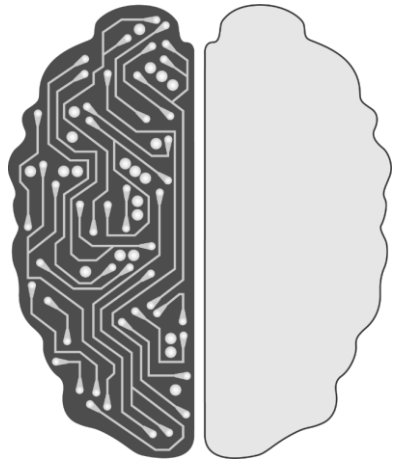
Introduction

Prof. Dr. Christina Bauer

christina.bauer@th-deg.de

Faculty of Computer Science



What do You Think? What do You know?

What is Machine Learning?

Machine Learning

- Machine learning is a science of getting computers to learn, without being explicitly programmed

- Spam filters
- Robots, e. g. for cleaning, health care, etc.
- Search algorithms
- Recommender systems
- Face recognition on photos
- …

Machine Learning

- AI dream: building machines as intelligent as a human
- One approach: learning algorithms that try to mimic how the human brain learns

- Grow out of work in AI
- Some more examples:
  - Database mining: Large datasets from the web/automation
    - Web click data, medical records, biology, engineering
  - Applications that cannot be programmed directly
    - Autonomous helicopters, handwriting recognition, Natural Language Processing (NLP), computer vision
  - Self-customizing programs
    - Amazon, Netflix product recommendations
  - Understanding human learning

Example Videos

- Cleaning Robot: https://youtu.be/XIPzSmwClJ8
- Some more examples: https://emerj.com/ai-sector-overviews/artificial-intelligence-home-robots-current-future-use-cases/

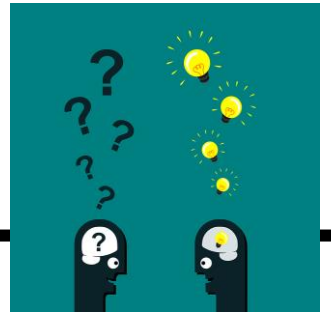
Definition Machine learning

- Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)
- Well-posed learning problem (Tom Mitchell, 1998):
  - **A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves the experience E.**

Checkers example:

- experience E: experience of having the program play tens of thousands of games
- task T: task of playing checkers, and
- performance measure P: the probability of winning the next game

QUESTION



A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves the experience E.

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam.

What is the task T?

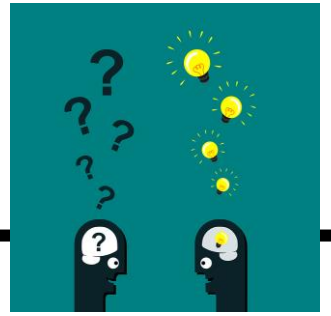
A: Classifying emails as spam or not spam.

B: Watching you label emails as spam or not spam.

C: The number (or fraction) of emails correctly classified as spam/not spam.

D: None of the above – this is not a machine learning problem.

QUESTION



A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves the experience E.

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam.

What is the task T?

A: Classifying emails as spam or not spam. - T

B: Watching you label emails as spam or not spam. - E

C: The number (or fraction) of emails correctly classified as spam/not spam. - P

D: None of the above – this is not a machine learning problem.

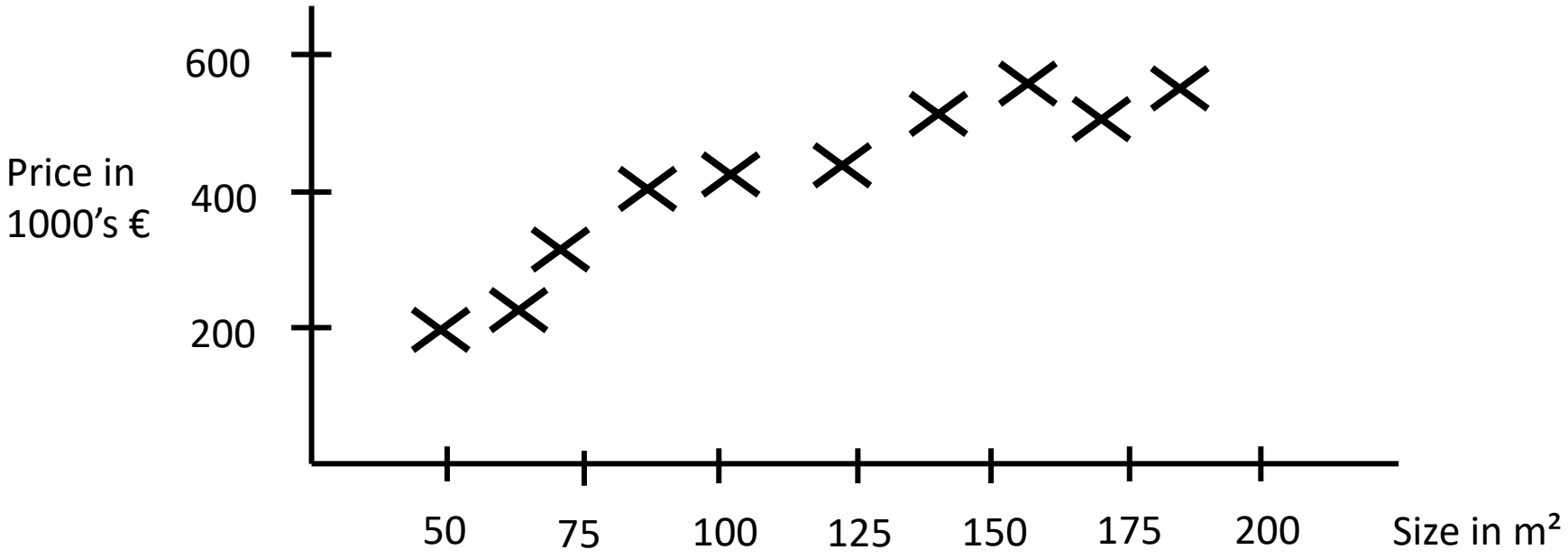
Machine Learning

- Machine Learning algorithms
  - Supervised learning
  - Unsupervised learning
- + reinforcement learning, recommender systems

- Practical advice for applying learning algorithms
- Talk about “how” but also “when” and “why”

