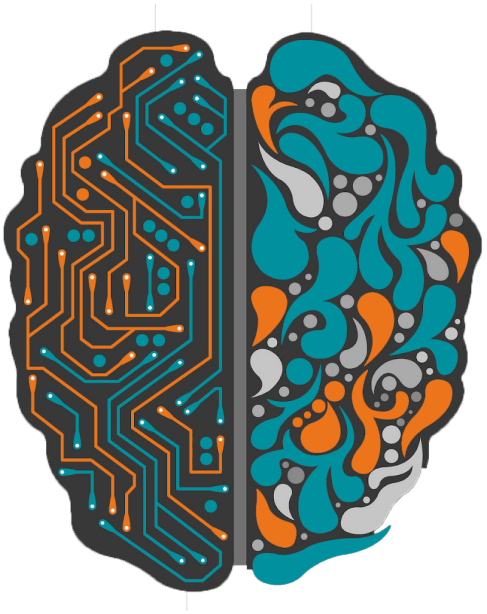
kahuna
managing security

Windesheim
Flevoland



Machine learning handbook

Het machine learning project

Vincent van Dijk
25 Oktober 2018

kahuna.nl
Amersfoort

1 Configuration page

1.1 Document

Date	Name	Extension	Version	Author	Status/Chance
27-09-2018	machine learning handbook	.gdoc	0.1	Vincent van Dijk	Concept
29-09-2018	machine learning handbook	.gdoc	0.2	Vincent van Dijk	Concept
01-10-2018	machine learning handbook	.gdoc	0.3	Vincent van Dijk	Concept
04-10-2018	machine learning handbook	.gdoc	0.4	Vincent van Dijk	Concept
05-10-2018	machine learning handbook	.gdoc	0.5	Vincent van Dijk	Concept
05-10-2018	machine learning learning	.gdoc	0.5.5	Vincent van Dijk	Changes in the chapters: cybersecurity
08-10-2018	machine learning handbook	.gdoc	0.6	Vincent van Dijk	Changes in the chapters: cybersecurity
09-10-2018	machine learning handbook	.gdoc	0.7	Vincent van Dijk	Changes in the chapters: cybersecurity and tools
23-10-2018	machine learning handbook	.gdoc	0.8	Vincent van Dijk	Added chapter: Input documentation. Changes in the chapters: tools and the read more of every chapter
25-10-2018	machine learning handbook	.gdoc	0.9	Vincent van Dijk	Added chapter: references. Changes in the chapters: tools and cybersecurity

1.2 Distribution management

Date	Name	Extension	Version	Send to
05-10-2018	machine learning handbook	geen	0.4	Issam Schinkel
05-10-2018	machine learning handbook v0.5	.pdf	0.5	Edwin Hennipman
08-10-2018	machine learning handbook	.pdf	0.5.5	Daniel Hoogenwerf
17-10-2018	machine learning handbook v0.7	.pdf	0.7	Edwin Hennipman
23-10-2018	machine learning handbook	.pdf	0.8	Issam Schinkel
25-10-2018	machine learning handbook	.pdf	0.9	Daniel Hoogenwerf
30-10-2018	machine learning handbook	geen	0.9	Daniel Hoogenwerf
31-10-2018	machine learning handbook	.pdf	0.9	Edwin Hennipman
31-10-2018	machine learning handbook	.pdf	0.9	Peter van der Post

1.3 Input documentation

Some of the most notable books I read and courses courses I followed before starting this project.

Name	Vorm	Author	Version
Deep Learning for Natural Language Processing	Ebook	Jason Brownlee	1.1
Deep learning with python	Ebook	Jason Brownlee	1.11
Long Short-Term Memory Networks With Python	Ebook	Jason Brownlee	1.2
Machine learning mastery with python	Ebook	Jason Brownlee	1.8
Introduction to time series forecasting with python	Ebook	Jason Brownlee	1.3
Data Science, Deep Learning and Machine Learning with Python	MOOC	Frank Kane	
Deep Learning A-Z™: Hands-On Artificial Neural Networks	MOOC	Kirill Eremenko and Hadelin de Ponteves	
Machine Learning A-Z™: Hands-On Python & R In Data Science	MOOC	Kirill Eremenko and Hadelin de Ponteves	
Artificial Intelligence A-Z™: Learn How To Build An AI	MOOC	Kirill Eremenko and Hadelin de Ponteves	
Python Machine learning	Book	Sebastian Raschka	1st edition

2 Table of contents

3 Introduction

3.1 Intro

Machine learning and artificial intelligence are two buzzwords that are meant to solve most of the problems in today's society (and blockchain will solve the leftovers). The reason that people think this is because of a lack of understanding machine learning or artificial intelligence. In this document I will try to enlighten the readers understanding about machine learning and artificial intelligence. You will have a clear understanding of what machine and artificial intelligence can do. And more importantly what it can NOT do.

3.2 Why

While working on the machine learning project within Kahuna I noticed there is a lack of knowledge about machine learning. This got reflected got reflected back by the stakeholders. This document is meant to enlighten the stakeholders so there is an in depth, clear understanding of the usage of machine learning within Kahuna.

3.3 Purpose

This document is meant to clear up all the misunderstandings that are around the buzz word machine learning. This document will also teach when to use and when not to use machine learning, the positives of machine learning, the negatives and the processes involved to develop machine learning models. We will also go into the use of machine learning in a cybersecurity landscape.

After reading this document you will be able to discuss machine learning on a high level with machine learning developers and data scientist.

3.4 Who

This document is written for every Kahuna employee that is interested in selling, using or consulting about Machine learning in a cyber security landscape.

3.5 Language

In contrast to the other documents in this project, that are written in dutch, this document will be written in english. This document will be written in english because I want this document to be distributable to all Kahuna employees, so also the ones in Spain or Belgium.

3.6 Prior knowledge

Before starting the machine learning project at Kahuna I have used machine learning for around a year. I ran one machine learning project at a traffic company. Further I have used machine learning for personal projects, like predicting cryptocurrency prices.

My experience with machine learning is mostly practical. I used machine learning as a tool to reach my project goals, I never used math to create my own machine learning algorithms.

4 Machine learning

Before I want to go too deep into machine learning I want to start off with this quote I found:
"Neural network AI is simple. So… Stop pretending you are a genius" (KDnuggets, 2018)

Machine learning and AI moved at a point where you don't have to be a genius to make good use of machine learning. Sadly learning about machine learning can still be challenging since most resources are not simplified enough. From my experience there are now two types of explanations. Explanations for statisticians / mathematicians and explanations for programmers.

Because of the simplicity where machine learning is getting at programmers can already start with actionable machine learning. These guides are getting very simpel, where are almost no math is required, only programming skills.

But this document is not a document for programmers. With this document you will be able to talk with machine learning developers about the use of machine learning.

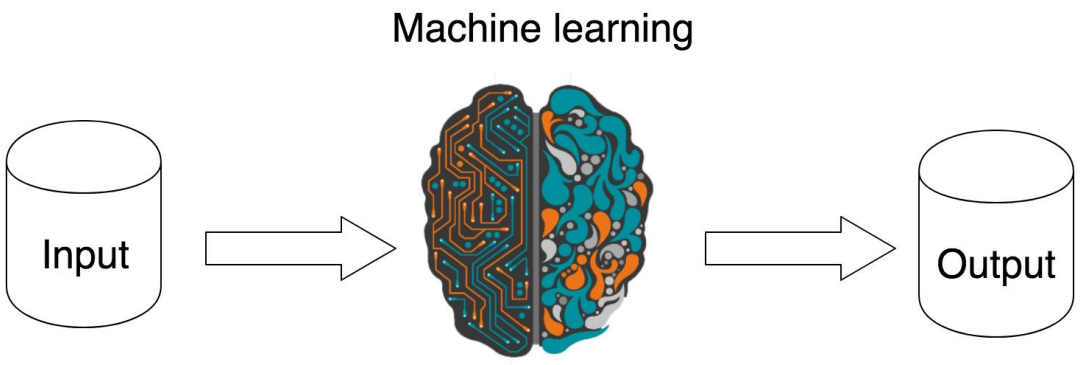
4.1 Definition

Defining the exact definition of machine learning is hard. I am not going to quote any definition from the internet since all the definition I came around won't make you any wiser.

