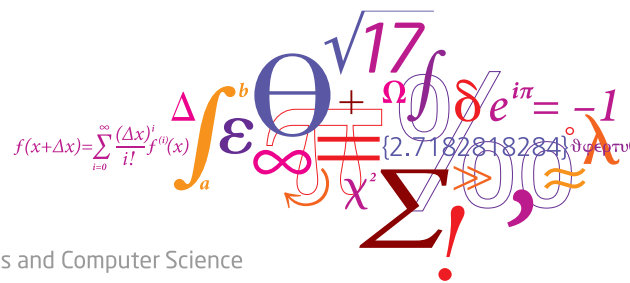
02450: Introduction to Machine Learning and Data Mining

Recap and discussion of the exam

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

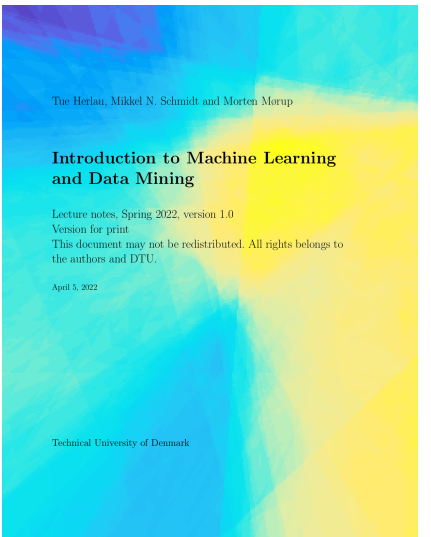
Today

Feedback Groups of the day:

Matthias Rötzer, Katrine Meldgård, Astrid Ingeborg Skiaker, Niccolò Moro, Gianluca Simone, Mario Rojo Vicente, Mathias Herløv Lund, Rita Rodrigues Saraiva, John Vinh Quang Tran, Fenfen Liu, Stine Robberstad, Jonas Martin Smidt, Anastasia-Danai Panagiotopoulou, Malthe Ørberg Pedersen, Jónas Mittún Peltonen, Javier Lopez Bort, Pablo Sánchez-Izquierdo Besora, Anna Miotto, Kaarel Siimut, Alberto Santi, Emil Laurberg Pedersen, Emilie Simone Muff, Jeremy Wilson, Jiaxin Wang, Charalampos Tsafaras, Giacomo Palù, Lau Thomsen, Weining Ren, Monica Nielsen, Jónas Ingi Valdimarsson, David van Scheppingen, Enrique Vidal Sanchez, Lennart Weigt, Line Winter, Florian Wolf, Xindi Wu, Jianan Yang, Mohammad Abu Yousuf, Anna Emilie Jennow Wedenborg

Reading material:

Chapter 1-Chapter 21



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

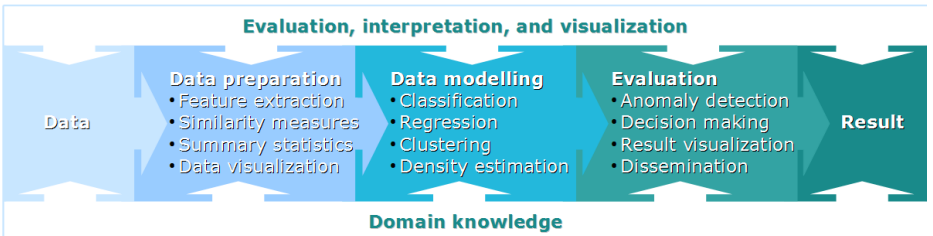
Unsupervised learning: Clustering and density estimation

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



Learning Objectives

- Remember key aspects taught in the course
- Practical information about the exam

What is Machine Learning (ML)?

Alan Turing
(1912-1954)

Alan Turing (1946)

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 Turing test**).

Proposed we should consider machines that were able to learn like children.

Arthur Samuel (1959)

Machine learning: **"Field of study that gives computers the ability to learn without being explicitly programmed"**

Samuels wrote a checkers playing program, had the program play 10000 games against itself and work out which board positions were good and bad depending on wins/losses.

