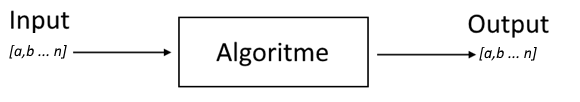
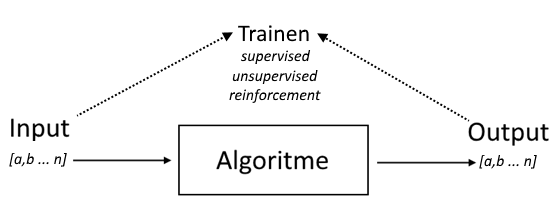
# Machine learning

Een zelflerend systeem is een algoritme gebaseerd op machine learning. Machine learning werd door Arthur Samuel, een pionier op dit gebied, gedefinieerd als: "A field of study that gives computers the ability to learn without being explicitly programmed.” [1]. In tegenstelling tot de eerder genoemde algoritmes is een zelflerend systeem in staat zichzelf te verbeteren. Hierdoor kan het taken uitvoeren waarbij reguliere algoritmes tekort schieten. Welke taken dit betreft, zullen we in de derde deelvraag behandelen.

Figuur Schematische weergave van een zelflerend systeem

In figuur \*\*\*\* is een schematische weergave van een zelflerend systeem afgebeeld. Bepaalde input data gaat het systeem in en bepaalde output data komt het systeem uit. De input en output data bestaat uit één of meerdere getallen. Als de input simpelweg een reeks getallen betreft, zal dit direct als input gebruikt kunnen worden. In het geval dat de input uit een ander datatype bestaat, zoals een plaatje, zal dit omgezet moeten worden in een reeks getallen voordat het in een zelflerend systeem gebruikt kan worden. Het algoritme zal deze getallen bewerken tot de gewenste output. Deze output wordt eveneens in getallen gegeven. Waar nodig zullen deze getallen dus weer moeten worden omgezet tot het gewenste datatype.

Er zijn veel verschillende algoritmes die gebruikt kunnen worden voor een zelflerend systeem. Elk algoritme heeft voor- en nadelen en is geschikt voor andere doeleinden. Een aantal van deze algoritmes zullen we in de tweede deelvraag behandelen.

Training

Figuur Schematische weergave van een zelflerend systeem

Een zelflerend systeem begint in de meeste gevallen zonder enige kennis van de data. Om de gewenste output te kunnen produceren is het dus nodig om het systeem eerst input data te geven zodat het kan leren. Dit proces wordt het ***trainen*** genoemd. Voor het trainen van een zelflerend systeem is training data nodig. Deze data moet gelijk of gelijkwaardig zijn aan de *echte* data. De training data kan in veel verschillende vormen voorkomen en de manier van trainen is afhankelijk van de vorm van de (training) data. In figuur \*\*\*\* is te zien dat het trainen los staat van het algoritme. Dit verschil zullen we in de volgende deelvraag wat duidelijker maken.   
Er zijn drie prominente manieren waarop een zelflerend systeem getraind kan worden: **supervised**, **unsupervised** en **reinforcement learning**.

Supervised Learning

In het geval van supervised learning heb je te maken met **labeled** training data. Anders gezegd: van een bepaalde input is de gewenste output al bekend. Een klassiek voorbeeld van een labeled dataset is een dataset van huisprijzen en huiseigenschappen (zie figuur xx)

|  |  |  |  |
| --- | --- | --- | --- |
| Huisprijs (output) | Huiseigenschappen (input) | | |
|  | Woonoppervlakte | Perceeloppervlakte | Aantal kamers |
| € 519.000 | 124 m² | 311 m² | 4 |
| € 569.000 | 133 m² | 309 m² | 5 |
| € 569.500 | 170 m² | 310 m² | 6 |

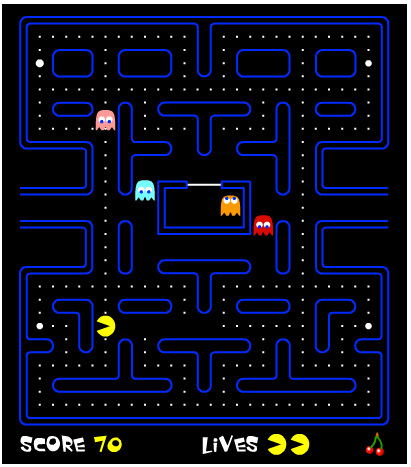
Figuur xx Labeled dataset   
Bron: http://www.funda.nl/koop/huizen/   
  
