we also have BB03B room 26.

02450: Introduction to Machine Learning and Data Mining

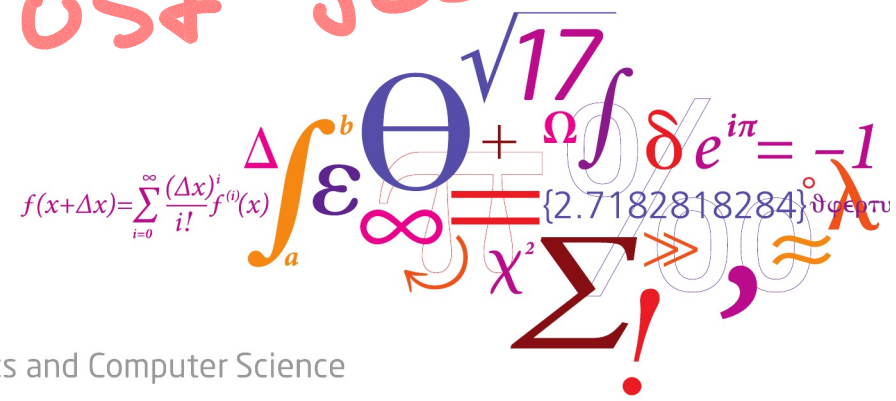
Introduction

tuhe@dtu.dk & 44+48

Tue Herlau

DTU Compute, Technical University of Denmark (DTU)

Please adjust sound volume!

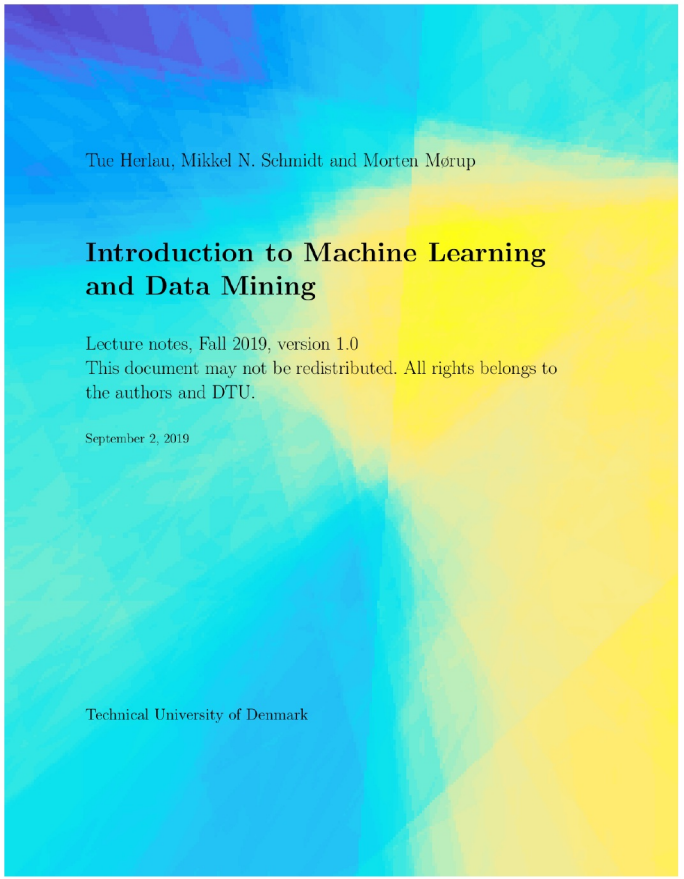


DTU Compute
Department of Applied Mathematics and Computer Science

Reading Material

Reading material:

Chapter 1



Lecture Schedule

1. **Introduction**
   3 September: C1

Data: Feature extraction, and visualization

2. Data, feature extraction and PCA
   10 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   17 September: C4, C5
4. Probability densities and data Visualization
   24 September: C6, C7

Supervised learning: Classification and regression

5. Decision trees and linear regression
   1 October: C8, C9 **(Project 1 due before 13:00)**
6. Overfitting, cross-validation and Nearest Neighbor (Note: Tentative)
   8 October: TBA
7. Bayes, Naive Bayes and performance evaluation (Note: Tentative)
   22 October: TBA
8. Artificial Neural Networks and Bias/Variance
   29 October: C14, C15
9. AUC and ensemble methods
   5 November: C16, C17

Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    12 November: C18 **(Project 2 due before 13:00)**
11. Mixture models and density estimation
    19 November: C19, C20
12. Association mining
    26 November: C21

Recap

13. Recap and discussion of the exam
    3 December: C1-C21 **(Project 3 due before 13:00)**

Piazza online help: `https://piazza.com/dtu.dk/fall2019/02450`
Videos/streaming of lectures: `https://video.dtu.dk`

Course homepage

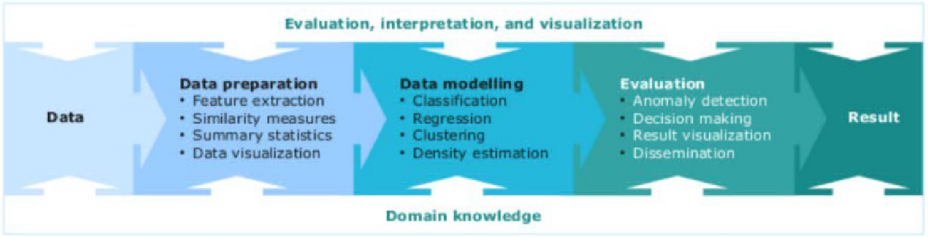
- Go to web page for syllabus and homework problems
  `http://www.compute.dtu.dk/courses/02450`

Section for Cognitive Systems
DTU Compute

02450 Introduction to Machine Learning and Data Mining

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.



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 obejctives), and make them applicable for a broad spectrum of engineering problems in e.g. biomedical engineering, chemistry, electrical engineering, and informatics.

Resources

Location


Morten Mørup

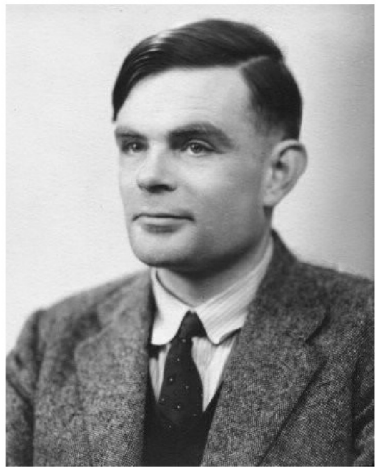

Tue Herlau

Plan for today:

- What is machine learning?
- Why do we learn different methods?
- Impact of machine learning
- This course
- Break
- Lecture 1, basic terminology
- Special announcement about the Neural student organization

Alan Turing (1946)

- Proved universal computation
- 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



Alan Turing
(1912-1954)

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
    - Work out which board positions were good and bad depending on wins/losses

Arthur Samuel
(1901-1990)

Tom Michell (1999)

- ***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."**
  - The checkers example,
    - E = 10000 games
    - T = playing checkers
    - P = if you win or not

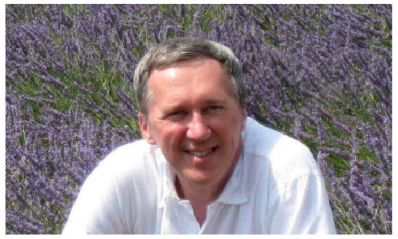
Tom Michell

Quiz 01: Machine learning definition

Please answer this quiz on Piazza:

https://piazza.com/dtu.dk/fall2019/02450

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. Diagnosing images of eyes

D. Physiological information about cataracts (genetic markers, disease progression, etc.)

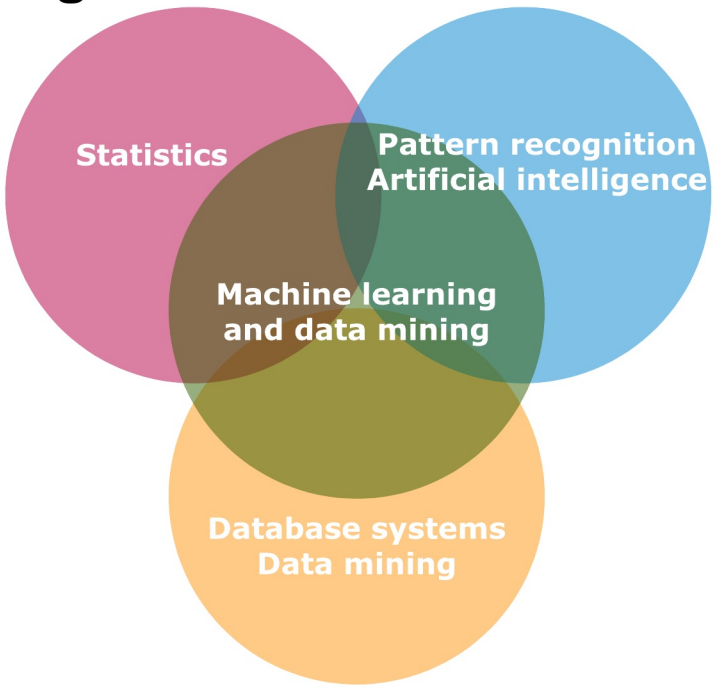
E. Don't know.

