UNIT 1

Overview of Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

BIOINF

HOW TO SOLVE THESE TASKS?

- Finding solutions of a system of equations
- Prediction of trajectory of a space shuttle
- Diagnosis whether a patient has a certain disease
- Prediction of outcome of election
- Recognition of handwritten characters
- Prediction of function of protein from its amino acid sequence

EXPLICIT MODELS

- Traditional disciplines like physics, chemistry, and biology are usually aiming at *exact explicit models*, i.e. to know how (and why) things work in a particular way; then a solution to a new problem can be found *deductively* using explicit knowledge
- That goal, however, is sometimes too difficult to achieve; reasons may be computational complexity, insufficient knowledge, insufficient information, etc.

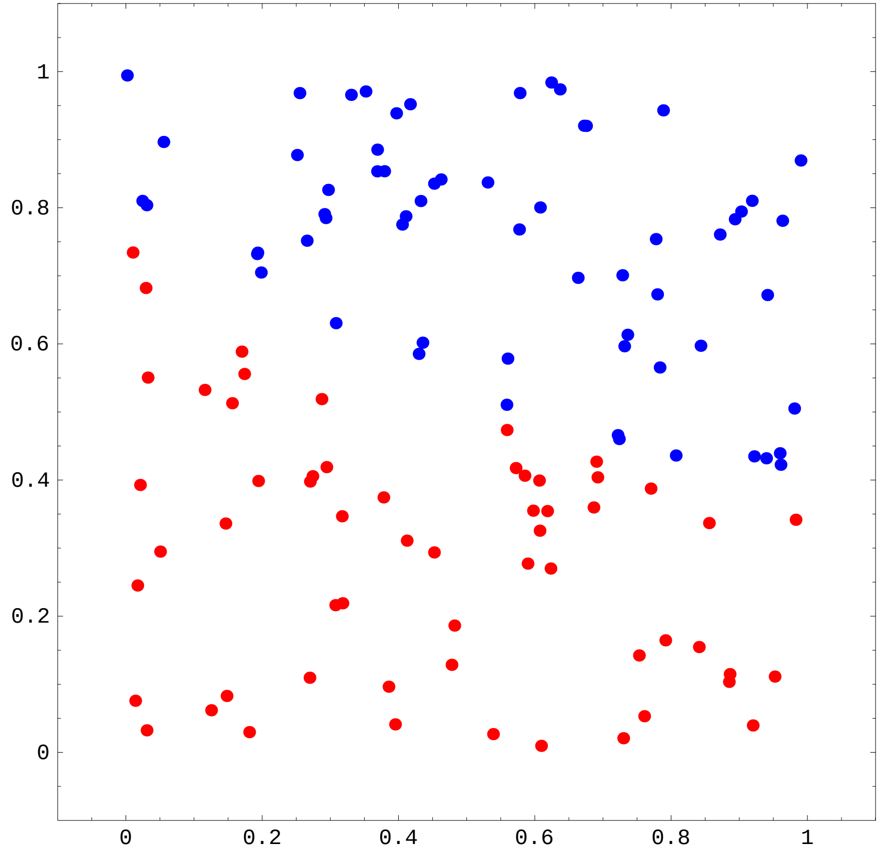
MACHINE LEARNING $=$ INDUCTIVE LEARNING

- Machine learning tries to elicit models/knowledge from *previously observed data* with the following two main goals:
  1. Getting insight
  2. Being able to predict future outcomes
- Putting it simple, machine learning is about *learning from data* (often called *inductive learning*).

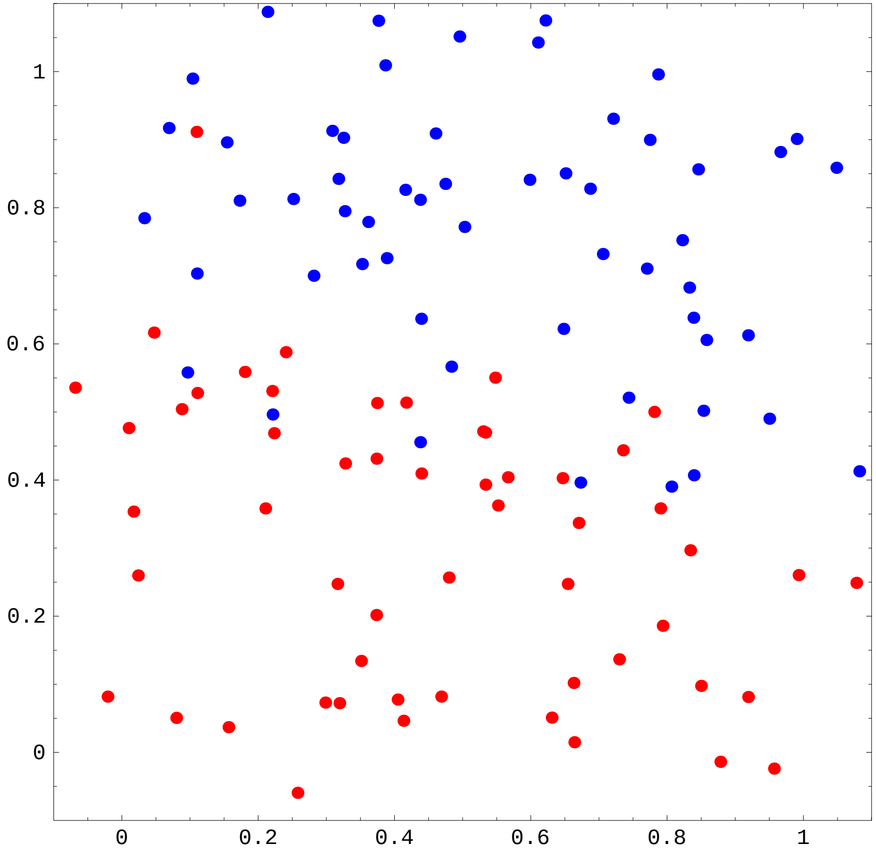
WHAT DO WE SEE HERE?