From my experience with talking with people about machine learning, I found that the two words "machine" and "learning" covers too much ground. Machine learning can imply that you as a person teach a machine to follow your commands. Programming is like this, with code, the language of computers you tell the machine what to do, the machine learns. But this is not really machine learning.

But if that is not machine learning, what is? Machine learning is the ability of a machine to learn on his own. You as a person only have to define the input and the output. (In some machine learning techniques you don't even need to specify the output, this will be handled in a future chapter).

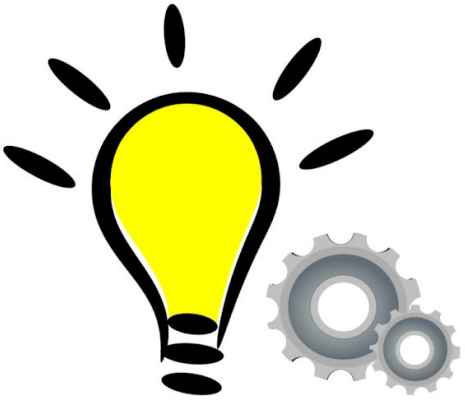
When the input and output is defined then the machine learning will use that data to learn a connection. The "learning of the connection" is machine learning. When machine learning is learning the connection it is called training. The machine learning algorithm is training himself to learn the connection. This can be hard to understand just from reading the text. That is why I prepared an example.



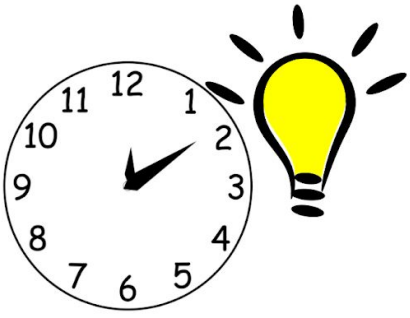
4.2 Example

John is a programmer that recently bought his own house

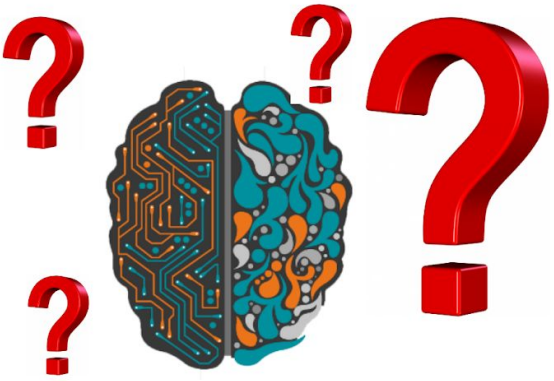
Because John is a programmer he wants to automate the lights in his house

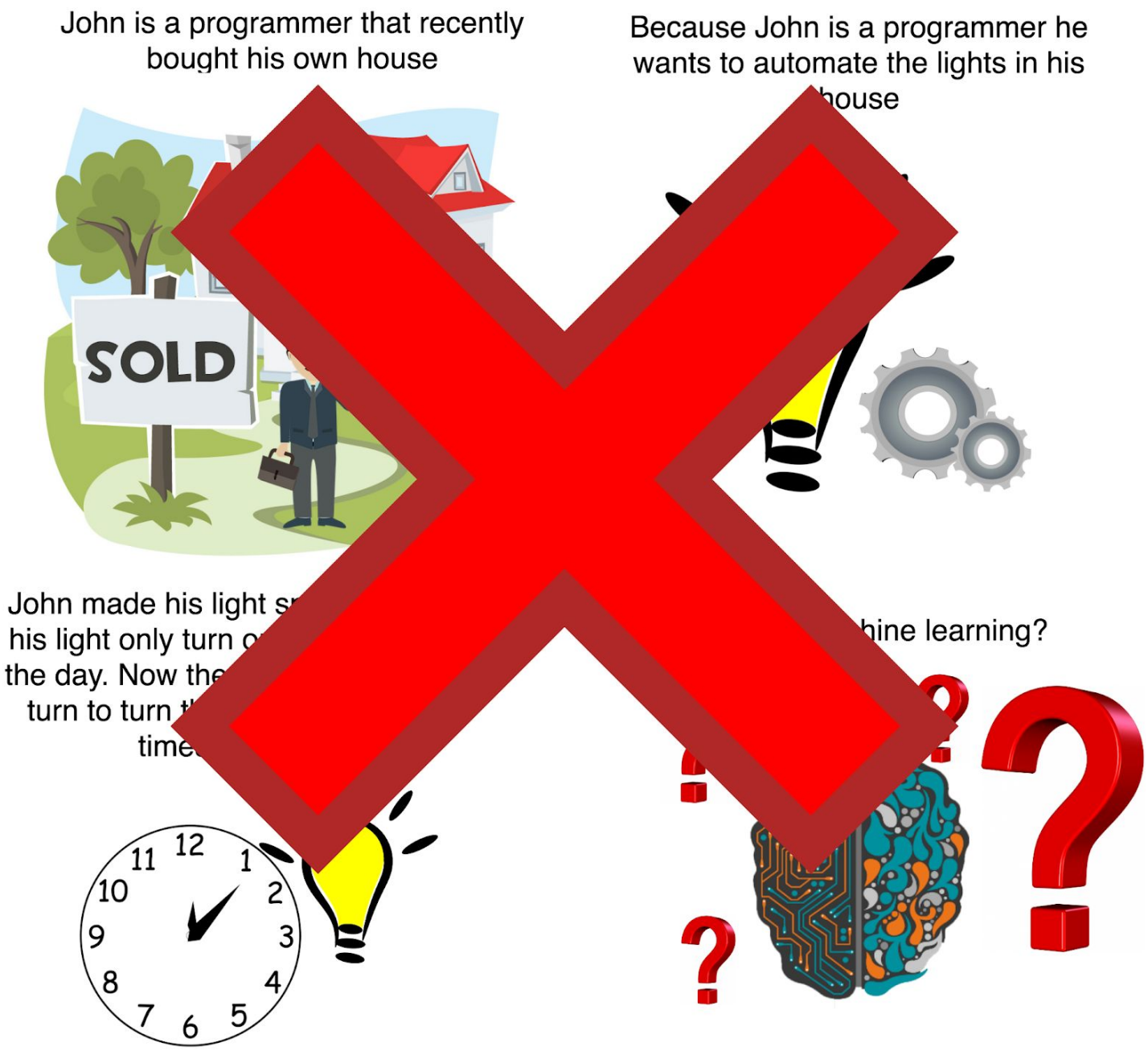


John made his light smart. He made his light only turn on certain times of the day. Now the machine learned to turn to turn the light on at specific times of the day



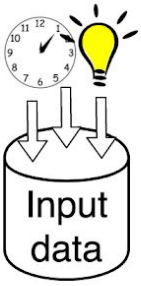
Is this machine learning?



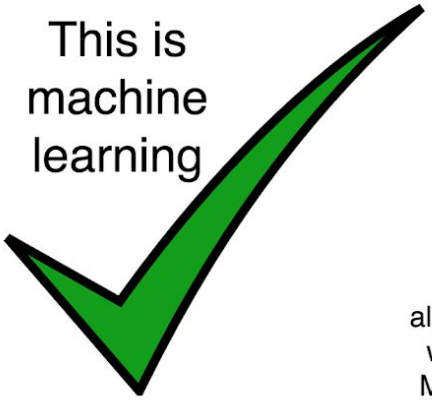

John is a programmer that recently bought his own house
Because John is a programmer he wants to automate the lights in his house
SOLD
John made his light s
his light only turn o
the day. Now the
turn to turn t
time
hine learning?