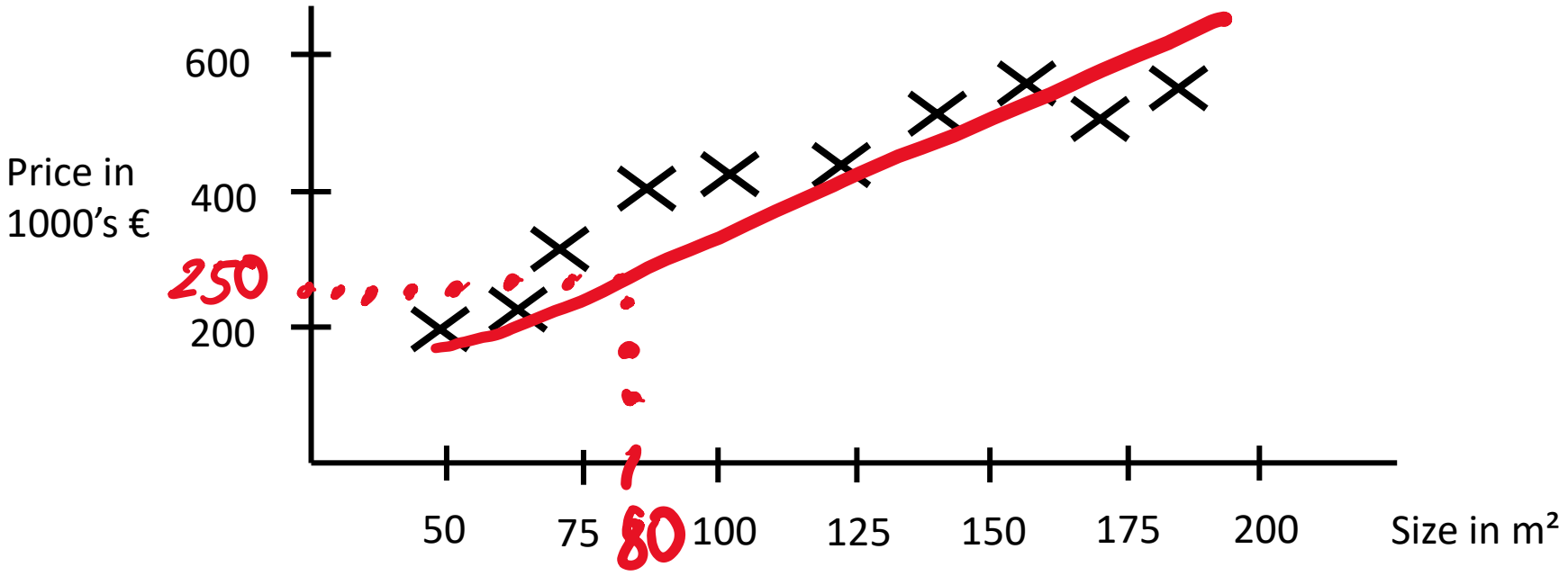
SUPERVISED LEARNING

House price prediction



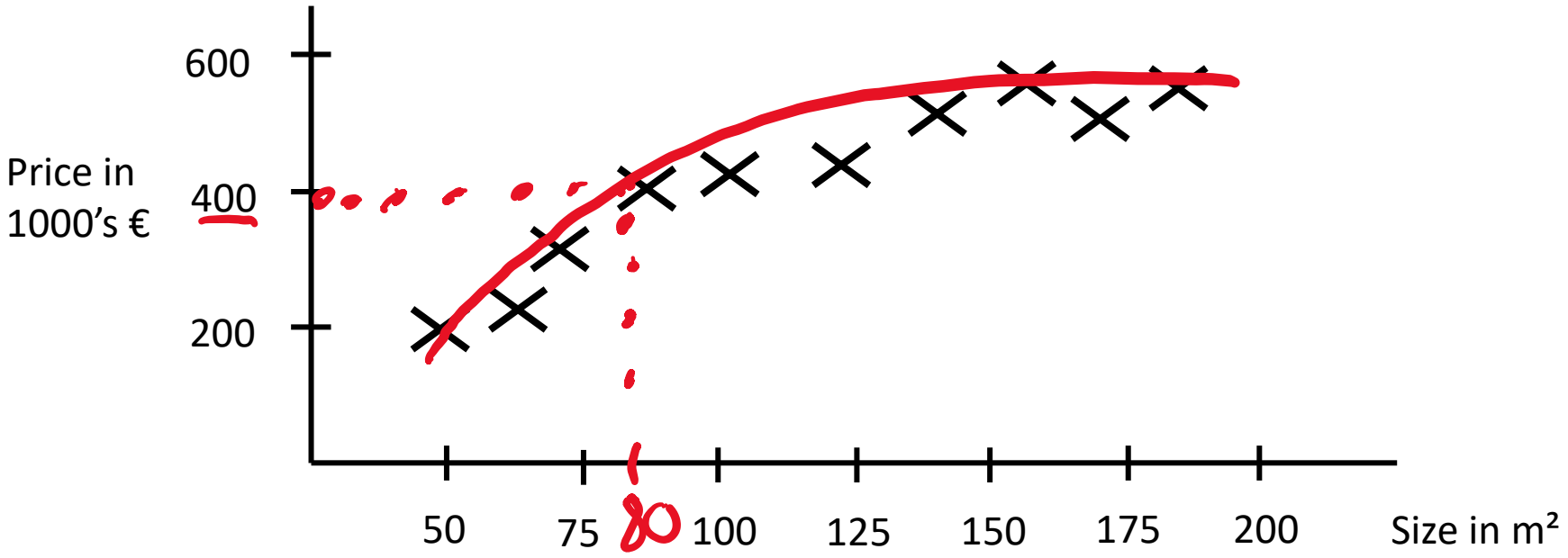
SUPERVISED LEARNING – LINEAR

House price prediction

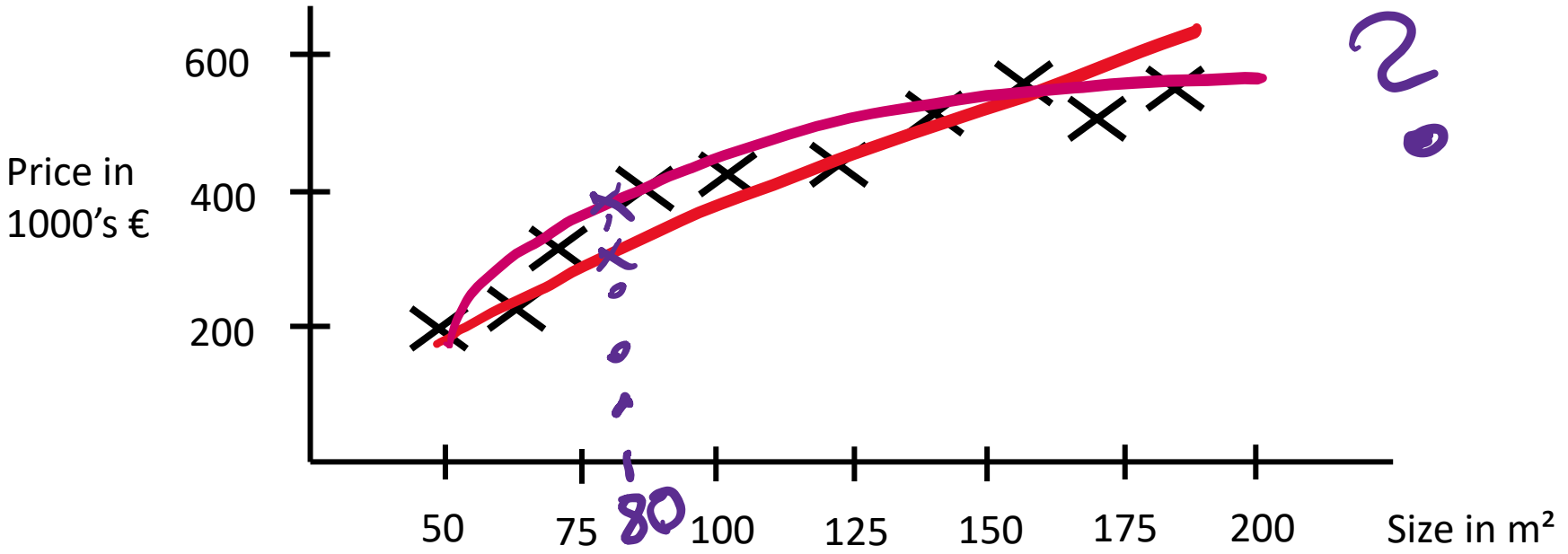


SUPERVISED LEARNING – SECOND-ORDER POLYNOMIAL