0.843475	0.709216	−1
0.408987	0.47037	+1
0.734759	0.645298	−1
0.972187	0.0802574	+1
0.90267	0.327633	−1
0.807075	0.872155	−1
0.240068	0.801159	−1
0.206602	0.562109	+1
0.581611	0.335561	+1
0.944329	0.026344	+1
0.569412	0.30145	+1
0.552694	0.864825	−1
0.700995	0.517267	−1
0.209818	0.342484	+1
0.94141	0.928017	−1
0.148546	0.198177	+1
0.872544	0.50608	−1
0.371062	0.272064	+1
...	...	...

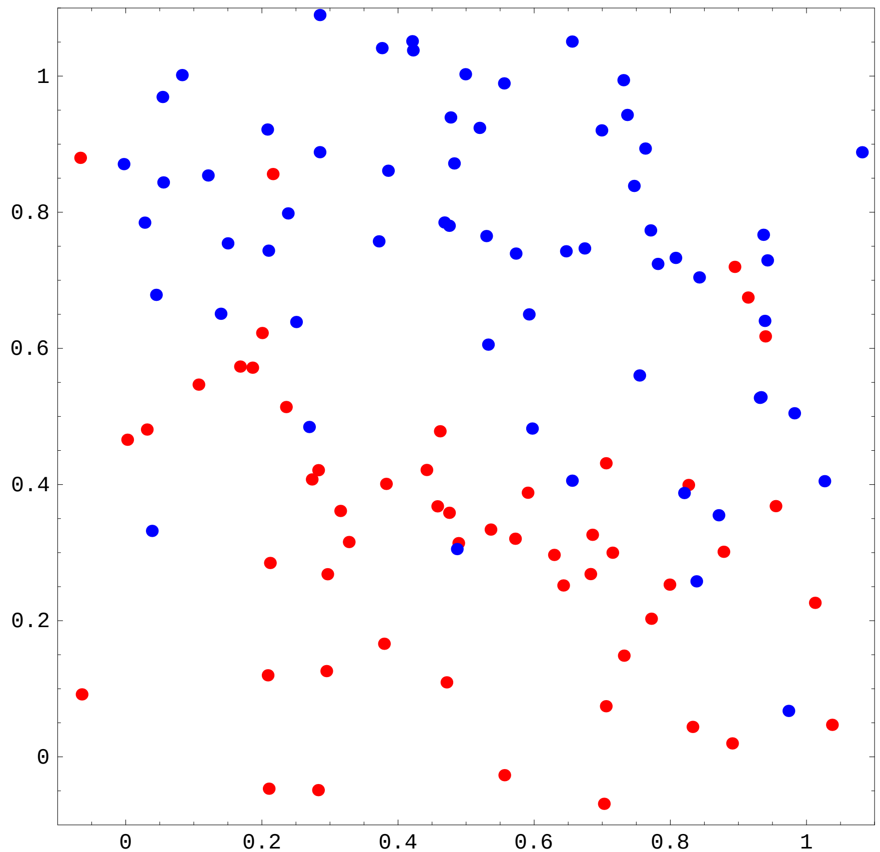
AND HERE?



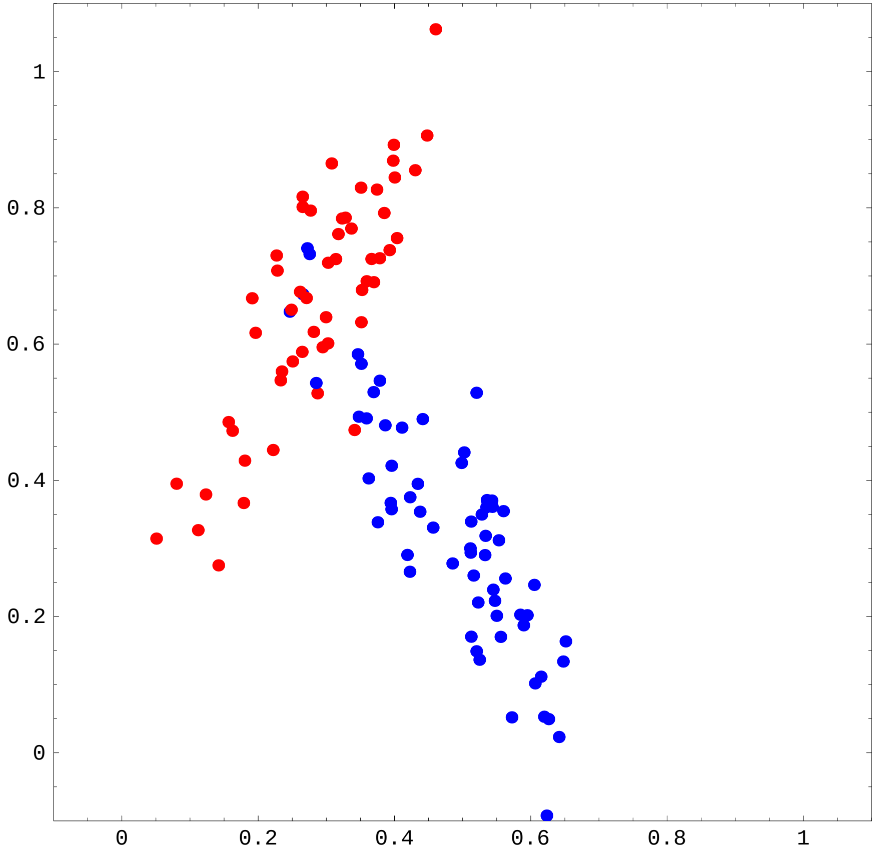
AND HERE?