John gives collects a big dataset that looks like this:

Timestamp	Light (0-100)
12:30	10
12:45	12
13:00	11

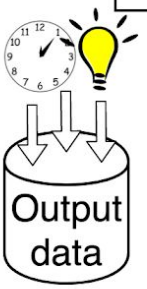


This is machine learning

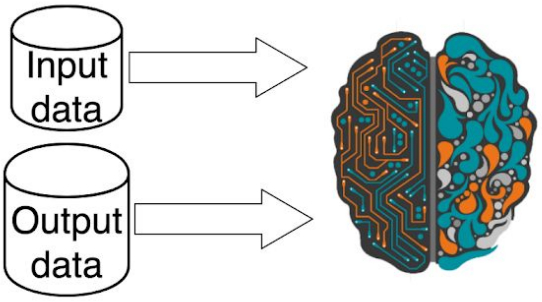


Then John collects a dataset with the output data:

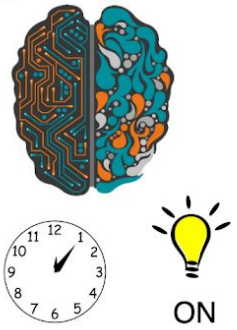
Timestamp	John wanted the light
12:30	on
12:45	off
13:00	on



Now it is machine learning time. John puts the datasets into the machine learning algorithm



The machine learning algorithm learned when you want to have the light on. Machine learning can now manage the light for John



4.3 Advantages and disadvantages

4.3.1 Advantages

- **Machine learning has the ability to learn connections in data that no human can ever make**
- Machine learning has the ability to learn itself to perform actions when certain conditions are met (the conditions when to use which action can be learned by machine learning)
- The performance is often higher than the older statistical models

4.3.2 Disadvantages

- The exact inner workings of the algorithms are a mystery. When a machine learning model is trained it is even almost impossible for mathematicians that created the algorithms to know what the model is doing
- Often requires very big datasets
- It is hard to validate the results
- The quality of the data needs to be extremely high

5 What can machine learning do?

Machine learning has more than one purpose and this makes it harder to define exactly what it can and can't do. That is why I collected a few example and abilities of machine learning. This way you can get a feel for what machine learning can do.

Note: Machine learning is always based on data. Without the necessary data or way to create the necessary data machine learning can't do anything.

- Learn connections between two datasets
- Predict future data values (example: predicting the weather of tomorrow)
- Create groups out of big unknown datasets (this is called clustering)
- Learn the path of victory (example: machine learning algorithm playing chess)
- Create recommendations based on your past interests
- Learn the connection between a message (like a twitter post) and the human emotion behind the message (happy, sad, etc)
- Detect items in images or videos. (face recognition, animal recognition, etc.)
- Learn to do actions that require small steps. This steps can not be longer than a second. (example: learning a robot to walk)
- Baselining data streams
- Detecting anomalies outside a baseline

6 What can machine learning NOT do?

Over a short period of time (it has been around a year since I got first introduced to machine learning) I heard quite a lot of comments about the usage of machine learning. This made it clear to me that people are not aware of the limitations of machine learning. These are some of the limitations of machine learning where you should be aware of.

- On the fly applying of machine learning without creating and researching datasets and machine learning models/algorithms
- Predict values to far ahead in the future
- Train machine learning on a limited dataset. You most cases you need A LOT of data
- Machine learning can only learn on repeatable data
- Start a business
- Give a human readable result without the interactions of humans. Humans are the only ones that can attach meaning to the output data
- **Without the right clean data a machine learning model can't do anything. Humans need to clean the data, remove biases and make sure that the data is high quality**
- Machine learning models/algorithms that are tuned and made for a dataset can NOT be used on another data set. The input data always needs to be identical to the data where the machine learning model is trained with

Note: this means that in theory machine learning can replace people that do short term repeatable work.

7 Types

There are three types of machine learning. You got supervised machine learning, unsupervised machine learning and reinforcement learning. In this chapter I will dive in into the different types. I will explain what the type does, when it gets used and what the up- and downside are.

7.1 Supervised machine learning

Supervised machine learning is the most used type of machine learning. Supervised machine learning is literally what the name says: “supervised”. This means that that you have control over the input AND the output of the model. But maybe you already noticed, in machine learning there is no such thing as full control.

7.1.1 Use cases examples

- Predicting stock prices
- Face recognition (we are given a face and answer who is it)
- Drug discovery (we are given a compound and we answer if it is a drug or not)
- Predicting for much money a user will spend in our shop based on his characteristic
- Predicting power consumption in the next month

7.1.2 Classification and regression

Supervised machine learning can be used in two modes. One is called classification and the other is called regression.

With classification you are using supervised machine learning to predict classifications. An example can be that you predict if a person is male or female based on their interests and hobbies. You classify the output data to be either male, female or any other gender that the person relates to.

With regression you are trying to predict values. For example you want to predict the the temperature in your house. Then you are trying to predict a value, in this case this value is temperature.

7.1.3 Advantages

- Supervised machine learning is powerful way to predict

7.1.4 Disadvantages

- Requires labeled data. This means that meaning of the data needs to be well defined

7.1.5 Read more about supervised machine learning

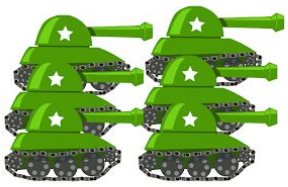
Supervised machine learning explained with linear regression as an example. (Some math is included)

- https://medium.com/machine-learning-for-humans/supervised-learning-740383a2feab

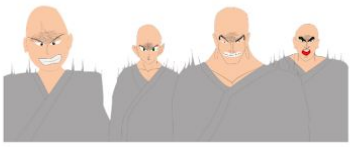
7.1.6 Example

Professor Daniel doesn't like people. He thinks they are too annoying

Because of this he wants to collect an army. An army to fight the world.

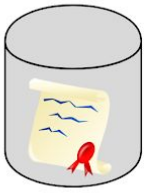


This army needs to be full with angry people



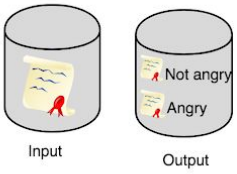
To recruit effieciently he wants to predict the if a person is angry based on his CV

Professor Daniel prepares a input dataset with all data from CV's



Then to make his output data he manually labels each person he got a CV from with angry or not angry

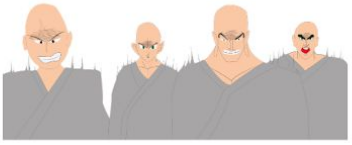
He now has an input and an output dataset



Now he will train the machine learning model

After training he has a model that predicts with 95% accuracy if a person is angry or not based on his CV

He can now efficiently recruit angry people by just putting CV's into the machine learning model



7.2 Unsupervised machine learning

With unsupervised machine learning we are letting go off even more control. Unsupervised machine learning does not even need a output dataset.

Unsupervised machine learning is a type of machine learning where you let go of your assumptions and supervision. The machine learning algorithm will interpret everything for you. This gives the machine learning algorithm the ability to find hidden structures in the datasets you give it.

7.2.1 Use cases examples

- Given set of images of stars, do they form some distinguishable types of stars?
- Given users activity on our website - are there distinguishable usage scenarios that we can find?
- We have record of a valid engine parameters and need a method to alarm as that it starts to behave "weird" (even though we do not know from the past what kind of "weird" we are looking for)
- We have recordings from camera of usual people behaviour, we want method to alarm as that "something unusual is happening" (without specifying what)
- We have set of high-dimensional data (like patients records) and want to visualize it (high dimensional data is a dataset with large number of attributes or features.)

7.2.2 Advantages

- Unsupervised machine learning learning is very good at finding patterns or groups in data

7.2.3 Disadvantages

- It is difficult to interpret the results

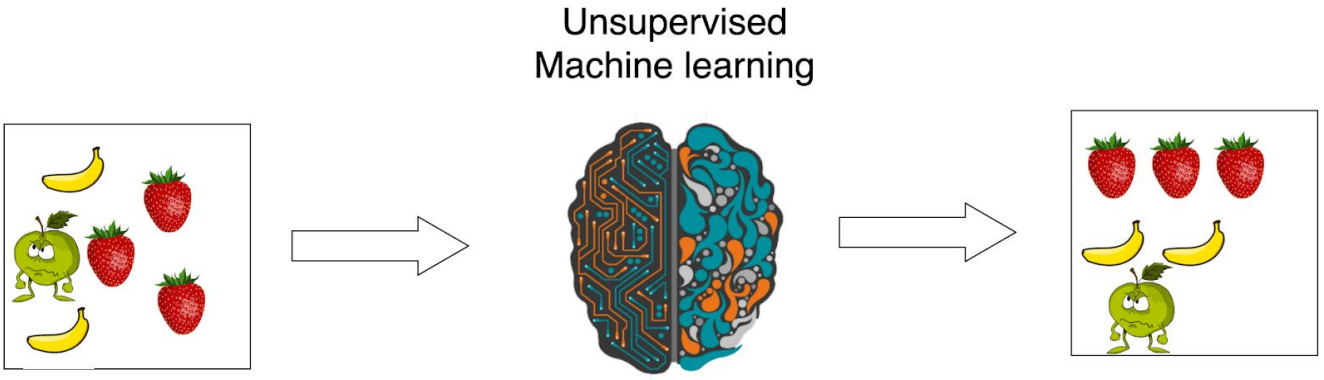
7.2.4 Clustering and dimension reduction

Unsupervised machine learning is mostly used for clustering or dimension reduction.