Bij de training dataset van figuur \*\*\*\* is de gegeven input de huiseigenschappen en de gewenste output de huisprijs. Het systeem wordt met deze dataset getraind. Hierdoor leert het een output te produceren die steeds dichter bij de gewenste output ligt. Als er een verband bestaat tussen de huiseigenschappen en de huisprijs, wat waarschijnlijk het geval is, zal het zelflerende systeem na genoeg trainen in staat zijn zelf bij nieuwe huiseigenschappen een huisprijs te voorspellen. [2]

## Unsupervised Learning

Unsupervised learning kan gebruikt worden bij een **unlabeled** dataset ofwel, een dataset waarbij de data niet geclassificeerd is en er geen gewenste output bekend is. Als je een dataset hebt van heel veel niet-geordende foto’s is het niet mogelijk om dit te classificeren. Als een deel van de dataset gelabeld wordt, zal met behulp van supervised learning de rest van de dataset geclassificeerd kunnen worden. Dit is echter in veel gevallen niet mogelijk, bijvoorbeeld doordat de dataset enorm groot is of er zodanig veel verschillende groepen bestaan dat het menselijk niet mogelijk is ook maar een deel te labelen. Ook kan het zo zijn dat men niet weet of er een verband aanwezig is.   
Kortom: unsupervised learning wordt gebruikt voor het classificeren van data, zonder dat er groepen vooraf gedefinieerd zijn. Met behulp van deze vorm van training zullen in een grote dataset verbanden kunnen worden ontdekt, die men misschien niet zonder hulp had kunnen achterhalen.[3]

## Reinforcement Learning

Reinforcement learning is een zeer specifieke soort van leren. Er is bij deze vorm van learning geen dataset met input data, maar er is een bepaalde **context**. In deze context is een omgeving waarin zich een **agent** bevindt*.* Een agent is een object dat bepaalde opdrachten kan uitvoeren. Door de agent bij bepaalde acties pluspunten of minpunten te geven kun je bepaald gedrag bevorderen.



Figuur Pacman   
Bron: https://www.flickr.com/photos/methodshop/4865516413

In figuur \*\*\*\* is het spel Pac-Man te zien. Op dit spel zou reinforcement learning toegepast kunnen worden. De agent is hierbij pacman, dit is namelijk een object dat bepaalde opdrachten kan uitvoeren, zoals: *beweeg naar links*. De context is hierbij het level, ofwel: de positie van de muren (de blauwe obstakels), de posities van de *ghosts (de gekleurde vijanden)*, de posities van de *pac-dots* (de kleine stipjes) en de posities van de *power-pellets* (de grotere stipjes). [4] Het eten van de pac-dots is positief, het geraakt worden door de ghosts is negatief. Door reinforcement learing toe te passen op het spel zal de agent steeds beter worden in het spelen van het spel.

# Normaliseren van data

Zoals ook in figuur xx te zien is, kunnen verschillende inputs erg van elkaar verschillen qua grootte. Zo zal het aantal kamers nooit in de buurt komen van het oppervlak. Uiteindelijk zou dit een probleem kunnen veroorzaken bij de berekeningen van het systeem. Een groot getal zou namelijk een veel groter aandeel kunnen hebben alleen omdat het getal zoveel groter is. Het is daarom gebruikelijk de inputs te **normaliseren**. Dit houdt in dat de inputs zullen veranderen in een getal met een waarde binnen een bepaald gebied zodat alle verschillende inputs eerlijk met elkaar vergeleken kunen worden. Je zou bijvoorbeeld voor de oppervlaktes kunnen stellen dat alle waardes tussen 100 m2 en 1000 m2 zullen liggen. Aan een input van 500 m2 zou je dan een waarde van 5 kunnen geven.

# Conclusie

Zelflerende computersystemen zijn algoritmes gebaseerd op machine learning. Een zelflerend systeem verschilt van reguliere algoritmes zoals breadth-first search en depth-first search doordat ze in staat zijn zichzelf te verbeteren.

# Bronnen

[1] Bron: <https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf>Geraadpleegd op: 21-5-2017  
Laatst gewijzigd op: niet bekend

[2] Bron: <https://www.coursera.org/learn/machine-learning/lecture/1VkCb/supervised-learning>  
Geraadpleegd op: 21-5-2017  
Laatst gewijzigd op: niet bekend

[3] Bron: <https://www.coursera.org/learn/machine-learning/lecture/olRZo/unsupervised-learning>  
Geraadpleegd op: 21-5-2017  
Laatst gewijzigd op: niet bekend

[4] Bron: <http://pacman.wikia.com>  
Geraadpleegd op: 27-5-2017  
Laatst gewijzigd op: niet bekend