House price prediction



Supervised learning - House price prediction



Supervised learning

"Right answers" are given

→ Data set of prices, e. g. 50/200

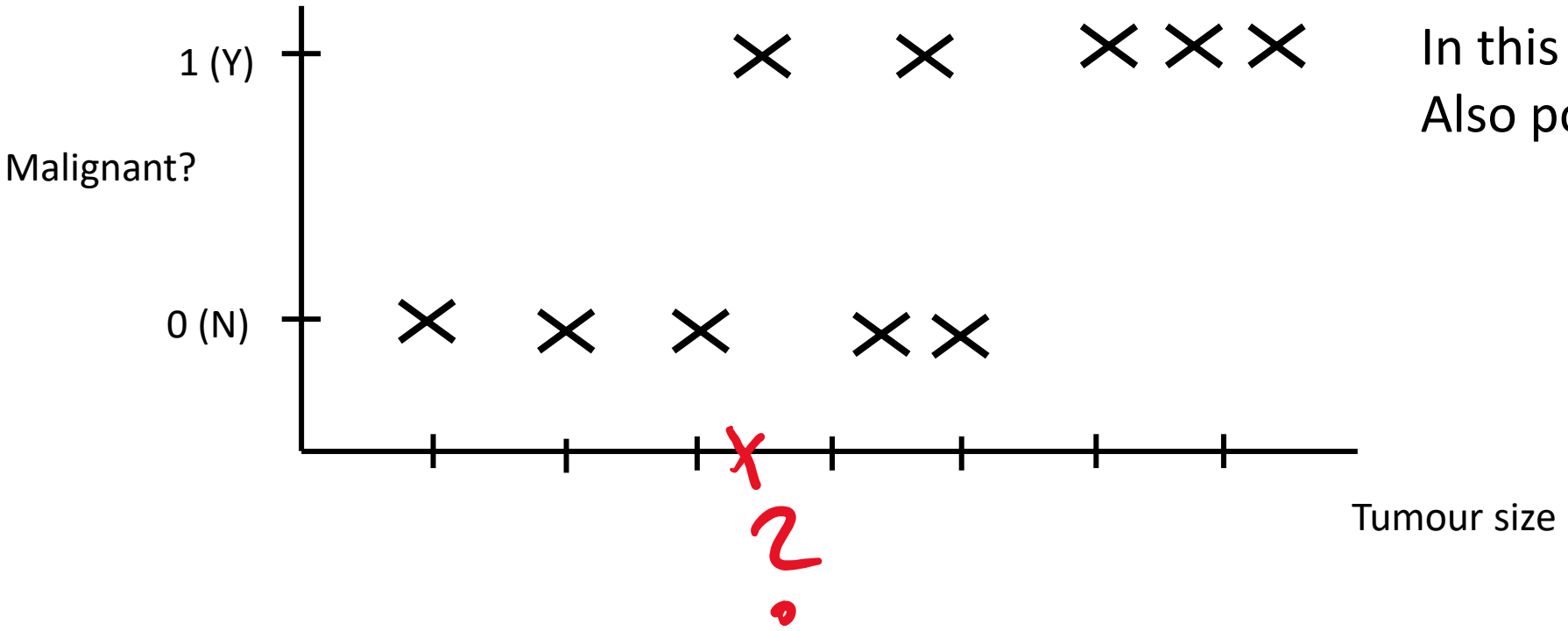
→ Algorithm produces more of the right answers

