# Introduction to Machine Learning

*Facebook: Inna Lougoum*

# Definition

At a high level, machine learning is a process in which the computer solves problems and makes decisions in a similar way that humans do.

*Facebook: Inna Lougoum*

# Is Machine Learning easy ?

Machine learning is easy! You do not need to have a heavy math knowledge or a heavy programming background to understand it. What you need is common sense, a good visual intuition, and a desire to learn and to apply these methods to anything that you are passionate about and where you want to make an improvement in the world

## Is Machine learning everywhere ?

Anywhere there is a job that requires repetition, that requires looking at data and gathering conclusions, machine learning can help.

Just to name a few applications of machine learning:recommendation systems, image recognition, text processing, self-driving cars, spam recognition, anything.

## Is it hard ?

Machine learning requires imagination, creativity, and a visual mind. This is all. It helps a lot if we know mathematics, but the formulas are not required. It helps if we know how to code, but nowadays, there are many packages and tools that help us use machine learning with minimal coding.

All you need is an idea of how to apply it to something, and some knowledge about how to handle data.

*Facebook: Inna Lougoum*

# But what exactly is machine learning ?

If we wanted to make a computer perform a task, we had to write a program, namely, a whole set of instructions for the computer to follow. This is good for simple tasks, but how do we get a computer to, for example, identify what is on an image? For example, is there a car on it, is there a person on it. For these kind of tasks, all we can do is give the computer lots of images, and make it learn attributes about them, that will help it recognize them. This is machine learning, it is teaching computers how to do something by experience, rather than by instructions. It is the equivalent of when, as humans, we take decisions based on our intuition, which is based on previous experience. In a way, machine learning is about teaching the computer how to think like a human.

***Facebook: Inna Lougoum***

# How humans and computers make decisions ?

Machine learning is about computers making decisions based on experience. In the same way that humans make decisions based on previous experiences, computers can make decisions based on previous data. The rules computers use to make decisions are called models.

## What is Data ?

Data is simply information. Any time we have a table with information, we have data. Normally, each row is a data point.

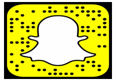
Let’s say, for example, that we have a dataset of users for recommending them Apps. In this case, each row represents a different user. Each user is described then, by certain features.

## What is features ?

Features are simply the columns of the table. In our recommendation example, the features may be gender, age etc. This is what describes our data. Some features are special, though, and we call them labels like Apps in our example.

*Facebook: Inna Lougoum*

# Recommending Apps

| Gender | Age | App |
|---|---|---|
| F | 15 | |
| F | 25 | |
| M | 32 |  |
| F | 40 | |
| M | 12 | |
| M | 14 | |

