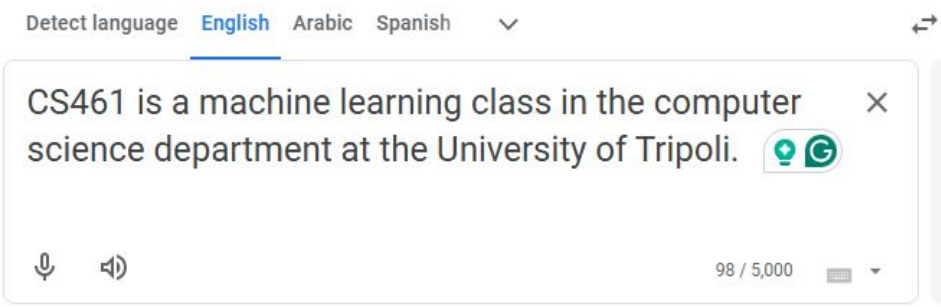
# Introduction to machine learning

Ali Aburas, PhD

**Slide credit:Andrew Ng**

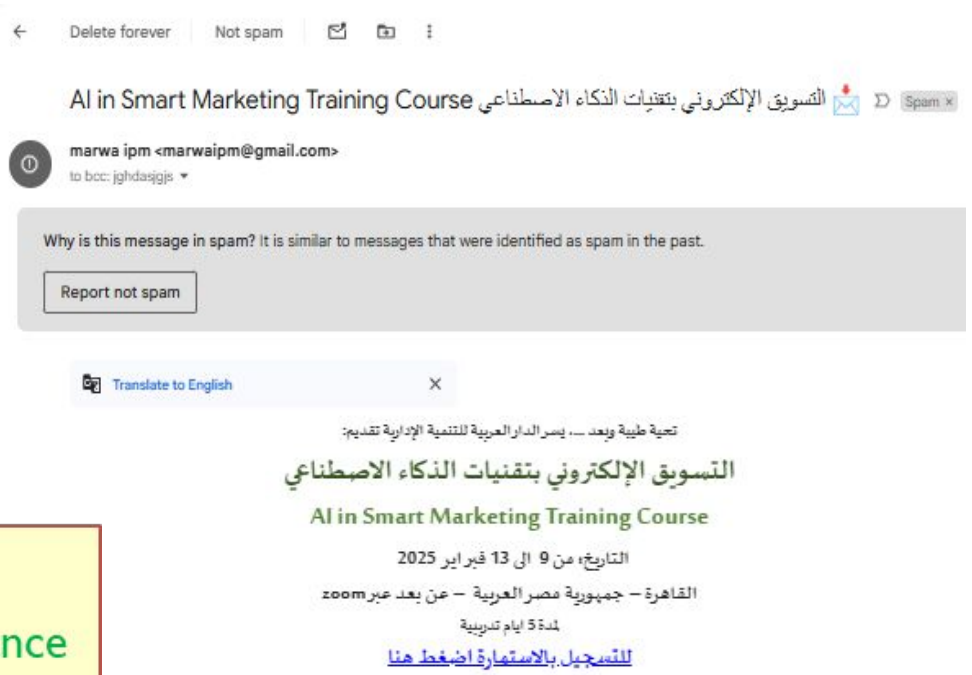
# Machine Learning is Everywhere





# Applications: Spam Detection

- This is a **binary classification task**: Assign **one of two labels (i.e. yes/no) to the input** (here, an email message)
- Classification requires a **model (a classifier)** to determine which label to assign to items.
- In this class, we **study algorithms and techniques to learn such models** from data.