image obtained from https://www.pexels.com/photo/beautiful-blue-eyes-close-up-dhyamis-kleber-609549/

Solution:

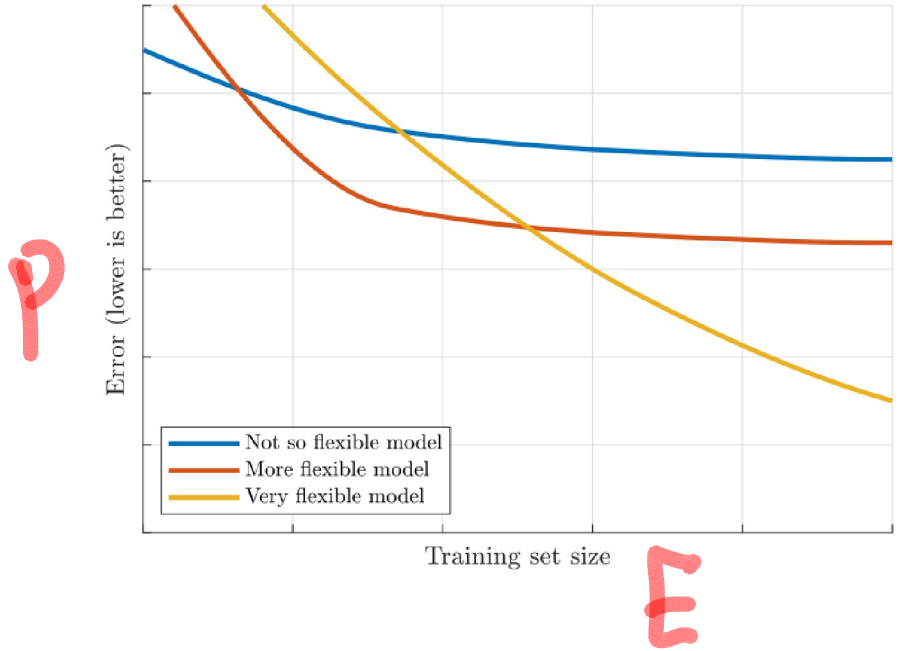
The correct answer is 2, the images of eyes along their diagnosis. In a normal machine-learning setting, the system would learn a mapping from images to disease labels based on a large set of examples of this mapping.

Machine-learning as a research area

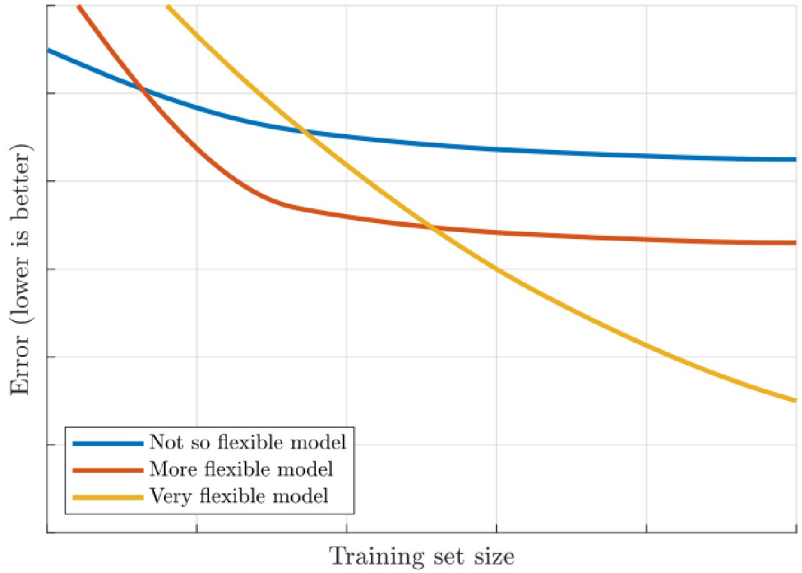


- Focus on a *learning algorithms* (rather than search, pathfinding, etc.)
- De-emphasis of explicit knowledge representations, etc.
- Gradual improvements (training time, amount of data)
- *General* algorithms (or algorithmic ideas)

Machine-learning as a research area



Machine-learning as a research area



models.

K-nearest neighbour methods
(Chapter 10)

Random forrest
(Chapter 15)

Linear/logistic regression
(Chapter 7)

Decicion tree
(Chapter 8)

Small neural network
(Chapter 13)

Large neural network

Less complex

More complex

Regularization
(Chapter 12)

Bagging/boosting
(Chapter 15)

Can machines really compete against humans?

- Humans have wisdom
- Humans have intuition
- Humans are moral

Image source: https://pixabay.com/da/meditation-\%C3\%A5ndelig-yoga-meditere-1384758/

Human abilities examined

source: https://www.boredpanda.com/dangerous-distracted-driving/?utm_source=google&utm_medium=organic&utm_campaign=organic

Man vs. Machine

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, Dota2 1v1: OpenAI deep reinforcement learning bot beats top professionals in 1v1 DOTA

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

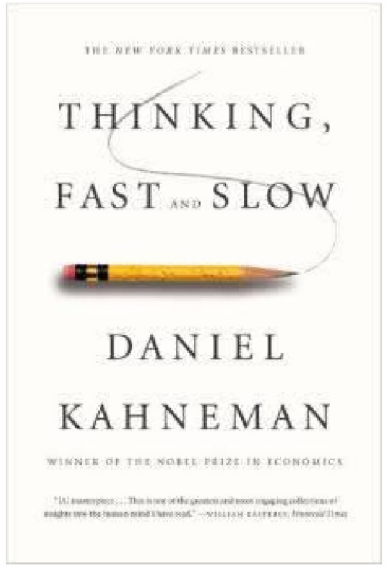
2016, Geoguessing Google PlaNet win 28 of 50 rounds; median localization error of 1132km vs. 2321km

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 stayes, 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 miscallaneous 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