Features

Labels

first datapoint

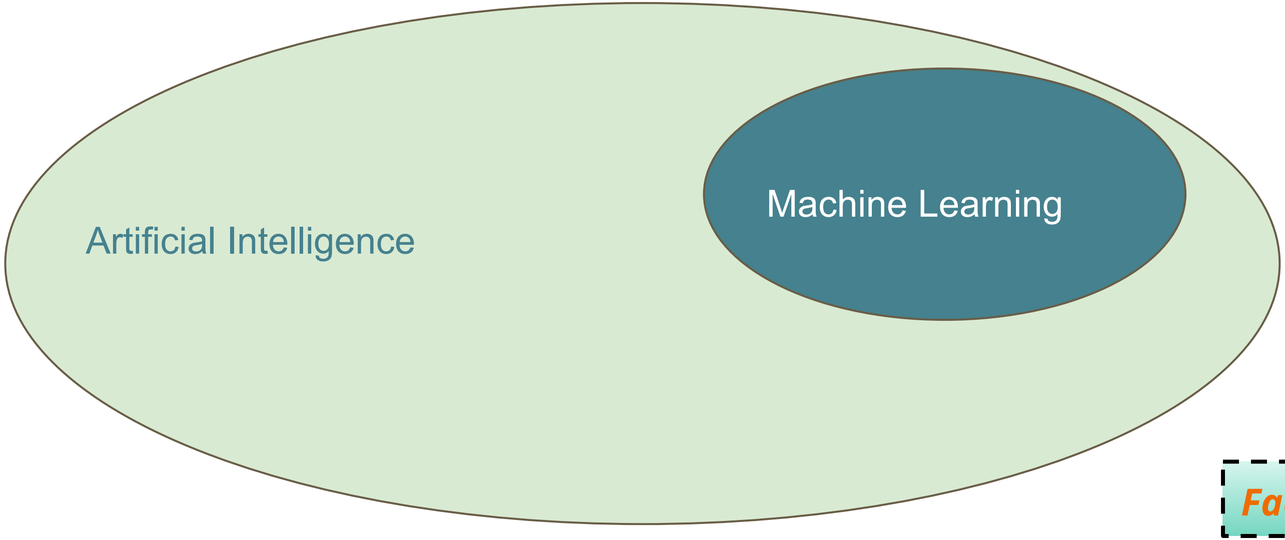
second datapoint

third datapoint

***Facebook: Inna Lougoum***

# What is the difference between artificial intelligence and machine learning?

First thing, machine learning is a part of artificial intelligence. So anytime we are doing machine learning, we are also doing artificial intelligence.



*Facebook: Inna Lougoum*

# how to teach the computer to make decisions ?

When we think of how to teach the computer to make decisions, we first think of how we as human make decisions. There are mainly two ways we use to make most decisions:

1. By using reasoning and logic
2. By using our experience.

Artificial intelligence is the name given to the process in which the computer makes decisions, mimicking a human. So in short, points 1 and 2 form artificial intelligence.

Machine learning is when we only focus on point 2. Namely, when the computer makes decisions based on experience. And experience has a fancy term in computer lingo: data. Thus, machine learning is when the computer makes decisions, based on previous data.

*Facebook: Inna Lougoum*

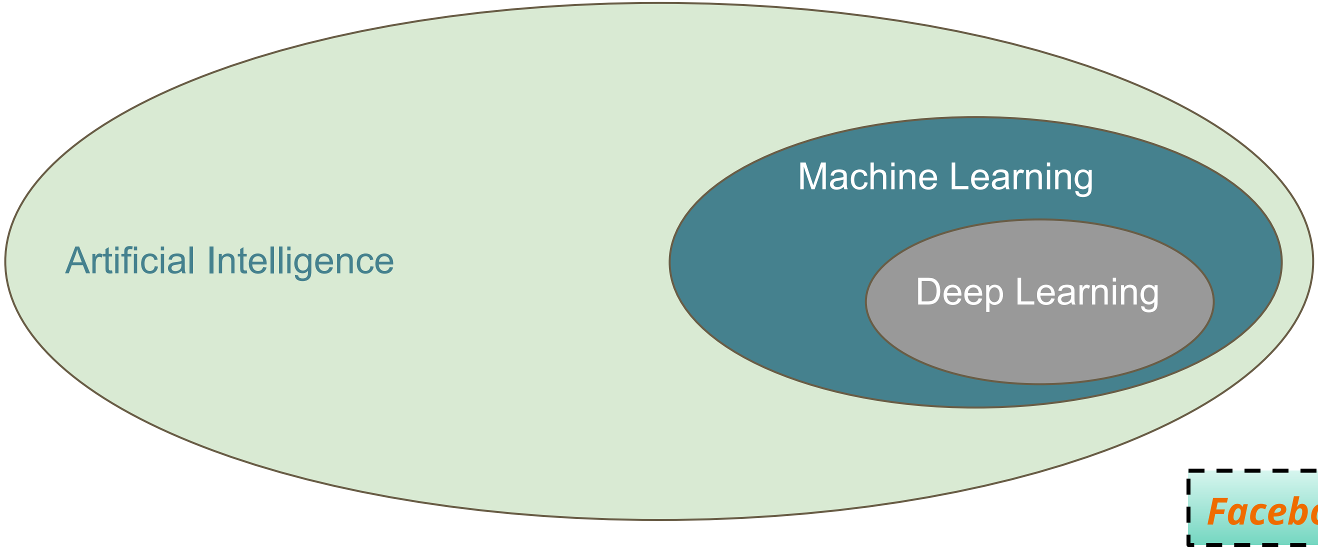
A small example would be how Google maps finds a path between point A and point B. There are several approaches, for example the following:

1. Looking into all the possible roads, measuring the distances, adding them up in all possible ways, and finding which combination of roads gives us the shortest path between points A and B.
2. Watching many cars go through the road for days and days, recording which cars get there in less time, and finding patterns on what their routes where.

As you can see, approach 1 uses logic and reasoning, whereas approach 2 uses previous data. Therefore, approach 2 is machine learning. Approaches 1 and 2 are both artificial intelligence.