Regression problem

Predict continuous value output

SUPERVISED LEARNING - CLASSIFICATION

Cancer



Classification problem

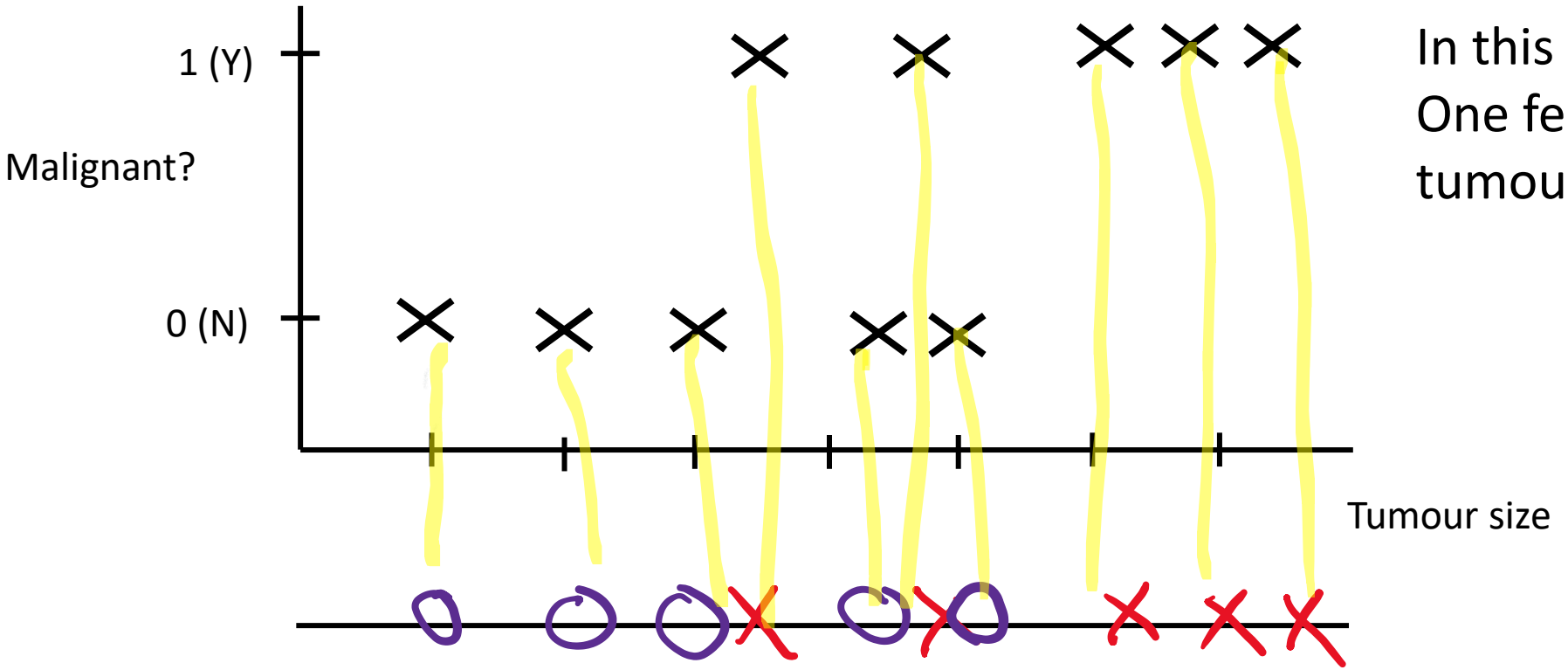
Discrete valued output

In this example: 0 or 1

Also possible: 0,1,2,3 etc.

SUPERVISED LEARNING - CLASSIFICATION

Cancer



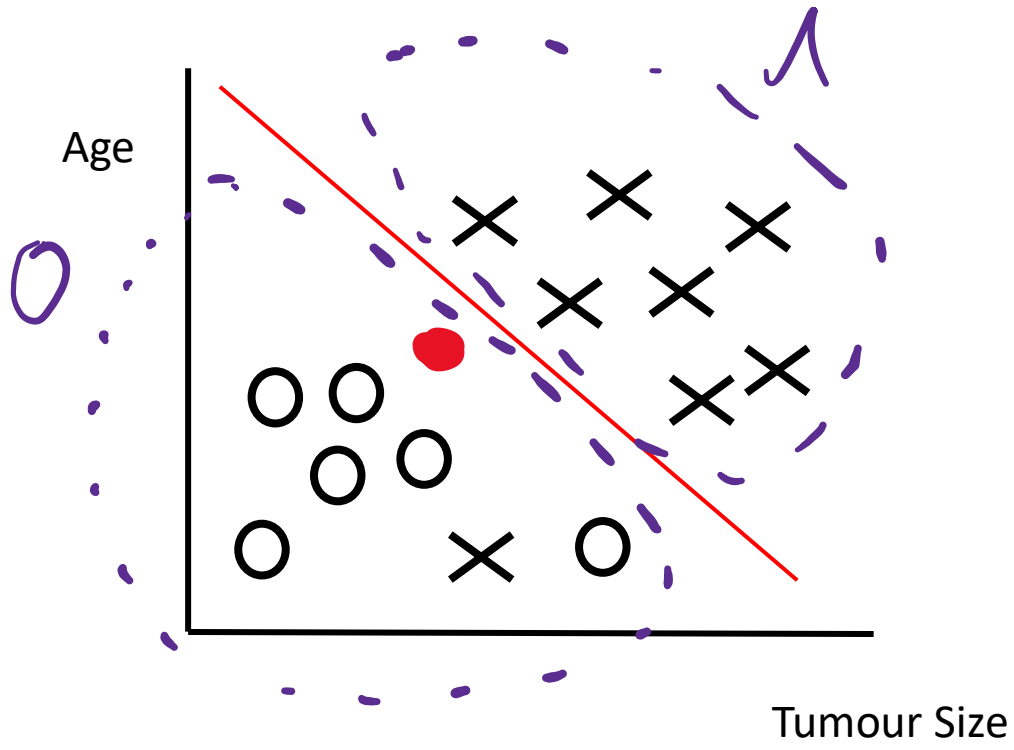
Classification problem

Discrete valued output

In this example: 0 or 1

One feature = attribute: tumour size

SUPERVISED LEARNING - CLASSIFICATION



Classification problem

Discrete valued output

In this example: 0 or 1

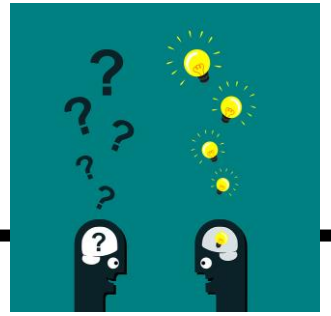
Two features:

age and tumour size

Other possible features:

- Clump thickness
- Uniformity of cell size
- Uniformity of cell shape
- Etc.
- For a lot of problems you want to deal with an infinite number of features → Support Vector Machine

QUESTION



You are running a company and you want to develop learning algorithms to address the following problems:

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will be sold over the next 3 month.

Problem 2: You would like software to examine individual costumer accounts, and for each account decide if it has been hacked.

Should you treat these as classification or as regression problems?

