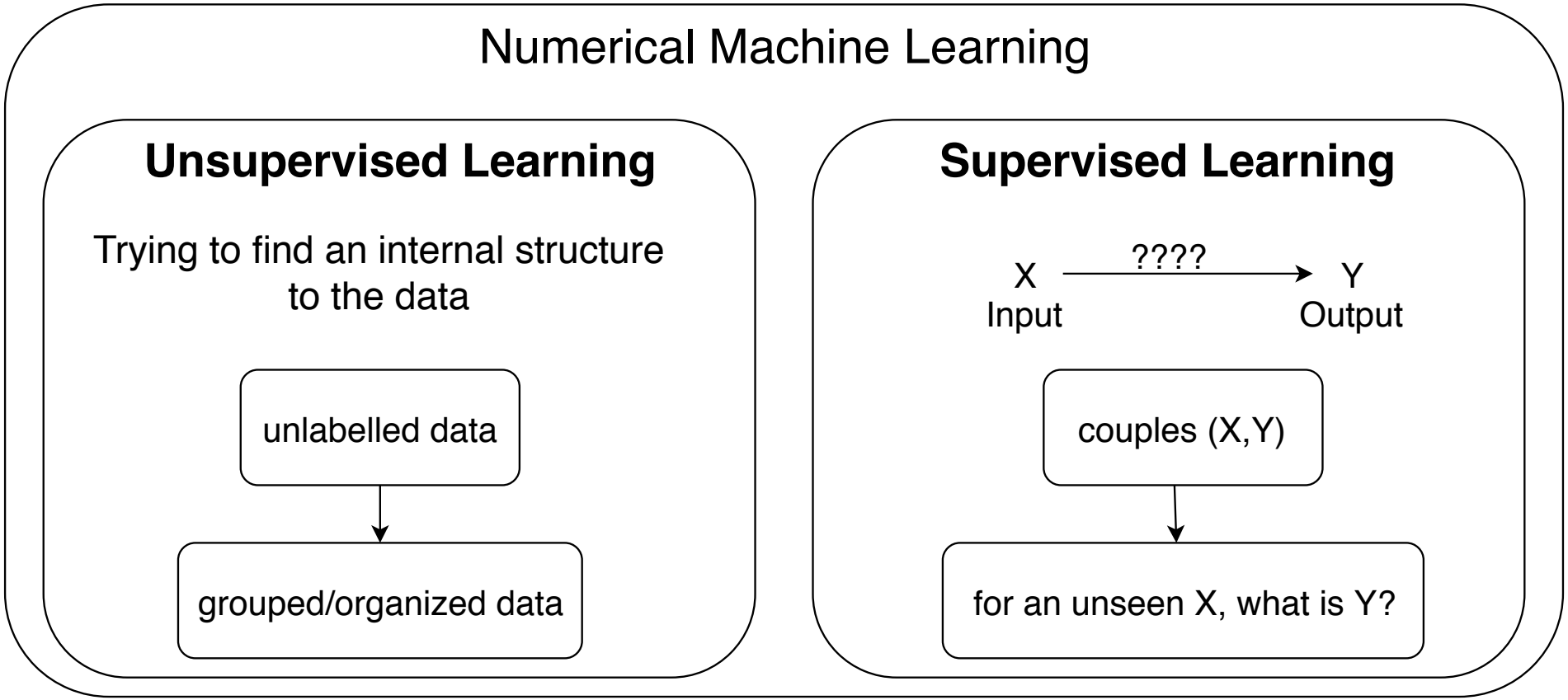
A bunch of words with loose definitions



AI & ML


Artificial intelligence
Doing equal or better than human reasoning with a machine
Machine Learning
Modelling reality from data to predict an outcome or identify patterns
Symbolic
Using human readable concepts
1950-1980
if color is red and texture is silky then apple else lemon
Numeric
Using statistical tools
1980-present
y = 2x2 -16x + 7
if y > 0 then apple else lemon

