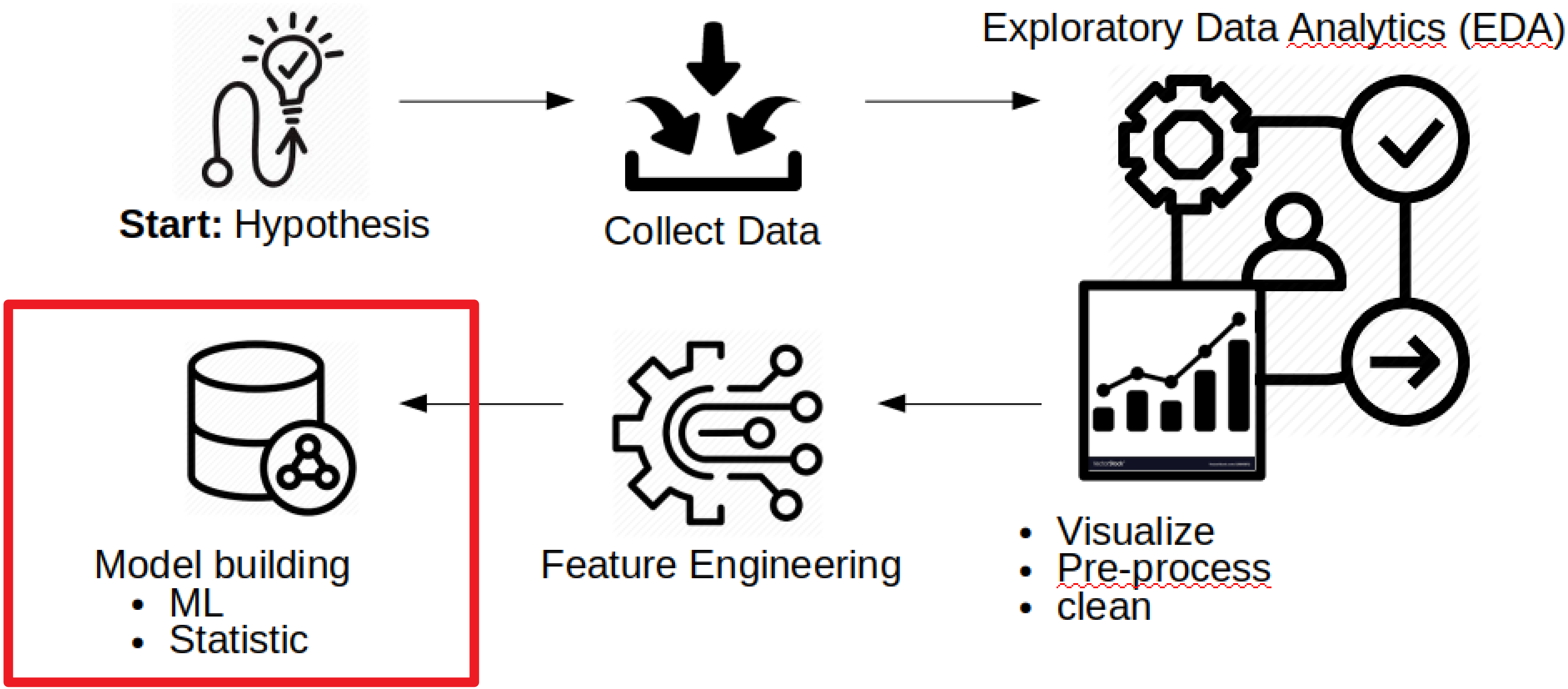
Hochschule Offenburg
offenburg.university