Emperical Increased availability of large/good datasets

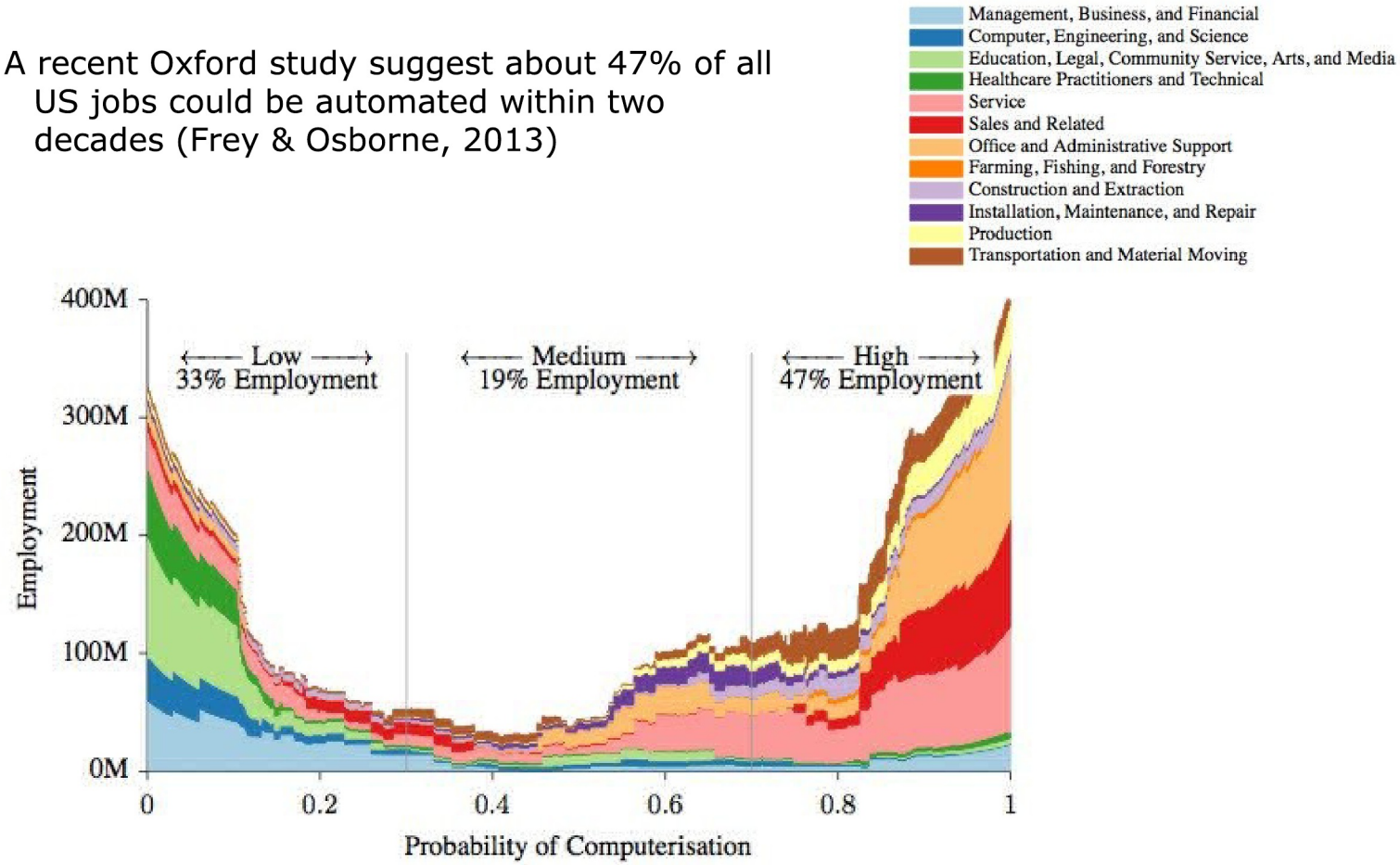
Technological Faster computers

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)



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

ML

Machine learning can, in principle, learn 3

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

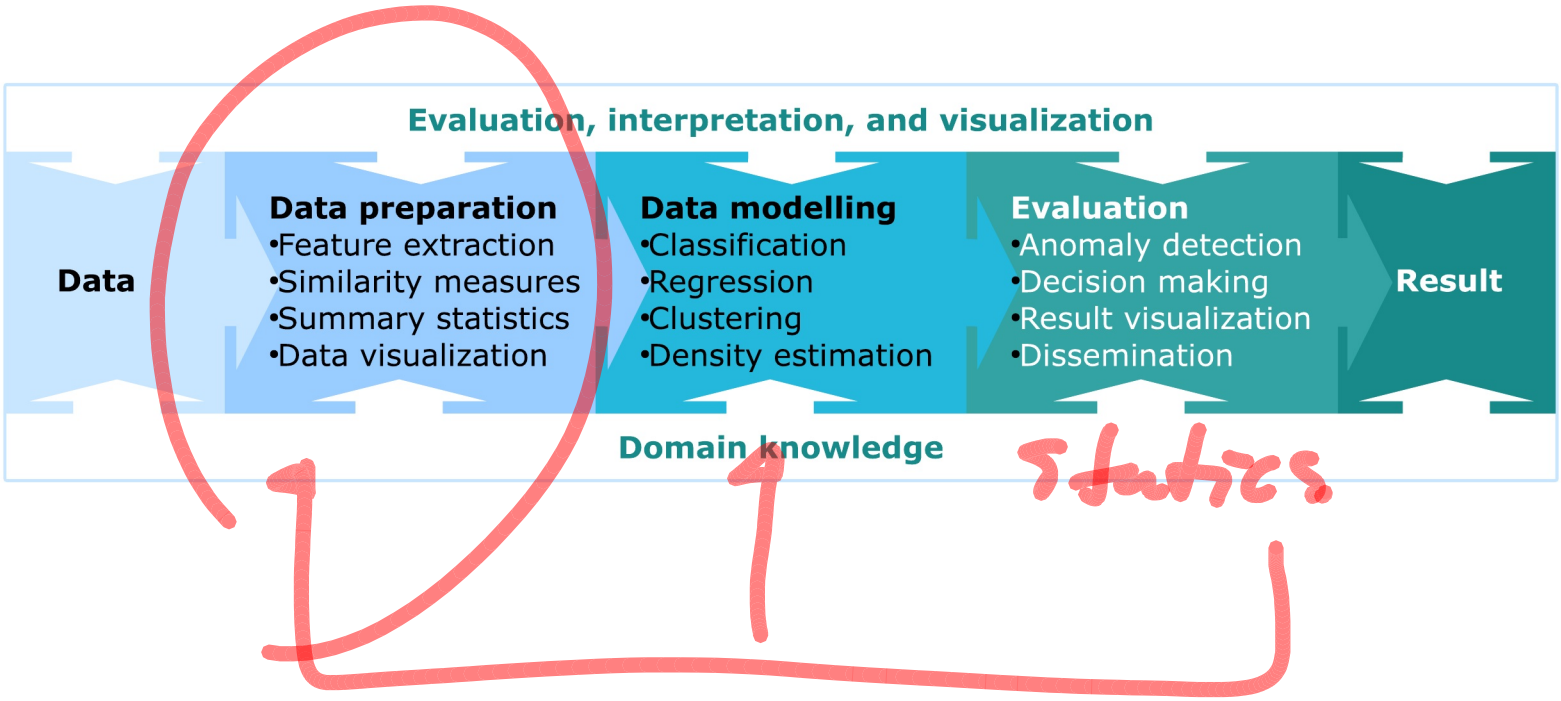
Elizabeth I: "Thou aimest high, Master Lee. Consider thou what the invention could do to my poor subjects. It would assuredly bring to them ruin by depriving them of employment, thus making them beggars."

> *"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

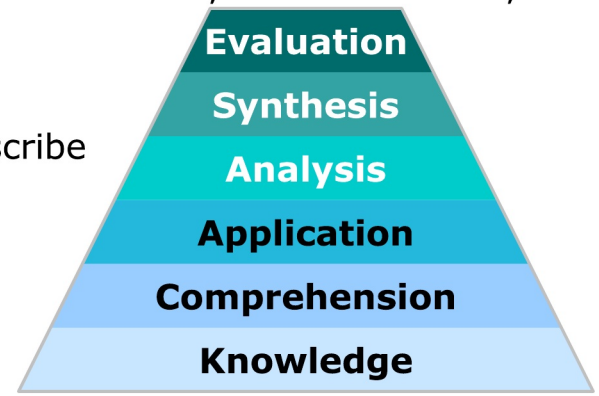
- We don't know what will happen
- Plausibly, many tasks will become so cheap humans will no longer perform them. Macroeconomics suggests:
  - Essential, non-automated tasks will both become more valuable, inhibit progress
  - Humans will do fewer jobs, play a relatively smaller role in the economy (the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

