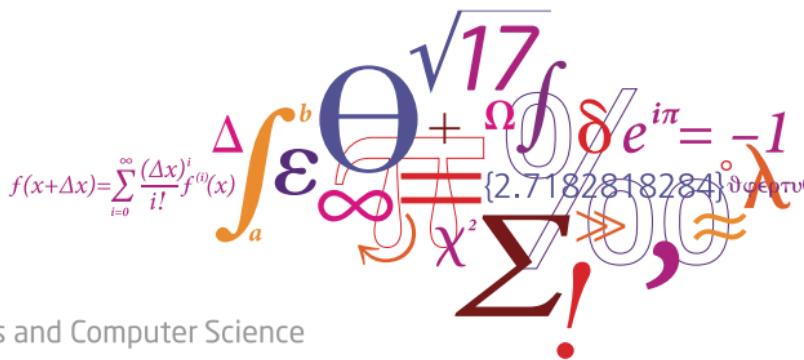


02450: Introduction to Machine Learning and Data Mining

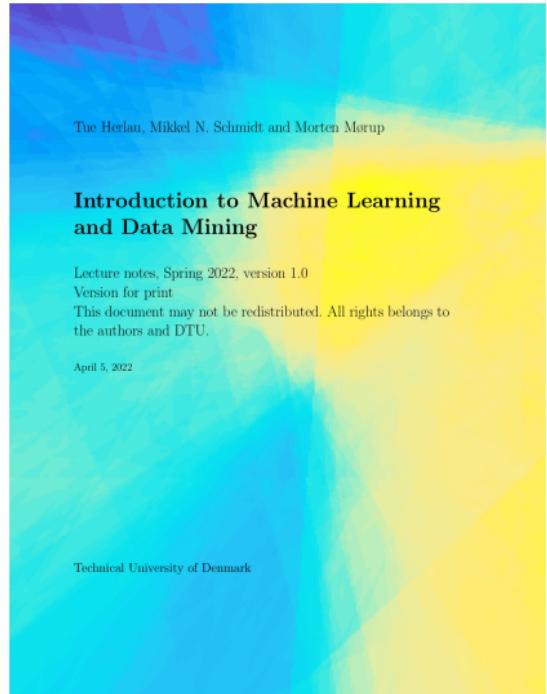
## Introduction

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



## Reading material: Chapter 1



# Lecture Schedule

## 1 Introduction

31 January: C1

Data: Feature extraction, and visualization

## 2 Data, feature extraction and PCA

7 February: C2, C3

## 3 Measures of similarity, summary statistics and probabilities

14 February: C4, C5

## 4 Probability densities and data visualization

21 February: C6, C7

Supervised learning: Classification and regression

## 5 Decision trees and linear regression

28 February: C8, C9

## 6 Overfitting, cross-validation and Nearest Neighbor

7 March: C10, C12 (Project 1 due before 13:00)

## 7 Performance evaluation, Bayes, and Naive Bayes

14 March: C11, C13

## 8 Artificial Neural Networks and Bias/Variance

21 March: C14, C15

## 9 AUC and ensemble methods

28 March: C16, C17

Unsupervised learning: Clustering and density estimation

## 10 K-means and hierarchical clustering

11 April: C18

## 11 Mixture models and density estimation

18 April: C19, C20 (Project 2 due before 13:00)

## 12 Association mining

25 April: C21

Recap

## 13 Recap and discussion of the exam

2 May: C1-C21

Online 24/7 help: Discussion Forum/Piazza  
Streaming & Videos: <https://panopto.dtu.dk/>  
Online exercises: MS Teams

# Public homepage

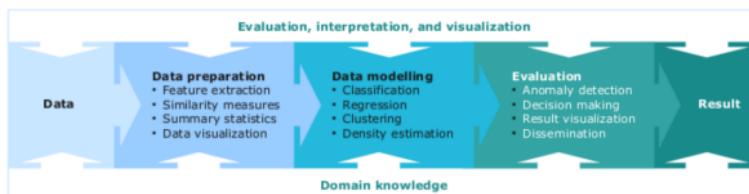
- Syllabus/homework also available on public homepage  
<http://compute.dtu.dk/courses/02450>  
and on DTU Learn (02450syllabus\_and\_practicals.pdf)



## 02450 Introduction to Machine Learning and Data Mining - Spring 2023

### Machine learning and data mining

The course is designed around a data modeling framework shown in the figure. Each lecture/assignment will focus on an aspect of the data modeling framework.



Bjørn Sand Jensen



Jes Frellsen

We emphasize the holistic view of modeling in order to motivate and stress the relevance of individual components and building blocks, disseminate the obtained competence (see the course learning objectives), and make them applicable for a broad spectrum of engineering problems in e.g. biomedical engineering, chemistry, electrical engineering, and informatics.

### Resources

#### DTU Learn

If you are enrolled in the course you can access material and participate in the course through the DTU Learn homepage.

#### Lectures

The lectures will take place in Building 116 auditorium 081 (overflow space available in Building 116, South and East lobby area where you can stream the lecture) on Tuesdays from 13:00-15:00.

#### Exercises

Exercises will take place after lectures Tuesdays from 15:00-17:00.

Please bring a laptop computer for the exercises. The exercises will be available in Matlab, R, and Python and we recommend selecting a language you are familiar with. If you

## Plan for today:

- What is machine learning?
- Why do we learn different methods?
- Impact of machine learning
- This course
- Pre-test
- Break
- Lecture 1, basic terminology
- Exercises in your favourite programming language (15:00–17:00)

## Alan Turing (1946)

Alan Turing  
(1912-1954)



- Universal computing
- Proposed machines should learn like children

*We are not in a position to answer if a machine can think because the terms machine and think are undefined. Rather we should ask if a machine can imitate a human*  
(the imitation game)

## Arthur Samuel (1959)

Arthur Samuel  
(1901-1990)



- Samuels wrote a checkers playing program
  - Program played 10000 games against itself
  - Learned value of each board position by considering the resulting score

**Machine learning:** *(The) field of study that gives computers the ability to learn without being explicitly programmed*

# Tom Michell (1999)

**A well-posed learning problem:** A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at the tasks in  $T$ , as measured by  $P$ , improves with experience  $E$

- Checkers example

- $E$ : Playing 10'000 games
- $T$ : Playing checkers
- $P$ : Win or loose



Tom Michell

# Quiz 01: Machine learning definition

Please answer this quiz on DTU Learn:

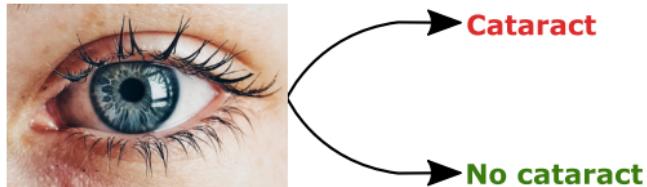
<https://learn.inside.dtu.dk/d2l/home/145394>