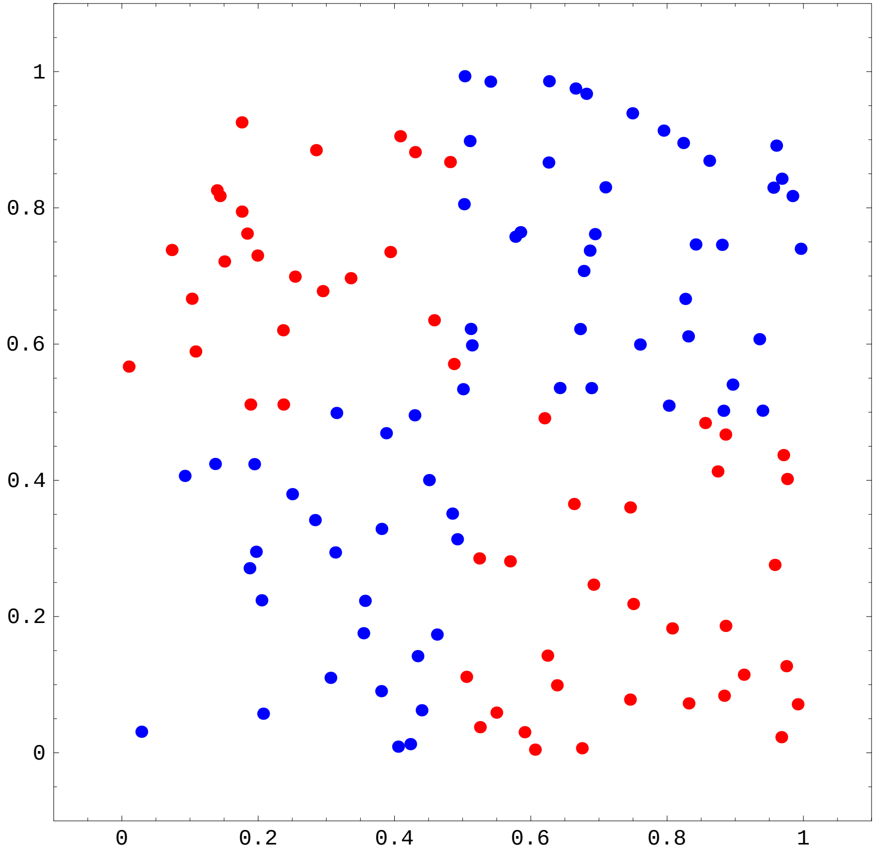
AND HERE?



AND HERE?



AND HERE?



AND HERE?

0.99516	0.890813	0.933726	0.793397	0.826405	0.236946	−1
0.853206	0.611647	0.317486	0.633609	0.411492	0.985231	+1
0.387494	0.459847	0.815049	0.394526	0.678227	0.031886	−1
0.733515	0.640438	1.19068	0.639685	0.0793674	0.160503	+1
0.274817	0.261054	1.20056	0.689895	0.401913	0.277955	−1
0.329943	0.241299	0.848705	0.721673	0.973852	0.795238	−1
0.334784	0.350487	0.315131	0.928277	0.816343	0.558292	−1
0.481578	0.738839	0.0925513	0.294667	0.612725	0.573062	−1
0.0940846	0.278992	0.451819	0.900141	0.220497	0.541176	+1
0.360569	0.638554	1.0307	0.260456	0.00658296	0.380672	+1
0.0857518	0.3775	0.386551	0.570562	0.15437	0.102717	+1
0.755808	0.1362	0.544536	0.848888	0.874862	0.307479	−1
0.421025	0.785714	0.449038	0.920612	0.420418	0.749187	−1
0.939446	0.0468747	0.15846	0.625944	0.198894	0.176125	+1
0.845362	0.767883	0.824993	0.725803	0.808218	0.63495	−1
0.484793	0.129329	0.0783719	0.465347	0.291457	0.254278	+1
0.399041	0.751829	0.763511	0.894785	0.47902	0.15156	−1
0.643232	0.615629	0.430261	0.0458972	0.446513	0.844081	+1
...	...	...	...	...	...	...

SUPERVISED VS. UNSUPERVISED MACHINE LEARNING

Supervised ML: an explicit target (output) value is given for each (input) data item; the goal is to identify the relationship between input and output

Unsupervised ML: no target value is given, the goal is to identify structure in the data

SUPERVISED MACHINE LEARNING

Classification: the output value is a class label
Regression: the output value is numerical

Supervised ML is sometimes called *predictive modeling*. This is due to the fact that the goal is most often to predict the output value for future input values.