Machine learning and data mining pipeline of this course



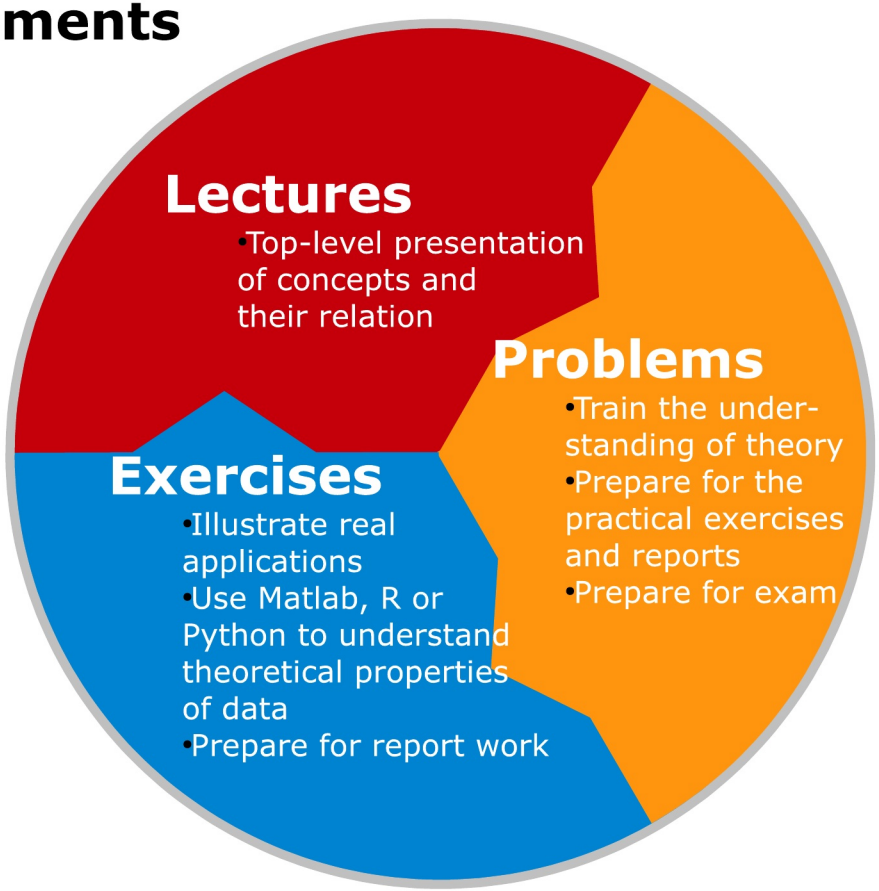
Learning objectives

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)

Course elements



Group Learning

- You should form groups of 2-3 students (max 3 students for the reports)
  - **During class** You are encouraged to sit together and discuss the online Piazza questions
  - **During exercises** Each exercise consist of a conceptual multiple-choice question from a previous exam as well as computer exercises relevant for the report work. Please only spend about 15 minutes on the multiple-choice question.
- If you want to work alone for the reports ask me for permission

+ piazza

Piazza online help

- `https://piazza.com/dtu.dk/fall2019/02450`
- Use 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?
- I mark questions as answered to indicate there is a response in the UI; please post follow-ups!

Examination

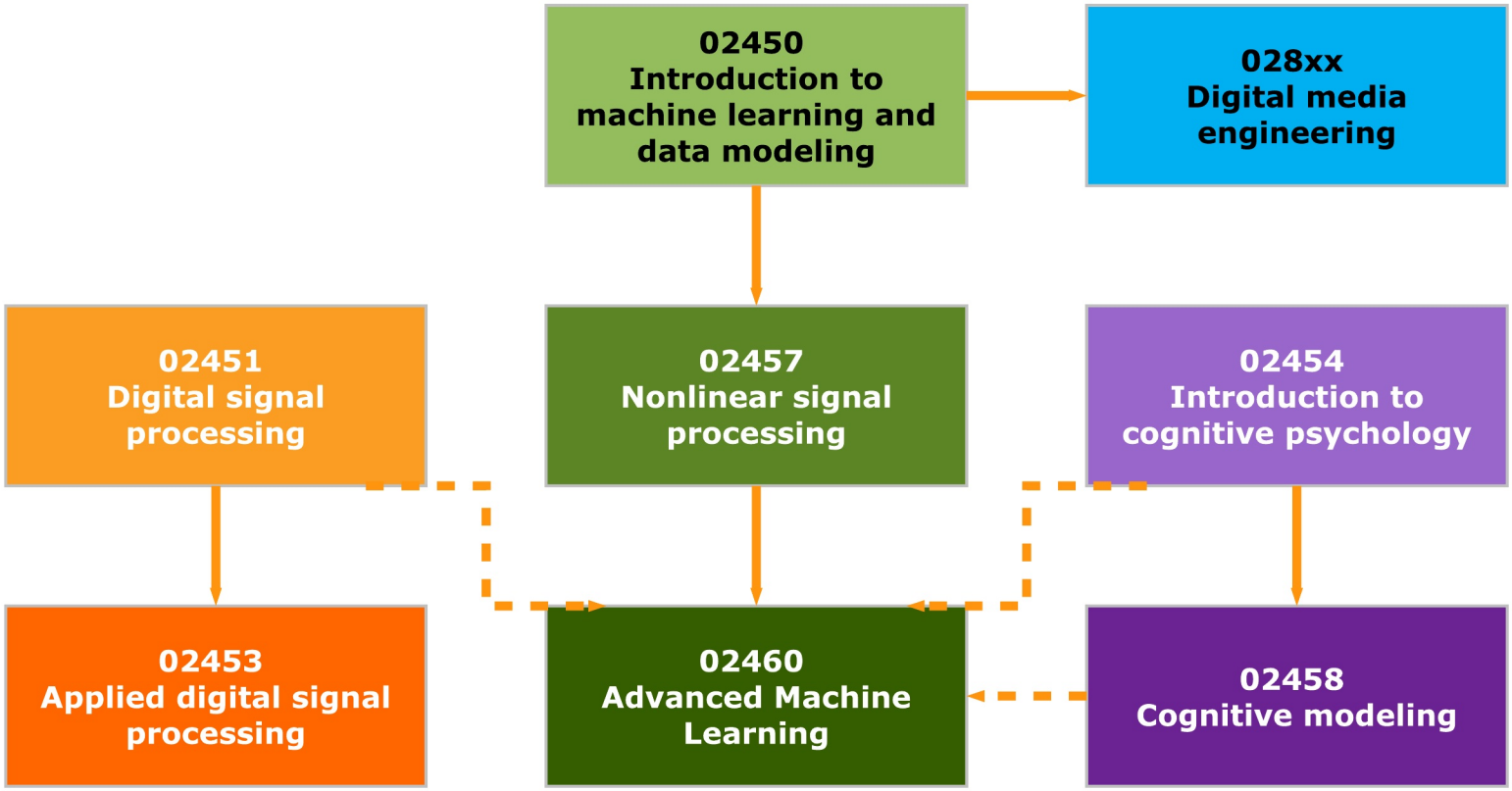
To test all learning objectives the exam is split into two parts:

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

  Report 1 October 1, *Data: Feature extraction, and visualization*
  Report 2 November 12, *Supervised learning: Classification and regression*
  Report 3 December 3, *Unsupervised learning: Clustering and density estimation*

- Final grade based on overall assessment of reports and written exam. The written exam is weighted more than the reports

Relevant courses at Section for Cognitive Systems



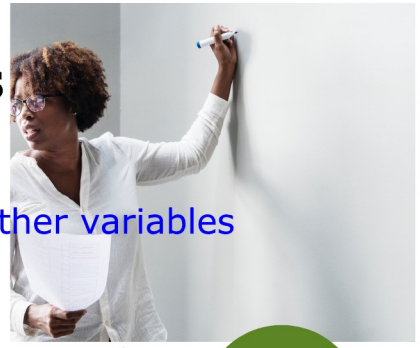
http://www.compute.dtu.dk/english/research/Cogsys/Courses
http://www.dtu.dk/english/education/msc/programmes/
→ Mathematical modelling and computation
→ Machine learning and signal processing

Pretest

The purpose of the pretest is to assess the students background and academic level in order to allow the teachers to adjust the presentation of the course material as well as measure student learning. This pretest will not be graded and will not influence exam results in any way. We do not expect you to be able to answer all questions.

campusnet.
not graded.
lecture part 2: 14:05

Data Mining and Machine Learning Tasks



Predictive tasks (Supervised learning)

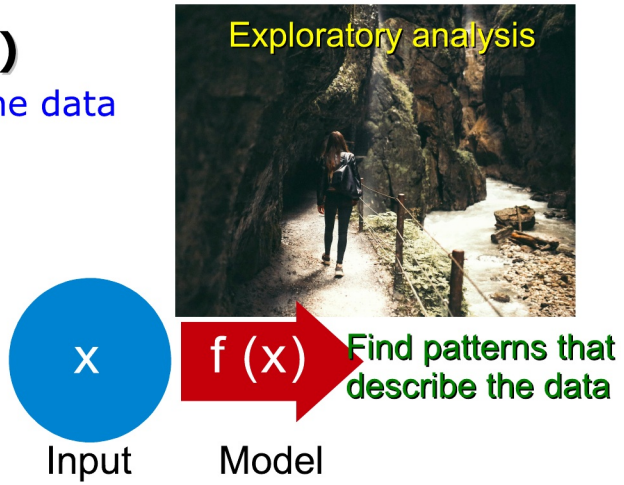
– Use some variables to predict unknown or future values of other variables