**-Well posed learning problem:** "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."



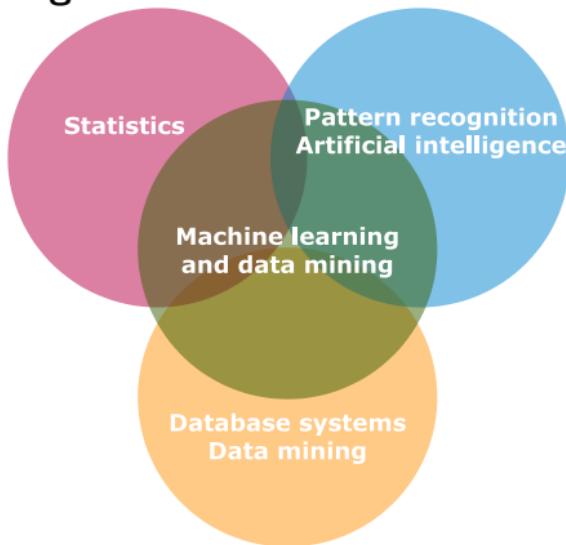
Tom Michell

Suppose a program watches as you label images of eyes as containing evidence of cataracts (clouding of the lens) or not, thereby learning to diagnose new examples. What is the experience E?



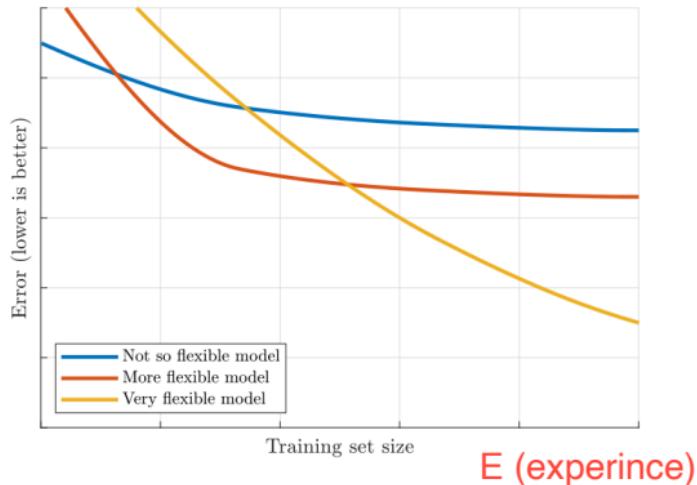
- A. The number of correctly diagnosed patients
- B. A database containing images of eyes with their labels
- C. The errors the program commits when trying to label the images.
- D. Physiological information about cataracts (genetic markers, disease progression, etc.)
- E. Don't know.

# Machine-learning as a research area



- Focus on a *learning algorithms* (rather than search, pathfinding, etc.)
- De-emphasize explicit knowledge representations, etc.
- Gradual improvements (training time, amount of data)
- *General* algorithms (or algorithmic ideas)

# Machine-learning as a research area



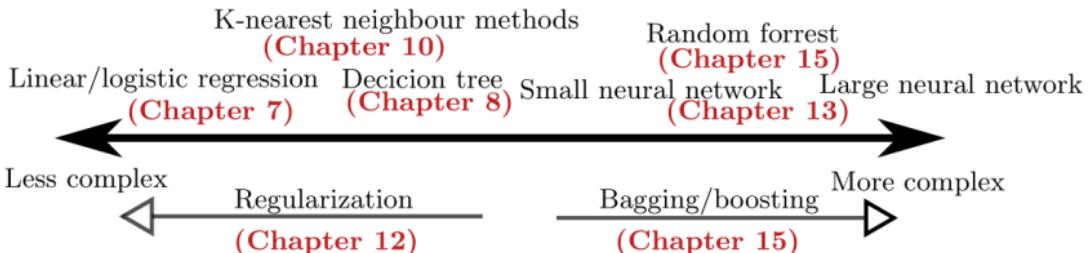
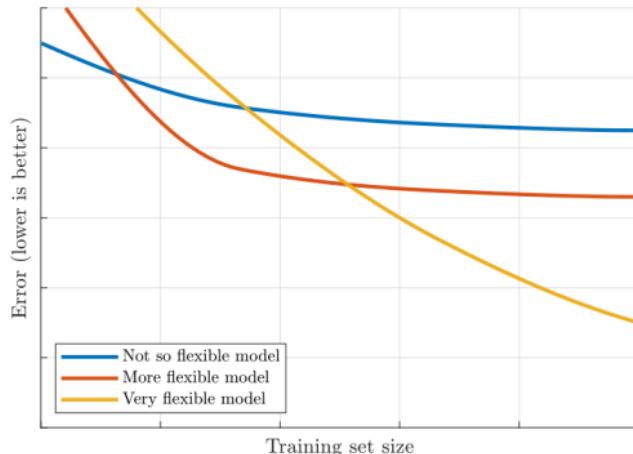
Examples:

blue: linear

red: 2. grads polynomium

yellow: complex graf

# Machine-learning as a research area



# Man vs. Machine

2022, ChatGPT

2021, alphafold Outperforms all state-of-art expert systems for protein structure prediction.

2020, Breast cancer Outperforms radiologists in breast-cancer detection (Nature)

2019, Lung cancer Outperform six doctors with a 5% reduction in false negatives (Deepmind / Nature Medicine)

2019, Starcraft 1v1: OpenAI deep reinforcement learning exhibit high-level performance in SC2

2018, BERT: Superhuman performance on the SQuADv1.1 wikipedia question-answer task

2018, alphago: superhuman chess/go learned from scratch

2017, Texas hold'em no limit: Libratus (Carnegie Mellon) beats top professional

2017, Go: Superhuman Go by reinforcement learning + imitation of expert games

2016, libreading : Superhuman libreading from Oxford and Google Deepmind

2016, conversational speech: Microsoft research demonstrate superhuman speech recognition

2016, Geoguessing Google PlaNet win 28 of 50 rounds; median localization error of 1132km vs. 2321km

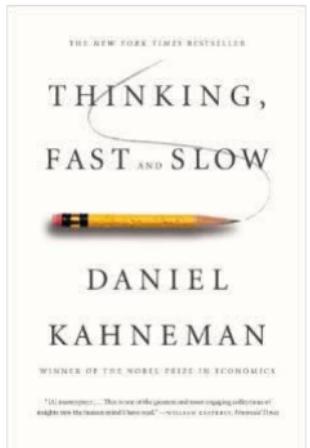
2015, closed-world image recognition Microsoft report error of 4.94% on ImageNet vs. 5.1% for top-human labeler

2015, Atari Google Deepmind obtain better-than-expert human performance on many Atari video games

List inspired by:

<https://finnaarupnielsen.wordpress.com/2015/03/15/status-on-human-vs-machines/>

# Human abilities examined



The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred (...) About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgement. No exception has been convincingly documented.