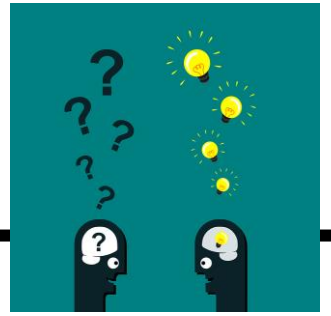
A: Treat both as classification problems.

B: Problem 1: classification 2: regression

C: Problem 1: regression 2: classification

D: Treat both as regression problems.

QUESTION



You are running a company and you want to develop learning algorithms to address the following problems:

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will be sold over the next 3 month.

Problem 2: You would like software to examine individual costumer accounts, and for each account decide if it has been hacked.

Should you treat these as classification or as regression problems?

A: Treat both as classification problems.

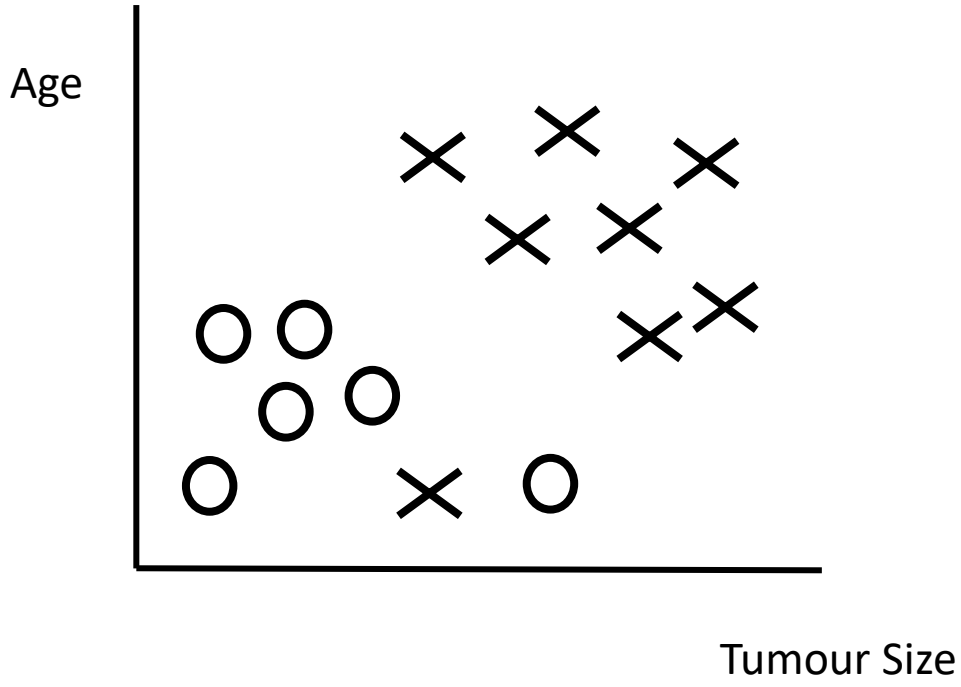
B: Problem 1: classification 2: regression

C: **Problem 1: regression (e. g. sold items: 100000)**
2: classification (hacked: 1 not hacked 0)

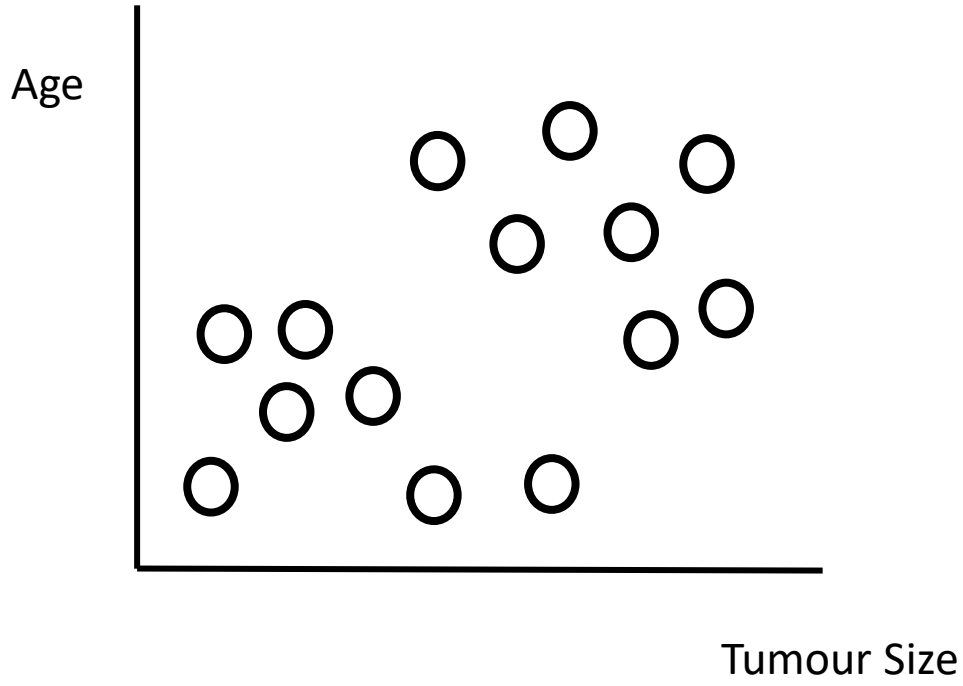
D: Treat both as regression problems.