Machine Learning I

Introduction and Overview

Machine Learning I



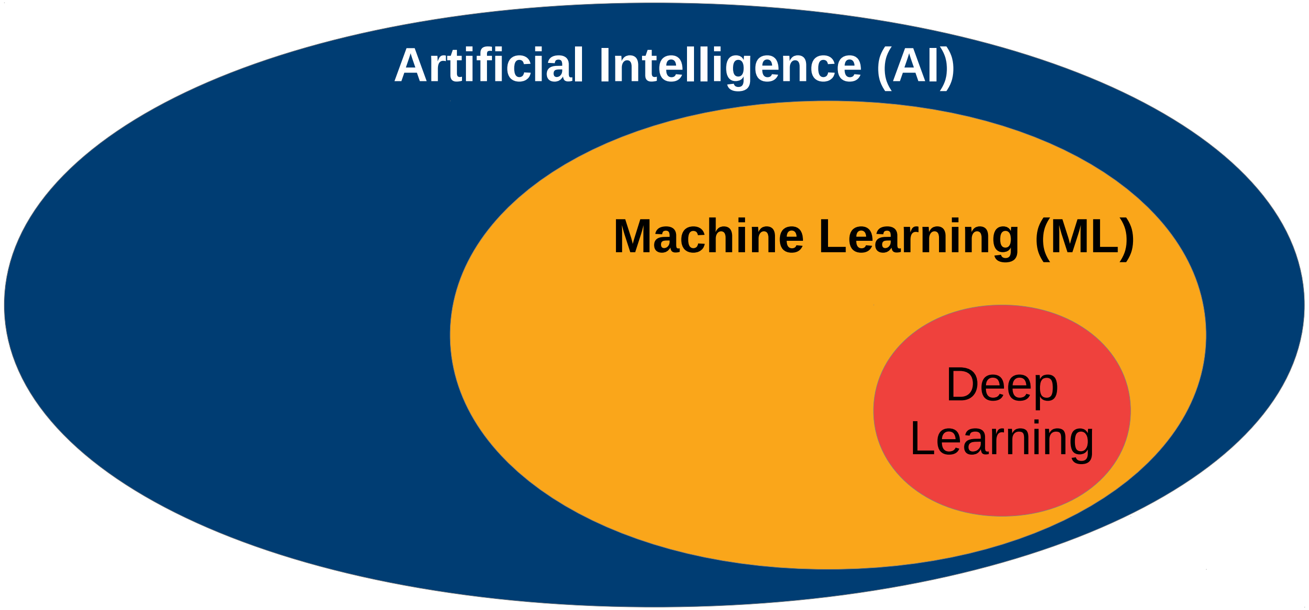
Hochschule Offenburg
offenburg.university

Outline

- Introduction to ML
  - Basic Definitions an Terminology
    - Supervised Learning
    - Generalization and Overfitting
    - Unsupervised Learning