Arthur Samuel
(1901-1990)

https://commons.wikimedia.org/wiki/File:This_is_the_photo_of_Arthur_Samuel.jpg

Tom Mitchell (1999) (http://www.cs.cmu.edu/~tom/)

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."**

For checkers we have:

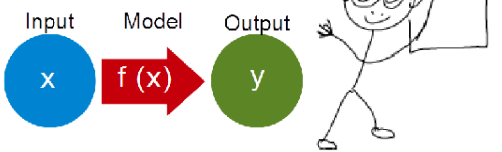
E = 10000 games

T = Playing checkers

P = If you win or not

What can ML be used for

Supervised learning

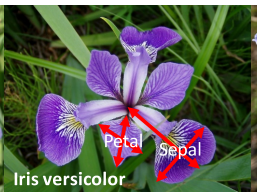


y: different classes ⇒ classification problem
y: continuous values ⇒ regression problem

Unsupervised learning

Density estimation
Association mining
Clustering

Example: Fisher's Iris data

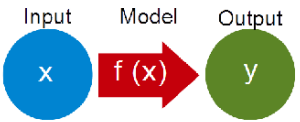


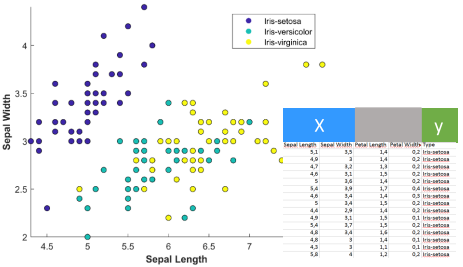
https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg
https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_versicolor_3.jpg
https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_virginica.jpg

Ronald Fisher
(1890 - 1962)

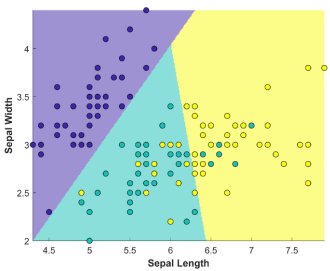
Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa
4,8	3,4	1,6	0,2	Iris-setosa
4,8	3	1,4	0,1	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa

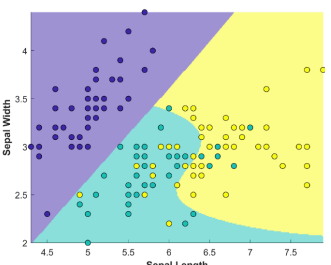
Supervised learning

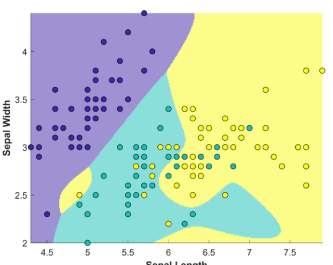




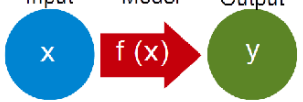
y: different classes ⇒ classification problem