UNSUPERVISED MACHINE LEARNING

Projection methods: down-projection of data to lower-dimensional space in order to concentrate on the essence of the data

Clustering: grouping of similar data objects

Density estimation: estimate the probability distribution of the data

Generative model: building a model that produces data that are distributed as the observed data

MISCELLANEOUS TOPICS

Reinforcement learning: learning by feedback from the environment in an online process

Feature extraction: computation of features from data prior to machine learning (e.g. signal and image processing)

Feature selection: selection of those features that are relevant/sufficient to solve a given learning task

Feature construction: construction of new features as part of the learning process

TERMINOLOGY

Model: the specific relationship/representation we are aiming at

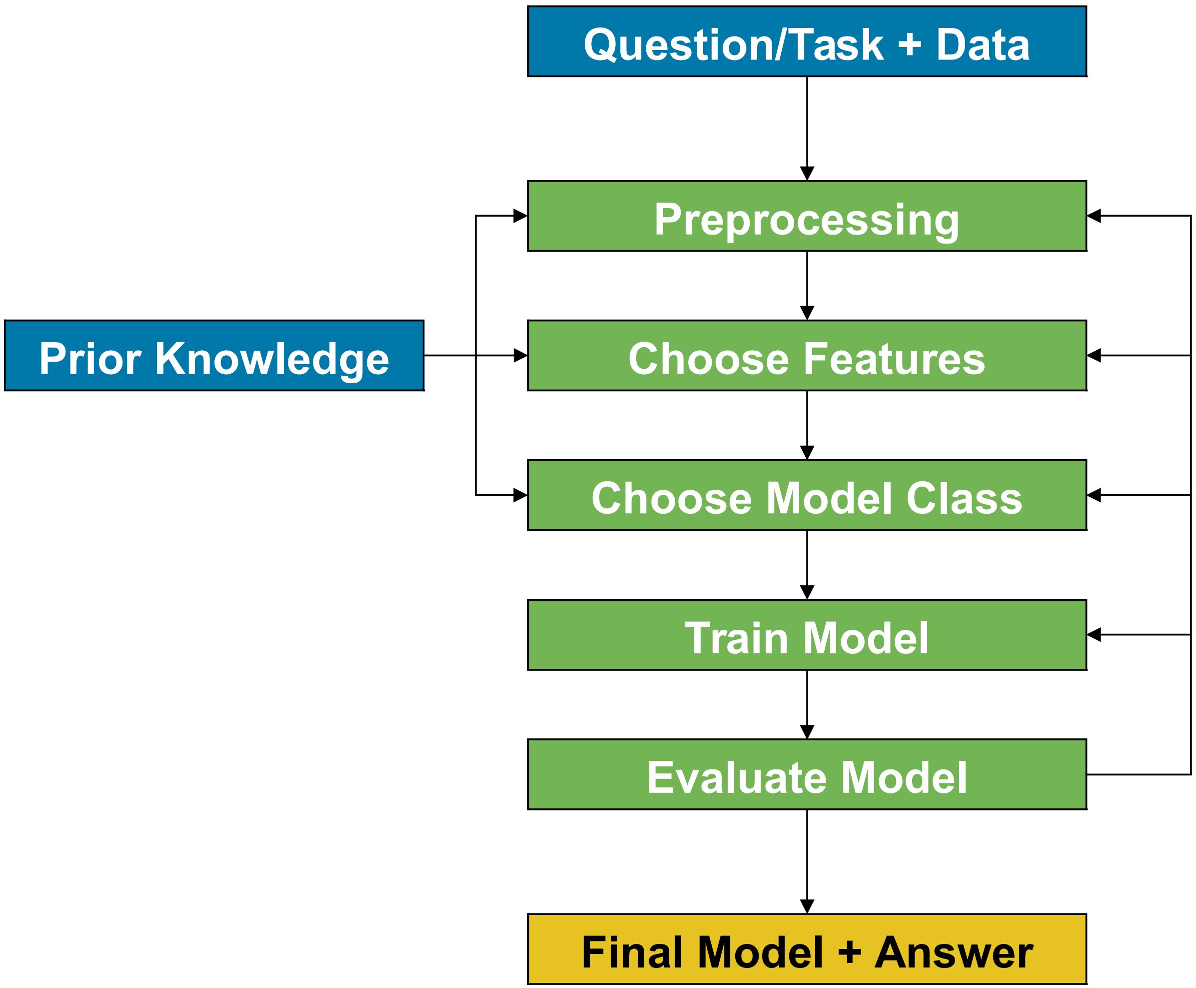
Model class: the class of models in which we search for the model