Clustering

Clustering is a technique to find groups in datasets. The machine learning algorithm will try to make sence of the dataset and make groups (clusters) of similar data groups.

Example



Remember that in real case scenarios you often don't really know what the clusters are. These clusters are machine defined clusters.

7.3 Read more about clustering

Another good explanation on subject of clustering.

- https://towardsdatascience.com/clustering-unsupervised-learning-788b215b074b

7.4 Dimension reduction

A dataset has features. These features are most often represented as columns in a table. Like this:

Name (feature 1)	Last name (feature 2)	City (feature 3)
John	Book	Amsterdam

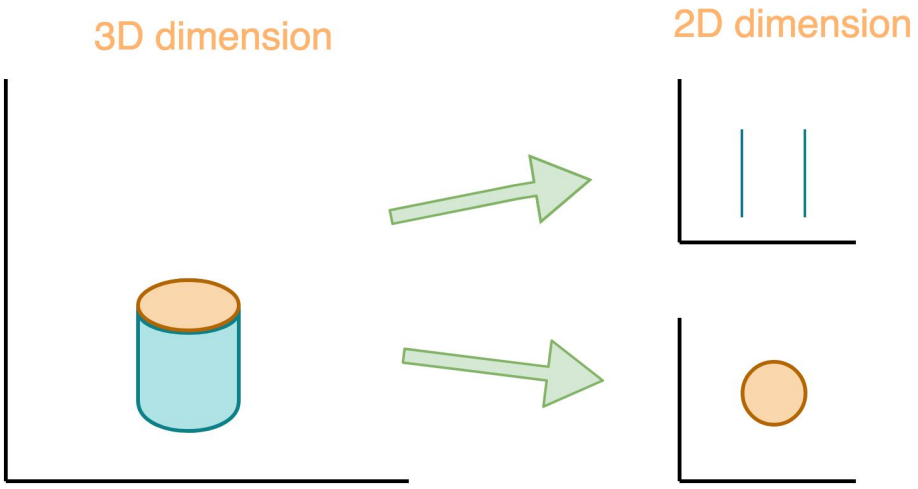
In this table there are three features. For simplicity sake we are gonna call these three features dimensions. So this will be a three dimensional dataset. With dimensional reduction we want to reduce the dimension. Most often this needs to be done to make data more interpretable.

To reduce the dimension of this dataset we can combine the name and last name. The dimension is now reduce from three dimensions to two dimensions. Now we have performed dimension reduction.

Name (feature 1 and 2)	City (feature 3)
John Book	Amsterdam

But in this little dataset that was most likely not needed but in the real world there are often scenarios where there are hundreds or thousands of features. Reducing the dimension in big datasets is problematic. Luckily unsupervised machine learning can be used to perform dimensional reduction.

Graphical Example



Read more about dimension reduction

More information about dimension reduction.

- https://www.quora.com/What-do-we-mean-by-high-dimensional-data
- https://data-warehouses.net/glossary/datadimensions.html
- https://en.wikipedia.org/wiki/Dimensionality_reduction
- https://www.quora.com/What-is-Dimension-reduction-in-machine-learning

7.5 Read more about supervised and unsupervised machine learning

Different explanations about the difference between supervised and unsupervised machine learning.

- https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms
- https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ (This website has some great books on machine learning)

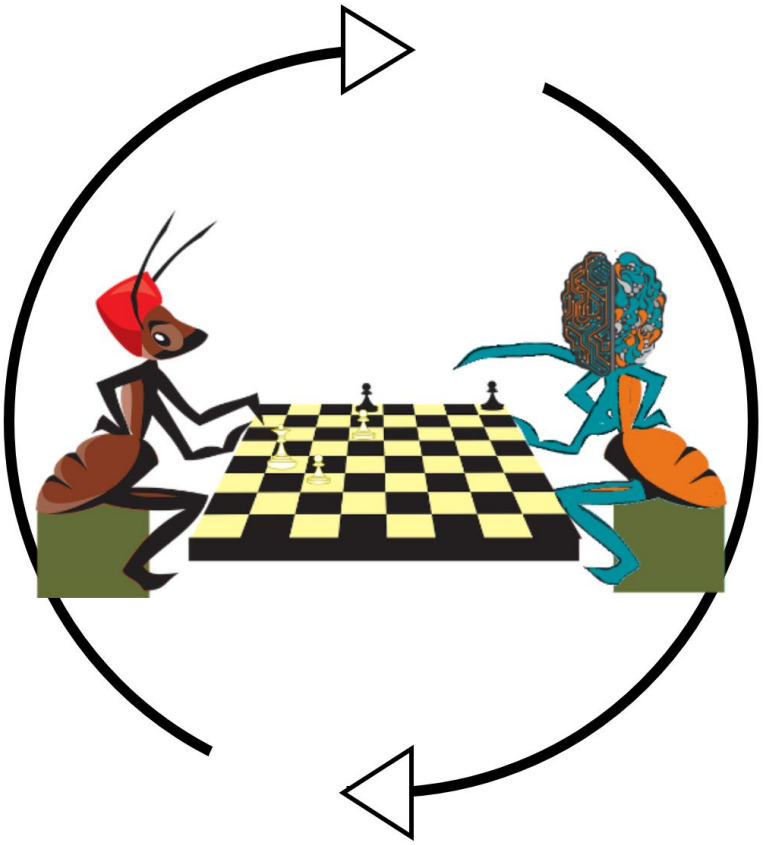
7.6 Reinforcement learning

“Reinforcement learning is a type of Machine Learning algorithms which allows software agents and machines to automatically determine the ideal behavior within a specific context, to maximize its performance.” (Rohit Akiwatkar, 2017)

This means that reinforcement learning is a type of machine learning that gives machines the power to find the “ideal behavior”.

This ideal behavior can be learned by setting a goal, a goal like winning a chess game. The machine learning algorithm will train until it finds the “ideal behavior” to win the game.

Reinforcement learning keeps playing the game over and over till the machine knows how to win



7.6.1 Use cases

- Learning a machine to play a game
- Learn a car to drive automatically
- Inventory management
- Personalised advertising
- Traffic Light Control
- Learning robots to walk

7.6.2 Advantages

- Learns to react on his environment

7.6.3 Disadvantages

- Requires a well set up environment
- Requires a very strict environment
- Requires a lot of resources (way more than other machine learning techniques)
- The machine learning algorithm needs to have a goal. This goal needs to be very specific, like winning a chess game

7.6.4 Example videos

Some videos of reinforcement learning in action.

- *https://www.youtube.com/watch?v=qv6UVOQ0F44*
- *https://www.youtube.com/watch?v=UVE0rxcffYo*
- *https://www.youtube.com/watch?v=0g9SlVdv1PY*

7.6.5 Read more

Read more about the use of reinforcement learning and what it can do.