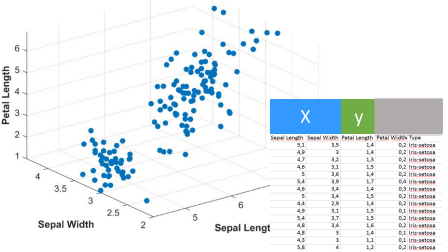


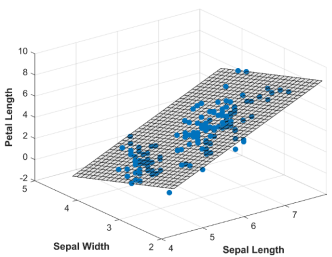


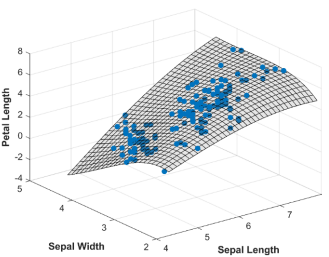
Supervised learning

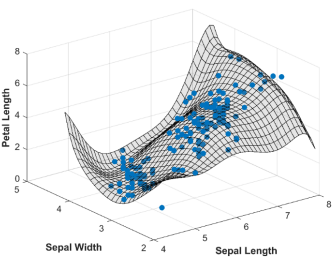


y: continuous values ⇒ regression problem

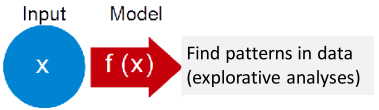


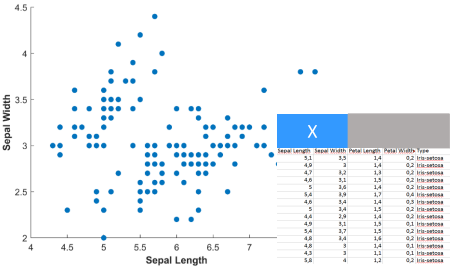




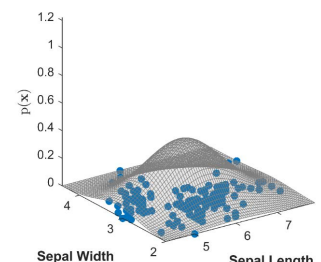


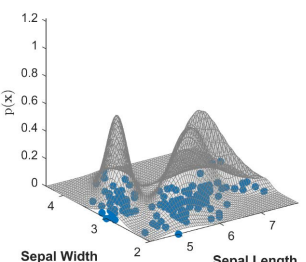
Unsupervised learning

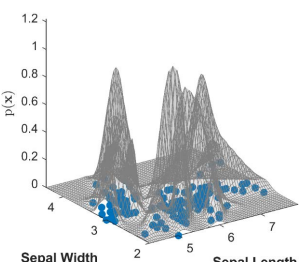




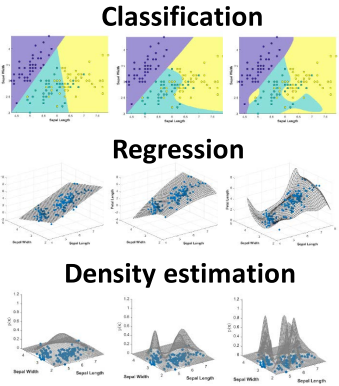
Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa
4.3	3	1.1	0.1	Iris-setosa
5.8	4	1.2	0.2	Iris-setosa





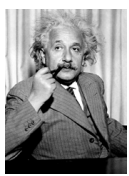


Control of model complexity



Lex Parsimoniae,
Law of parsimony
Given two models with same predictive performance, the simpler model is preferred over the more complex model (paraphrased)

William of Ockham
(1288-1347)



"Everything should be made as simple as possible, but not simpler"

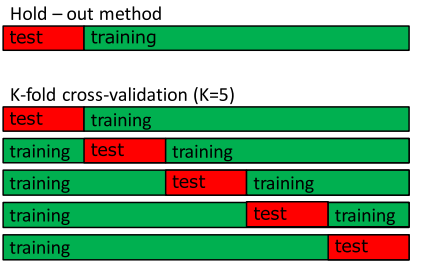
Albert Einstein
(1879 - 1955)

Machine learning in one sentence:

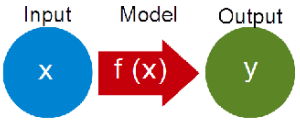
The aim of machine learning is to minimize the generalization error.

Generalisation error:
The extend to which a machine learning method on average will fail when evaluated on an infinite amount of (test) data.

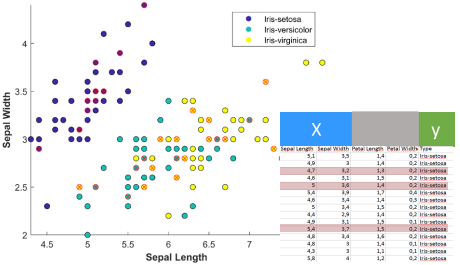
Cross-validation:
A framework for quantifying the generalization error from the available data.