Unsupervised Learning

supervised

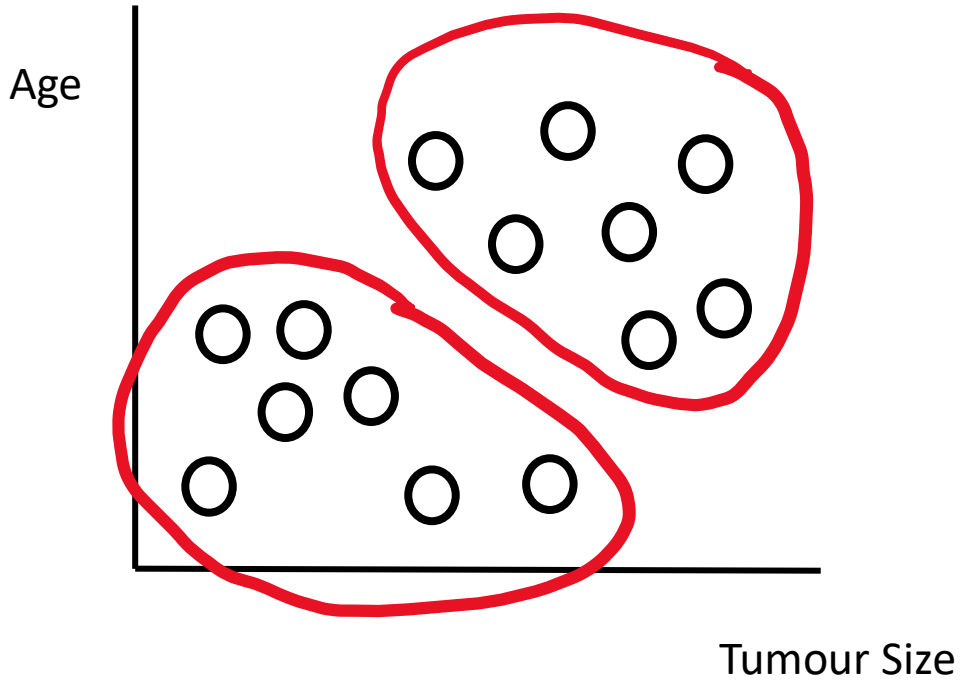


unsupervised



Unsupervised Learning

Clustering algorithm



Unsupervised Learning

Hotel quarantine: 'It'll cost us thousands and we'll be miles from home'

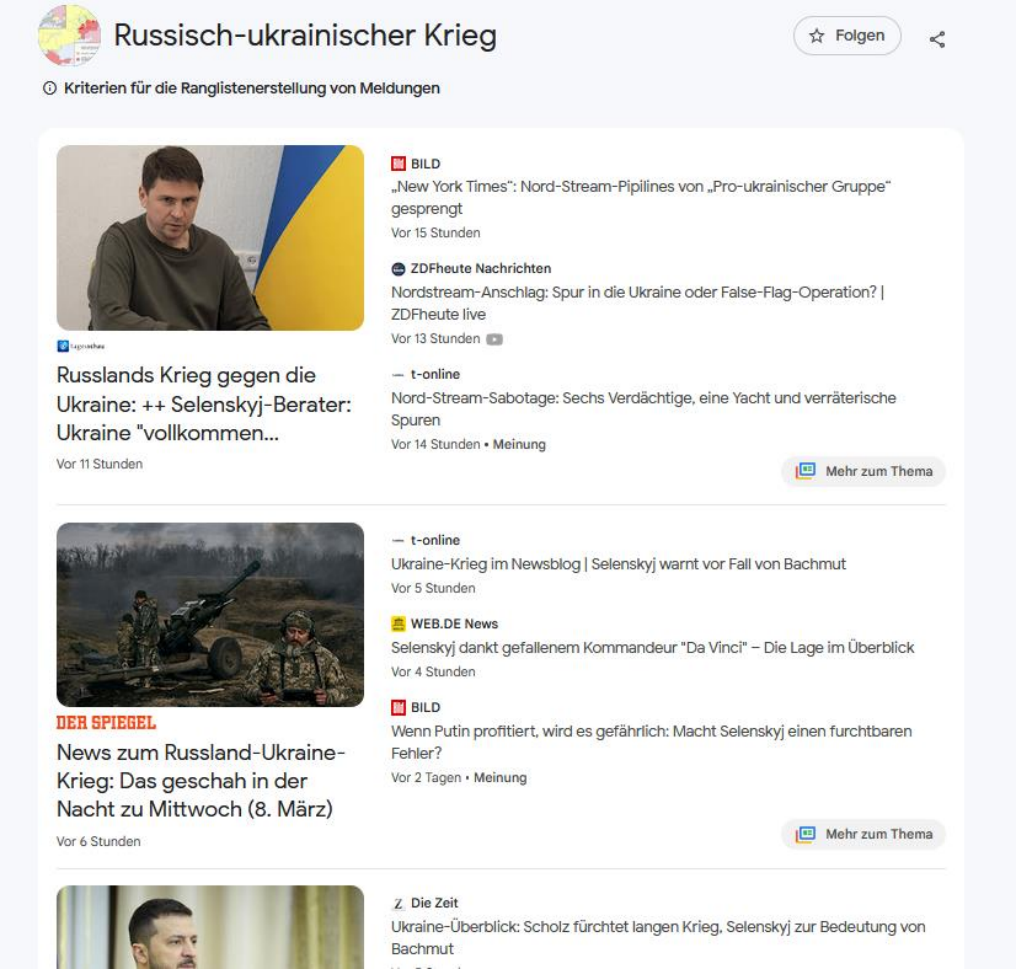
BBC News · 2 hours ago

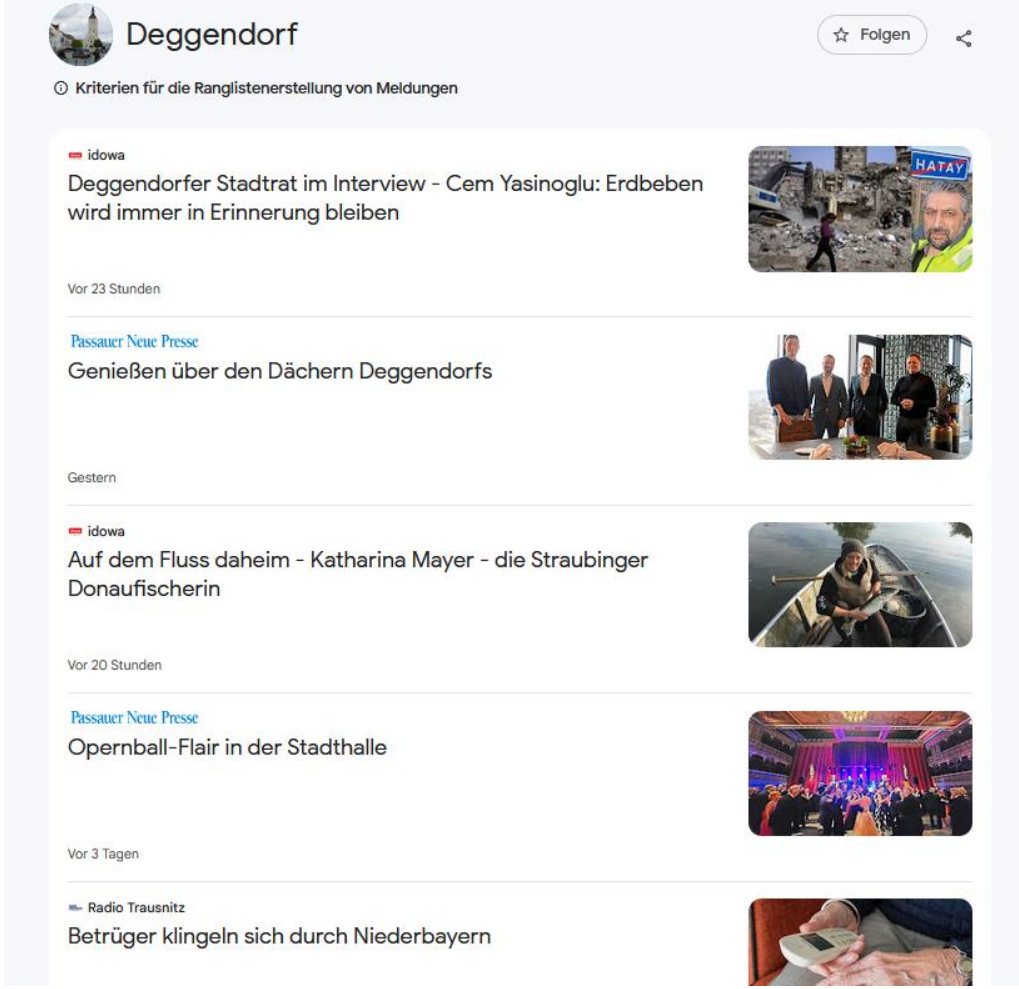
- **Inside a quarantine hotel on Heathrow's 'Isolation Row'**
  The Independent · 3 hours ago
- **Coronavirus in the UK: Quarantine loophole still exists just hours before hotel policy begins, says Michael Matheson**
  Edinburgh News · 22 hours ago
- **Hotel quarantine is another example of too little too late – it's all up to immigration officials now**
  The Independent · Yesterday · Opinion
- **Covid vaccine rollout 'an unbelievable effort' - Johnson**
  BBC News · 15 hours ago