Supervised VS Unsupervised



Unsupervised Learning example

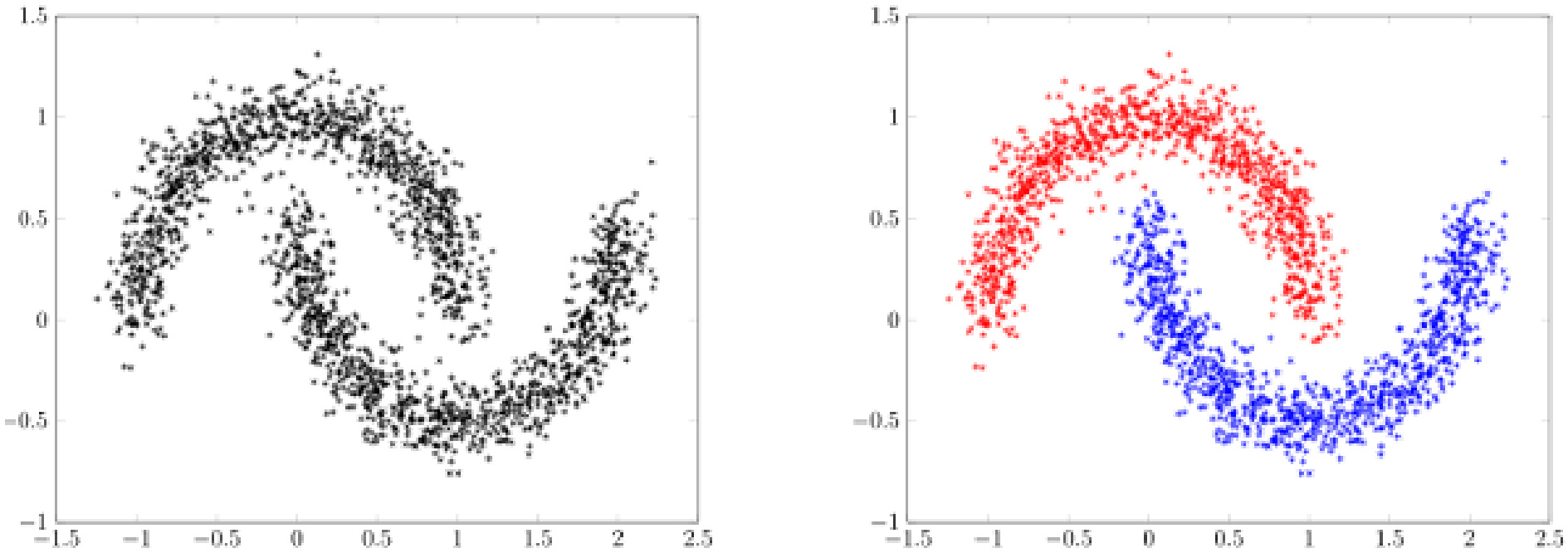
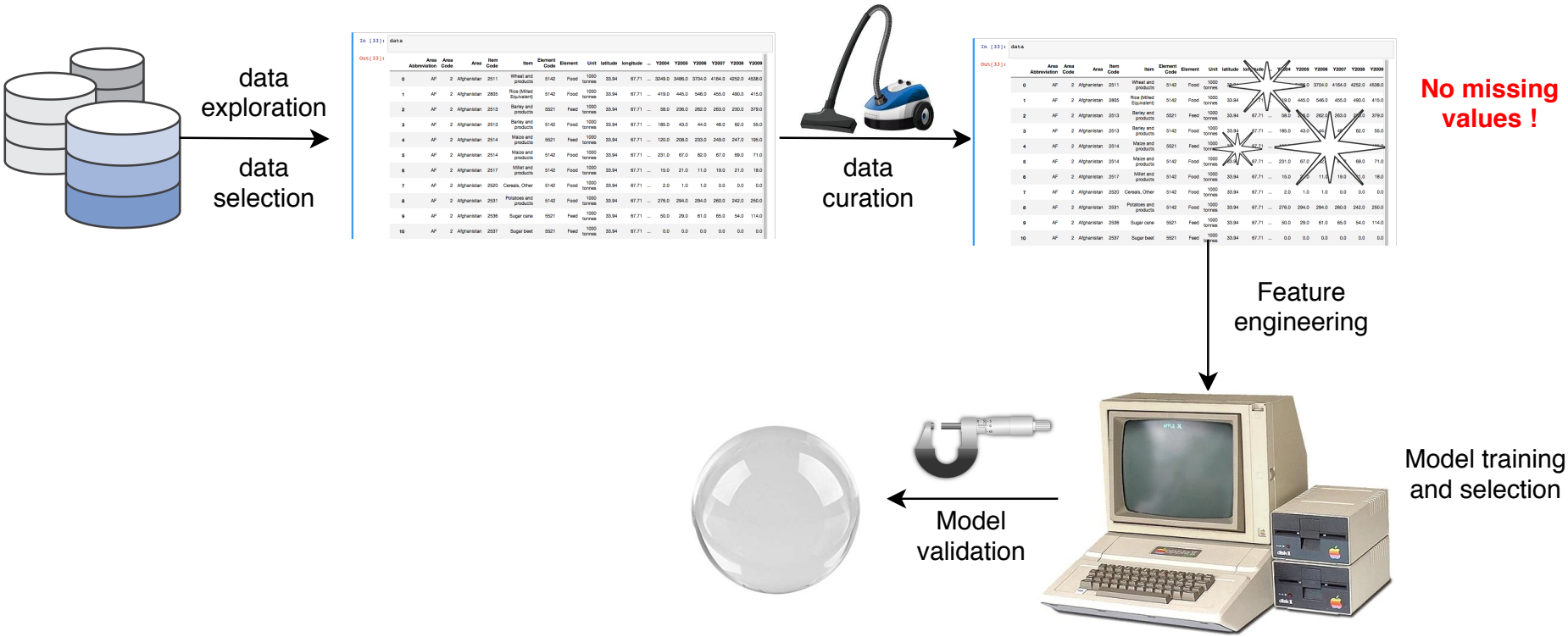


Figure: Spectral clustering

Data mining


data exploration
data selection
data curation
No missing values !
Feature engineering
Model training and selection
Model validation

Classification VS Regression

Supervised Learning

Goal: finding the hidden link between input and output

X → Y

Classification
Output = label
→ Wheat

Regression
Output = continuous quantity
→ 7.2 t/ha

Regression task examples

Data	Prediction
soil samples from top soil layer	organic matiere content
forces acting on a chisel and speed	soil moisture
zoometric measurements of the animals before slaughter	weight for beef cattle 150 days 2 to 222 d. before slaughter
vegetation indices, spectral bands of red band NIR	estimation of grassland biomass