Introduction to ML

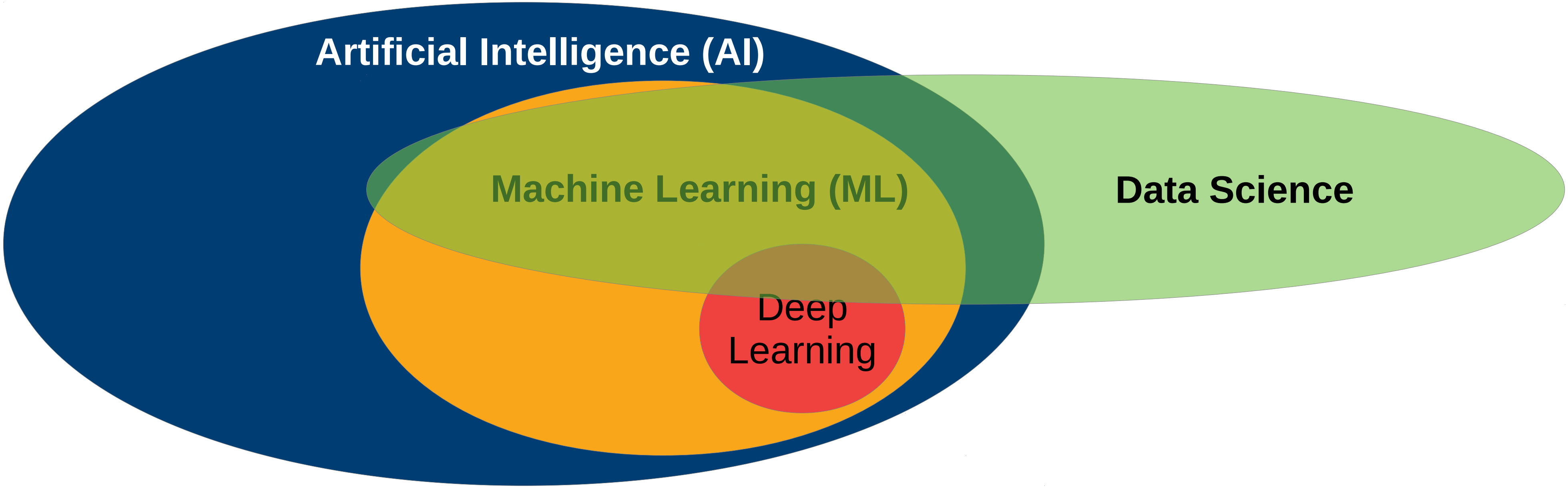
Research and Application Fields



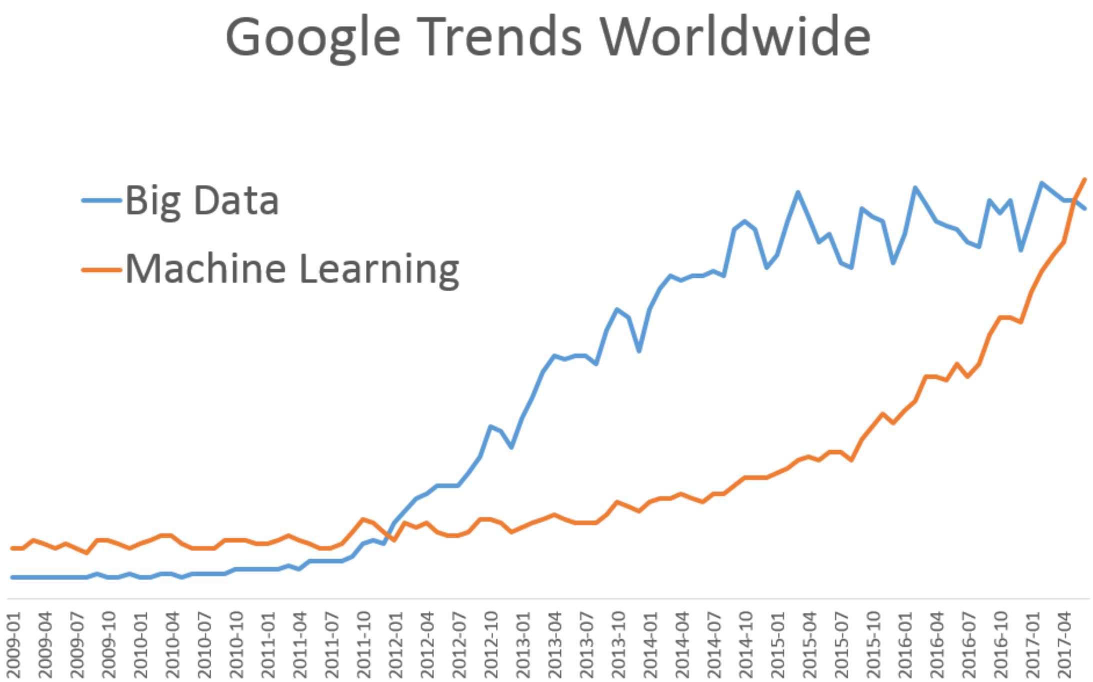
Introduction to ML

Hochschule Offenburg
offenburg.university

Research and Application Fields



The ML Hype


Google Trends Worldwide
Big Data
Machine Learning
2009-01
2009-04
2009-07
2009-10
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01
2012-04
2012-07
2012-10
2013-01
2013-04
2013-07
2013-10
2014-01
2014-04
2014-07
2014-10
2015-01
2015-04
2015-07
2015-10
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04