The range of predicted outcomes has expanded to cover medical variables such as longevity of cancer patients, the length of hospital stays, the diagnosis of cardiac disease, and the susceptibility of babies to sudden infant death syndrome; economic measures such as the prospect of success for new businesses, the evaluation of credit risks by banks, and the future career satisfaction of workers; questions of interest to government agencies, including assessment of the suitability of foster parents, the odds of recidivism among juvenile offenders, and the likelihood of other forms of violent behaviour; and the miscellaneous outcomes such as the evaluation of scientific presentations, the winners of football games, and the future prices of Bordeaux wine. Each of these domains entails a significant degree of uncertainty and unpredictability. We describe them as “low-validity environments.”. In every case, the accuracy of experts was matched or exceeded by a simple algorithm. (Kahneman 2011)

# Why now?

**Scientific** Advances in algorithmic ideas

**Empirical** Increased availability of large/good datasets

**Technological** Faster computers

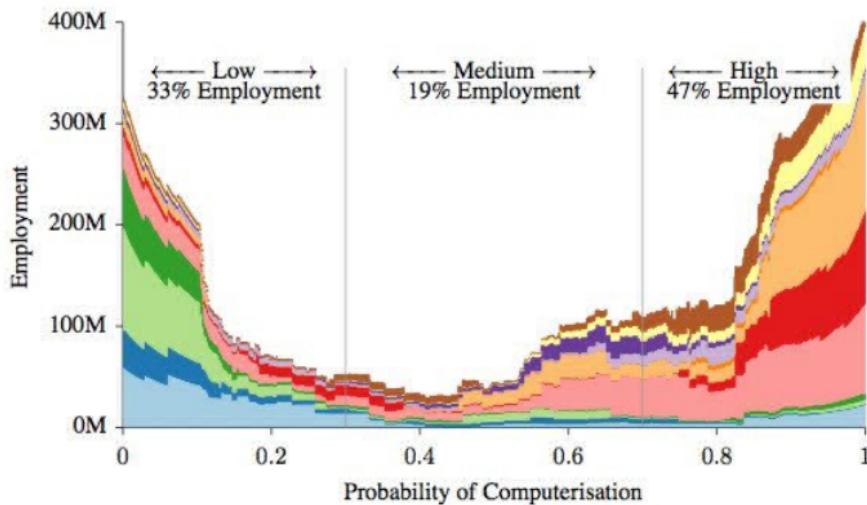
**Social** Libraries which automate routine tasks; increased sharing of code, etc.

**Economical** Greatly increased resource allocation

# Machine learning as a disruptive technology

- A recent Oxford study suggest about 47% of all US jobs could be automated within two decades (Frey & Osborne, 2013)

Management, Business, and Financial
Computer, Engineering, and Science
Education, Legal, Community Service, Arts, and Media
Healthcare Practitioners and Technical
Service
Sales and Related
Office and Administrative Support
Farming, Fishing, and Forestry
Construction and Extraction
Installation, Maintenance, and Repair
Production
Transportation and Material Moving



# Economical impact I

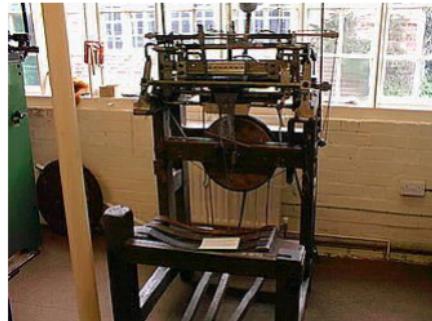
- Basic economics: When things become cheap, we will use it in more places
- Example: Microprocessors perform computations
- Microprocessors did not change the world because people did a lot of computations in 1950, but because **nearly everything can at least partially be turned into a computation problem** (bookkeeping, telephony, photography, entertainment, navigation, design, education, economics, science, etc.)
- We should not ask what situations **are as of now** a machine-learning problem, but which **can be turned into one**

## Economical impact II

Many jobs match this description

- ① Recognize what situation you are in
- ② Collect relevant data
- ③ Given data about situation make a **prediction** such as: (i) outcome of performing a given action in the situation or (ii) which action is appropriate
- ④ Perform action
- ⑤ Repeat ...

Machine learning can, in principle, learn 3



[https://en.wikipedia.org/wiki/William\\_Lee\\_\(inventor\)](https://en.wikipedia.org/wiki/William_Lee_(inventor))

William Lee (1563–1614) was an English clergyman and inventor who devised the first stocking frame knitting machine in 1589

Elizabeth I: "Thou aimest high, Master Lee. Consider thou what the invention could do to my poor subjects. It would assuredly bring to them ruin by depriving them of employment, thus making them beggars."

*"our discovery of means of economising the use of labour can outrun the pace at which we can find new uses for labour, as Keynes (1933) pointed out.*

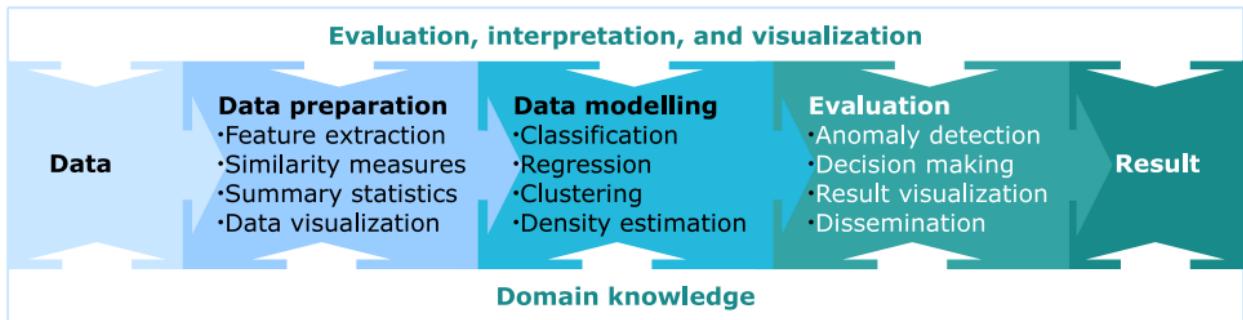
***The reason why human labour has prevailed relates to its ability to adopt and acquire new skills by means of education (Goldin and Katz, 2009). Yet as computerisation enters more cognitive domains this will become increasingly challenging (Brynjolfsson and McAfee, 2011)."***

(Taken from Frey & Osborne, 2013)