Classification task examples

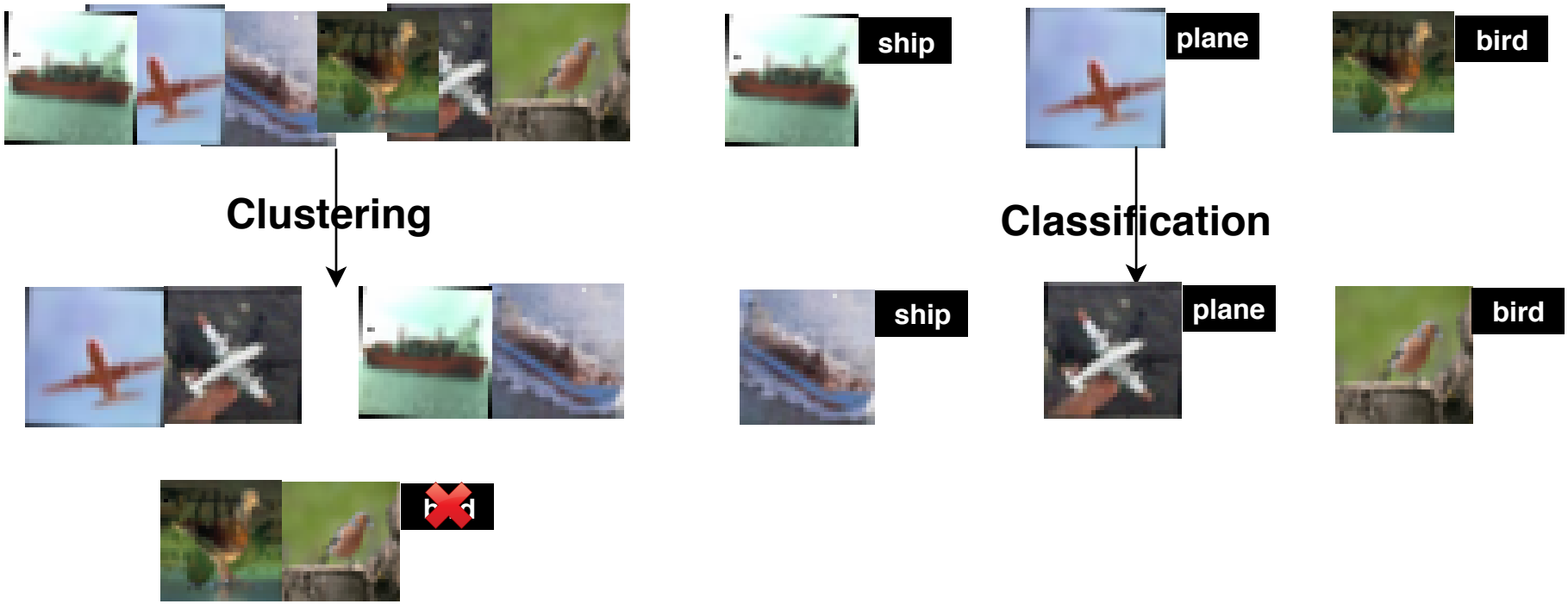
Data	**Prediction**
dairy cow movements and rumination activity	pregnancy
Spectral features from hyperspectral imaging	weed vs *Zea mays*
1553 color pig face images	pig face recognition
20 chemical components from rice samples	geographical origin of the sample

Algorithm examples

Classification	**Regression**
(Regularized) Logistic Regression,	(Regularized) Linear Regression,
Linear Discriminant Analysis,	Support Vector Regressor,
Support Vector Machine,	Regression Trees,
Classification Trees,	Boosted Regression Trees,
Boosted Classification Trees,	Random Forest,
Random Forest,	(Deep) Neural Networks
(Deep) Neural Networks	

Classification ≠ Clustering

Clustering VS Classification



Unlabeled data for training

Finding an internal structure

Unsupervised

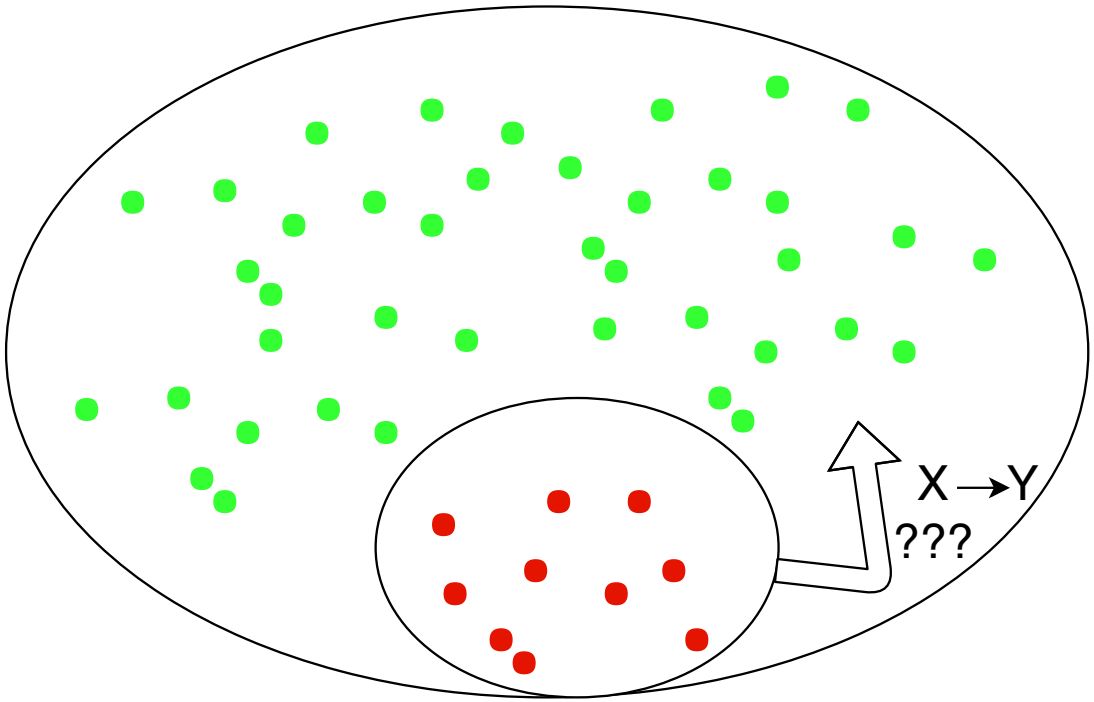
Labelled data for training

What is the label for an unseen object ?