- **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

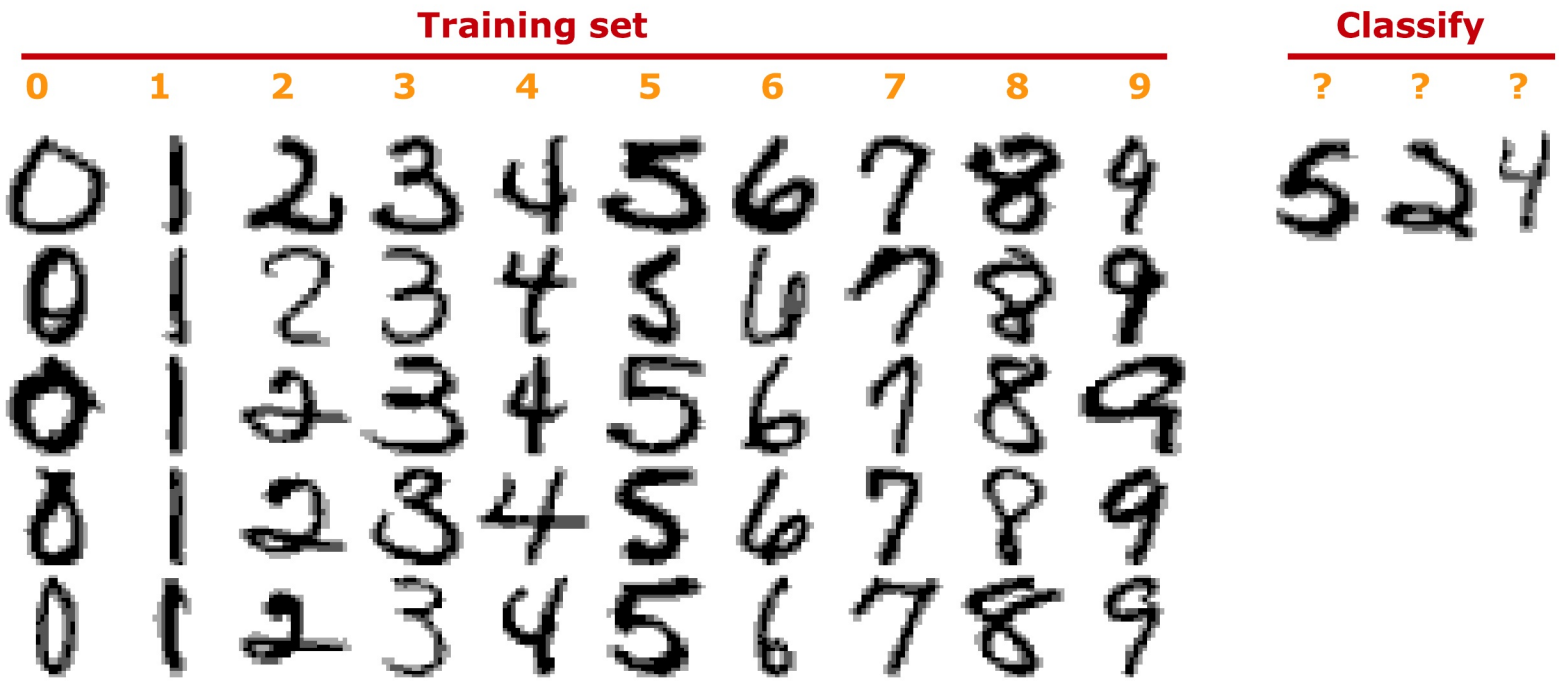
- **Clustering**
  – Discover group structure in data
- **Association rule discovery**
  – Discover how data objects relate to each other
- **Anomaly detection**
  – Find data objects that are abnormal



Classification: Definition

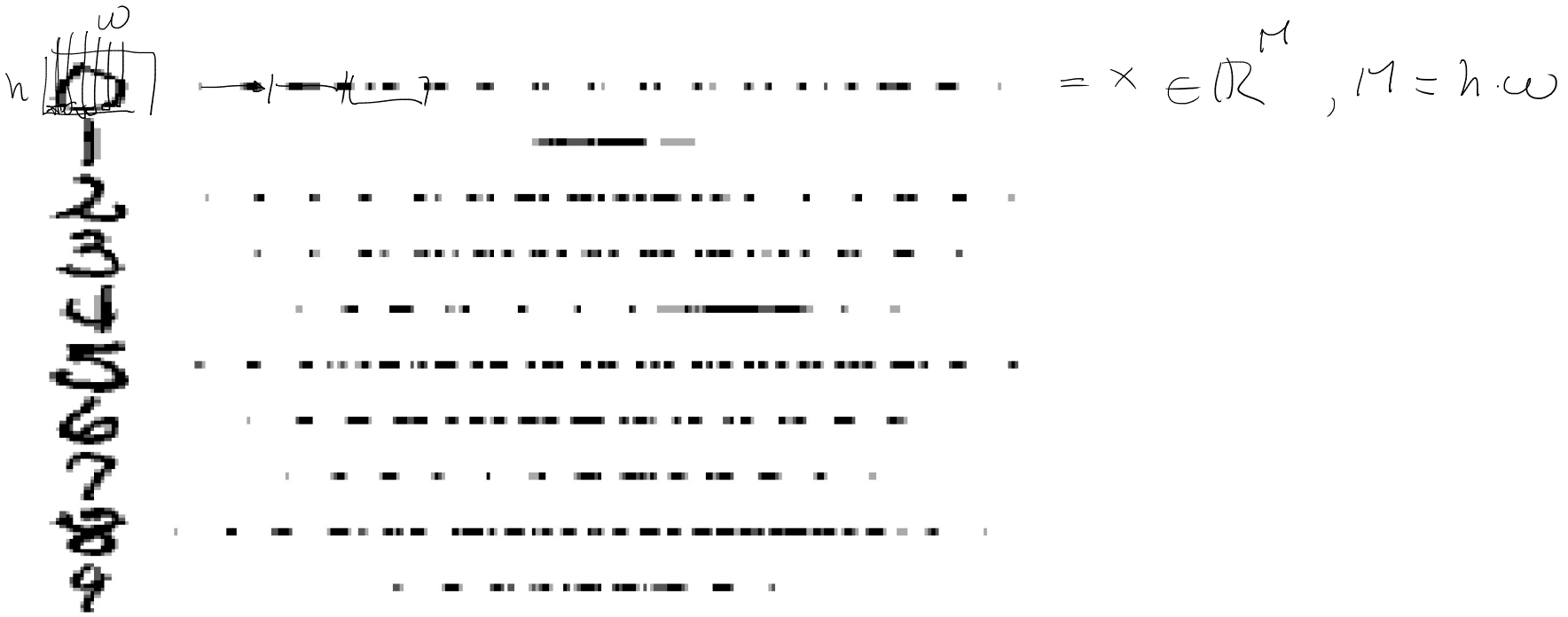
- Given a collection of data objects (**training set**)
  - Each object has associated a number of features X
  - Each object belongs to a certain class y
- Define a **model** for the class given the other features $f(x) \to y$
- Goal: Assign a class label to a **previously unseen object**