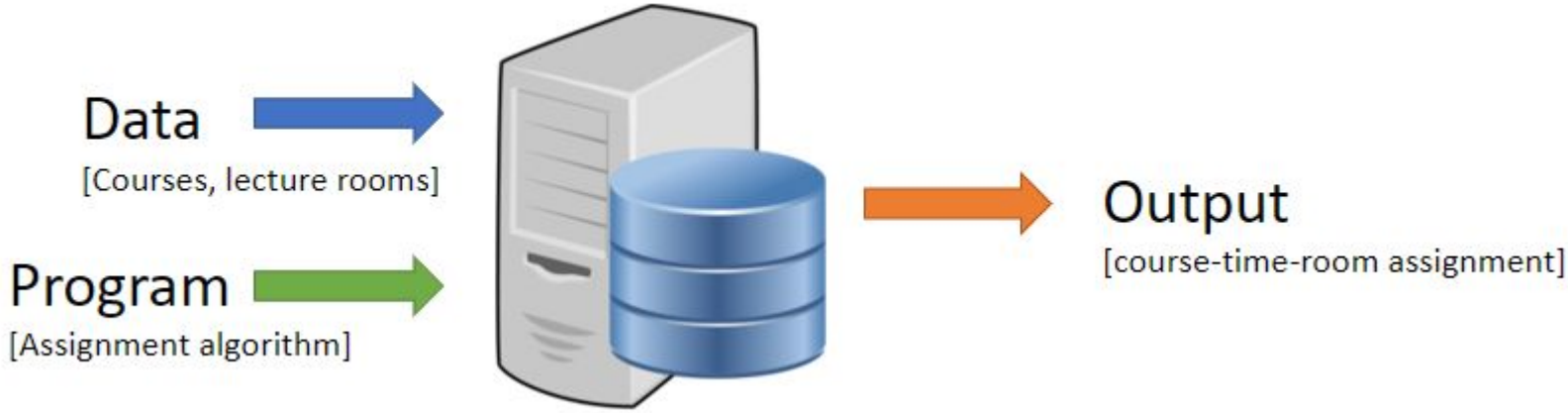



| Documents | Labels |
|---|---|
| Documents | Politics, Sports, Finance |
| Sentences | Positive, Negative |
| Phrases | Person, Location |
| Images | Cats, Dogs, Snakes |
| Medical records | Admit again soon/Not |

# Some More Involved Examples

- Driving:
  - https://www.youtube.com/watch?v=_1MHGUC_BzQ
  - E.g., go to 11:48
- Objects:
  - https://www.youtube.com/watch?v=_1MHGUC_BzQ
  - Go to 1:41
- Tesla Accidents:
  - https://www.youtube.com/watch?v=FVgkWii5JdM
  - Go to 1:45

# Traditional CS




# Machine learning


Example by Dr. Kilian Weinberger

# Learning

– Learning is at the core of
  - Understanding High Level Cognition
  - Performing knowledge intensive inferences
  - Building adaptive, intelligent systems
  - Dealing with messy, real world data
  - Analytics

– Learning has multiple purposes
  - Knowledge Acquisition
  - Integration of various knowledge sources to ensure robust behavior
  - Adaptation (human, systems)
  - Decision Making (Predictions)

# Learning = Generalization

- The ability to perform a task in a situation which has never been encountered before
- The learner has to be able to classify items it has never seen before.


**How can we expect a program to make predictions on items it has never seen before? What should this program rely on?**

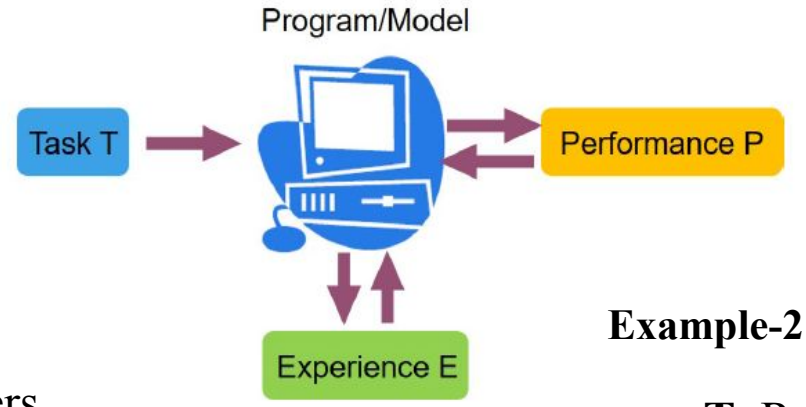
# What is Machine Learning?