View Full coverage

news.google.com/ - 2021

UNSUPERVISED LEARNING



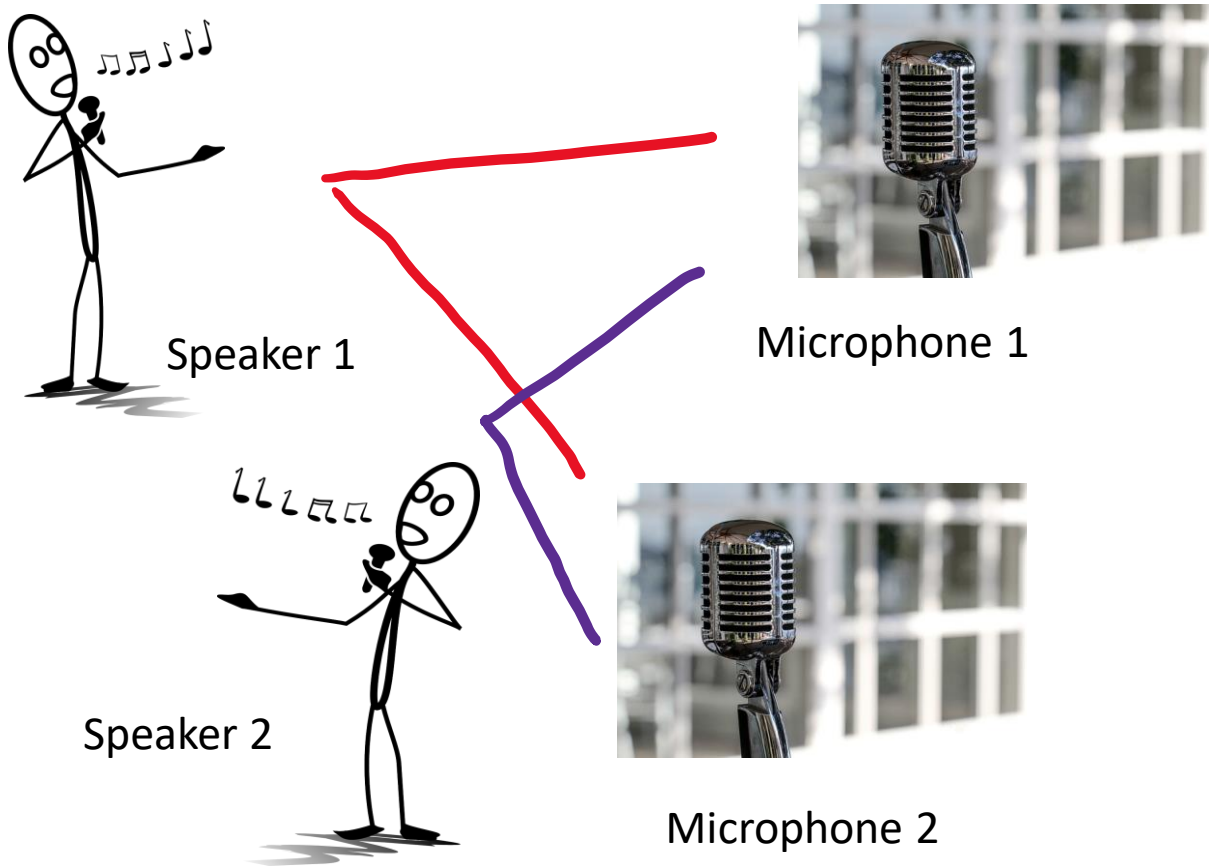


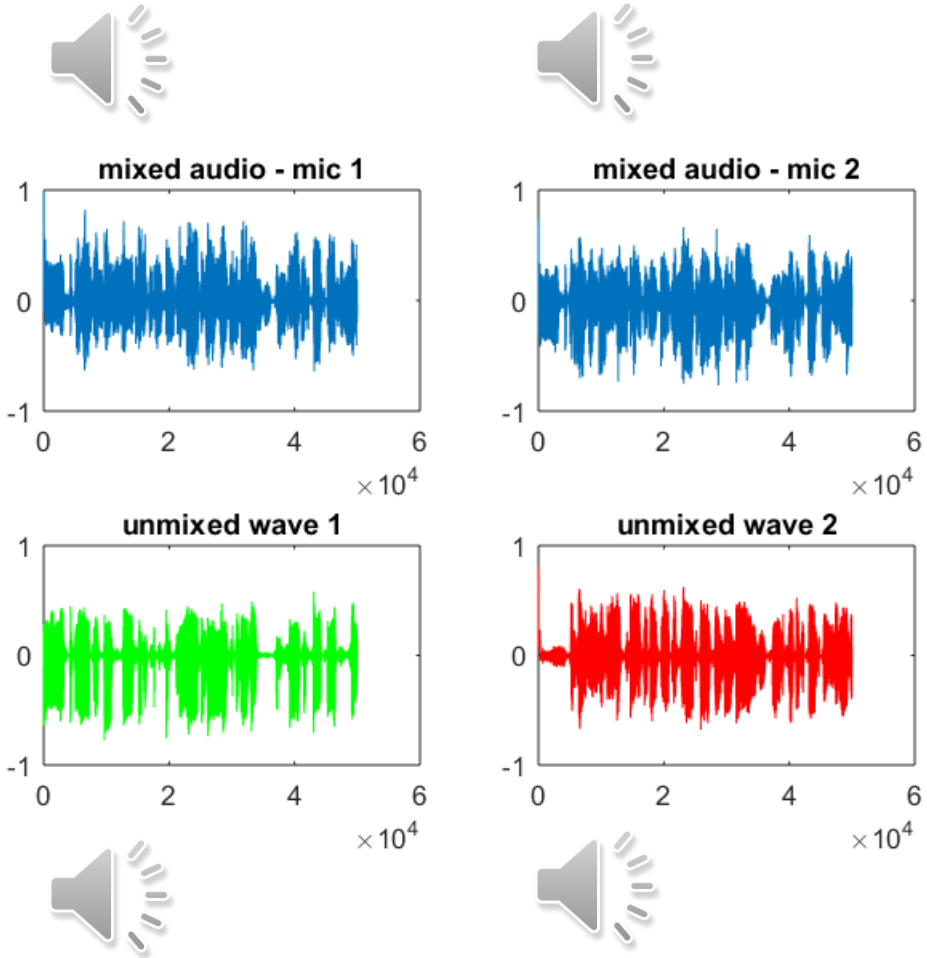
news.google.com/ - 2023

Unsupervised Learning

- Data centre management: Organize computer cluster
- Social science: Social network analysis
- Market segmentation – based on costumer data
- Astronomical data analysis – e. g. “how are galaxies formed”

THE COCKTAIL PARTY PROBLEM AND ALGORITHM





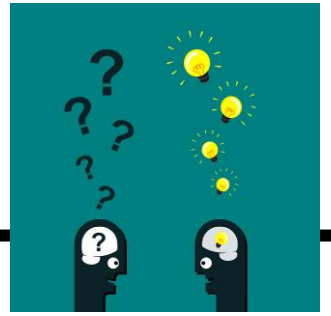
Audio-Files:
https://cnl.salk.edu/~tewon/Blind/blind_audio.html

THE COCKTAIL PARTY PROBLEM

```
E. g.: Octave: [W,s,v]=svd((repmat(sum(x.*x,1),size(x,1),1).*x)*x‘);
```

Common programming languages for machine learning: Python, Matlab, R

QUESTION



With of the following problems would you address using an unsupervised learning algorithm? (more than one can apply)

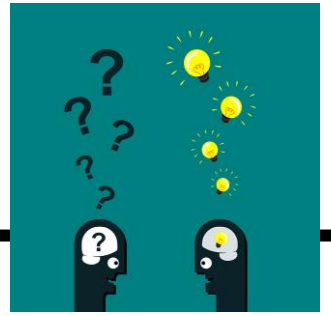
A: Given email labelled as spam/not spam.

B: Given a set of news articles found on the web, group them into a set of articles about the same story.

C: Given a database of costumer data, automatically discover market segments and group customers into different market segments.

D: Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

QUESTION



With of the following problems would you address using an unsupervised learning algorithm? (more than one can apply)

A: Given email labelled as spam/not spam. **– labelled data**

B: Given a set of news articles found on the web, group them into a set of articles about the same story. – e. g. Google news

C: Given a database of costumer data, automatically discover market segments and group customers into different market segments.

D: Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not. **– labelled data**

Wrap up – Machine learning Definition

- Two definitions of Machine Learning are offered. Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.
- Tom Mitchell provides a more modern definition: "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."
- Example: playing checkers.
- E = the experience of playing many games of checkers
- T = the task of playing checkers.
- P = the probability that the program will win the next game.

WRAP UP – SUPERVISED LEARNING

- In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.
- Supervised learning problems are categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

WRAP UP – UNSUPERVISED LEARNING

- Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.
- We can derive this structure by clustering the data based on relationships among the variables in the data.
- With unsupervised learning there is no feedback based on the prediction results.

Quiz – Question 1

A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. What would be a reasonable choice for P?

A: The probability of it correctly predicting a future date's weather.

B: The process of the algorithm examining a large amount of historical weather data.

C: The weather prediction task.