- https://top.quora.com/What-is-reinforcement-learning
- https://www.analyticsvidhya.com/blog/2017/01/introduction-to-reinforcement-learning-implementation/

7.7 Deep learning

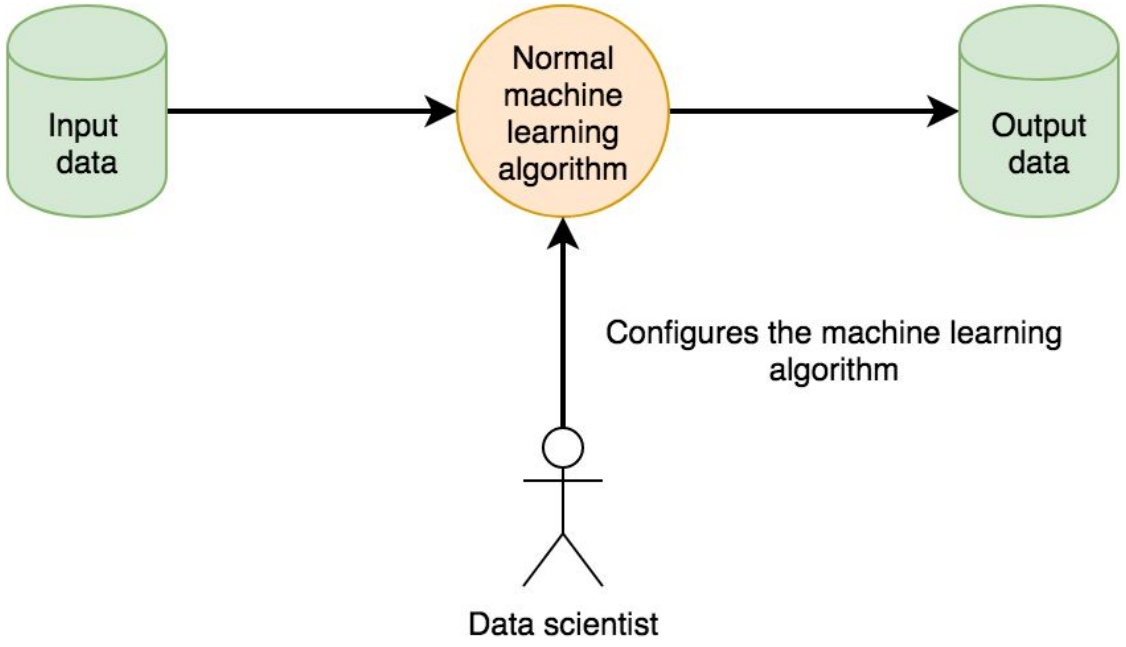
Deep learning can be used for supervised, unsupervised and reinforcement learning. So deep learning is not really a different machine learning. Still because of the differences between normal machine learning and deep learning often gets classified as a type.

Most often when people think about machine learning they think of this artificial brain. Normal machine learning is not really like an artificial brain. Deep learning on the other hand is very much based on the brain.

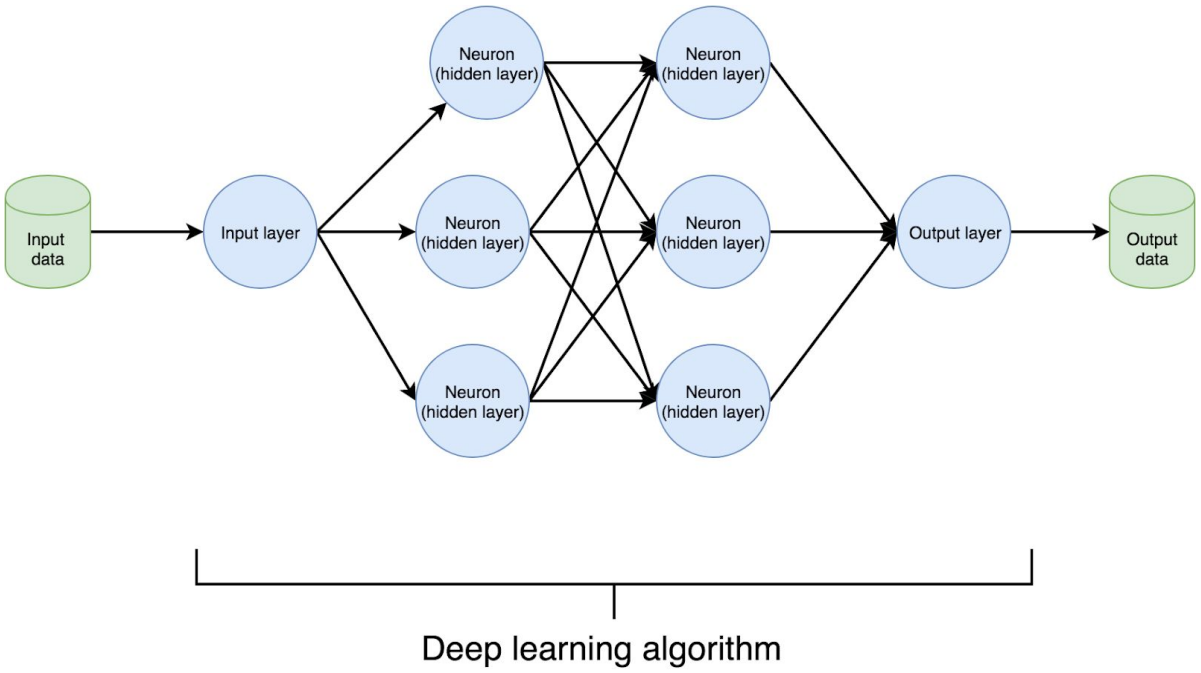
7.7.1 Understanding deep learning

To understand the power of deep learning you need to understand some inner workings of deep learning as well as some of the inner workings of normal machine learning. I will explain these inner workings with two images.

Normal machine learning



Deep learning



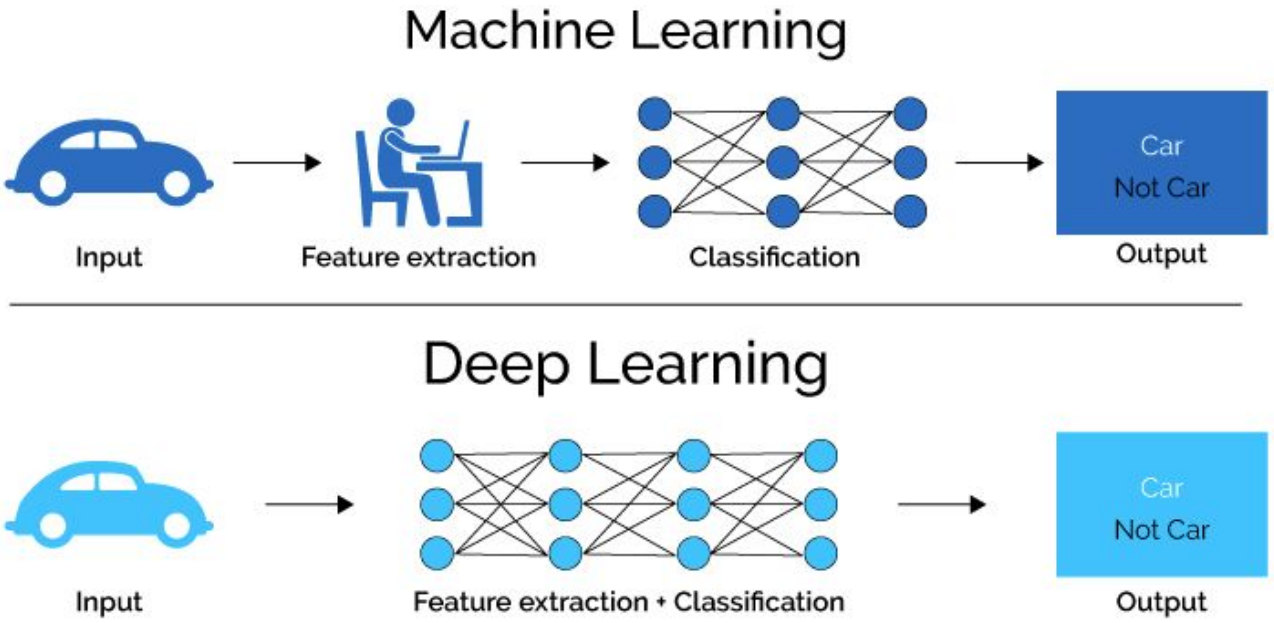
Understanding

If you look at the two images you can see a difference. With normal machine learning you need to configure the model. The algorithm configuration gives the user more control over the algorithm but at the same time leaves part of the job over to the data scientist.
Deep learning changes that. Deep learning takes even more control away from the data scientist.