# What about deep learning?

Deep learning is the most commonly used type of machine learning. The reason is simply that it works really well.This term applies to every type of machine learning that uses Neural Networks.So in other words, deep learning is simply a part of machine learning, which in turn is a part of artificial intelligence.



*Facebook: Inna Lougoum*

# How do humans think?

When we humans need to make a decision based on our experience, we normally use the following framework:

1. We remember past situations that were similar.
2. We formulate a general rule.
3. We use this rule to predict what will happen if we take a certain decision.

For example, if the question is: “Will it rain today?”, the process to make a guess will be the following:

1. We remember that last week it rained most of the days.
2. We formulate that in this place, it rains most of the time.
3. We predict that today it will probably rain.

We may be right or wrong, but at least, we are trying to make an accurate prediction.

*Facebook: Inna Lougoum*

# Example 1: An annoying email friend

We have a friend called Almardi, who likes to send us a lot of email. In particular, a lot of his emails are spam, in the form of chain letters, and we are starting to get a bit annoyed at him. It is Saturday, and we just got a notification of an email from him. Can we guess if it is spam or not without looking at the email?

SPAM AND HAM :

Spam is the common term used for junk or unwanted email, such as chain letters, promotions, and so on.

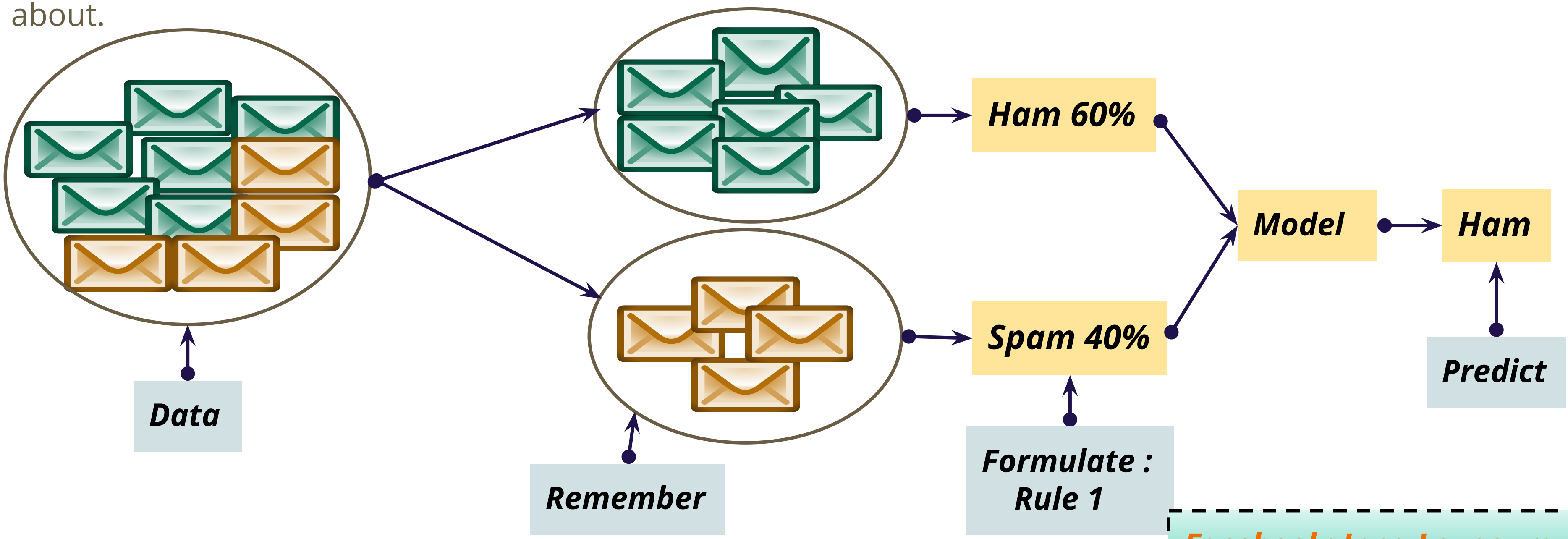
the term 'ham' is used to refer to non-spam emails.

For this, we use the remember-formulate-predict method.