## Economical impact III

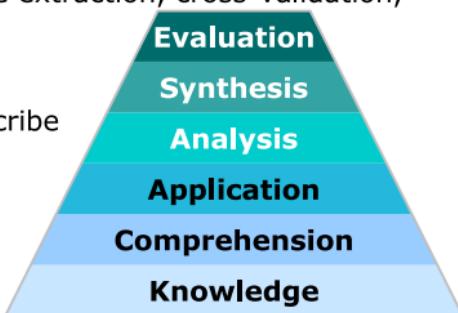
- We don't know what will happen
- Plausibly, many tasks will become so cheap humans will no longer perform them
  - Essential, non-automated tasks will both become more valuable, inhibit progress
  - Humans will do fewer jobs, play a relatively smaller role in the economy (the share of capital will increase relative to labor)
  - The most **destructive** forms of automation is when tasks are only **slightly** better done by machines

# Machine learning and data mining pipeline of this course



# Learning objectives

1. Describe the major steps involved in data modeling from preparing the data, modeling the data to evaluating and disseminating the results.  
*(Knowledge)*
2. Discuss key machine learning concepts such as feature extraction, cross-validation, generalization and over-fitting, prediction and curse of dimensionality.  
*(Comprehension)*
3. Sketch how the data modeling methods work and describe their assumptions and limitations.  
*(Knowledge and Comprehension)*
4. Match practical problems to standard data modeling problems such as regression, classification, density estimation, clustering and association mining.  
*(Comprehension and Application)*
5. Apply the data modeling framework to a broad range of application domains in medical engineering, bio-informatics, chemistry, electrical engineering and computer science.  
*(Application)*
6. Compute the results of the data modeling framework by use of Matlab, R or Python.  
*(Application)*
7. Use visualization techniques and statistics to evaluate model performance, identify patterns and data issues.  
*(Analysis)*
8. Combine and modify data modeling tools in order to analyze a data set of their own and disseminate the results of the analysis.  
*(Application, Analysis, Synthesis and Evaluation)*



# Assesment

The assesment consists of two components

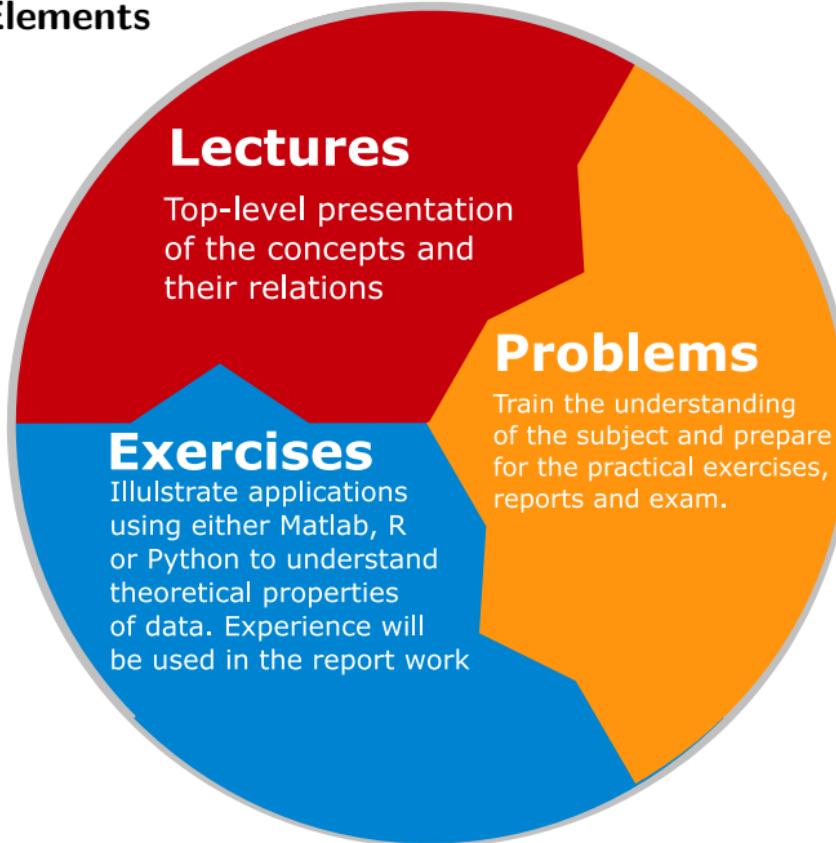
- 4 hours written multiple choice exam (homework problems are from previous exams)
- 2 project reports

**Report 1** March 7, *Data: Feature extraction, and visualization*

**Report 2** April 18, *Supervised learning: Classification and regression*

- Final grade based on overall assessment of reports and written exam. The written exam is weighted more than the reports ( $\approx 75\%$  vs  $\approx 25\%$ )

# Course Elements



## Course format

### Lecture session [2 hours per week]

- Physically - main auditorium 116-81. [275 seats]
- Online - live webcast via Panopto (via [Video & Streaming](#) on DTU Learn).
- Offline - lectures recorded and uploaded to Panopto a few days after the lecture

### Exercise session [2 hours per week]

- Physically - find rooms & information on DTU Learn (recommended).
- Online - interaction with TAs using (chat/video/audio) via Microsoft Teams channels.

# Detailed structure and activities

- **Workload:** DTU's **nominal** workload for a 5 ECTS course is **9 hours per week** during the 13-week period and 140 hours in total.

- **Structure and study-activities :**

- Lecture session [2 hours per week]:
  - **High-level lectures** including quizzes and in-class discussions (formative, i.e. not assessed)
- Supervised **exercises** (formative) [2 hours per week]
  - Focus on the central aspects of the exercises, not the project work.
  - TAs will go over some aspect of the exercise/homework around 16:45 (starting in week 2).
- Self-study, preparation & project work [5 hours per week]
  - We assign **readings, homeworks, quizzes** and provide **old exams** to help you study more effectively (not assessed).
  - **Project work (assessed)**, you apply the taught theory to your own data.
  - Online help/assistance on any aspect of the course via Piazza.
- **Exam (assessed)** [4h + preparation (19h)]

# Group Learning

- We encourage group learning during lectures, exercises and project work.
  - **During lectures** You are encouraged to work together and discuss the online Quiz questions.
  - **During exercises** Each exercise consists of computer exercises relevant for the report work as well as a conceptual multiple-choice question from a previous exam. Please only spend about 15 minutes on the multiple-choice question.
  - **Project reports** Submitted as group work (3 people) based on your own dataset.
- For the project reports, you have to register your group at DTU Learn > My Course > Groups (target is 3 persons per group).