D: None of these.

Quiz – Question 1

A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. What would be a reasonable choice for P?

A: The probability of it correctly predicting a future date's weather.

B: The process of the algorithm examining a large amount of historical weather data.

C: The weather prediction task.

D: None of these.

Quiz – Question 2

Suppose you are working on weather prediction, and you would like to predict whether or not it will be raining at 5pm tomorrow. You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

- Regression
- Classification

Quiz – Question 2

Suppose you are working on weather prediction, and you would like to predict whether or not it will be raining at 5pm tomorrow. You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

- ○ Regression
- ○ Classification

Quiz – Question 3

Suppose you are working on stock market prediction, and you would like to predict the price of a particular stock tomorrow (measured in euros). You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

- ○ Regression
- ○ Classification

Quiz – Question 3

Suppose you are working on stock market prediction, and you would like to predict the price of a particular stock tomorrow (measured in euros). You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

- ○ Regression
- ○ Classification

Quiz – Question 4

Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

A: Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

B: Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).

C: Given historical data of children's ages and heights, predict children's height as a function of their age.

D: Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".

Quiz – Question 4

Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

A: Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.

B: Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).

C: Given historical data of children's ages and heights, predict children's height as a function of their age.

D: Take a collection of 1000 essays written on the US Economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow "similar" or "related".

Quiz – Question 5

Which of these is a reasonable definition of machine learning?

A: Machine learning is the field of allowing robots to act intelligently.

B: Machine learning learns from labelled data.

C: Machine learning is the science of programming computers.

D: Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Quiz – Question 5

Which of these is a reasonable definition of machine learning?

A: Machine learning is the field of allowing robots to act intelligently.

B: Machine learning learns from labelled data.

C: Machine learning is the science of programming computers.

D: Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Recommended Material

Andrew Ng

- https://see.stanford.edu/Course/CS229

Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009): The Elements of Statistical Learning. Springer, New York