Hochschule Offenburg
offenburg.university

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Introduction to ML

Hochschule Offenburg
offenburg.university

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning

- Labeled data
- Direct and quantitative evaluation

- Learn model from „ground truth“ examples
- Predict unseen examples

Introduction to ML

Supervised Learning

Basic Notation:

Data is given as tuples

$$(X, Y) := \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Where X is the actual **data** (sample) and y the associated **label.**

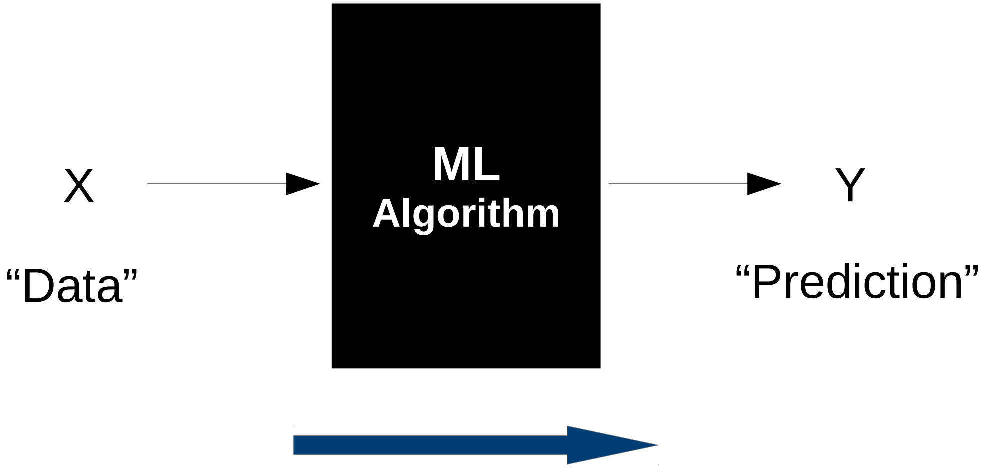
For most ML algorithms **(many Deep Learning algorithms are an exception)**

$$x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$$

The data has to be represented as vectors and the labels are scalars.

Introduction to ML

Supervised Learning as a Black Box



ML algorithms "learns" ***mapping*** from input to output by example tuples

Introduction to ML

Supervised Learning: Example: Classification

ML

"coffee mug"

X: Image

Classification: Y discrete
semantic class label

ML algorithms "learns" ***mapping*** from input to output by example tuples

Introduction to ML

Supervised Learning: Example: Classification

X: Image

Vector

ML

“coffee mug”

Scalar

Classification: Y discrete
semantic class label

ML algorithms “learns” ***mapping*** from input to output by example tuples

Function f(x) → y

Introduction to ML

Supervised Learning: Example: Classification

X: Image

ML

Y: {“coffee mug”, “non coffee mug}

Supervised Learning: Example: Classification

X: Image

pixel by pixel →

$$\begin{pmatrix} 5 \\ 8 \\ 15 \\ 23 \\ 21 \\ 30 \\ 28 \\ \ldots \\ 5 \end{pmatrix}$$