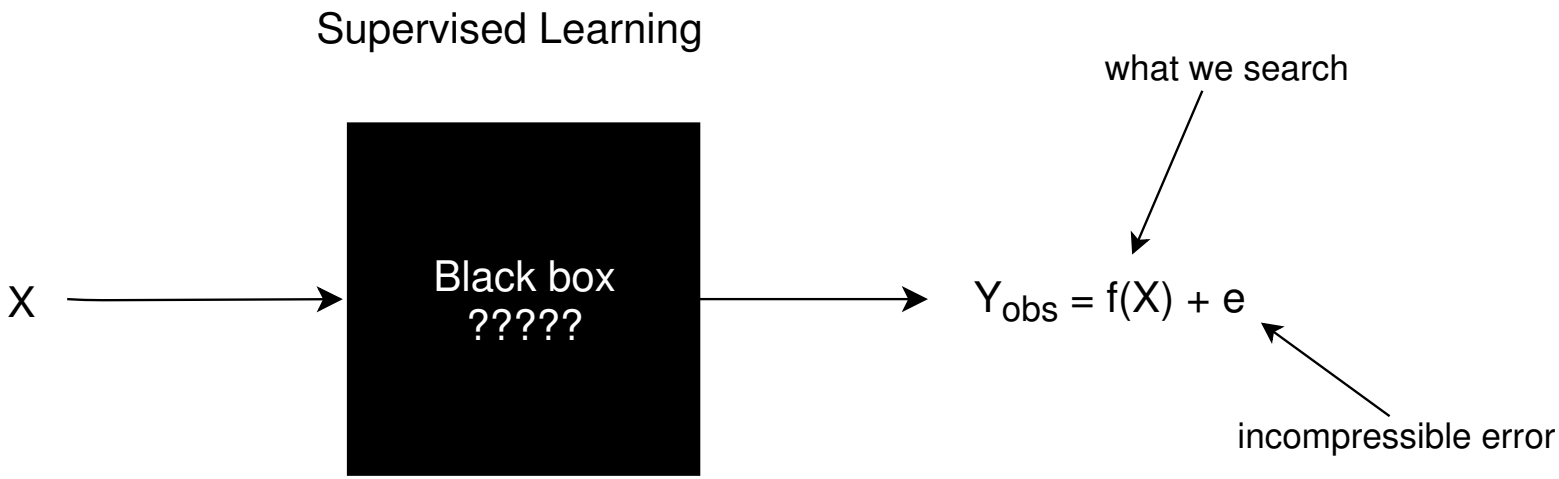
Supervised

Problematic

Limited sample to learn general patterns for population