Classification: Example



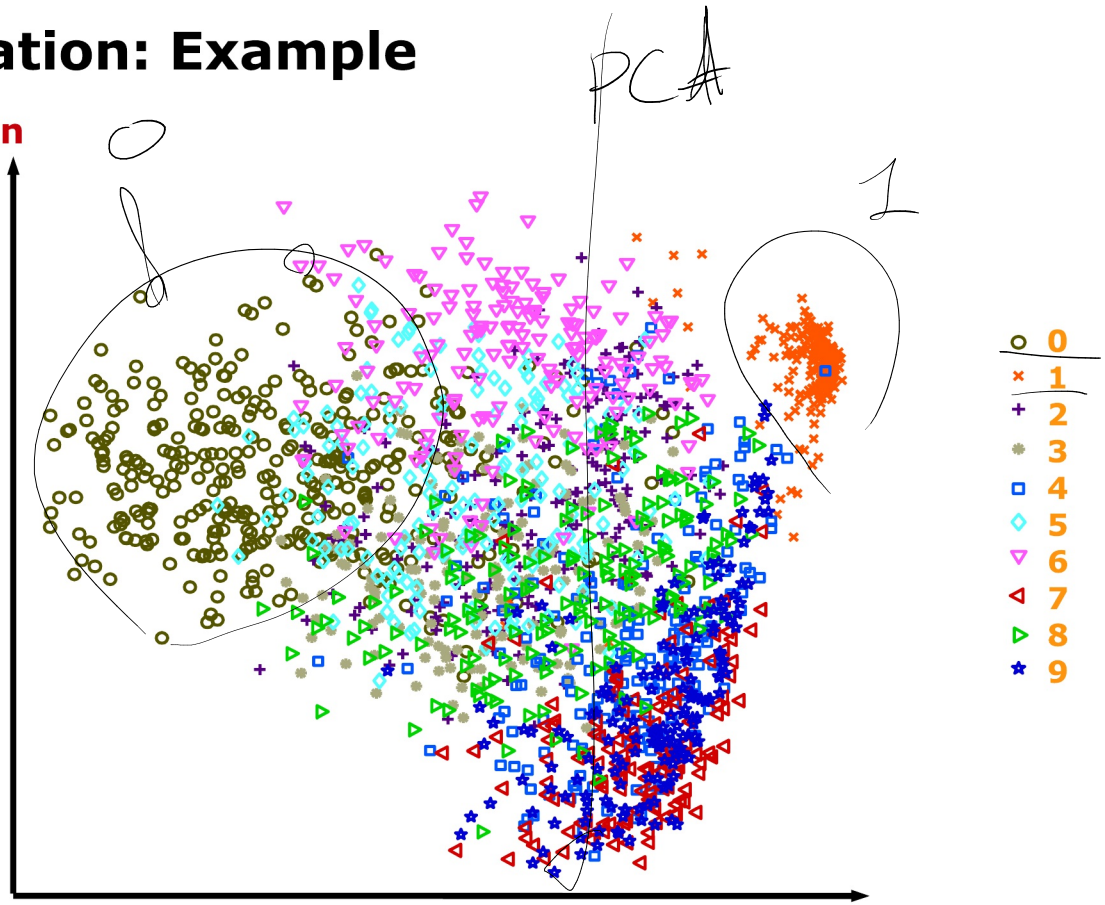
Classification: Example

Data representation

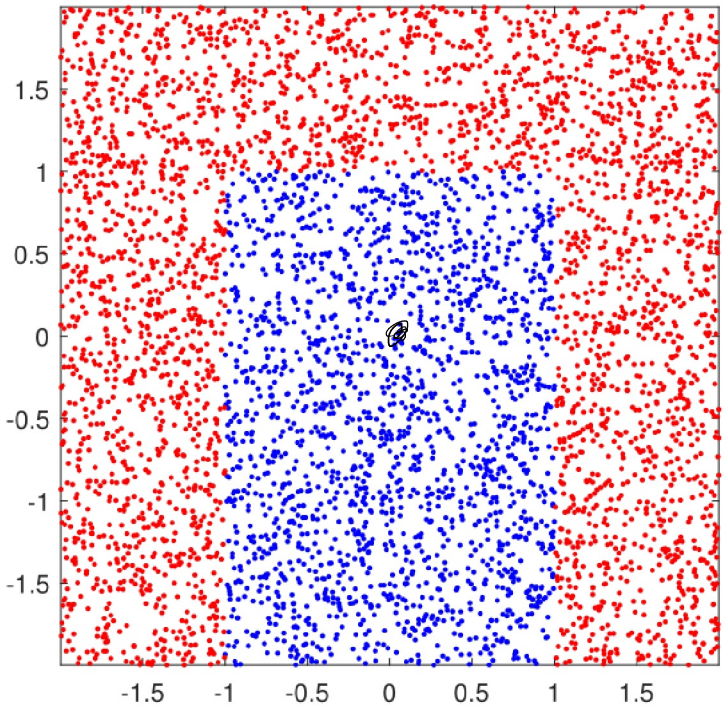


Classification: Example

Visualization



Quiz 02 (Piazza): Decision rules



The figures 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:

$$z_1 = \max\{0, x - 1\} + \max\{0, -1 - x\}$$
$$z_2 = \max\{0, -1 + y\}$$

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

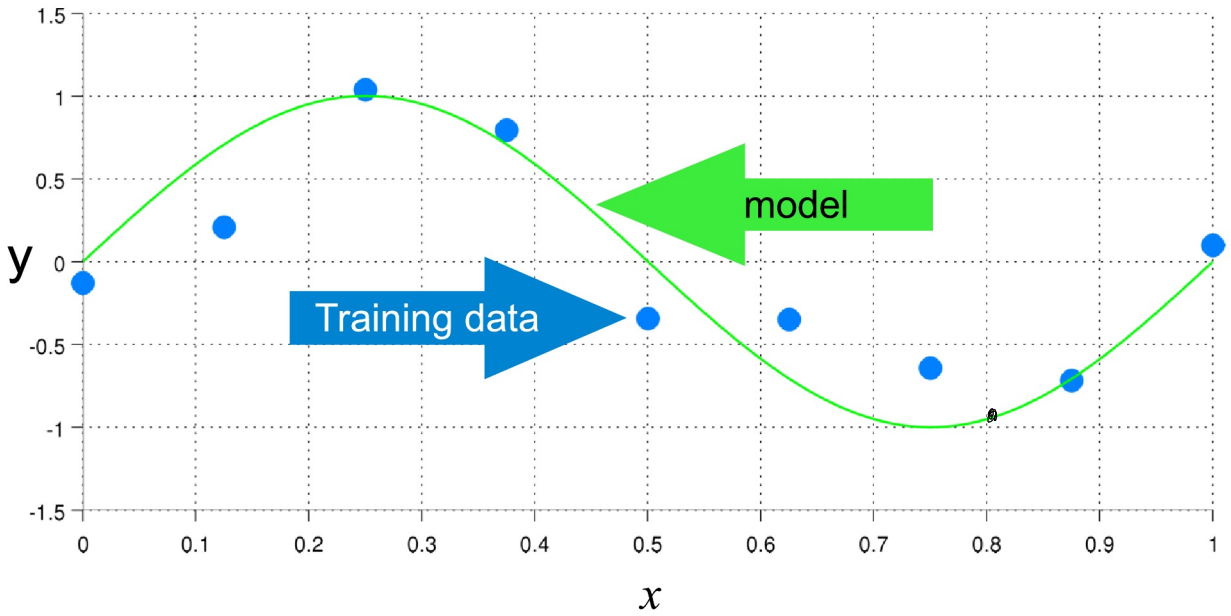
Solution:

The first option will give the correct boundary. To see this, check out $(x, y) = 0$. In this case $z_1 = 0$ and $z_2 = 0$. Therefore, only the first option gives the correct answer.

Actually, this decision rule is a (slightly obscured) example of a neural network with RELU units. We will learn more about neural networks later in this course.

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

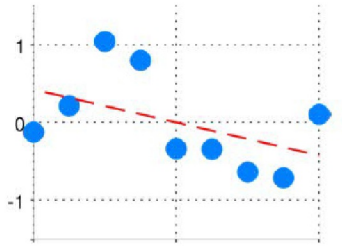
- Predict **sales amounts** of new product based on
  - advertising expenditure
- Predict **wind velocity** as a function of
  - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
  - previous index time series and market indicators

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

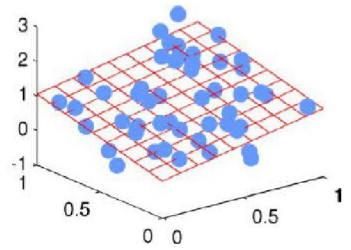
1-dimensional inputs

$$f(x) = w_0 + w_1 x$$

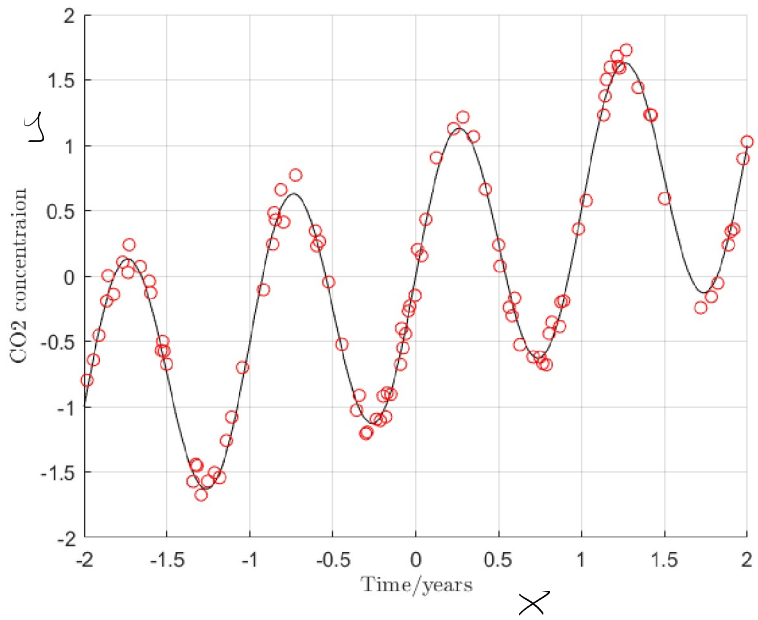


2-dimensional inputs

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2$$



Quiz 03 (Piazza): Regression



The figures shows an example regression problem where the CO2 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 CO2 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. $y = -0.5x + \sin(2\pi x)$~~