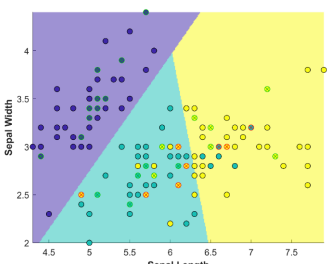


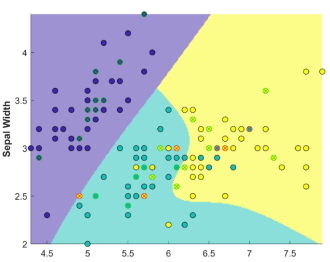
Supervised learning

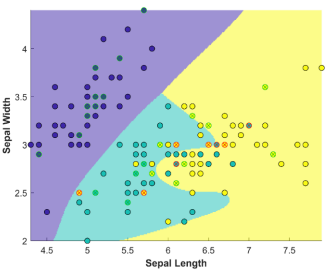


y: different classes ⇒ classification problem

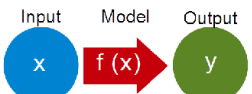




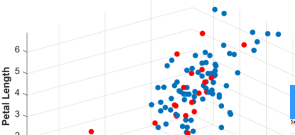


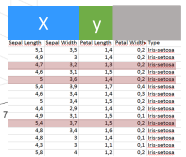


Supervised learning

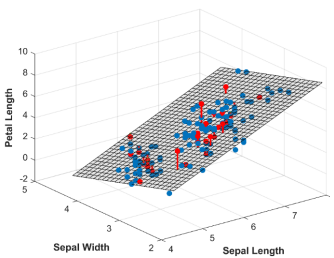


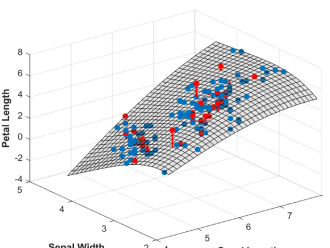
y: continuous values ⇒ regression problem

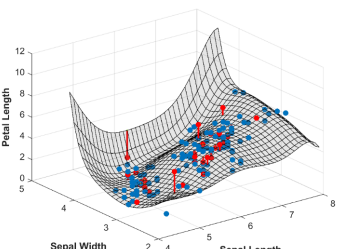




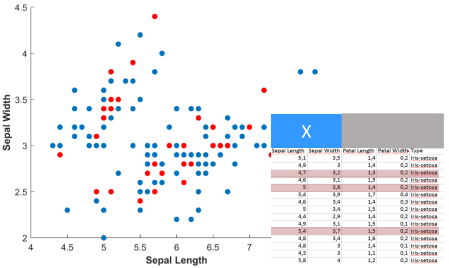
Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa
4,8	3,4	1,6	0,2	Iris-setosa
4,8	3	1,4	0,1	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa



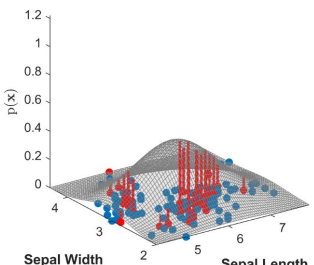


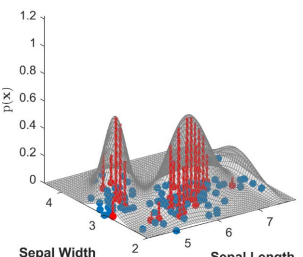


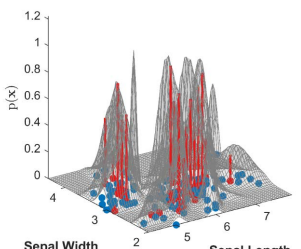
Unsupervised learning



Sepal Length	Sepal Width	Petal Length	Petal Width	Type
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa
4,8	3,4	1,6	0,2	Iris-setosa
4,8	3	1,4	0,1	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa

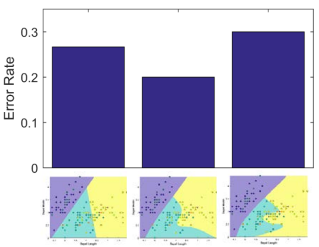




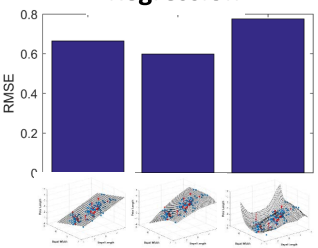


Experience = Amount of data (number of observations)
Task = Classification/Regression/Density estimation
Performance =How well we can predict classes/output values/where data occur

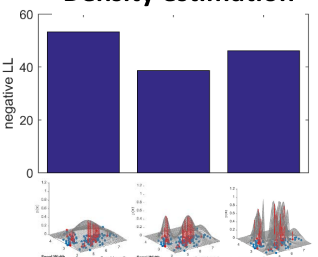
Classification

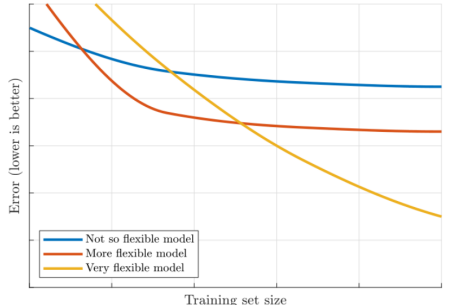


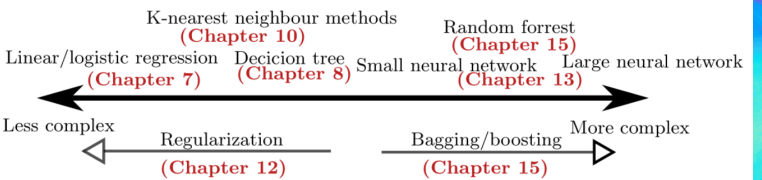
Regression

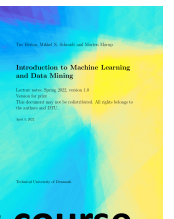


Density estimation

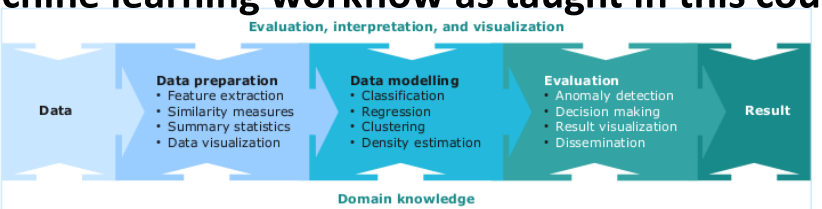








Machine learning workflow as taught in this course



Feedback/Improvements

- How to improve the hybrid setup (lectures and exercises)
- The AV setup was not always optimal (Panopto and the AV in the room)
- Lectures/book: "Too much maths" vs. "too little maths".
- Exercises: "too much programming", "not enough programming".
- Should the exercies be mandatory (perhaps even graded)?
- More explicit information about the dataset selection (e.g., a fixed dataset for all)?

Ideas welcome!

What should I do, if I want more Machine Learning?

This course (02450) is recommended prerequisite for:

- 02456 Deep learning
- 02471 Machine learning for signal processing
- 02477 Bayesian machine learning
- 02460 Advanced Machine Learning (currently being restructured)
- Special courses (typically on advanced topics).
- Projects: BSc, MSc and PhD.

Assesment

The grade is based on an "overall assesment" of the projects (approx. 1/4) and the written exam (approx. 3/4).