## Online help

- <https://piazza.com/dtu.dk/spring2023/02450> (Sign up!)
- Use Discussion Forum (i.e. Piazza) for 24/7 help
- Ensure everyone have access to the same information
  - **Bad: very general questions, i.e. can you explain GMM?**  
Good: Here is what I understand, but I don't get equation ...
  - **Bad: error without context, i.e. I tried to do a PCA but I get an error the matrix has the wrong size?**  
Good: What language are you using and what is the error; code which produced the mistake; what did you try to accomplish?
  - **Bad: Can you explain the PCA question in the Fall 2009 exam?**  
Good: Insert screenshot of fall 2009 exam, explain what part of the solution is unclear

## Subsequent ML courses

This course (02450) is recommended prerequisite for:

- 02456 Deep learning
- 02471 Machine learning for signal processing
- 02477 Bayesian machine learning
- 02460 Advanced Machine Learning (also recommended prerequisites 02456 / 02471 / 02477)

More courses will probably be added soon.

# Pretest

- The purpose is to assess your background in order to adjust the presentation and measure your learning.
- Some of the questions will be hard and may seem unfair. Do not Google it. We want to know, if you know.
- **Not part of your assessment. We never look at individual results.**
- Start during the break and finish during the exercise session.

Go to: DTU Learn > 02450 > Quizzes > Pretest

# Data Mining and Machine Learning Tasks

## Predictive tasks (Supervised learning)

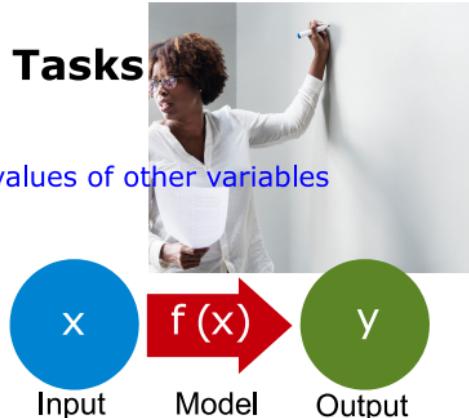
- Use some variables to predict unknown or future values of other variables

- **Classification**

- Discrete output  
(Determine which class a new data object belongs to)

- **Regression**

- Continuous output  
(Determine the output value from the input variables)



## Descriptive tasks (Unsupervised learning)

- Find human-interpretable patterns that describe the data

- **Clustering**

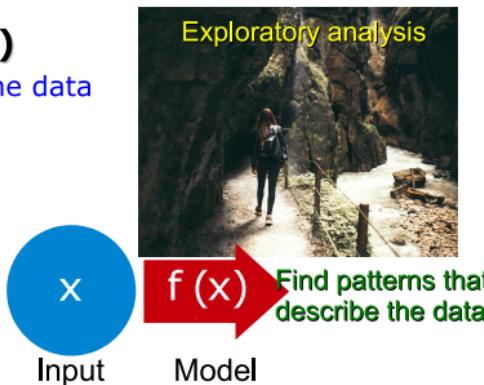
- Discover group structure in data

- **Association rule discovery**

- Discover how data objects relate to each other

- **Anomaly detection**

- Find data objects that are abnormal



# Classification: Definition

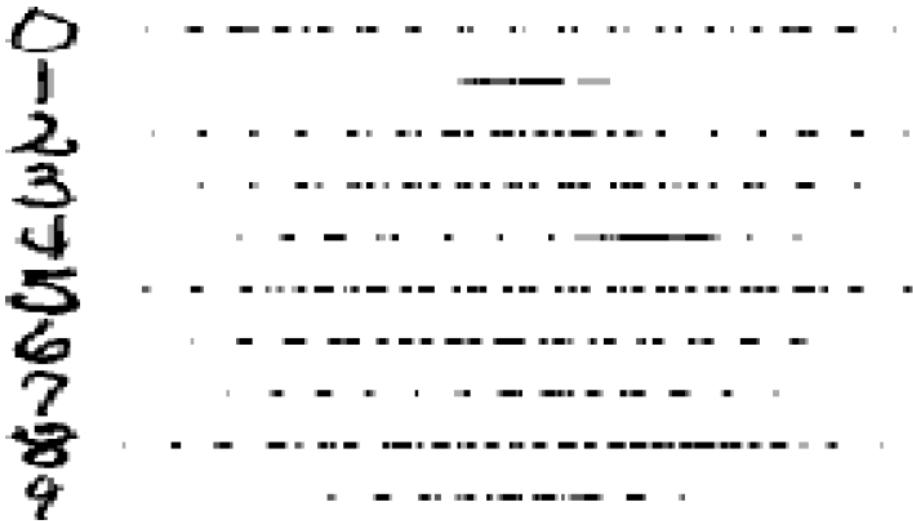
- Given a collection of data objects (**training set**)
  - Each object has associated a number of features
  - Each object belongs to a certain class
- Define a **model** for the class given the other features
- Goal: Assign a class label to a **previously unseen object**

## Classification: Example

Training set										Classify		
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	5	2	4
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	?	?	?