$$X \in \mathbb{R}^n$$

→ **ML** →

Y': {-1,1} $Y' \in \mathbb{R}$

1: “coffee mug”,
-1: “non coffee mug

Introduction to ML

Supervised Learning: Example: Classification

ML Model: function *f* separating mugs from rest

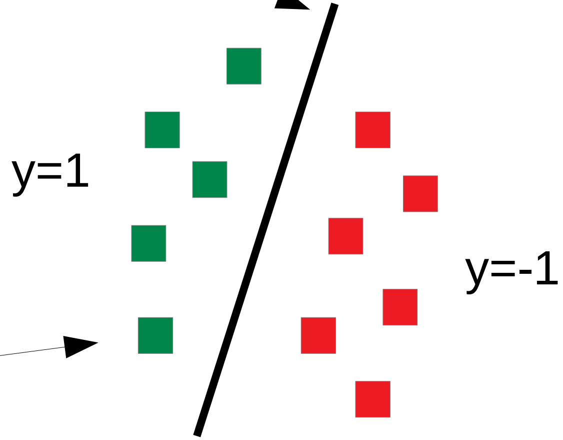
X

$$\begin{pmatrix} 5 \\ 8 \\ 1 \\ 5 \\ 2 \\ 3 \\ 0 \\ 2 \\ 8 \\ \ldots \\ 5 \end{pmatrix}$$

$X \in \mathbb{R}^n$

ML

Y': {-1,1}



$X \in \mathbb{R}^n$

Introduction to ML

Supervised Learning: Example: Classification

LEARNING: approximate „best“ *f* for the given data

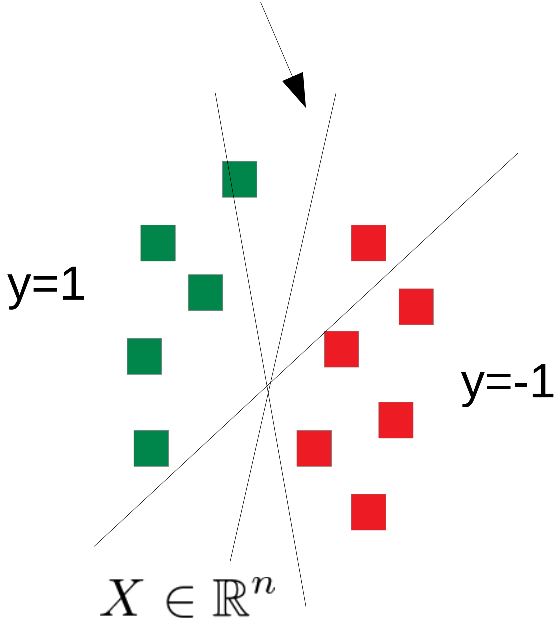
X

$$\begin{pmatrix} 5 \\ 8 \\ 1 \\ 5 \\ 2 \\ 3 \\ 0 \\ 2 \\ 8 \\ \dots \\ 5 \end{pmatrix}$$

$X \in \mathbb{R}^n$

ML

Y’: {-1,1}



Supervised Learning: Example: Classification

LEARNING: optimization problem:

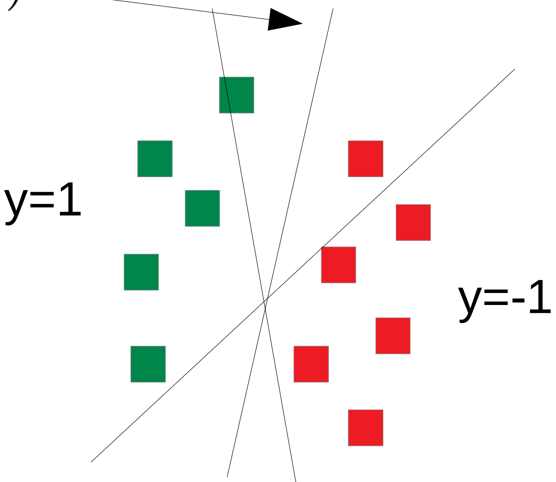
$$min(\|f(X,Y),Y'\|)$$

X

$(5, 8, 1, 5, 2, 3, 0, 2, 8, \dots, 5)^T$

$X \in \mathbb{R}^n$

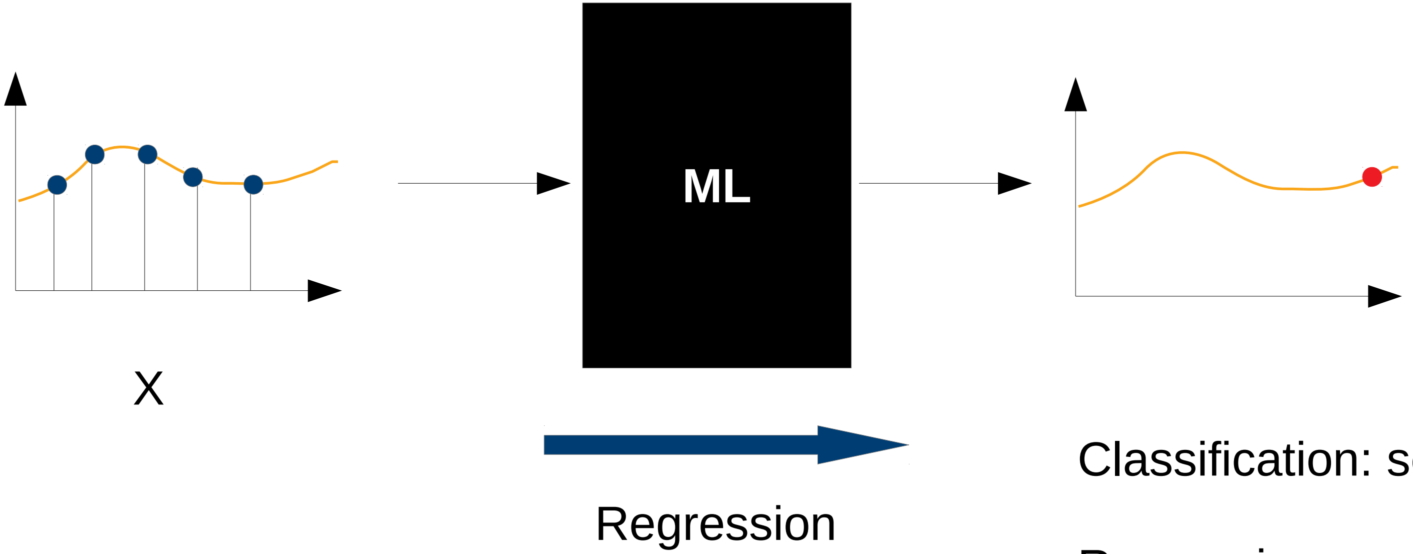
→ **ML** → Y': {-1,1}



$X \in \mathbb{R}^n$

Introduction to ML

Supervised Learning: Example: Regression



Y is **continuous**

Classification: separation function

Regression: continuous mapping
→ harder problem

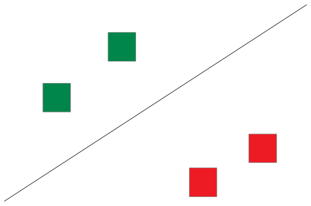
Hochschule Offenburg
offenburg.university

Challenges of Supervised Learning

- Not only need data – also need to have Y → human annotation
  - Getting “enough” labeled data is expensive
  - Sometimes impossible

UNDERFITTING

$$min(\|f(X,Y),Y'\|)$$



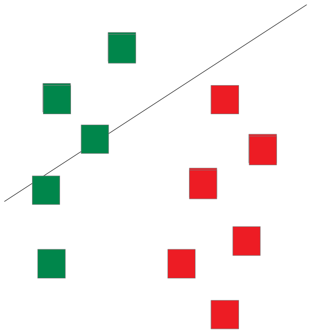
Training model
On little data

Challenges of Supervised Learning

- Not only need data – also need to have Y → human annotation
  - Getting “enough” labeled data is expensive
  - Sometimes impossible