The problem of supervised learning



A good model has an error close to e **on unseen points**

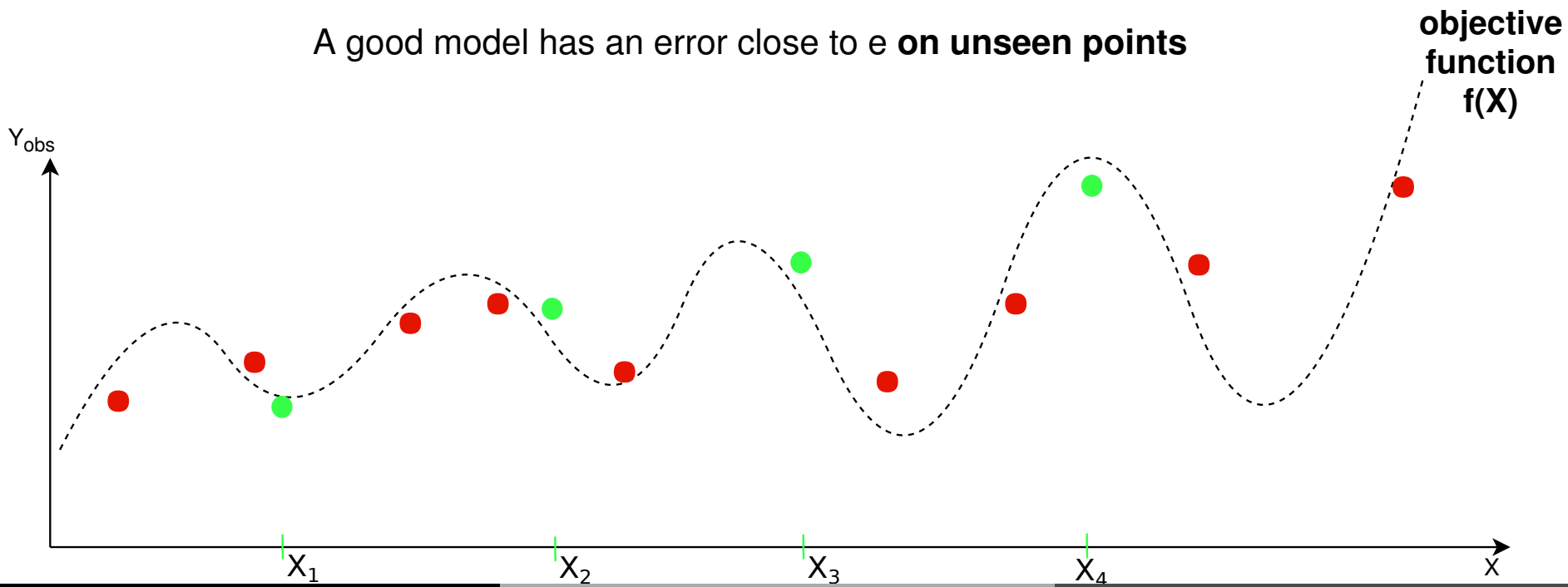