First let us remember, say, the last 10 emails that we got from Amardi. We remember that 4 of them were spam, and the other 6 were ham. From this information, we can formulate the following rule:

# A very simple machine learning model

Rule 1: 4 out of every 10 emails that Almardi sends us are spam. This rule will be our model. Note, this rule does not need to be true. our prediction may be wrong. We may open the email and realize that it is spam. But we have made the prediction to the best of our knowledge. This is what machine learning is all about.



*Facebook: Inna Lougoum*

# Can we do better ?

But you may be thinking, 6 out of 10 is not enough confidence on the email being spam or ham, can we do better? Let's try to analyze the emails a little more. Let's see when Almardi sent the emails to see if we find a pattern.

*Facebook: Inna Lougoum*

# Example 2: A seasonal annoying email friend

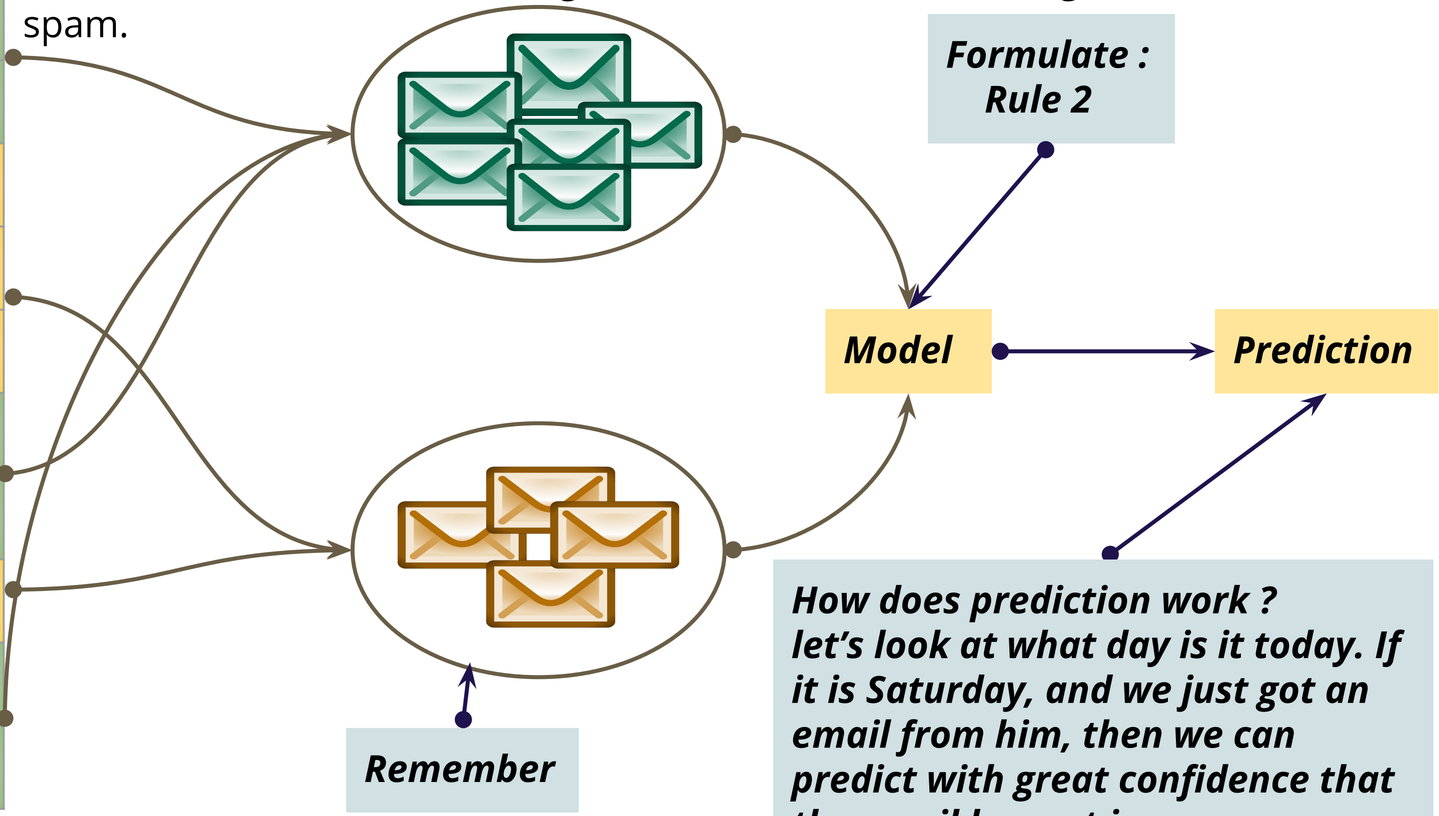
Let us look more carefully at the emails that Almardi sent us in the previous month. Let's look at what day he sent them. Here are the emails with dates, and information about being spam or ham:

Now things are different. Can you see a pattern? It seems that every email Almardi sent during the week, is ham, and every email he sent during the weekend is spam. This makes sense, maybe during the week he sends us work email, whereas during the weekend, he has time to send spam, and decides to roam free. So, we can formulate a more educated rule:

A slightly more complex machine learning model, done by a human ***Rule 2:*** Every email that Almardi sends during the week is ham, and during the weekend is spam.

| ***Day*** | ***Email*** |
|---|---|
| Monday | Ham |
| Tuesday | Ham |
| Saturday | Spam |
| Sunday | Spam |
| Sunday | Spam |
| Wednesday | Ham |
| Friday | Ham |
| Saturday | Spam |
| Tuesday | Ham |
| Thursday | Ham |

***Data***



***How does prediction work ? let’s look at what day is it today. If it is Saturday, and we just got an email from him, then we can predict with great confidence that the email he sent is spam***

***Facebook: Inna Lougoum***

Let’s give things names, in this case, our prediction was based on a feature. The feature was the day of the week, or more specifically, it is being a weekday or a day in the weekend. You can imagine that there are many more features that could indicate if an email is spam or ham. Can you think of some more? In the next paragraphs we’ll see a few more features.

# Example 3: Things are getting complicated!

Now, let's say we continue with this rule, and one day we see Almardi in the street, and he says "Why didn't you come to my wedding party?" We have no idea what he is talking about. It turns out last Sunday he sent us an invitation to his wedding party, and we missed it! Why did we miss it, because he sent it on the weekend. It seems that we need a better model. So let's go back to look at Almardi's emails, in the following table, this is our remember step. Now let's see if you can help me find a pattern.

What do we see? It seems that the large emails tend to be spam, while the smaller ones tend to not be spam. This makes sense, since maybe the spam ones have a large attachment.So, we can formulate the following rule:

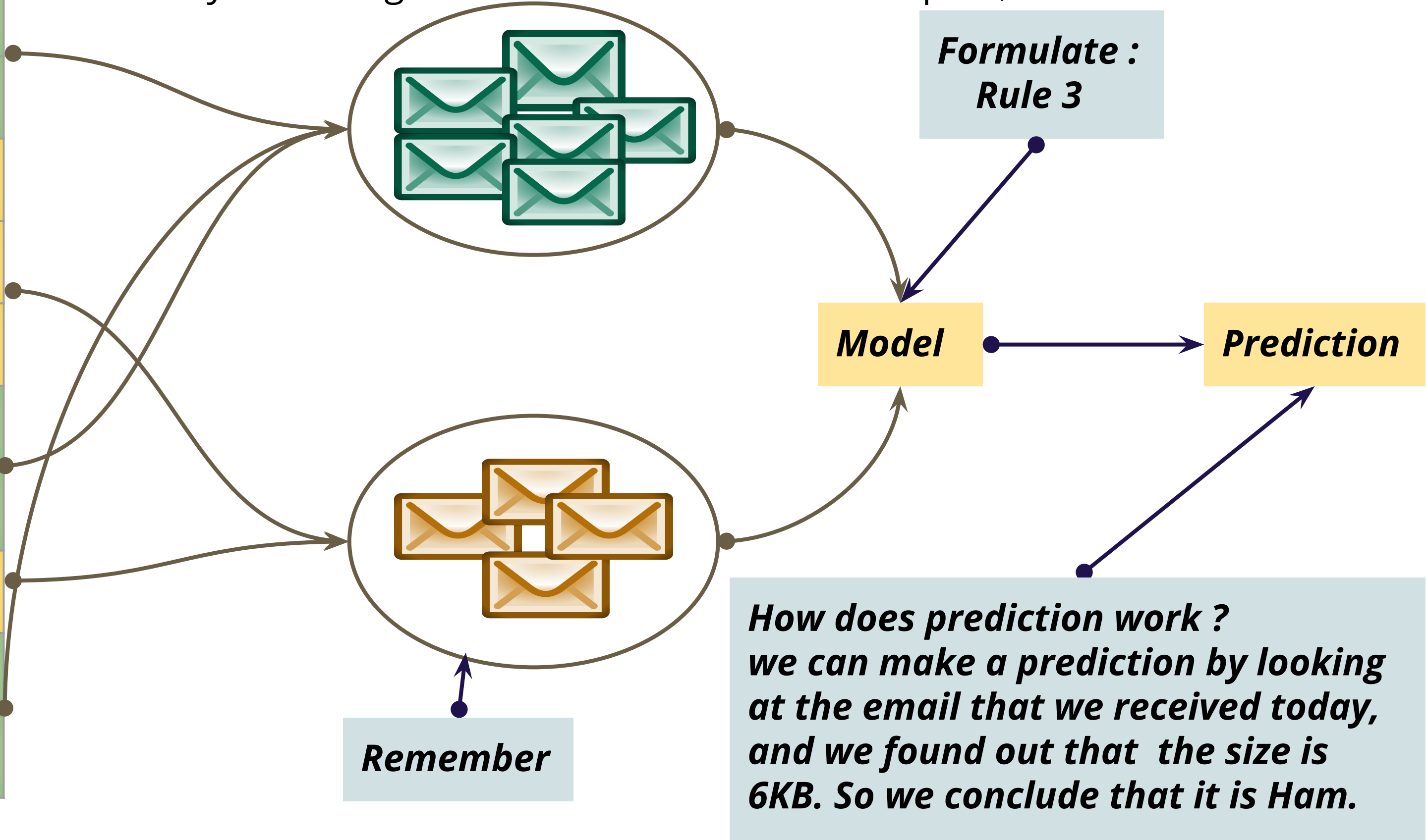
*Facebook: Inna Lougoum*

Another slightly more complex machine learning model, done by a human

***Rule3:*** Any email larger of size more than 12 KB is spam, otherwise it is ham.

| ***Size*** | ***Email*** |
|---|---|
| 1 KB | Ham |
| 12 KB | Ham |
| 16 KB | Spam |
| 20 KB | Spam |
| 18 KB | Spam |
| 3 KB | Ham |
| 5 KB | Ham |
| 25 KB | Spam |
| 1 KB | Ham |
| 3 KB | Ham |

***Data***



***Formulate : Rule 3***

***Model***

***Prediction***

***Remember***

***How does prediction work ?***
***we can make a prediction by looking at the email that we received today, and we found out that the size is 6KB. So we conclude that it is Ham.***

***Facebook: Inna Lougoum***

# Example 4: More?

Our two classifiers were good, since they rule out large emails and emails sent on the weekends. Each one of them uses exactly one of these two features. But what if we wanted a rule that worked with both features? Rules like the following may work:

1. ***Rule 4:*** If an email is larger than 10KB or it is sent on the weekend, then it is classified as spam. Otherwise, it is classified as ham.
2. ***Rule 5:*** If the email is sent during the week, then it must be larger than 15KB to be classified as spam. If it is sent during the weekend, then it must be larger than 5KB to be classified as spam. Otherwise, it is classified as ham.Or we can even get much more complicated.
3. ***Rule 6:*** Consider the number of the day, where Monday is 0, Tuesday is 1, Wednesday is 2, Thursday is 3, Friday is 4, Saturday is 5, and Sunday is 6. If we add the number of the day and the size of the email (in KB), and the result is 12 or more, then the email is classified as spam. Otherwise, it is classified as ham.

All of these are valid rules. And we can keep adding layers and layers of complexity. Now the question is, which is the best rule? This is where we may start needing the help of a computer.