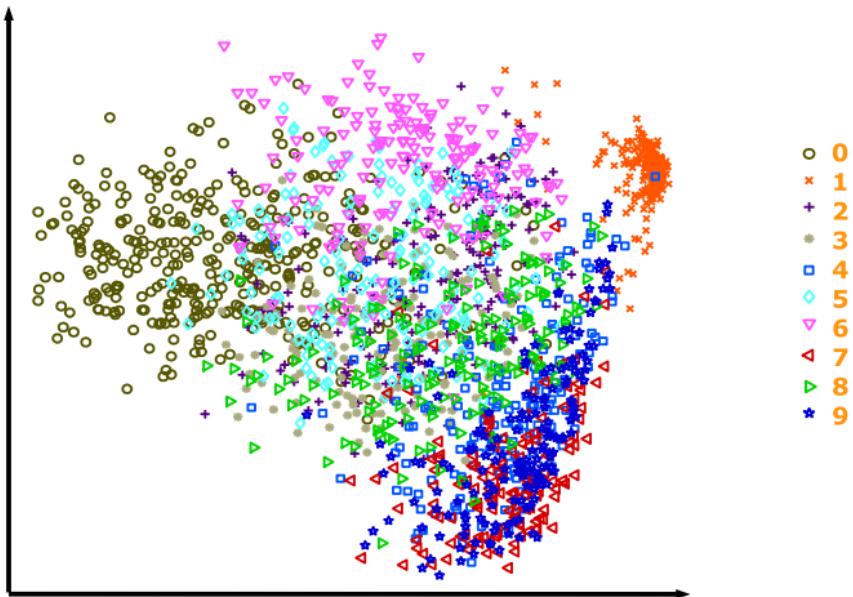
# Classification: Example

## Data representation

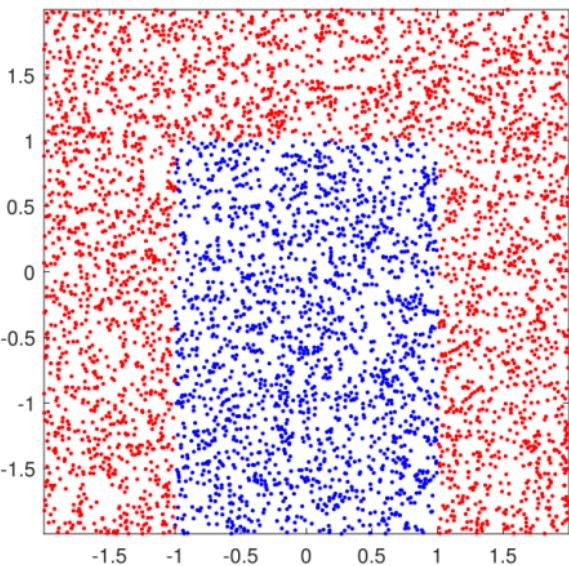


# Classification: Example

## Visualization



## Quiz 02 (DTU Learn): Decision rules



The figure shows an example classification problem consisting of a large number of observations  $(x, y)$  along with their class (red and blue).

A *decision rule* is just a function which takes an  $(x, y)$  coordinate and outputs either the red or the blue class. Suppose we define:

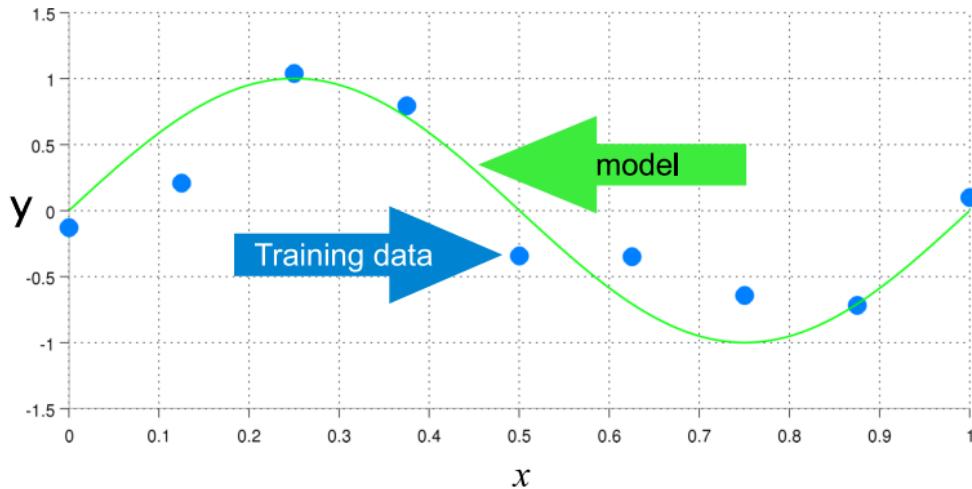
$$\begin{aligned}z_1 &= \max\{0, x - 1\} + \max\{0, -1 - x\} \\z_2 &= \max\{0, -1 + y\}\end{aligned}$$

Which of the following *decision rules* solve the problem?

- A. If  $z_1 = z_2 = 0$  classify as blue and otherwise as red
- B. If  $z_1 = z_2 = 1$  classify as blue and otherwise as red
- C. If  $z_1 = 1$  and  $z_2 = 0$  classify as blue and otherwise as red
- D. If  $z_1 = 0$  and  $z_2 = 1$  classify as blue and otherwise as red
- E. Don't know

# Regression: Definition

- Given a collection of data objects
  - Each object has associated a number of features
  - Each object has associated a **continuous valued variable**
- Define a **model** for the variable given the features
- Goal: Predict the value of the variable for a **previously unseen object**



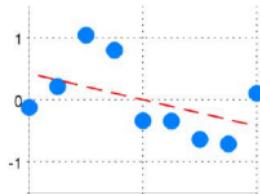
# Regression: Example

- Predict **sales amounts** of new product based on
  - advertising expenditure
- Predict **wind velocity** as a function of
  - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
  - previous index time series and market indicators

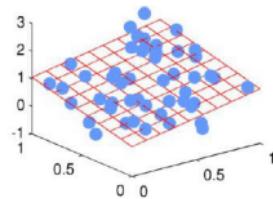
# Regression: Example

- Predict **sales amounts** of new product based on
  - advertising expenditure
- Predict **wind velocity** as a function of
  - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
  - previous index time series and market indicators

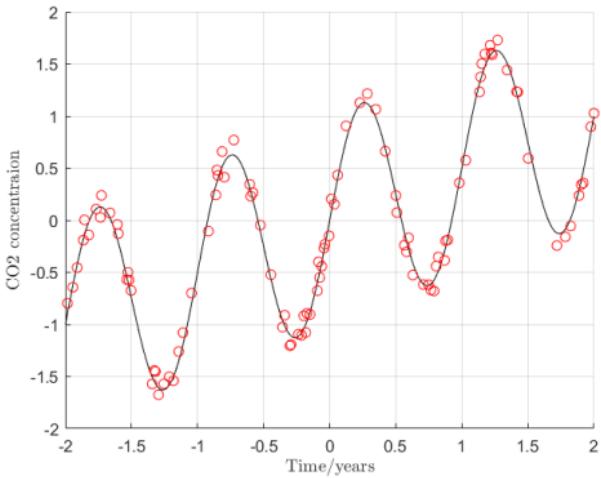
1-dimensional inputs  
 $f(x) = w_0 + w_1x$



2-dimensional inputs  
 $f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2$



## Quiz 03 (DTU Learn): Regression



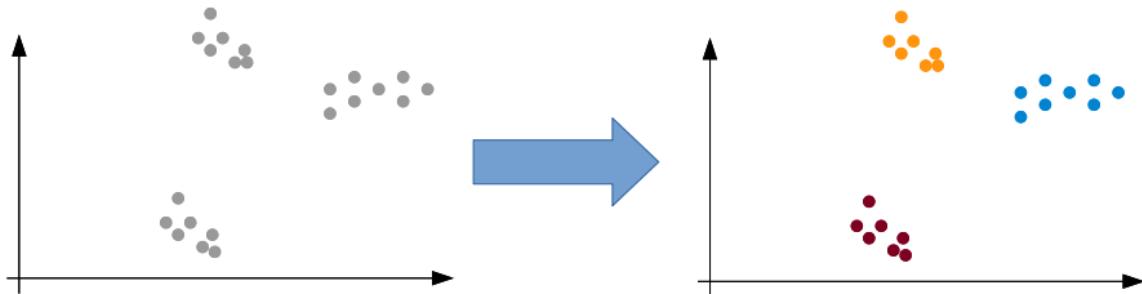
The figure shows an example regression problem where the CO<sub>2</sub> concentration is measured as a function of the time of year.