Illustration of overfitting

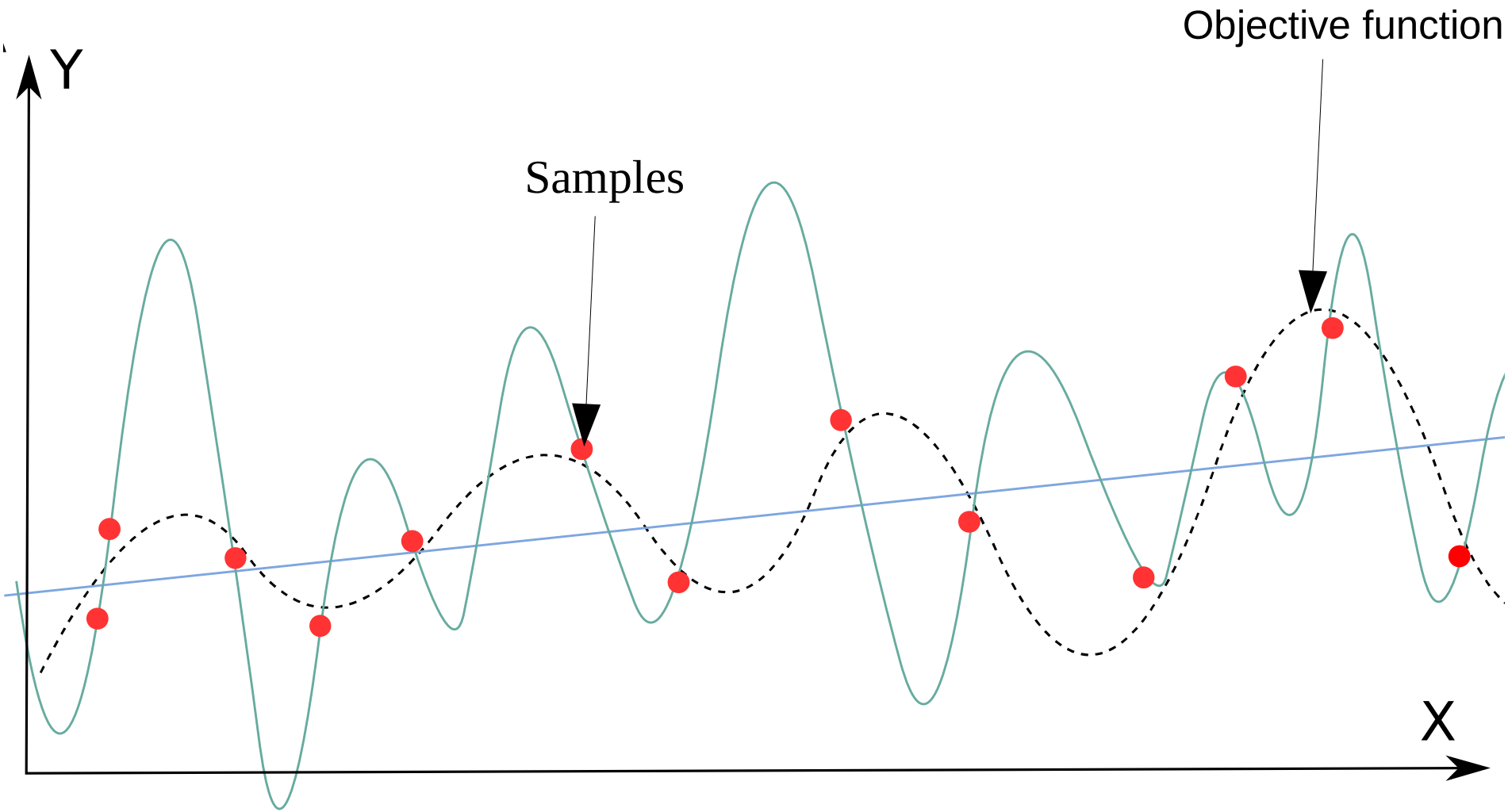


Illustration of overfitting