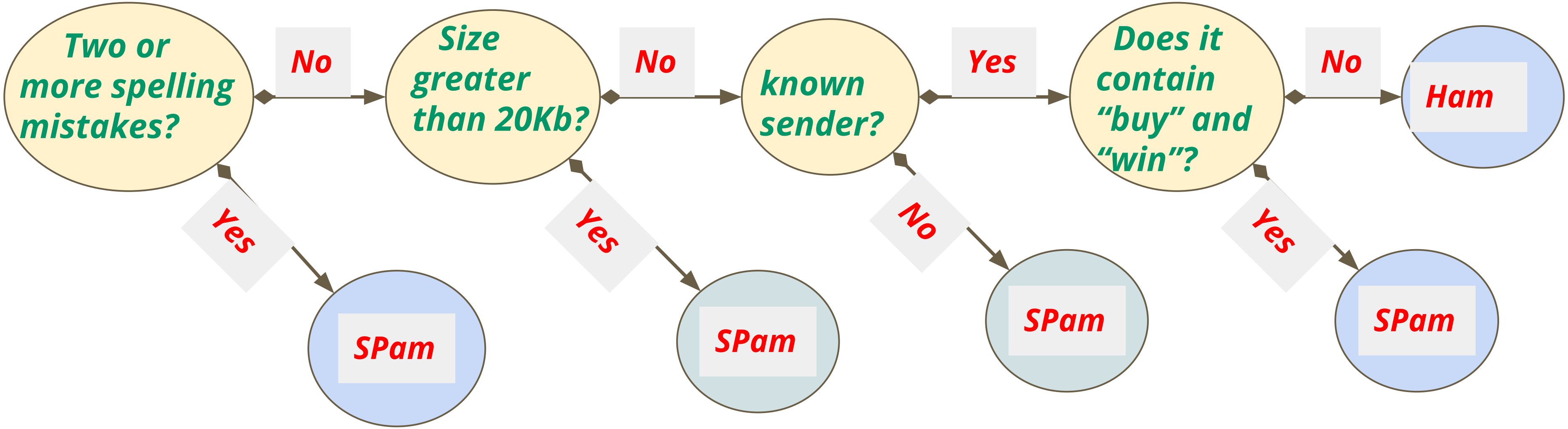
# How do computers think?

The goal is to make the computer think the way we think, namely, use the remember-formulate-predict framework. In a nutshell, here is what the computer does in each of the steps.

1. Remember: Look at a huge table of data.
2. Formulate: Go through many rules and formulas, and check which one fits the data best.
3. Predict: Use the rule to make predictions about future data.

This is not much different than what we did in the previous section. The great advancement here is that the computer can try building rules such as rules 4, 5, or 6, trying different numbers, different boundaries, and so on, until finding one that works best for the data. It can also do it if we have lots of columns. For example, we can make a spam classifier with features such as the sender, the date and time of day, the number of words, the number of spelling mistakes, the appearances of certain words such as “buy”, or similar words. A rule could easily look as follows:

Rule 7 :



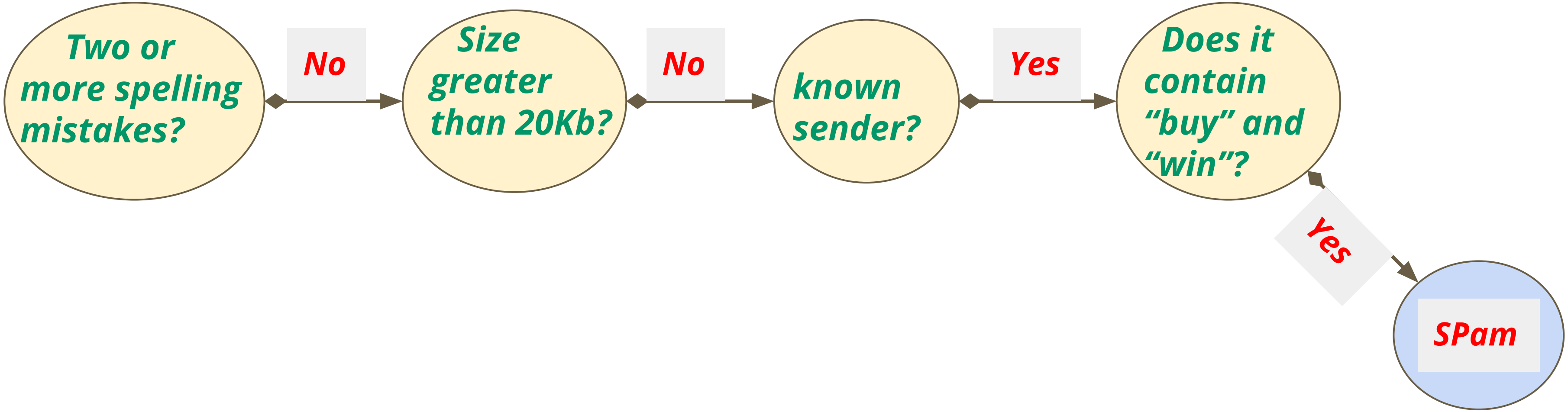
## It's Spam or Ham ?