Deep learning takes over with the brain like structure (the different layers with neuron) it has. This brain like structure makes the connections in the data.

Note: this does not mean that deep learning has the ability to do everything. Deep learning also needs to be setup, configured (different type of configuring, which we will not mention for simplicity sake) and tested

Another image (from a different perspective)



[What is the difference between deep learning and usual machine learning?]. Overgenomen van quora.com. https://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning door Krishna Srinivasan.

7.7.2 Use cases

- Can be used for supervised, unsupervised or reinforcement machine learning projects instead of normal machine learning to increase performance

7.7.3 Advantages

- Deep learning is very popular right night because of the increase results compared to normal machine learning models
- Deep learning has the ability to make better connection in datasets than normal machine learning
- Faster results. Using deep learning models can decrease the development time in een machine learning project because feature extraction can be skipped (depends on the situation)

7.7.4 Disadvantages

- Very prone to overfitting (Explanation of overfitting: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/)
- More resource intensive
- There is even less control over the machine learning algorithm

7.7.5 More deep learning

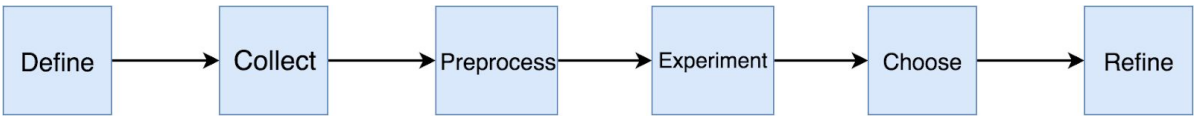
More explanations about deep learning.

- https://www.youtube.com/watch?v=vOppzHpvTiQ
- https://www.kdnuggets.com/2015/01/deep-learning-explanation-what-how-why.html

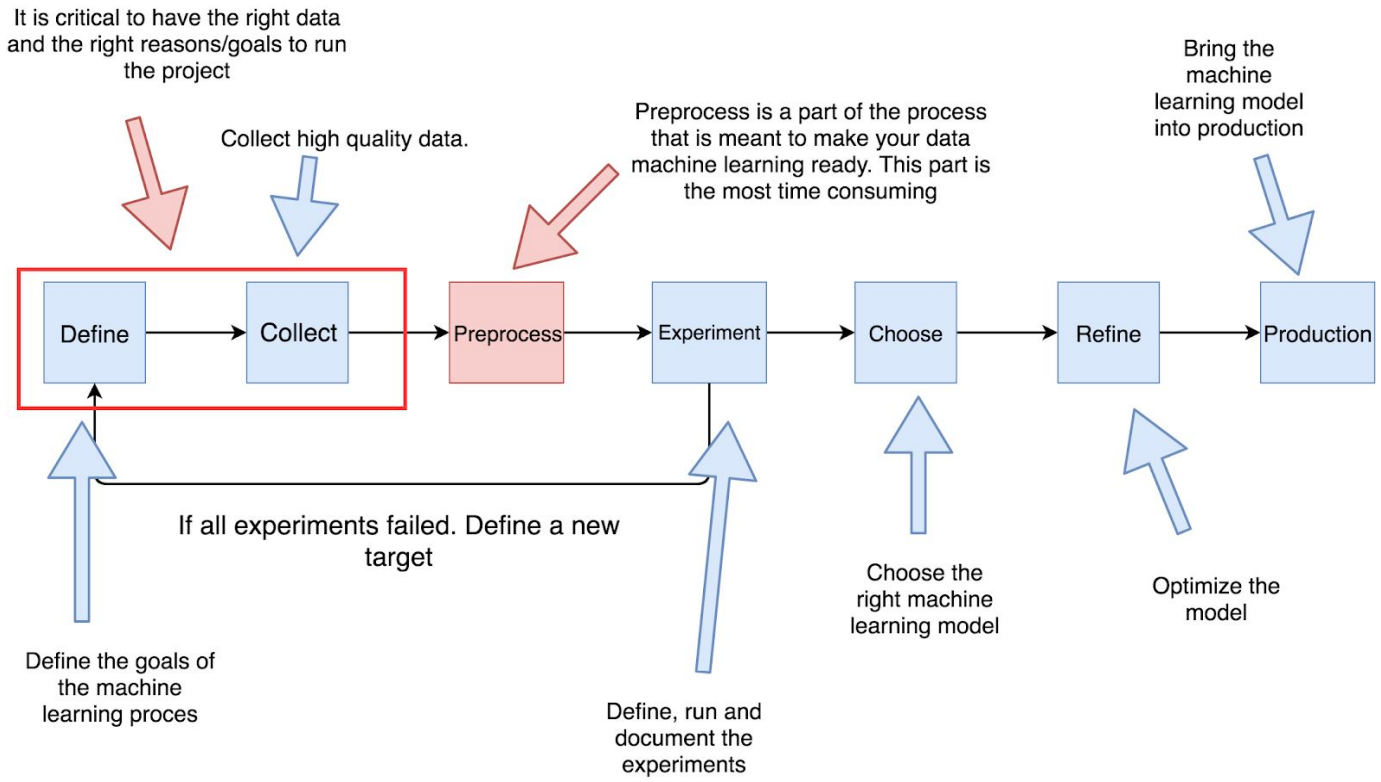
8 Process

To work with machine learning projects it is important to understand the processes of a machine learning project. These processes can differ a lot in different scenarios but it is still important to cover the main processes involved with machine learning.

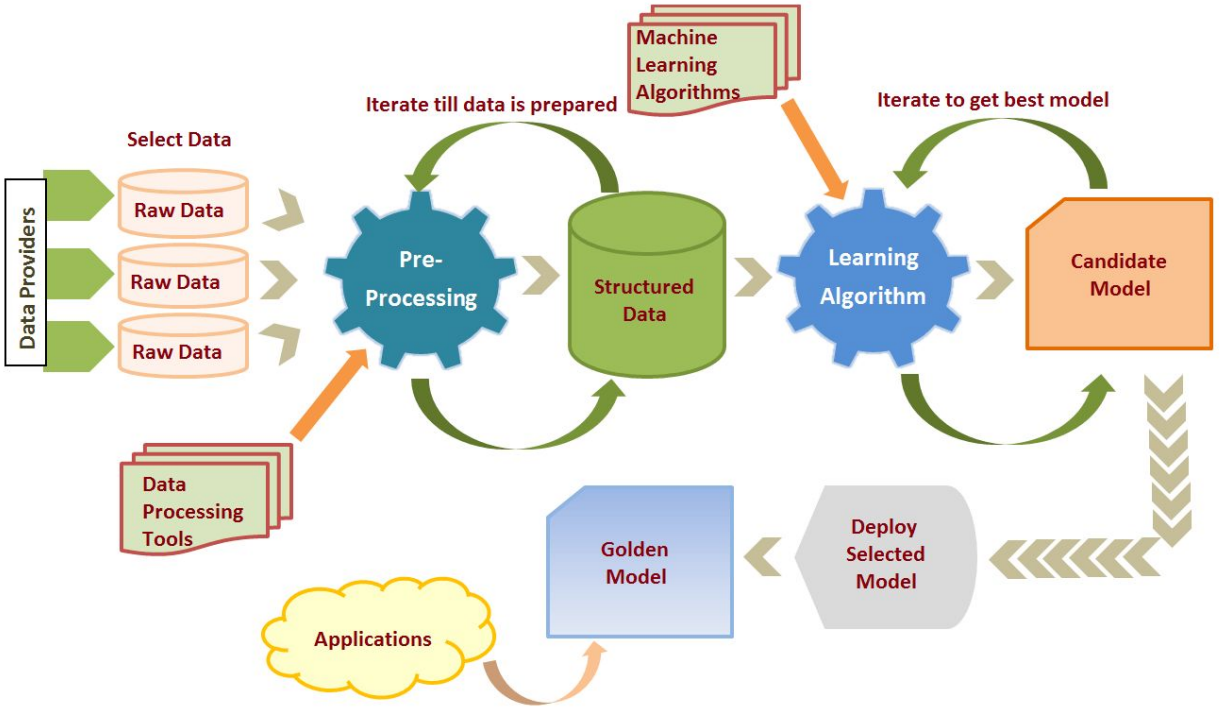
8.1 The main process



8.2 The main process (explained)



8.3 The technical view



[Machine Learning Process And Scenarios]. Overgenomen van elearningindustry.com. https://elearningindustry.com/machine-learning-process-and-scenarios door Akhil Mittal.