- Definition by Tom Mitchell (1998): Machine learning is the study of algorithms that
  - improve their performance P
  - on a task T
  - with experience E.

  A well-defined learning task is given by <P, T, E>

# What is Machine Learning ? (Contd.)

Improve on task **T**, with respect to performance metric **P**, based on experience **E**.



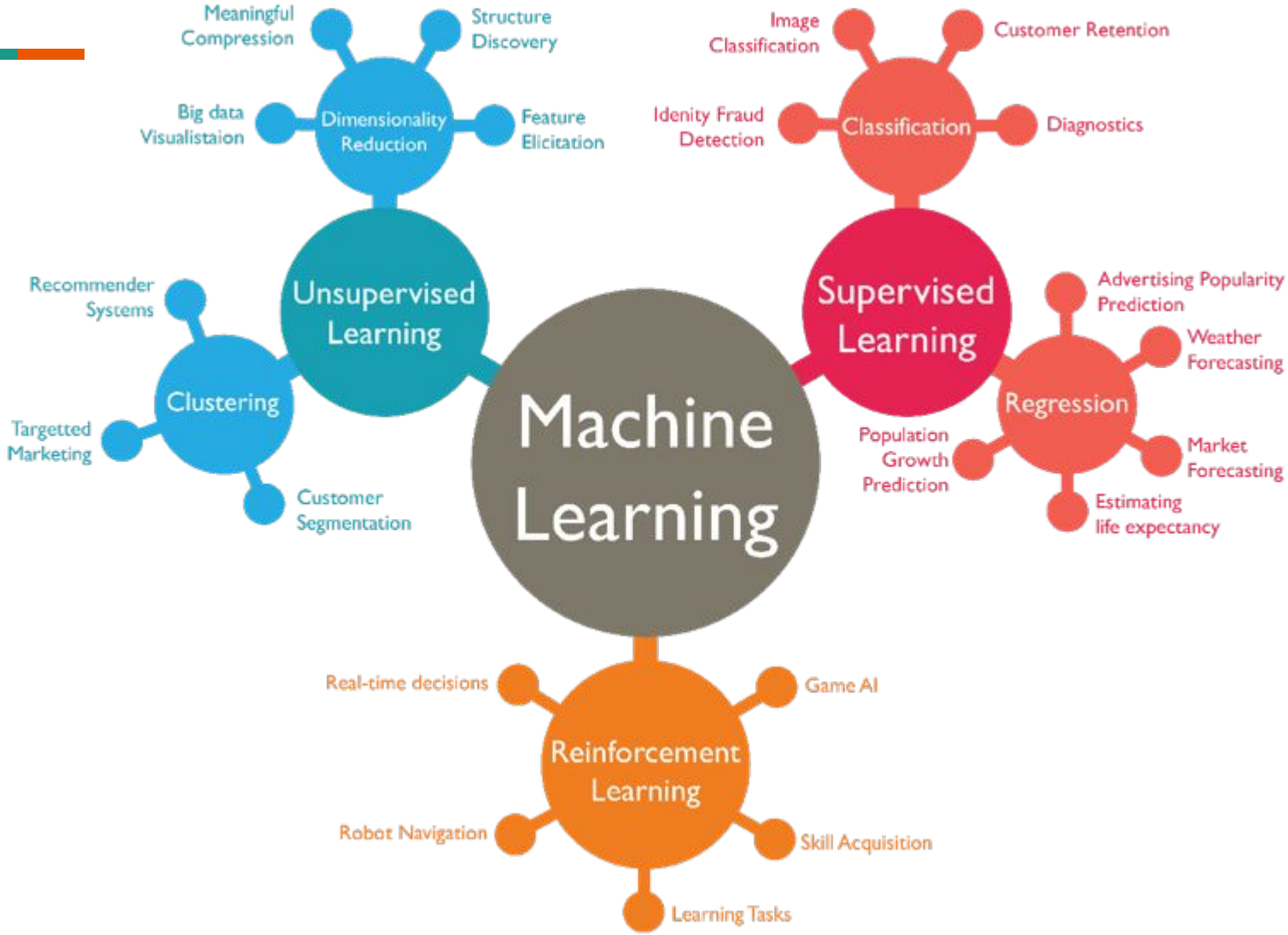

**Example-1**

- **T**: Playing checkers
- **P**: Percentage of games won against an arbitrary opponent
- **E**: Playing practice games against itself