UNDERFITTING

$$min(\|f(X,Y),Y'\|)$$



→ bad sampling
Of the data distribution

Introduction to ML

Challenges of Supervised Learning

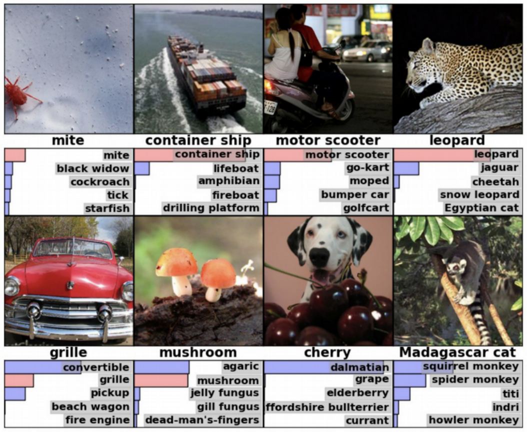
- Not only need data – also need to have Y → human annotation
  - Getting “enough” labeled data is expensive
  - Sometimes impossible

ImageNet Challenge

Example:

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Introduction to ML

Challenges of Supervised Learning

- Not only need data – also need to have Y → human annotation
  - Getting “enough” labeled data is expensive
  - Sometimes impossible

Example:

CITYSCAPES DATASET

Introduction to ML

Challenges of Supervised Learning

- Not only need data – also need to have Y → human annotation
  - Getting “enough” labeled data is expensive
  - Sometimes impossible
- Training data is **only a sample:** prediction must work on **all data** → **generalization**

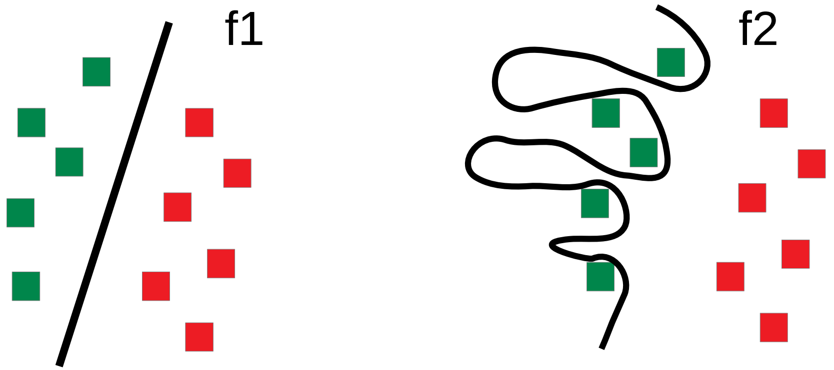
Introduction to ML

Challenges of Supervised Learning

- Training data is **only a sample:** prediction must work on **all data** → **generalization**

Which model is better?

$$min(\|f(X,Y),Y'\|)$$



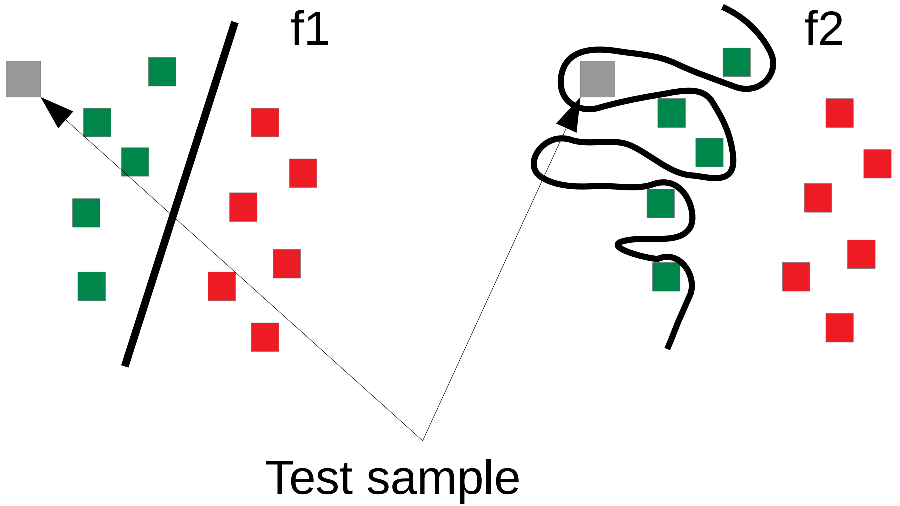
Introduction to ML

Challenges of Supervised Learning

- Training data is **only a sample:** prediction must work on **all data** → **generalization**