8.4 More about machine learning processes

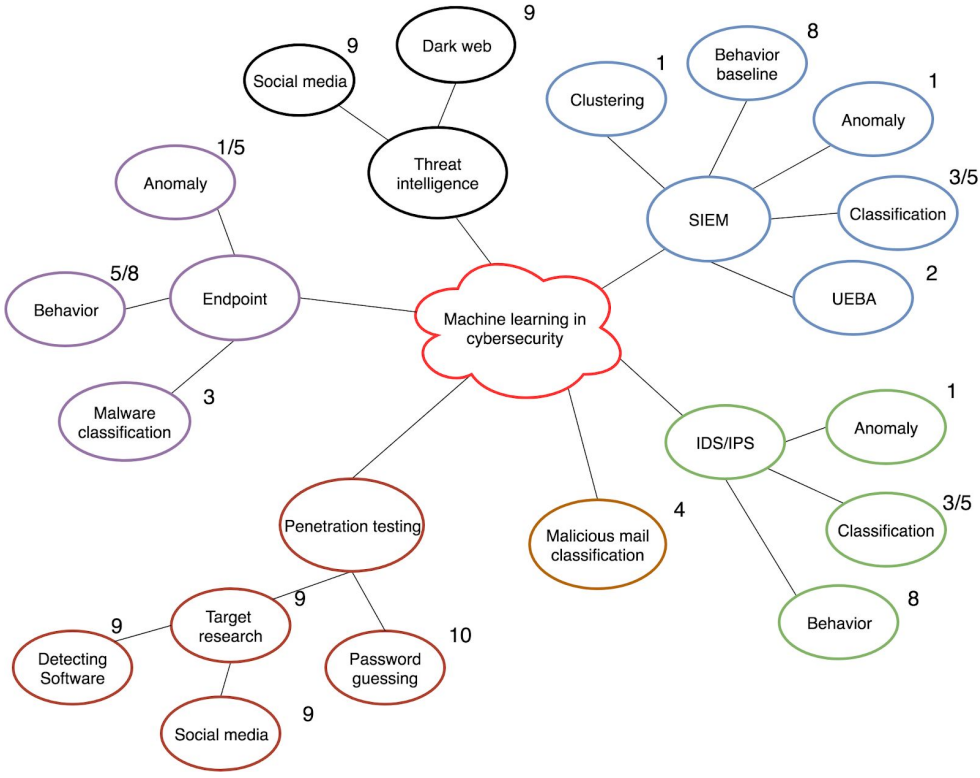
Read more about the different ways of viewing the machine learning processes.

- https://medium.com/sigmoidal/machine-learning-development-process-youve-got-it-wrong-396270e653f4
- https://elearningindustry.com/machine-learning-process-and-scenarios
- https://machinelearningmastery.com/process-for-working-through-machine-learning-problems/

9 Cybersecurity

9.1 Use cases

1. Anomaly detection
2. User behavior analytics
3. Malware detection
4. Spam mail detection
5. Host intrusion detection
6. Endpoint protection
7. Network intrusion detection
8. Behavior baselining
9. Information gathering
10. Attack tools



9.2 Use case examples

- Machine learning learns when an email is spam and when it is not
- Machine learning learns when a file is malicious and when it is not
- Detecting malicious network traffic based on anomalies in the network
- User activity baselining. Detect users that are acting out of their baseline, maybe they are compromised or have malicious intent
- Machine learning learns the behavior of malicious network traffic
- Finding patterns in network data

9.3 Advantages

What are the advantages of using machine learning in a cyber security environment.

- M**achine learning has the ability to detect dynamical instead of static**
- Machine learning has the ability to process way more information than human analysts are able to do
- Machine learning has the ability to detect patterns and characteristics of data that humans can not detect

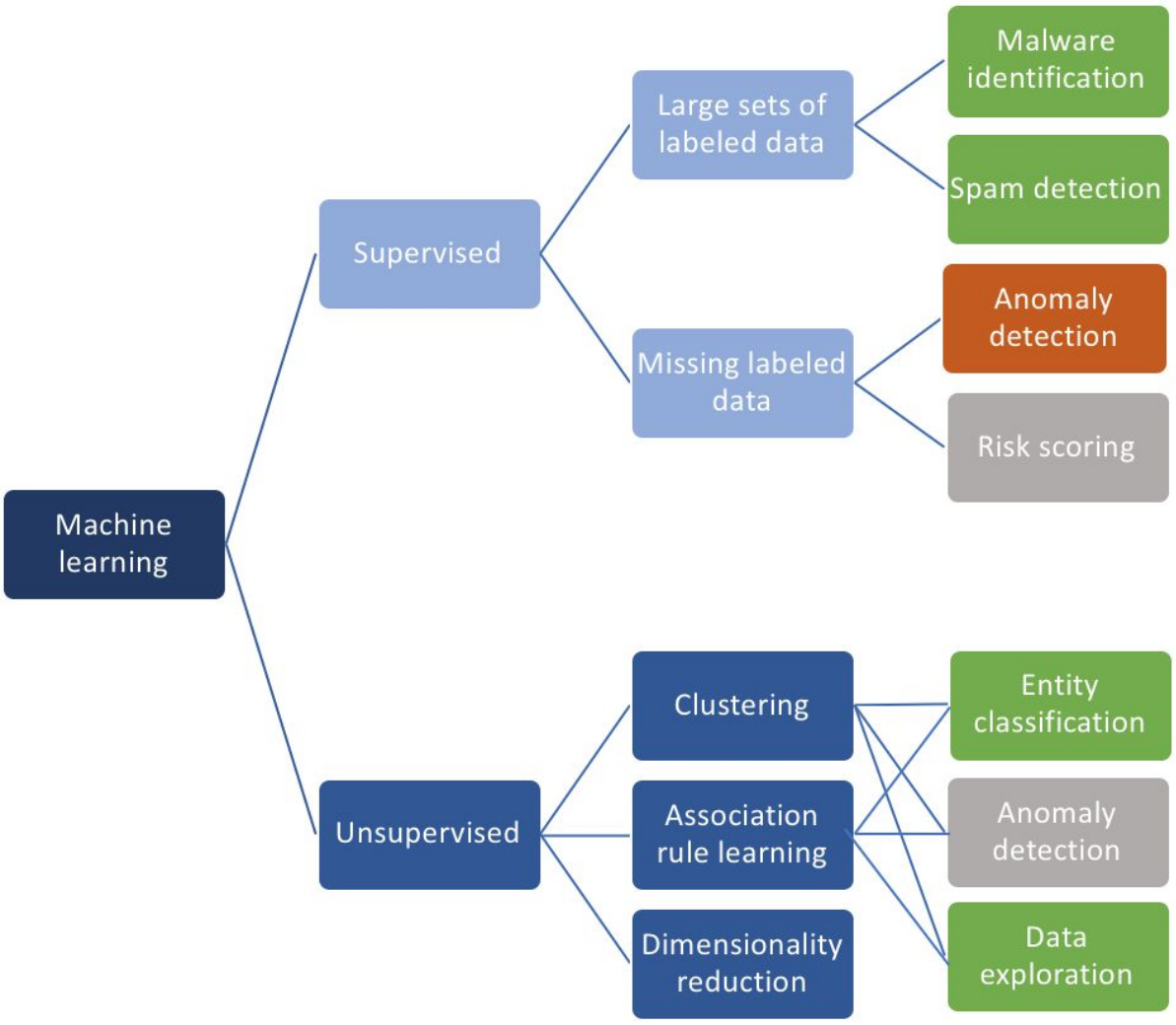
9.4 Disadvantages

What are the disadvantages of using machine learning in a cyber security environment.