| More than two spelling mistakes? | size greater than 20Kb? | known sende? | Does it contain "buy" and "win"? |
|---|---|---|---|
| 1 | 5 KB | Yes | Yes |

Facebook: Inna Lougoum

| More than two spelling mistakes? | size greater than 20Kb? | known sender? | Does it contain “buy” and “win”? |
|---|---|---|---|
| 1 | 5 KB | Yes | Yes |

## It’s Spam or Ham ?



Two or more spelling mistakes?
No
Size greater than 20Kb?
No
known sender?
Yes
Does it contain “buy” and “win”?
Yes
SPam

Facebook: Inna Lougoum

Rule 8:  If

**size + 10 x (*number of spelling mistakes*) - (*number of appearances of the word ‘mom’*) + 4 x (*number of appearances of the word ‘buy’*) > 10, then we classify the message as spam.**

**Otherwise we do not.**

Now the question is, which is the best rule? The quick answer is: The one that fits the data best. Although the real answer is: The one that generalizes best to new data. At the end of the day, we may end up with a very complicated rule, but the computer can formulate it and use it to make predictions very quickly. And now the question is: How to build the best model?

# Summary

1. Machine learning is easy! Anyone can do it, regardless of their background, all that is needed is a desire to learn, and great ideas to implement!
2. Machine learning is tremendously useful, and it is used in most disciplines. From science to technology to social problems and medicine, machine learning is making an impact, and will continue making it.
3. Machine learning is common sense, done by a computer. It mimics the ways humans think in order to make decisions fast and accurately.
4. Just like humans make decisions based on experience, computers can make decisions based on previous data. This is what machine learning is all about.

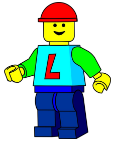
Machine learning uses the remember-formulate-predict framework, as follows:

1. ***Remember:*** Use previous data.
2. ***Formulate:*** Build a model, or a rule, for this data.
3. ***Predict:*** Use the model to make predictions about future data.

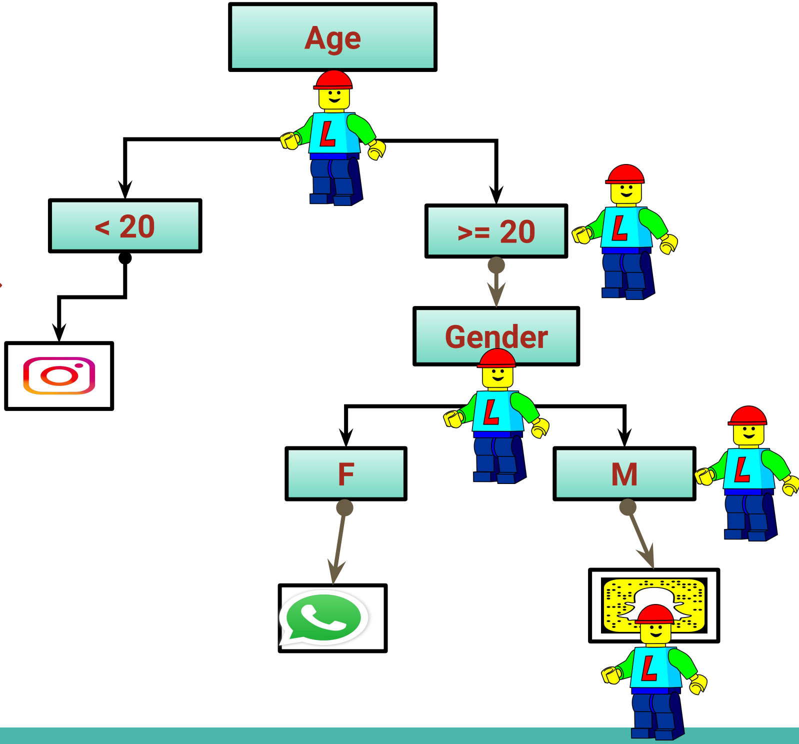
# Exercise : Recommending Apps

| Gender | Age | App |
|---|---|---|
| F | 15 | 📷 |
| F | 25 | 📞 |
| M | 32 | 👻 |
| F | 40 | 📞 |
| M | 12 | 📷 |
| M | 14 | 📷 |

| Gender | Age | App |
|---|---|---|
| M | 23 | ? |



Which App will be recommended to this new user ?



*Facebook: Inna Lougoum*