**Example-2**

- **T**: Recognizing handwritten words
- **P**: Percentage of words correctly classified
- **E**: Database of human-labeled images of handwritten words

# Types of Learning



# 1. 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.
- The thing we want to predict is called the **target** or the response variable
- Usually, we need **training data**
- 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.

# Supervised Classification Problems

- Given $(x_1, y_1), \ldots, (x_n, y_n)$, learn a function that predicts $y$ given $x$
- **Classification:** Labels $y$ are categories

- Cancer diagnosis (**Training Set**)

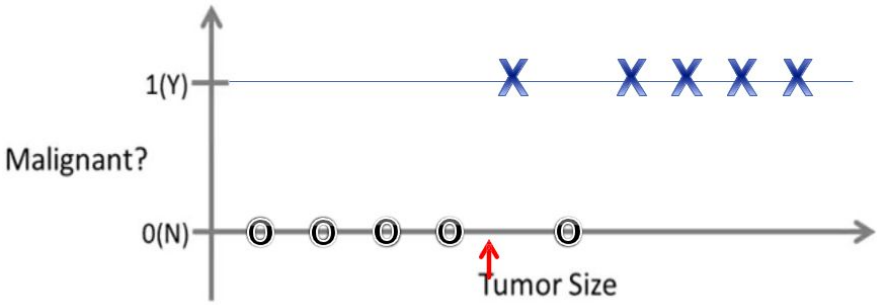
| Patient ID | # of Tumors | Avg Area | Avg Density | Diagnosis |
|---|---|---|---|---|
| 1 | 5 | 20 | 118 | Malignant |
| 2 | 3 | 15 | 130 | Benign |
| 3 | 7 | 10 | 52 | Benign |
| 4 | 2 | 30 | 100 | Malignant |

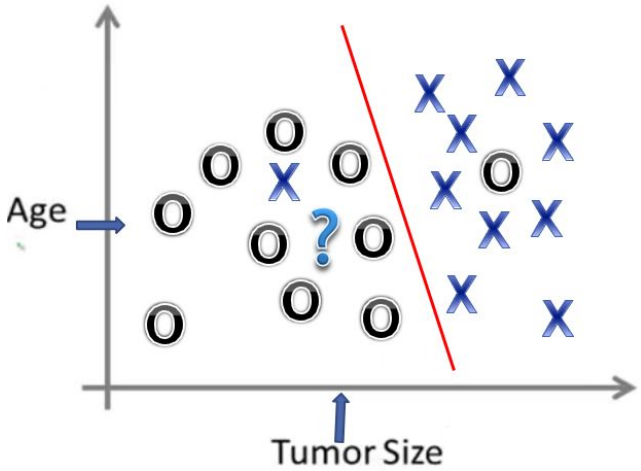
- Use the above training set to learn how to classify patients where diagnosis is not known (**Test Set**):

| Patient ID | # of Tumors | Avg Area | Avg Density | Diagnosis |
|---|---|---|---|---|
| 101 | 4 | 16 | 95 | ? |
| 102 | 9 | 22 | 125 | ? |
| 103 | 1 | 14 | 80 | ? |