Parameters: representations of concrete models inside the given model class

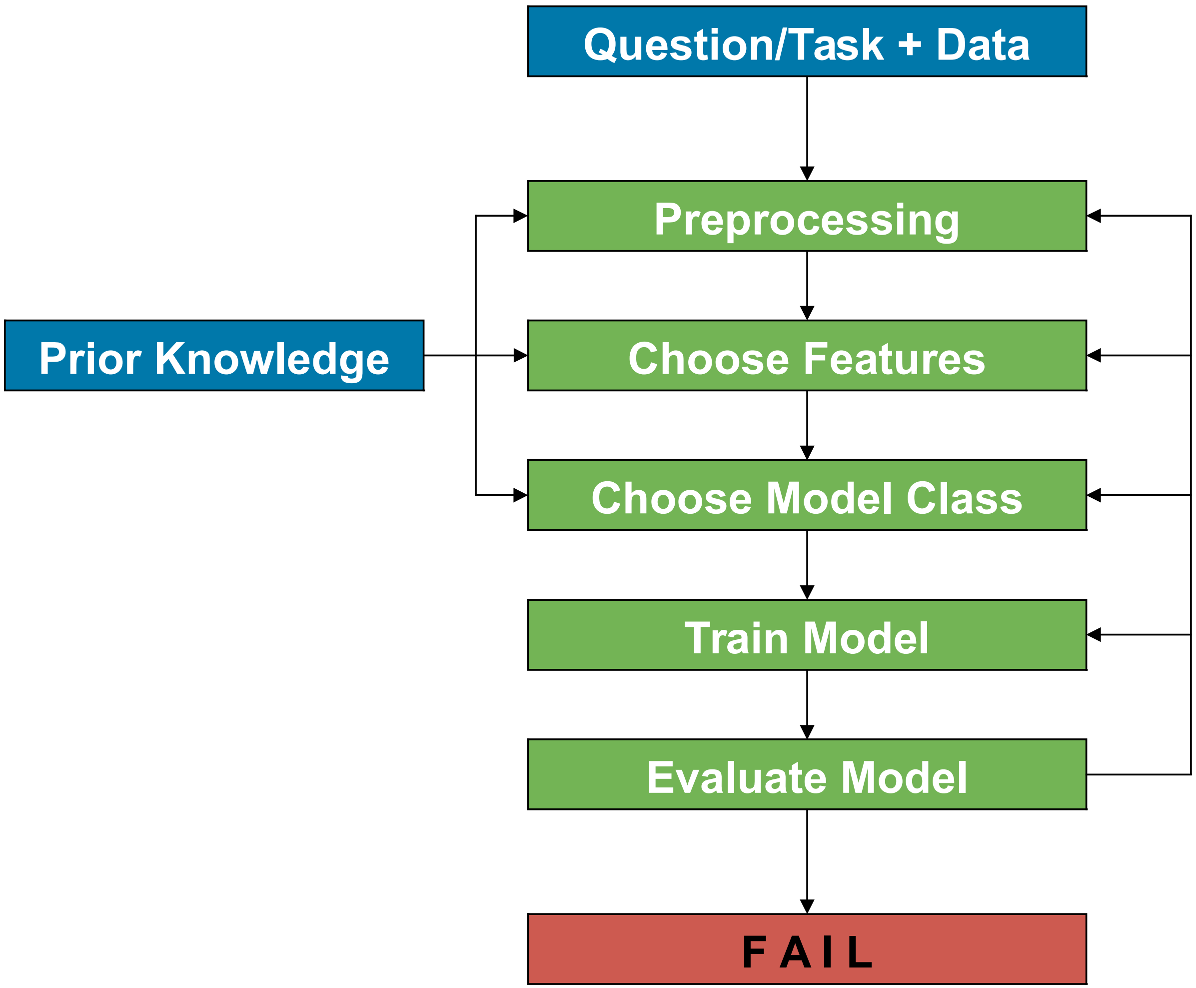
Model selection/training: process of finding that model from the model class that fits/explains the observed data in the best way

Hyperparameters: parameters controlling the model complexity or the training procedure

BASIC DATA ANALYSIS WORKFLOW



BASIC DATA ANALYSIS WORKFLOW



PARAMETRIC VS. NON-PARAMETRIC MODELS

Parametric Models: the models are parameterized with parameters outside or exceeding the data space

Non-Parametric Models: there is no specific underlying parameter model; data points/representatives themselves are the parameters fully describing the model

SOME WORDS OF ENTHUSIASM

- Machine learning methods are able to solve some tasks for which explicit models will never exist
- Machine learning methods have become standard tools in a variety of disciplines (e.g. signal and image processing, bioinformatics)

BUT . . . SOME WORDS OF CAUTION

- Machine learning is not a universal remedy
- Quality of models is depending on quality and quantity of data
- What cannot be measured/observed can never be identified by machine learning
- Machine learning complements explicit/deductive models instead of replacing them
- Machine learning is often applied in a naive way

SUPERVISED MACHINE LEARNING IN BIOINFORMATICS

- Protein secondary structure prediction
- Protein structure and function classification
- Gene recognition
- Splice site and alternative splice site recognition
- Gene selection and prediction of outcomes from gene expression data
- Prediction of nucleosome positions

GOALS OF THIS COURSE

- To understand the underlying principles of (supervised) machine learning
- To understand what can go wrong in supervised machine learning
- To be able to evaluate the quality of a model created by supervised machine learning
- To gain deeper insight to the fields of support vector machines, random forests, and neural networks

INTRODUCTORY EXAMPLE: FISH RECOGNITION

- Example borrowed from

  R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2nd edition. John Wiley & Sons, 2001. ISBN 0-471-05669-3.
- Automated system to sort fish in a fish-packing company: salmons must be distinguished from sea bass optically

- **Given:** a set of pictures with known fish, the training set
- **Goal:** automatically distinguish between salmons and sea bass for future pictures

TWO SAMPLE IMAGES

Salmon:

Sea bass:

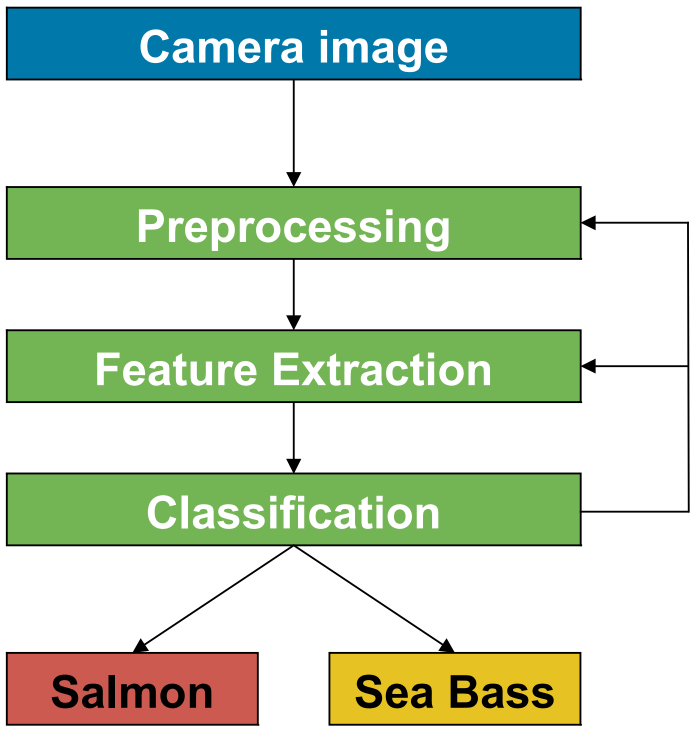
TWO SAMPLE IMAGES

Salmon:

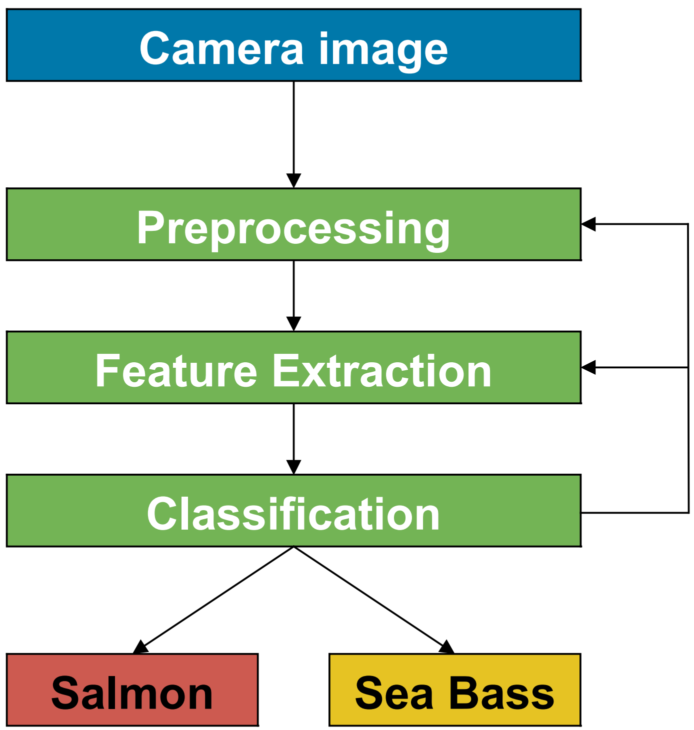
Sea bass:

How can we distinguish these two kinds of fish visually?

BASIC WORKFLOW



BASIC WORKFLOW

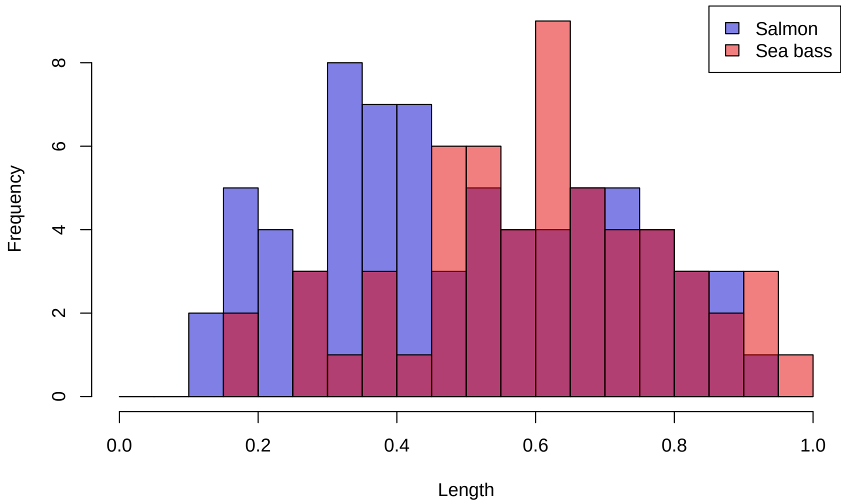


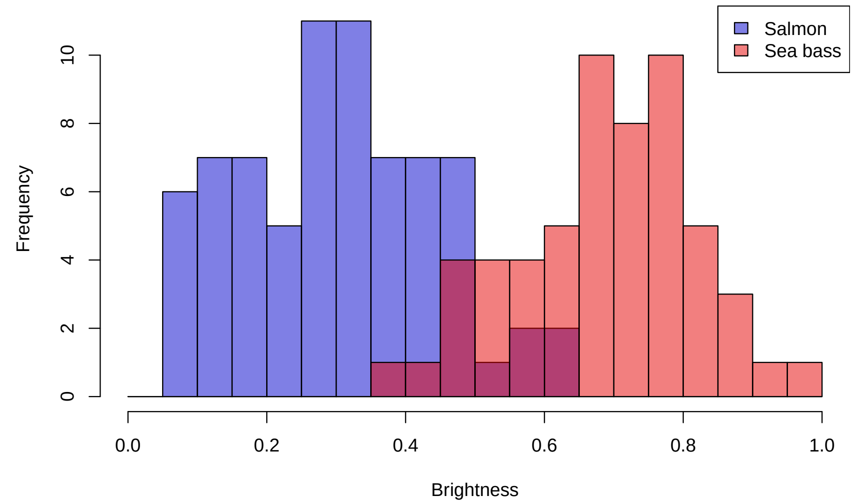
Preprocessing: contrast and brightness correction, segmentation, alignment