The written exam:

- When: 23rd May 2023, 9:00-13:00
- Location: DTU Campus (see `https://eksamensplan.dtu.dk`).
- Duration: 4 hours.
- Exam available at `https://eksamen.dtu.dk` (passcode provided by invigilators).
- You **must** bring a computer with wifi (and a charger).
- All aids permitted (**no internet**; except eksamen.dtu.dk).
  - Download the material you need to your computer (incl. Python/Matlab/R)
  - Contact the study admins if in doubt about what you can bring/use.
- Individual (NB AI-models such as ChatGPT disallowed c.f. DTU rules).
- Multiple choice (5 options; 1 correct answer; correct answer: $+3$; incorrect: **-1**, don't know: 0).
- **Fully electronic** (exam set and hand-in).
- Syllabus: See document on DTU Learn.

Exam electronic hand-in

- Available at same location as `.pdf` exam set
- Please hand in as a `.txt` file
- Keep it machine-readable
- Only write A, B, C, D or E (do not indicate multiple options, e.g. "A or C", it will be marked as E)
- Do **not** fill out `Documents/answers.txt` and upload `Downloads/answers.txt`

`answers.txt`:

```
ELECTRONIC HANDIN FOR THE 02450 EXAM

The exam is multiple choice.  All questions have four possible
answers marked by the letters A, B, C, and D as well as the answer
"Don't know" marked by the letter E. Correct answer gives 3 points,
wrong answer gives -1 point, and "Don't know" (E) gives 0 points.

The individual questions are answered by replacing the question
marks '?'  at each of the 27 questions with one of the letters A,
B, C, D, or E.

Please do not change the file format or write additional comments
asides your student number and your answers in terms of the letters
A,B,C,D or E for each of the questions.

Q01:  ?
Q02:  ?
Q03:  ?
Q04:  ?
Q05:  ?
Q06:  ?
Q07:  ?
Q08:  ?
Q09:  ?

...
```

Exam electronic hand-in

- Available at same location as `.pdf` exam set
- Please hand in as a `.txt` file
- Keep it machine-readable
- Only write A, B, C, D or E (do not indicate multiple options, e.g. "A or C", it will be marked as E)
- Do **not** fill out `Documents/answers.txt` and upload `Downloads/answers.txt`

`answers.txt`:

```
ELECTRONIC HANDIN FOR THE 02450 EXAM

The exam is multiple choice.  All questions have four possible
answers marked by the letters A, B, C, and D as well as the answer
"Don't know" marked by the letter E. Correct answer gives 3 points,
wrong answer gives -1 point, and "Don't know" (E) gives 0 points.

The individual questions are answered by replacing the question
marks '?'  at each of the 27 questions with one of the letters A,
B, C, D, or E.

Please do not change the file format or write additional comments
asides your student number and your answers in terms of the letters
A,B,C,D or E for each of the questions.

Q01:  A
Q02:  B
Q03:  E
Q04:  D
Q05:  E
Q06:  A
Q07:  B
Q08:  B
Q09:  C

...
```

... a few more comments/suggestions

- Read the whole exam set and prioritize the questions you are most comfortable with.
- Make rational decisions! Expected score for uniform random guess: $1/5$ * (+3 - 1 - 1 -1 + 0) = 0.
- Keep in mind the type and variability of questions from year to year! All questions are unique, and there are always unseen/new question types and variations.
- Practice the exam with a strickt 4 hour time limit **without** the solutions.
- Do the pratical exercies! They provide you with insights/intuitions into the methods and variations.

Rest of today... and beyond

Today

- Try the test exam!
  - Go to `https://eksamen.dtu.dk`
  - Passcode: 03699 (only valid for the test exam!)
  - Download the `.pdf` and `answers.txt` file..
  - Input answers in `answers.txt`
  - Figure out how to upload it
  - Available 2nd-14th May (solutions availble via DTU Learn on the 8th May).
  - Not formally evaluated

Assistance beyond week 13 - Piazza

- Please help each other (you can learn a lot from addressing questions from your fellow students)!
- Monitored sporatically by TAs/staff.
- Staff and TAs will not address/monitor new posts on Piazza after 16:00 on 19th May 2023.

Thank you!

- And best of luck with the exam!