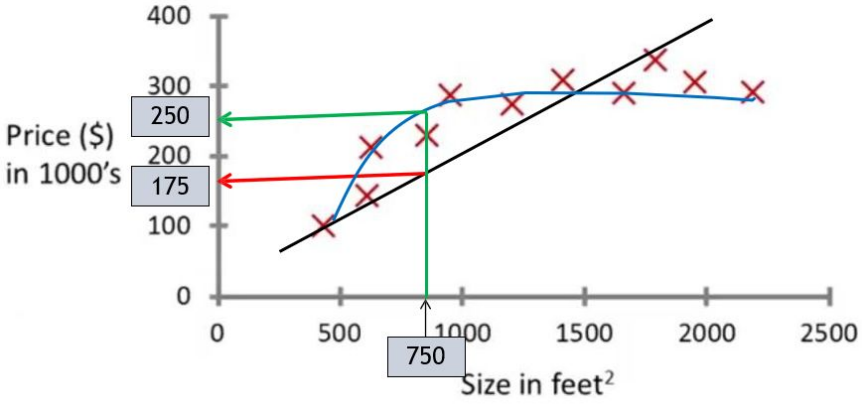
# Classification and Regression

- **Classification**: finding decision boundaries



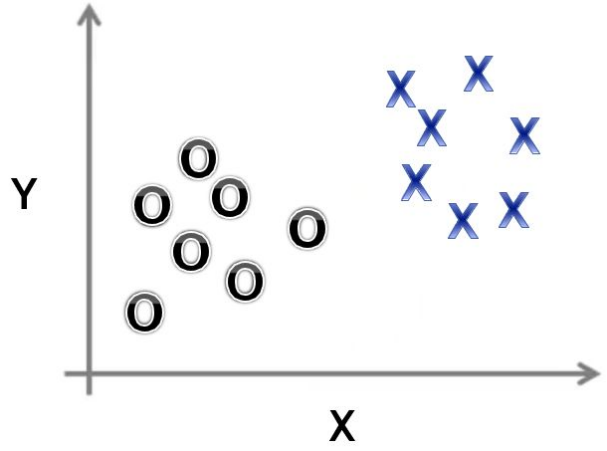



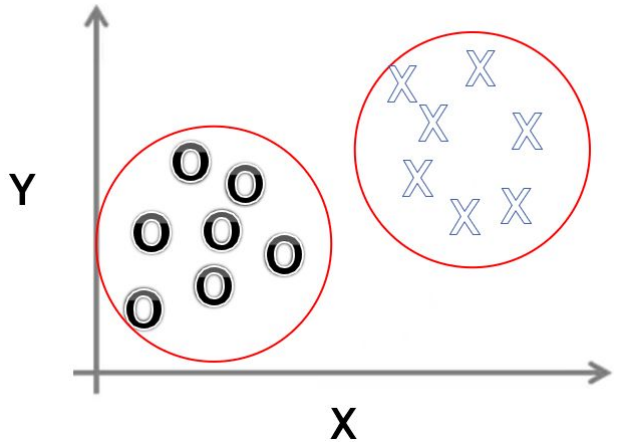

- **Regression**: fitting a curve/plane to data



# 2. Unsupervised Learning

- Unsupervised learning, on the other hand, 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, i.e., there is no teacher to correct you.