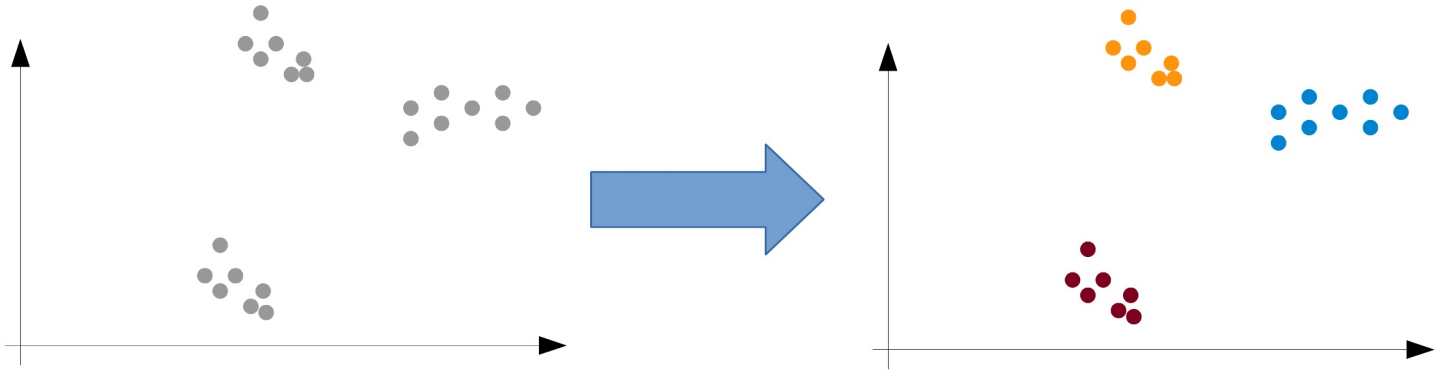
Solution:

The right answer is $y = 0.5x + \sin(2\pi x)$. The functions involving cosines can be ruled out since they have the wrong period; meanwhile the function has an upwards trend which means it must involve the term $0.5x$.

Clustering: Definition

- Given a collection of data objects X
  - 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

$x = (x_1, x_2)$



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 occurence of an item based on occurences of other items

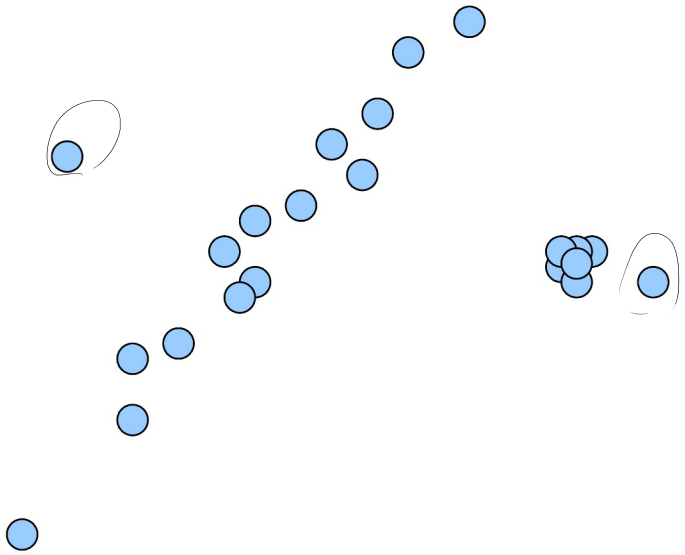
Association rule discovery: Example

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

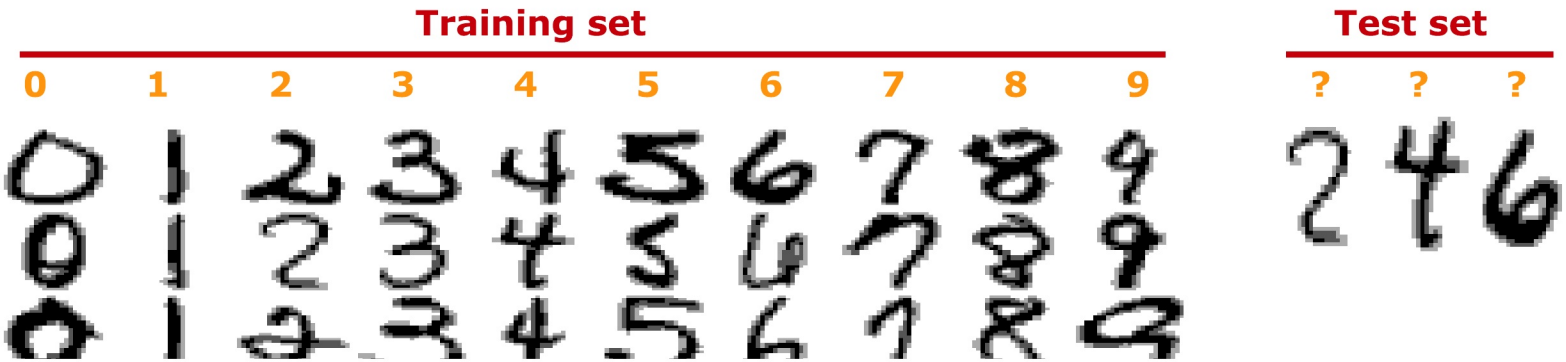
- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour



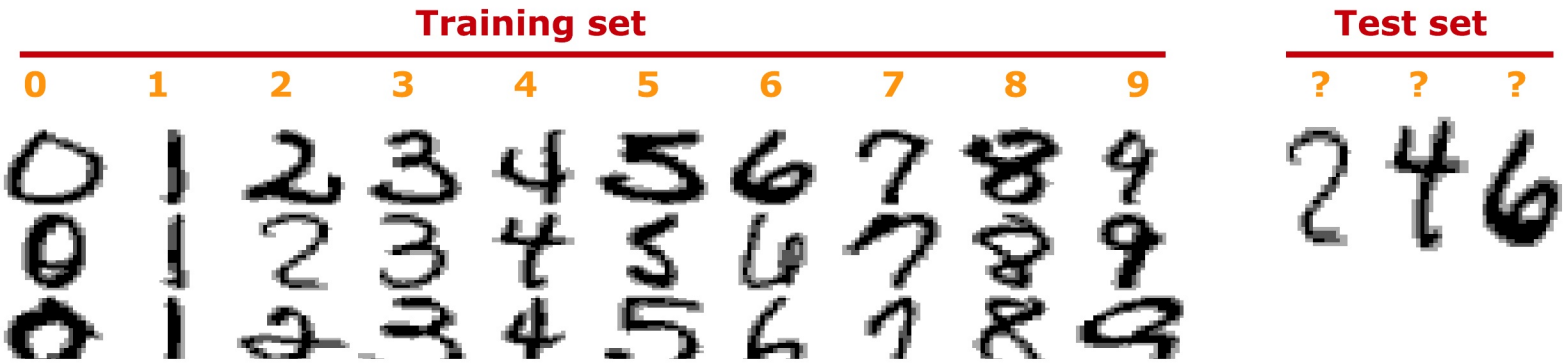
Anomaly detection: Example

- 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

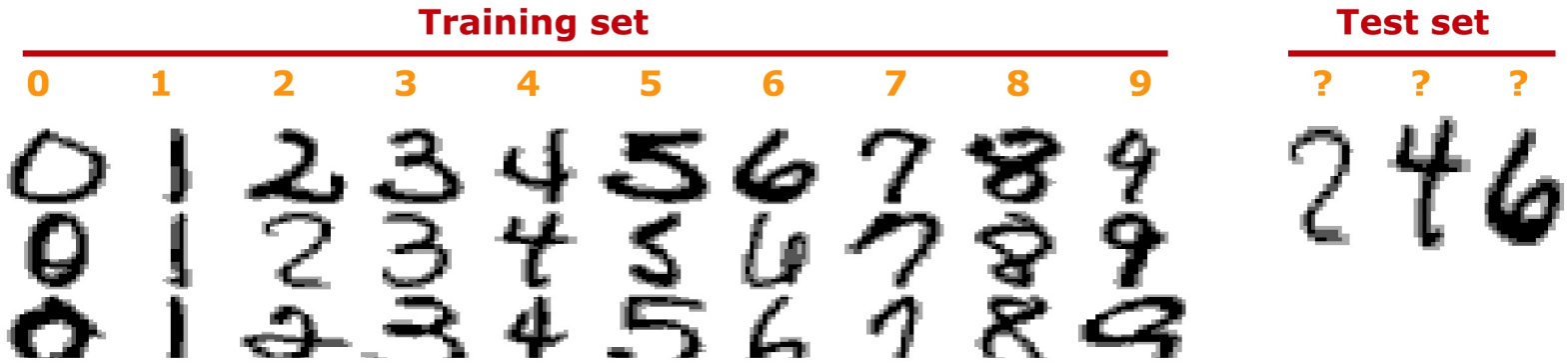


Models in machine learning

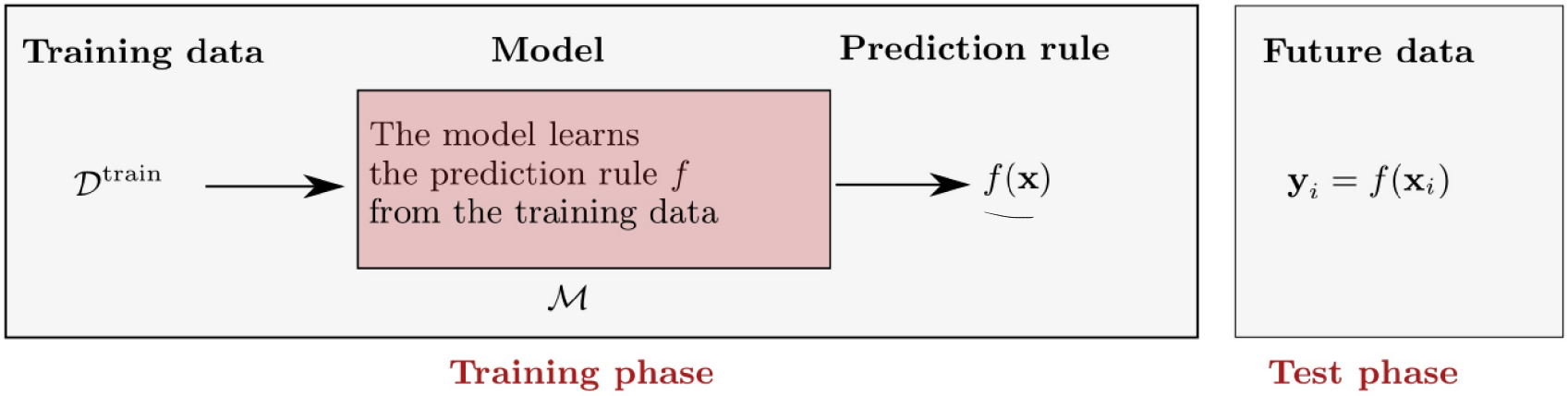


Classifying digits is a mapping $f : \mathbb{R}^M \rightarrow \{0, 1, \ldots, 9\}$

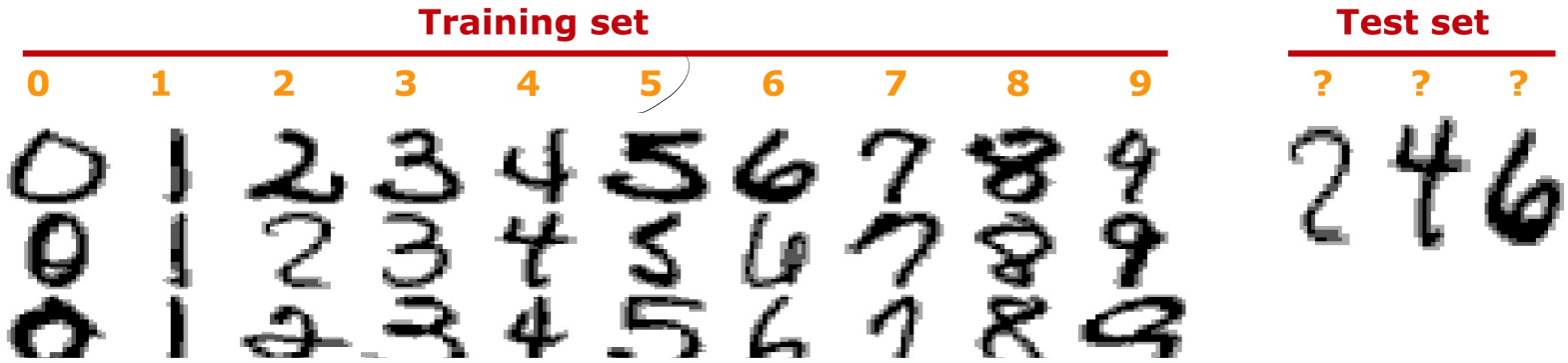
Models in machine learning



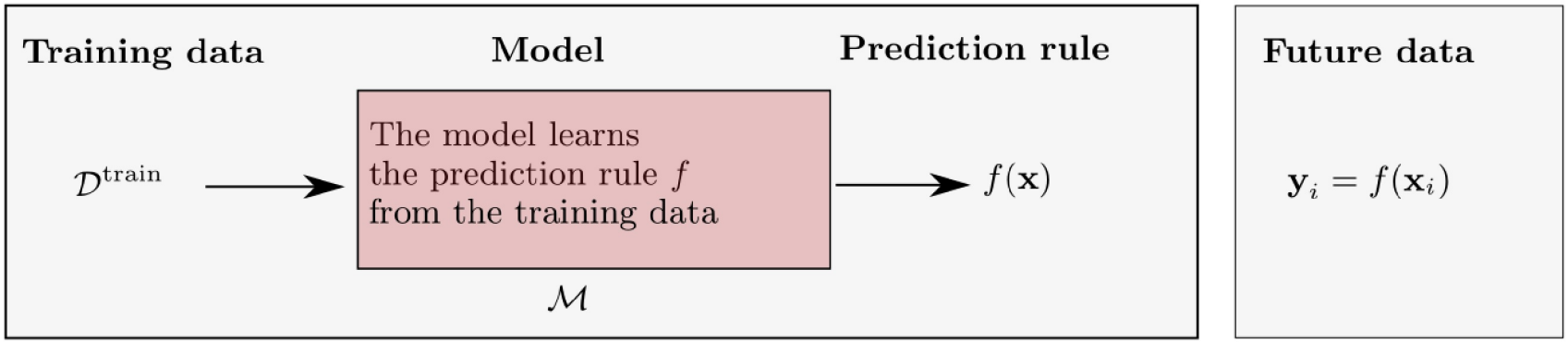
Classifying digits is a mapping $f : \mathbb{R}^M \rightarrow \{0, 1, \ldots, 9\}$



Models in machine learning



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



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

Exercises

We support Matlab, Python, and R

- Exercises today will guide you through installing your chosen environment
- If you have no experience with either, we recommend Matlab (to obtain Matlab see: `http://www.gbar.dtu.dk/faq/28-matlab`)

Rooms for exercises:

- building 303A, auditorium 42, (Python)
- building 303A, auditorium 41, (Python)
- building 308, student area on 1st floor (north), (Python)
- building 308, student area on 1st floor (south), (Python+Matlab)
- building 308, room 101, (Python)
- building 308, room 109, (Python+Matlab)
- building 308, room 117, (R)
- building 308, room 127, (Matlab)
- 303A/databar 048 (40), (Python)
- 303A/holdområde OEST (96), (Python)
- 303A/aud. 44 (96), (Python)

Exercises Today

- Follow exercise instructions available on campusnet
- Form groups (max 3 students) and select a dataset for the reports
- Please have your dataset approved and your group registered (talk with your instructor)

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/)