- The false positive rate has to be extremely low
- Labeling of data is very expensive. Since it requires expensive security experts
- Security data is very dynamic. Technologies change very fast!
- Environment specific. When a machine learning model is developed for one environment it does not work in another environment with the same data types. Each environment is different
- It is hard to interpret the machine learning output

9.5 Machine learning types used

This is an example on which machine learning types can be used on the use cases.



[AI and Machine Learning in Cyber Security]. Overgenomen van towardsdatascience.com. https://towardsdatascience.com/ai-and-machine-learning-in-cyber-security-d6fbee480af0 door Raffael Marty.

9.6 Read more

9.6.1 Pentesting

Some of example pentesting tools that use machine learning.

- https://www.kitploit.com/2016/11/deep-pwning-metasploit-for-machine.html
- https://www.kitploit.com/2018/05/gyoithon-growing-penetration-test-tool.html

9.6.2 Papers

Papers about the use of machine learning in a cyber security environment.

- https://www.researchgate.net/profile/Michael_Atighetchi/publication/281558896_Use_of_Machine_Learning_in_Big_Data_Analytics_for_Insider_Threat_Detection/links/55edf38c08ae0af8ee19dedb.pdf
- http://www.cse.psu.edu/~trj1/cse543-f16/docs/Axelsson.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7307098

10 Tools

The goal of this chapter is to give an insight into the different components needed for the building of machine learning models. These components will be enhanced with an overview of the software that is currently available.

10.1 Developing

There are a lot of different tools on the market that can be used to develop machine learning models. Every tool has there advantages and disadvantages. Because of this professionals may use a set of tools instead of just one. An developing tool is a tool that can be used to build, program or define machine learning models. These are the tools where are the machine learning magic is happening

In the table below you see an comparison between the different tools:

Tool	Graphical Interface	Program-ming	Community support	Business support	Ease of use	Amount of features	Ability to manipulate data	Price	Comment
R		X	High	None	Low	High	High	None	R can only be used for data science
Python		X	High	None	Medium	High	High	None	Most popular. Has the best machine learning modules in the market
Splunk	X		Medium	High	High	Medium	High	High	
MLjar	X		Low	High	High	Medium	Low	Low	Limited dataset size
Weka	X		Medium	None	Medium	Medium	Low	None	
Julia		X	Medium	None	Low	Medium	High	None	
Java		X	Medium	None	Low	High	High	None	Battle proven in production environments
Orange	X		Medium	None	High	Medium	Low	None	Drag and drop machine learning
Scala		X	Medium	None	Low	Medium	Low	None	
Matlab		X	High	Medium	Low	High	High	Low	
Microsoft Azure	X		Medium	High	High	Medium	High	Medium	Drag and drop machine learning
ELKI	X		Low	None	Medium	Low	Low	None	
KNIME	X		Medium	None	High	Medium	Low	None	Drag and drop machine learning

10.1.1 Programming vs Graphical

- Programming languages have more option for customizability
- There is more online help and tutorials for the use of machine learning in the programming languages
- Graphical interfaces have pre-build machine learning algorithms, preprocessing methods and performance statistics
- You often need to buy a licence to use the graphical interfaces

10.1.2 How to choose a tool

To find the right tool you need to do some research on your own. The table above will narrow the direction but without a good look into the exact features of each tool you won't succeed in choosing the right tool.

To choose a tool for development you need to pay attention to the following variables:

- Ease of use
- Integration with/into other systems
- Skill in programming
- Ease of documentation. Graphical tools often have an easy way to document what is happening. (Look at the tools that have a drag and drop interface)
- Time, how fast does the model needs to be developed
- Performance, what are the performance criteria? Aiming for an higher performance tends to be more expensive in money as well in time. Is the performance increase worth the effort?
- Features. Some tools have specialized features and some are decent at everything
- Support from the community. Better community support eases the development a lot
- Support from the business. Professional support has better quality and can take away responsibility over some problems in the developing process
- The ability to scale

10.2 Databases

Machine learning does not work without data. This data may be stored in various data sources. Databases are the most commonly used data source in an production environment. In some cases for the sake of prototyping datasets are created in the form of csv files or other files formats.

Since databases are the most used data source in machine learning I will go into the two different types of databases and how they affect machine learning.

10.2.1 Structured

Structured databases guarantee data structure. Which means that columns in datasets can not chance and it is possible to make relations between tables. They are good at keeping structure in your data but suffer in performance.

Examples

- Mysql
- PostgreSQL
- Sqlite
- Microsoft SQL server
- Oracle

More structured databases

- https://en.wikipedia.org/wiki/List_of_relational_database_management_systems

10.2.2 Unstructured

Unstructured databases do not guarantee data structure. The mapping of the data gets done automatically. Unstructured databases have better performance and deal better with unstructured data.

- MongoDB
- Elasticsearch
- HBbase
- Cassandra
- Neo4j
- MariaDB

More unstructured databases

- https://en.wikipedia.org/wiki/NoSQL

10.2.3 Designed for machine learning

There are some databases specifically designs for machine learning to increase the performance of machine learning.

Example

- MLDB

10.3 Processing

Machine learning requires a lot of processing power. Especially on very big data there needs to be a good infrastructure support the processing power needed by machine learning. There are a few ways to handle the infrastructure.

10.3.1 Tools

Some machine learning developing tools have the necessary processing power build into the tool. Splunk is one example. No extra software is needed to support machine learning with the needed processing power.

10.3.2 Cloud

Cloud providers that offer machine learning services provide the infrastructure to handle the amount of processing power needed. No extra setup is required.

Some examples of cloud providers with machine learning services:

- Google cloud
- Amazon
- Microsoft azure
- IBM watson
- BigML

10.3.3 Infrastructure

When you setup your own environment you need to create an environment that can handle data processing in parallel. Parallel processing is the key of handling big data. Currently Hadoop is the industry standard.

Examples of technologies that can handle parallel data processing:

- Hadoop
- Apache Storm
- Apache Flink
- Apache Mesos
- Pachyderm

Good explanation of Hadoop

- https://www.youtube.com/watch?v=DCaiZq3aBSc

11 References

Innoarchitech.com. Geraadpleegd op 2018, 25 oktober, van https://www.innoarchitech.com/machine-learning-an-in-depth-non-technical-guide/

Kdnuggets.com. Geraadpleegd op 2018, 04 oktober, van https://www.kdnuggets.com/2015/01/deep-learning-explanation-what-how-why.html

Kdnuggets.com. Geraadpleegd op 2018, 25 oktober, van https://www.kdnuggets.com/2017/01/machine-learning-cyber-security.html

Machinelearningmastery.com. (2016, 20 maart). Overfitting and Underfitting With Machine Learning Algorithms. Geraadpleegd op 2018, 04 oktober, van https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Mariusz Kierski. (2017, 3, March). Machine Learning development process—you've got it wrong. Geraadpleegd op 2018, 05 oktober, van https://medium.com/sigmoidal/machine-learning-development-process-youve-got-it-wrong-396270e653f4

Machinelearningmastery.com. (2014, 11 februari). Applied Machine Learning Process. Geraadpleegd op 2018, 05 oktober, van https://machinelearningmastery.com/process-for-working-through-machine-learning-problems/

Raffael Marty. (2018, 01, January). AI and Machine Learning in Cyber Security. Geraadpleegd op 2018, 08 oktober, van https://towardsdatascience.com/ai-and-machine-learning-in-cyber-security-d6fbee480af0

Researchgate.net. Geraadpleegd op 2018, 09 oktober, van https://www.researchgate.net/profile/Michael_Atighetchi/publication/281558896_Use_of_Machine_Learning_in_Big_Data_Analytics_for_Insider_Threat_Detection/links/55edf38c08ae0af8ee19dedb.pdf

Psu.edu. Geraadpleegd op 2018, 09 oktober, van http://www.cse.psu.edu/~trj1/cse543-f16/docs/Axelsson.pdf

Ieee.org. Geraadpleegd op 2018, 09 oktober, van https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=

Alexandra Leslie. (2018, 01, June). What is Hadoop Good For? (Best Uses, Alternatives, & Tools). Geraadpleegd op 2018, 25 oktober, van https://www.hostingadvice.com/how-to/what-is-hadoop/#improper-hadoop-uses

Cmu.edu. Geraadpleegd op 2018, 25 oktober, van https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=527973