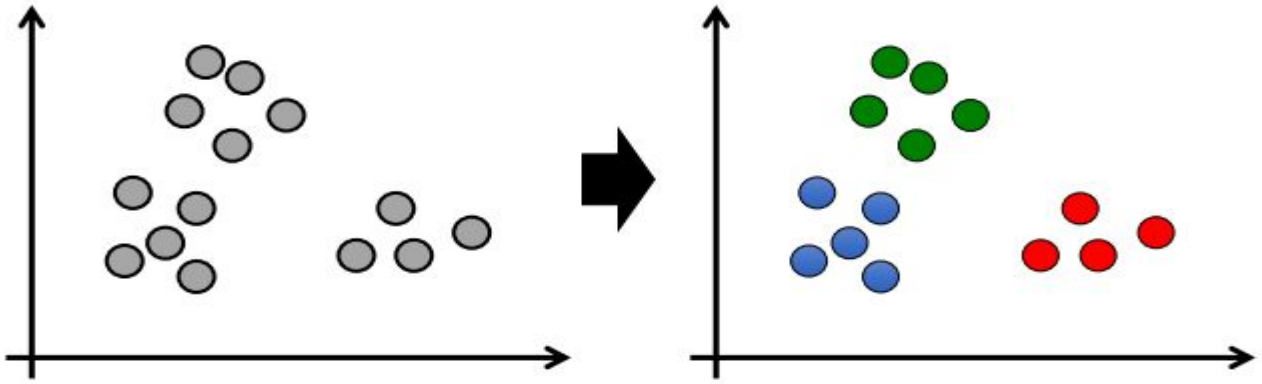


# 2. Unsupervised Learning

- For an unsupervised learning problem, we do not focus on prediction of any particular thing, but try to nd interesting aspects of the data
- For example, say you have a lot of data about your blog's visitors. You may want to run a **clustering algorithm** to try to detect groups of similar visitors

- Given $x_1, \ldots, x_n$ (no labels), output hidden structure in $x$'s
  - E.g., clustering



**An unlabeled training set for unsupervised learning**

**Clustering**

# Why Do We Need Machine Learning?

# Why Do We Need Machine Learning?

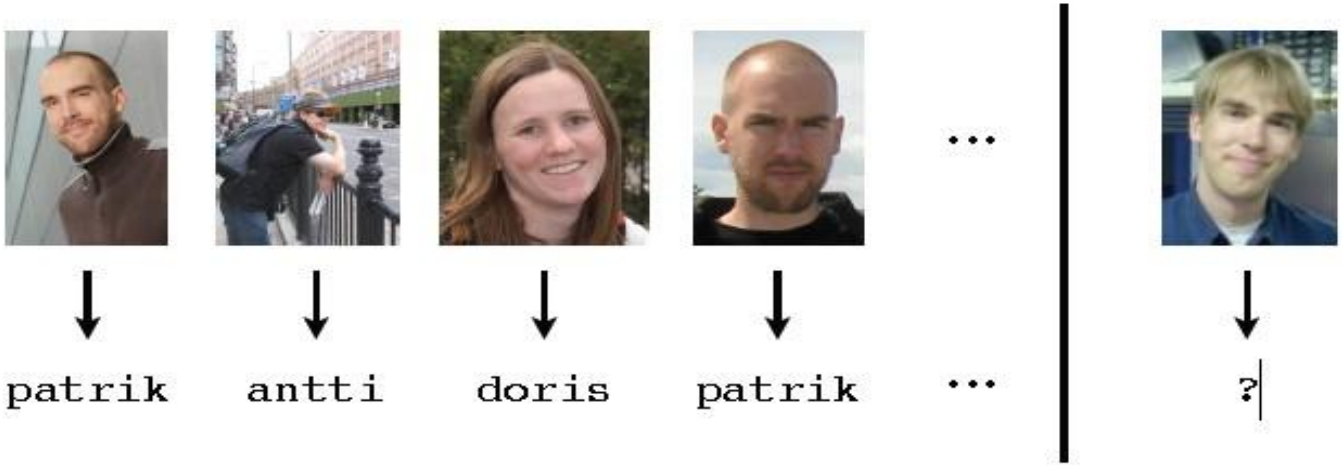
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task (knowledge engineering bottleneck)
- Develop systems that can automatically adapt and customize them- selves to individual users.
  - Personalized news or mail filter
  - Personalized tutoring
- Discover new knowledge from large databases (data mining)
  - Market basket analysis (e.g. diapers and cakes)
  - Medical information mining (e.g. migraines to calcium channel blockers to magnesium)
- Computational studies of learning may help us understand learning in humans and other biological organisms

# Example 3: Prediction of search queries

- ~~The programmer provides a standard dictionary~~ (words and expressions change!)
- Previous search queries are used as examples!