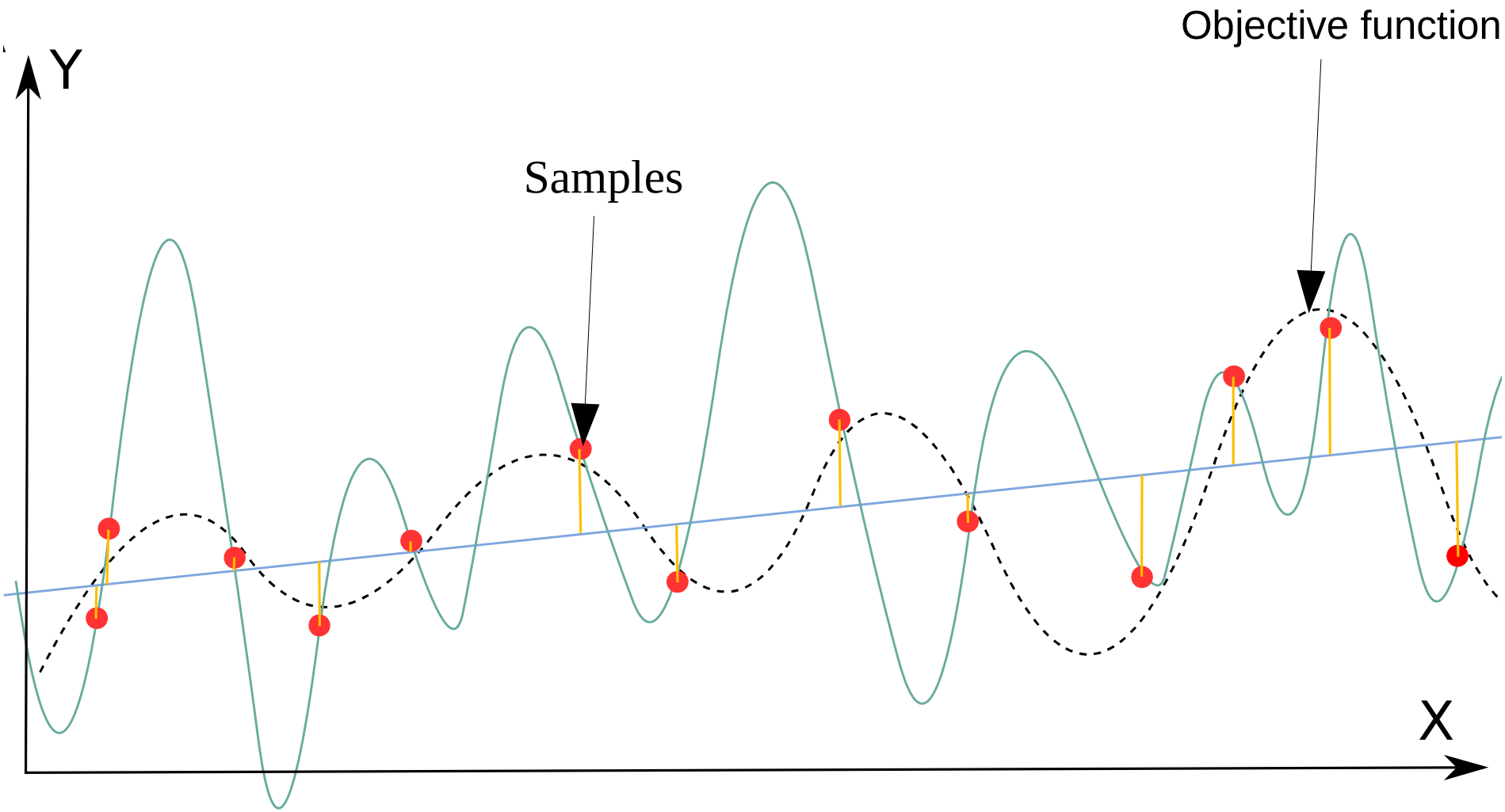
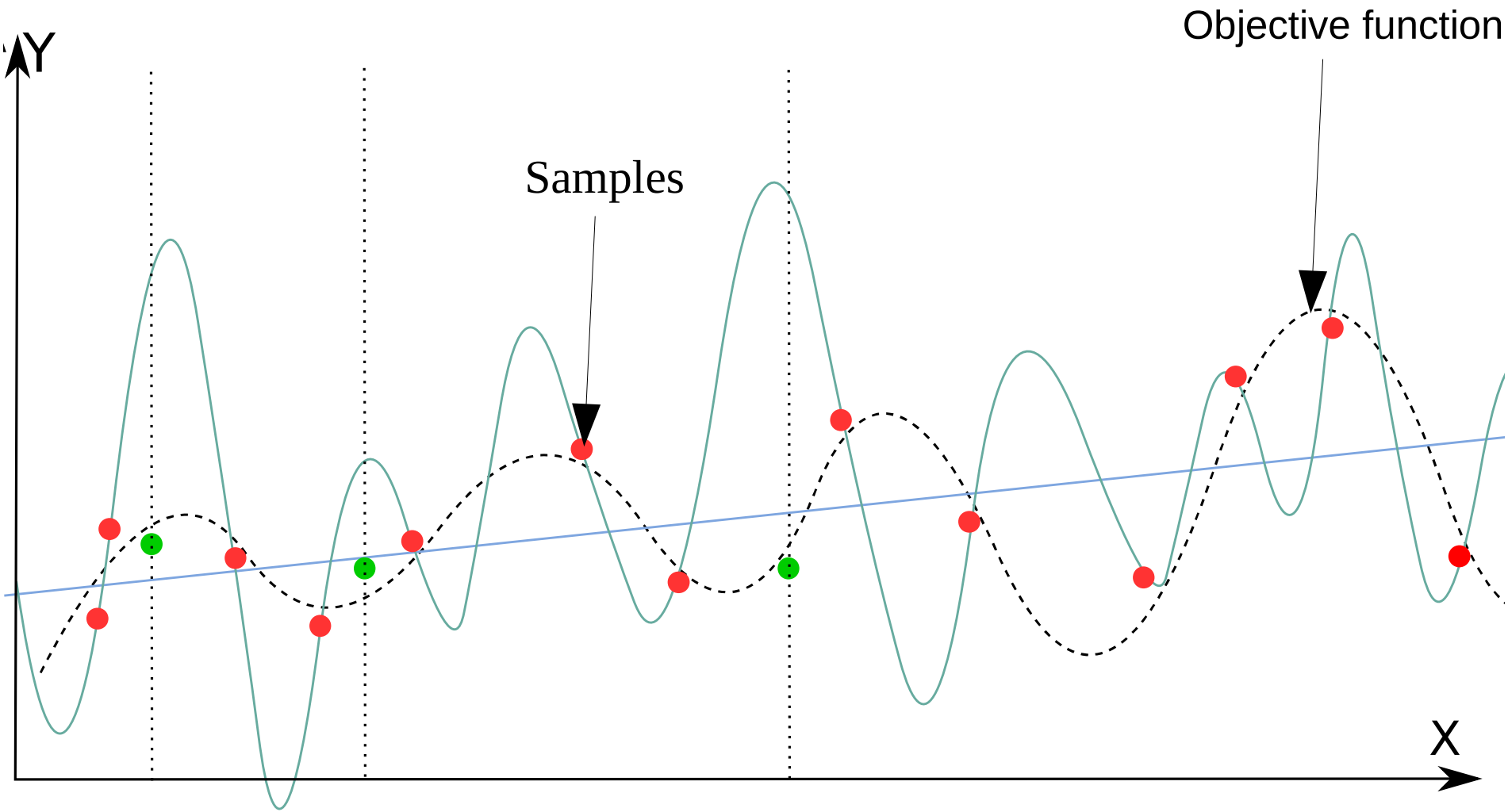
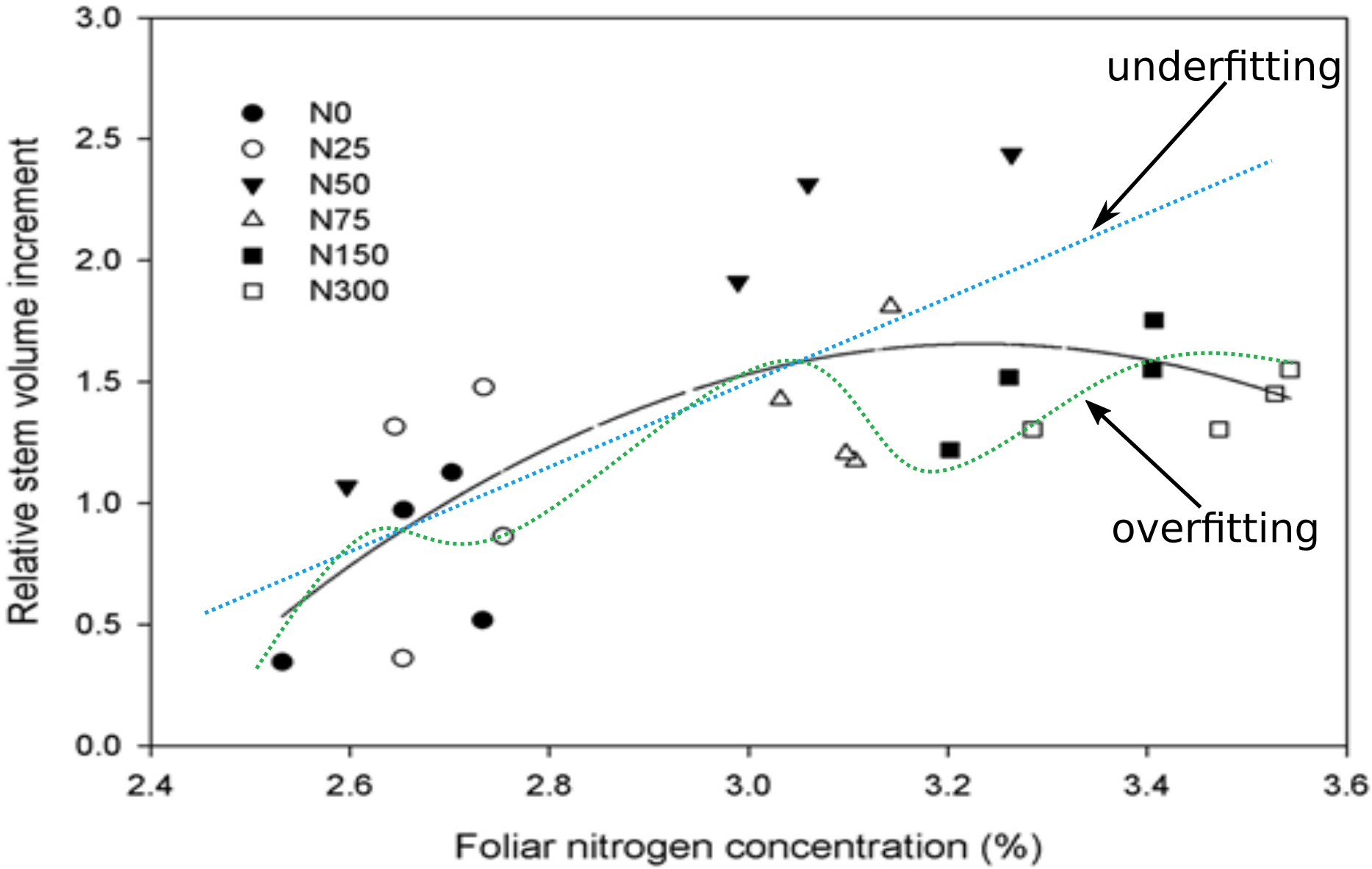