Features:

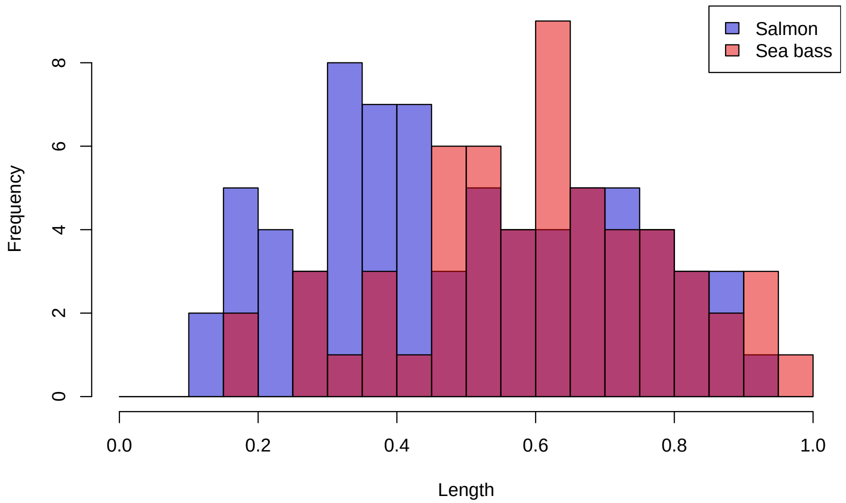
1. Length
2. Brightness

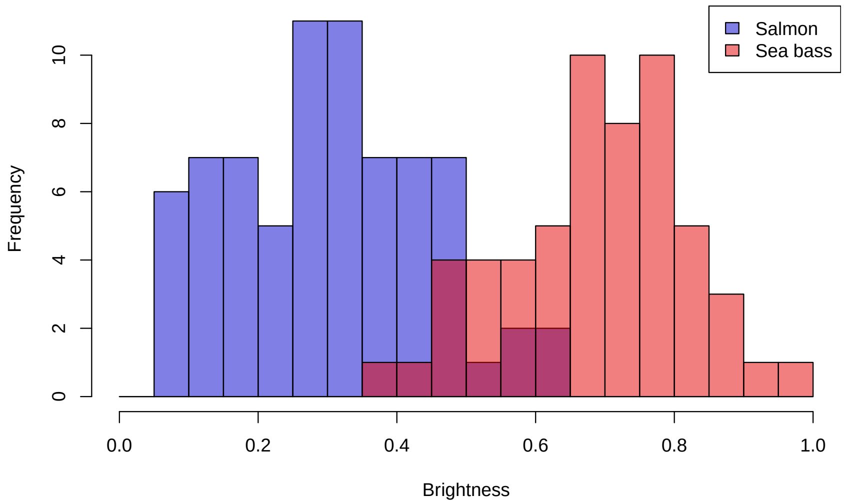
USING ONE FEATURE





USING ONE FEATURE

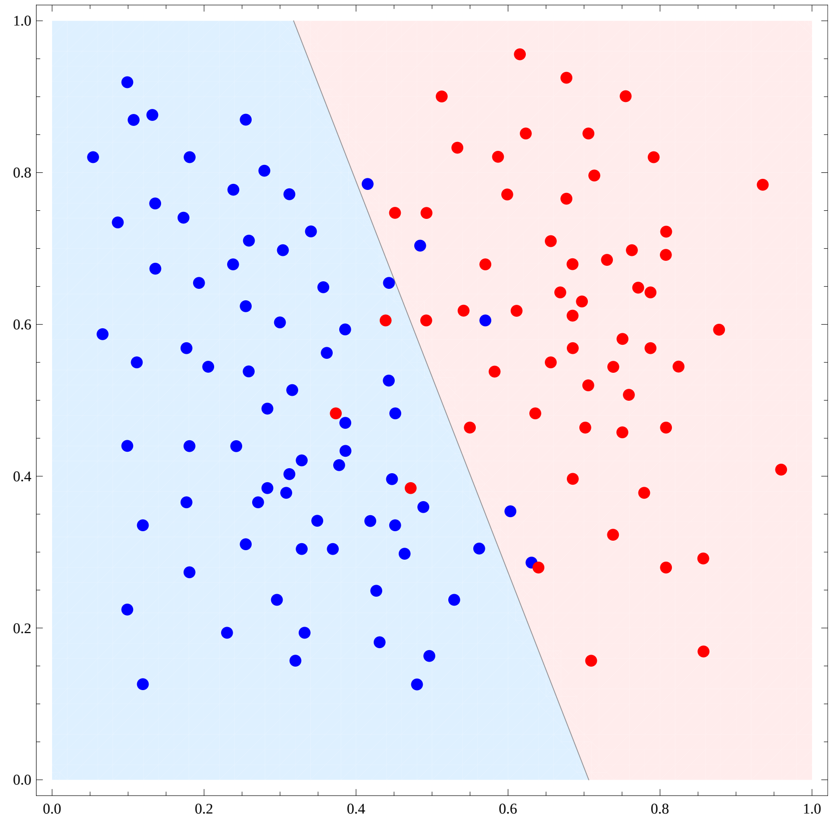




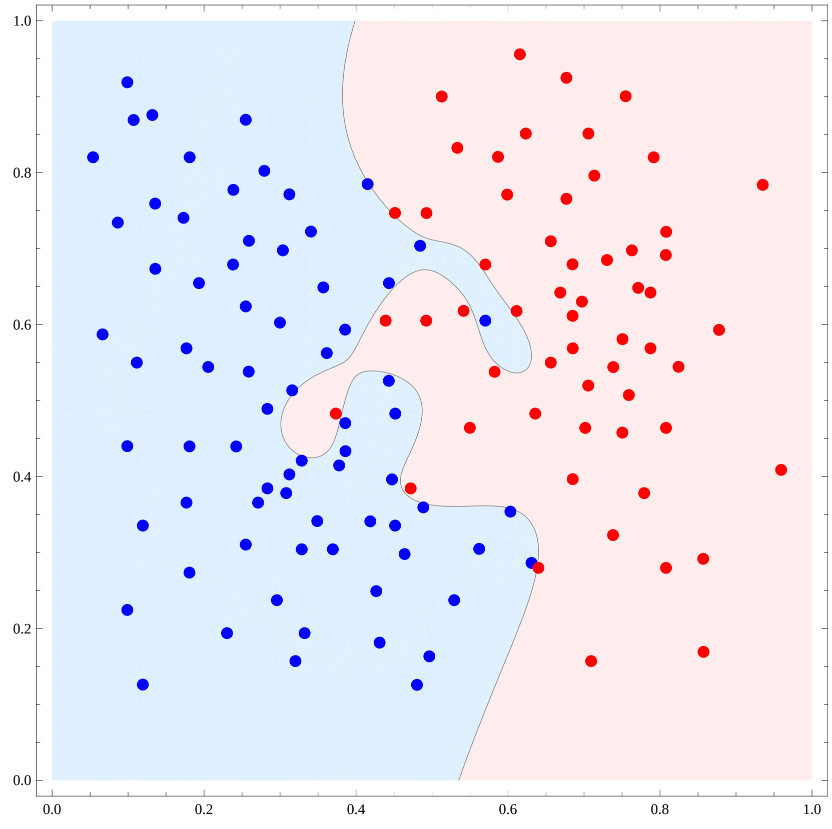
Questions:

1. Which is the better feature?
2. Where should we put the threshold?

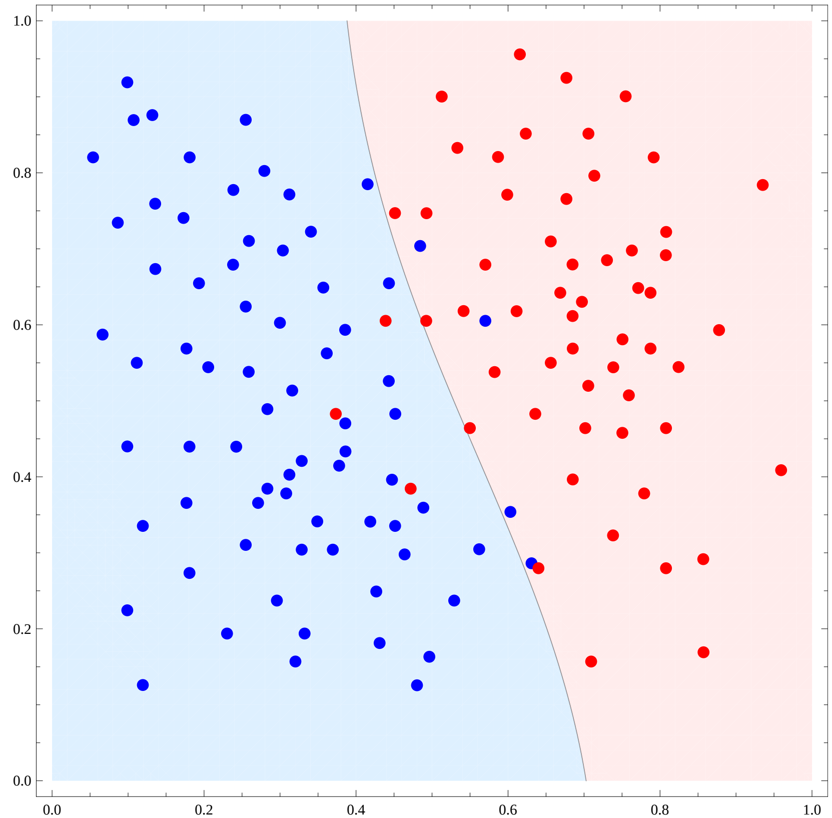
USING TWO FEATURES: LINEAR SEPARATION



USING TWO FEATURES: HIGHLY NONLINEAR SEPARATION



USING TWO FEATURES: MILDLY NONLINEAR SEPARATION



QUESTIONS

- Which is the best result and why?
- What is the best way to measure the quality of a classifier?
- Which methods for constructing classifiers are available?
- Is there a theoretical basis (instead of a purely intuitive one) to answer these questions?

QUESTIONS

- Which is the best result and why?
- What is the best way to measure the quality of a classifier?
- Which methods for constructing classifiers are available?
- Is there a theoretical basis (instead of a purely intuitive one) to answer these questions?

These questions will be the point of departure of this course.