Which model is better?

$$min(\|f(X,Y),Y'\|)$$



Hochschule Offenburg
offenburg.university

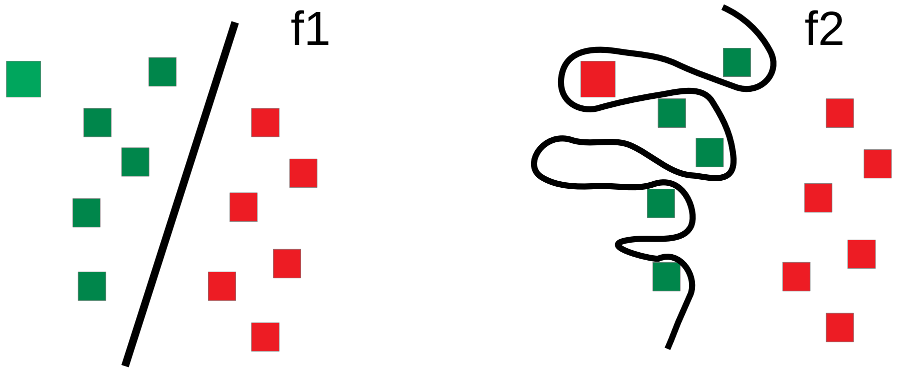
Challenges of Supervised Learning

- Training data is **only a sample:** prediction must work on **all data** → **generalization**

OVERFITTING

Model “to close” to train data

Very likely to happen in practice.
→ we need to work against this...



Introduction to ML

Data Preparation: Split into Train, Test, and Validate

A basic technique (we will learn more later) to at least detect overfitting is to split the available data into two or three subsets:

- Use unbiased **test set** for final evaluation of a model
- Use **train set** for model training
- **Validation set** (part of train set) can be used to optimize hyper parameters of the model

Caution: sets must be unbiased! (→ random sampling)
In practice it can be hard to guarantee clean train/test sets:
e.g. how to treat possible variance different data sources?
→ statistical analysis needed!

Introduction to ML

Basic evaluation (more techniques to come)

Train error: measure of how well the model predicts the given labels

$$Err_{train} := \frac{1}{|X_{train}|} \sum_{x_i \in X_{train}} |f(x_i) - y_i|$$

low train error is the necessary condition for a "good" model

Test error: **same as train error: low test error is the sufficient condition**

$$Err_{test} := \frac{1}{|X_{test}|} \sum_{x_i \in X_{test}} |f(x_i) - y_i|$$

Introduction to ML

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning

- NO Labeled data
- NO Direct and quantitative evaluation

- Explore structure of data

Introduction to ML

Hochschule Offenburg
offenburg.university

Unsupervised Learning

Data without “labels” $(x_1, x_2, \ldots, x_n)$

- **Clustering**
- **Outlier Detection (e.g. Defect or Intrusion detection)**

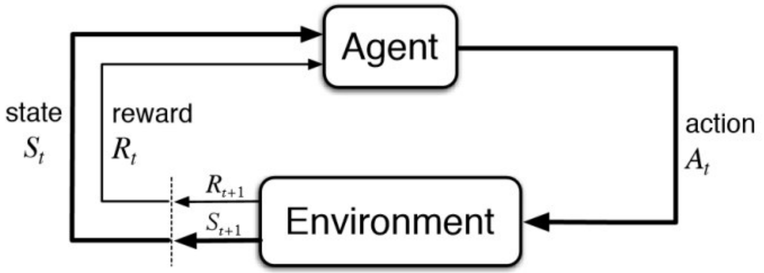
Introduction to ML

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning



- Learning decisions in an interactive environment
- State ↔ Action learning
- Game playing and robotics
- Hardly use in Data Science

ML in Python

Libraries used in this lecture:

Introduction in this week‘s lab

… introduction in Block 8