We wish to come up with a prediction rule  $y = f(x)$  where  $y$  is the relative CO<sub>2</sub> concentration and  $x$  is the time of year. Which of the following functions would be a good candidate?

- A.  $y = 0.5x + \cos(x)$
- B.  $y = -0.5x + \cos(x)$
- C.  $y = 0.5x + \sin(2\pi x)$
- D.  $y = -0.5x + \sin(2\pi x)$
- E. Don't know

# Clustering: Definition

- Given a collection of data objects
  - Each object has associated a number of features
  - A measure of **similarity** between objects is defined
- Goal: **Group the objects** into clusters such that
  - Objects within each cluster are similar
  - Objects in separate clusters are less similar



# Clustering: Example

## Document clustering

- Goal
  - Find groups of similar documents based on the words appearing in them

- Approach
  - Identify frequently occurring words in each document
  - Define a similarity measure based on the word frequencies
  - Perform clustering to find groups of documents

- Motivation
  - Use the clusters to relate a new document to existing documents
  - Better search algorithms: Return documents that are similar but do not have the exact search keywords

## Association rule discovery: Definition

- Given a set of **records**
  - Each containing a number of **items from a set**
- Goal: Produce dependency rules
  - Predict the occurrence of an item based on occurrences of other items

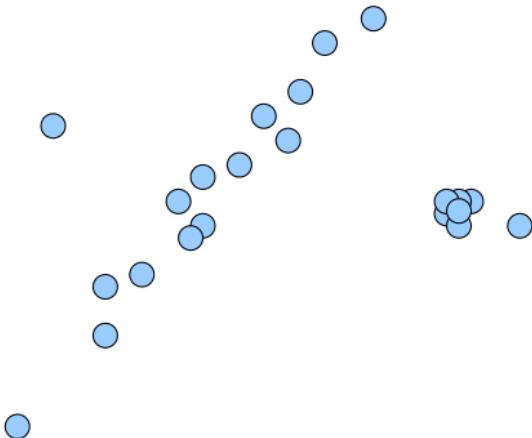
# Association rule discovery: Example

## Market basket analysis

Training set	Rules discovered
1.{Bread, Coke, Milk}	{Milk} ►{Coke}
2.{Beer, Bread}	{Diaper, Milk} ►{Beer}
3.{Beer, Coke, Diaper, Milk}	
4.{Beer, Bread, Diaper, Milk}	
5.{Coke, Milk}	

## Anomaly detection: Definition

- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour



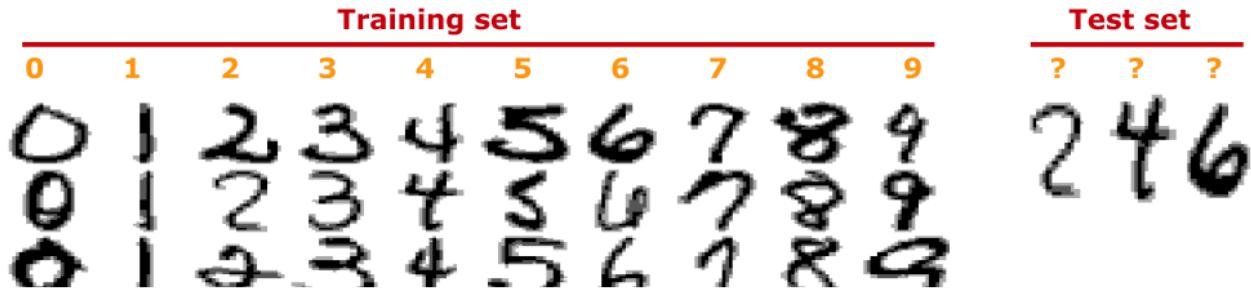
# Anomaly detection: Example

- Credit card **fraud detection**
  - Recognize dubious credit card transactions based on the transaction history of the card holder
- Detection of **outliers** in data measurements
  - Remove erroneous measurements due to misreading from an instrument
- **Fault detection** in system health monitoring
  - Detect when a wind turbine performs poorly due to ice coating on blades

# Models in machine learning

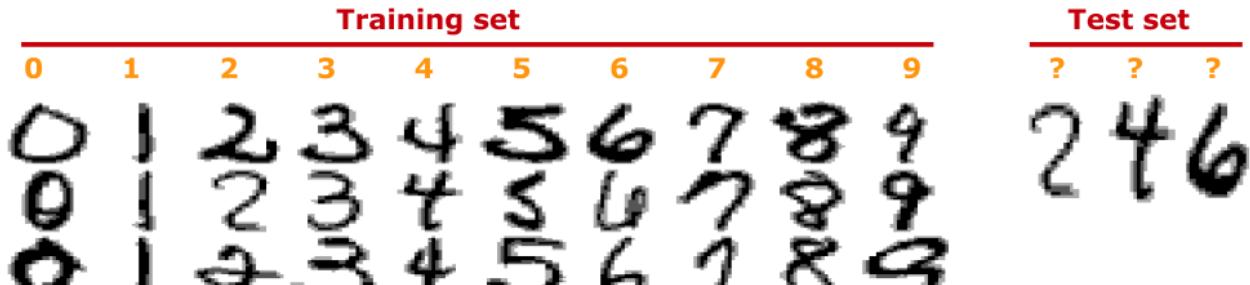
Training set										Test set		
0	1	2	3	4	5	6	7	8	9	?	?	?
0	1	2	3	4	5	6	7	8	9	2	4	6
0	1	2	3	4	5	6	7	8	9			
0	1	2	3	4	5	6	7	8	9			

# Models in machine learning

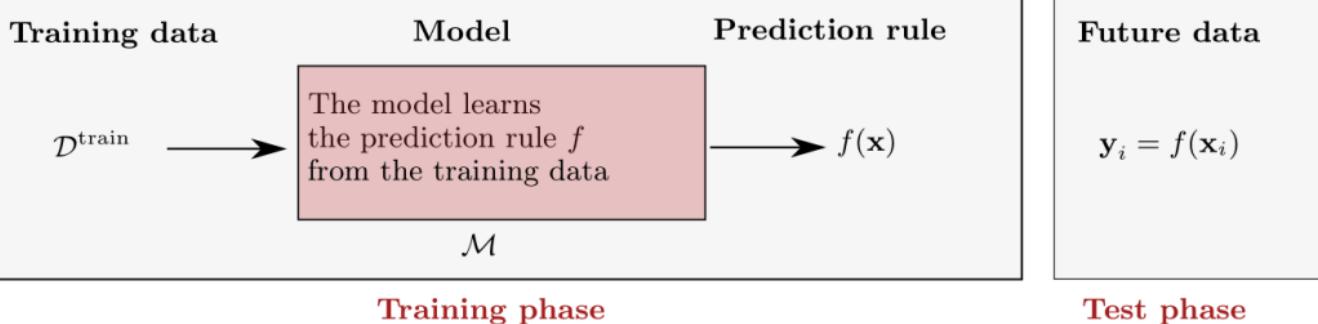


Classifying digits is a mapping  $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$