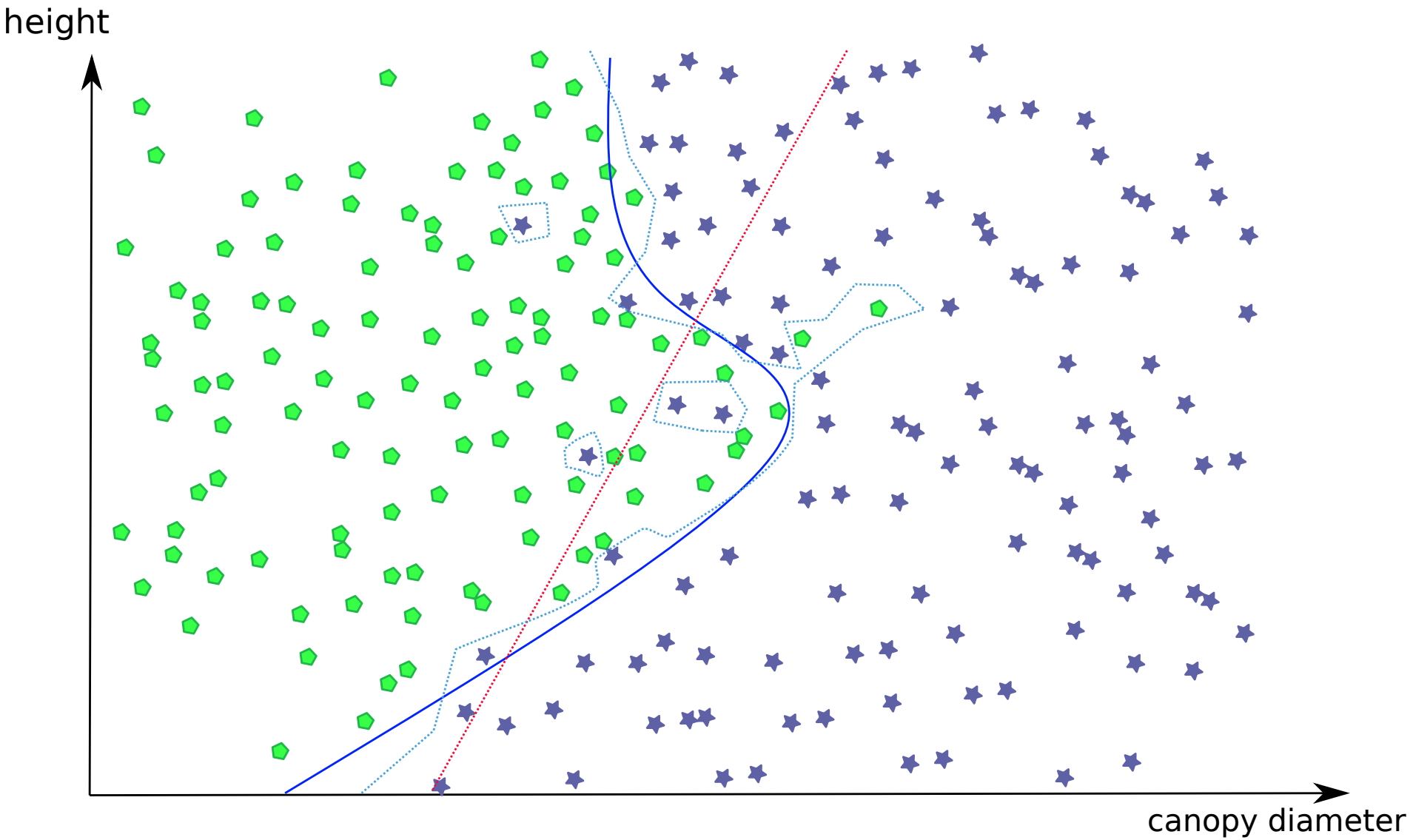
Illustration of overfitting



Underfitting-Overfitting regression example



Underfitting-Overfitting classifcation example



Bias Variance Tradeoff

How to correctly generalize what we learnt from samples?

$\hookrightarrow$ We need some methodology to avoid **overfitting** while having a sufficient model complexity

Does the best algorithm exist?

No-free-lunch-theroem *Wolpert, D.H., Macready, W.G. (1997)*

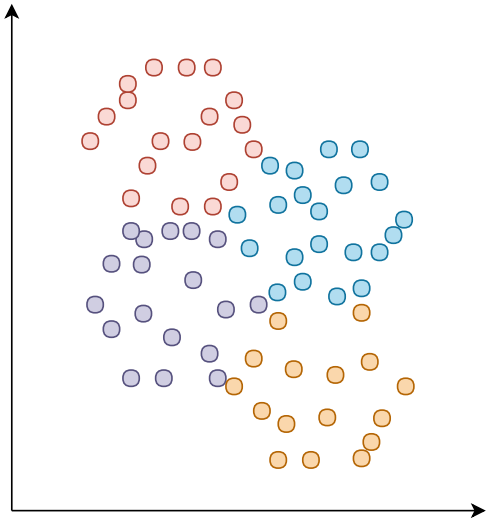
There is no best algorithm for general problem solving. Each algorithm has its own inductive bias. A particular bias is necessary to best solve a problem.

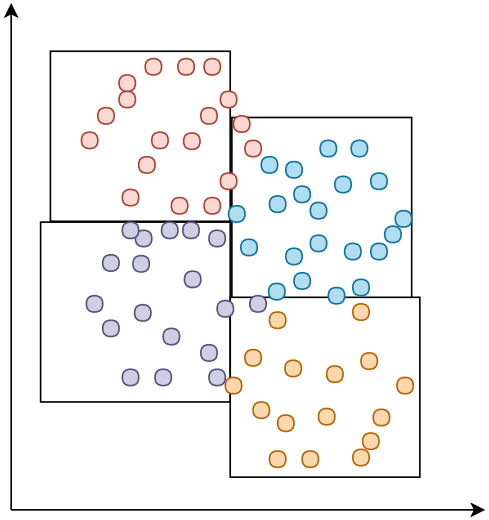
Inductive bias (1)

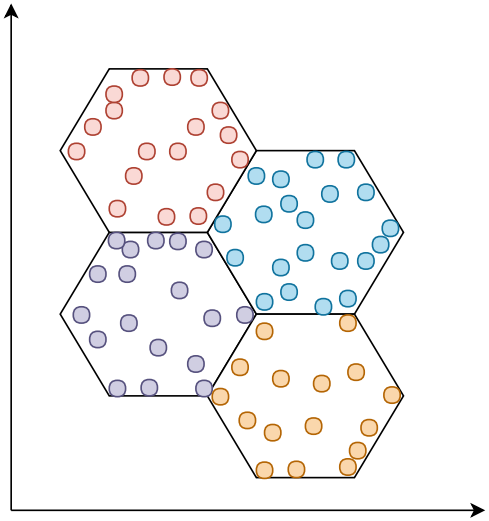
An inductive bias is a constraint limiting the space of solutions. It prioritizes one solution over others with assumptions. As a consequence it is easier to search a solution in that limited space.

Rote learner = no inductive bias

Inductive bias (2)







Deep Learning inductive bias

Is Deep Learning best for everything? No

Convolutional Neural Newtorks: spatial translation invariance inductive bias

Recurrent Neural Networks: temporal translation invariance inductive bias

$\hookrightarrow$ Best for image classification/regression, some time series problems, natural language processing

What did we learn?

→ Numeric ML = Unsupervised + Supervised Learning

→ Supervised Learning = Regression + Classification

→ Fundamental pb of Supervised Learning = How to generalize rules from samples to population? We have to avoid **overfitting**

→ There is no best algorithm for solving all problems

Q&A

Q&A