# Example 2 : face recognition

- Face recognition is hot (facebook, apple; security; . . . )
- ~~Programmer writes rules: "If short dark hair, big nose, then it is Mikko"~~ (impossible! how do we judge the size of the nose?!)
- The computer is shown many (`image, name`) example pairs, and the computer learns which features of the images are predictive (difficult, but not impossible)

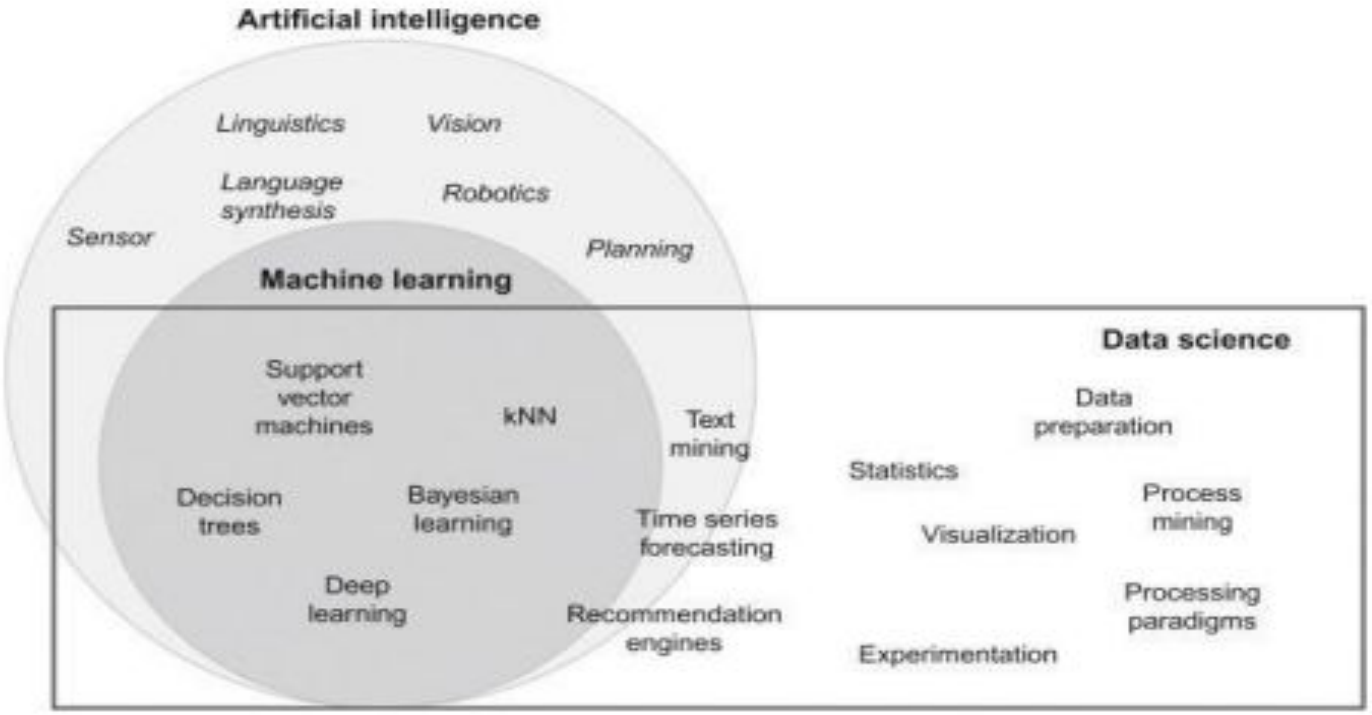


# AI vs. Machine Learning vs. Data Science

# AI vs. Machine Learning vs. Data Science

- **Artificial intelligence** (AI) is one of the many disciplines of Computer Science. The aim of AI is to make machines think like humans do to make decisions based on the data provided to the machine. You've probably interacted with AI even if you don't realize it—voice assistants like Siri and Alexa are founded on AI technology, as are customer service chatbots that pop up to help you navigate websites.
- **Machine learning** aims to make machines learn based on the data stored in it and make decisions. It is a subset of AI. Through machine learning, practitioners develop artificial intelligence through models that can "learn" from data patterns without human direction. The unmanageably huge volume and complexity of data (unmanageable by humans, anyway) that is now being generated has increased machine learning's potential, as well as the need for it.
- **Data science** is a process in which huge volumes of raw data is analysed and processed to extract meaningful information to use for various research and business purposes. It is a long process that starts from collecting the raw data (both structured and unstructured)

# AI vs. Machine Learning vs. Data Science



# Machine Learning Workflow