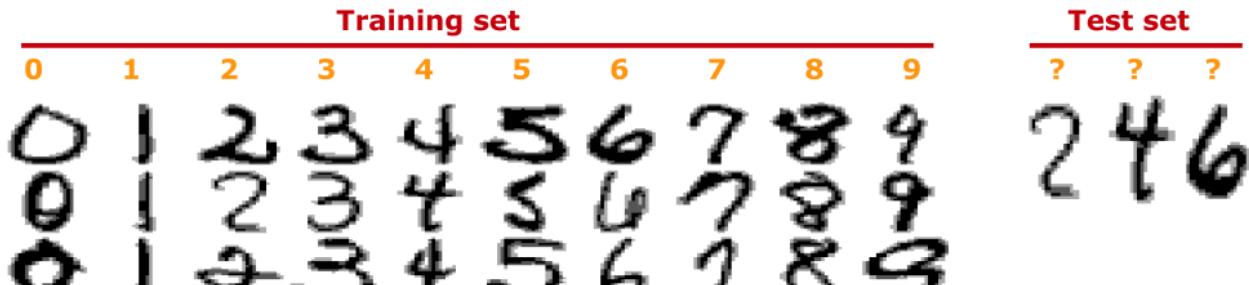
# Models in machine learning



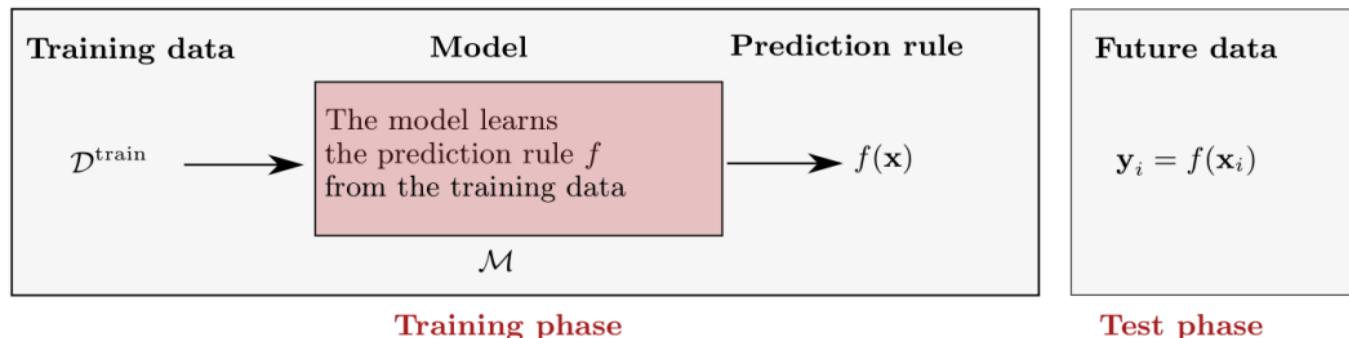
Classifying digits is a mapping  $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$



# Models in machine learning



Classifying digits is a mapping  $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$



How often the learned function  $f$  makes errors in the future is the **generalization error**

# Exercises

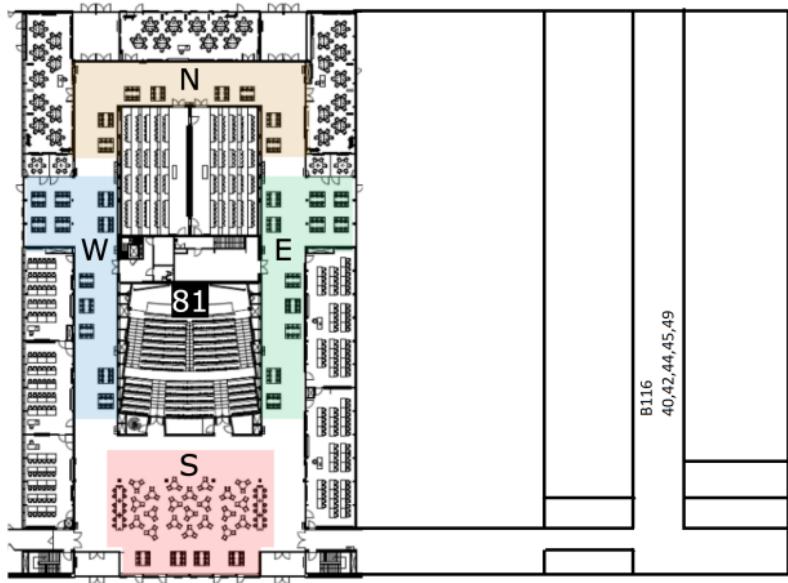
We support **Matlab**, **Python**, and **R**

- Exercise 0 guided you through installing your chosen environment
- If you have no experience with either, we recommend **Python**

Rooms for exercises:

- Building 116, Auditorium 81, (Python, R and Matlab)
- Building 116, Room 011 [online room], (Python,R,Matlab)
- Building 116, Room 013, (Python)
- Building 116, Room 015, (Python)
- Building 116, Room 040, (Python)
- Building 116, Room 042, (Python)
- Building 116, Room 044, (Python)
- Building 116, Room 045, (Python)
- Building 116, Room 049, (Python)

# Building 116



# Exercises Today

- Follow exercise instructions available on DTU learn (Exercise 1)
- Start forming groups (**target is 3 students per group**)
  - Find team members via the exercise session, Discussion Forum (i.e. Piazza), or other channels.
  - Unable to find a group (say by week 3)? Enter your info in MS Teams > General > Shared files > 02450missing\_a\_group.xlsx. We will then assign you to a group.
  - Once formed, register your group on DTU Learn.
- Start looking for a dataset and discuss the suitability of specific dataset with a TA.
  - Instructions for finding a dataset on DTU Learn > 02450 > Project descriptions > 02450finding\_a\_dataset\_for\_reports.pdf and the project 1 description.

## Resources

<https://www.mckinsey.com> Impact assessment of automation by McKinsey

(<https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy>)

<https://towardsdatascience.com> Another introduction to machine learning basics (<https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4>)

<https://www.economicsofai.com> conference on modelling economical impact of AI (<https://www.economicsofai.com/nber-conference-toronto-2017/>)

<https://deepmind.com> Obviously google-focused, but otherwise a great resource for what is hot right now (<https://deepmind.com/>)