Google News Initiative

7 LESSONS - 51 MIN

Introduction to Machine Learning

This course will introduce you to the potential of machine learning and help you reflect on how you can use it responsibly to enhance your journalism.

Created by:

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

JournalismAi

vrt NWS

Google News Initiative

LESSON 1

Machine Learning, journalism and you

How machine learning is entering your personal and professional life.

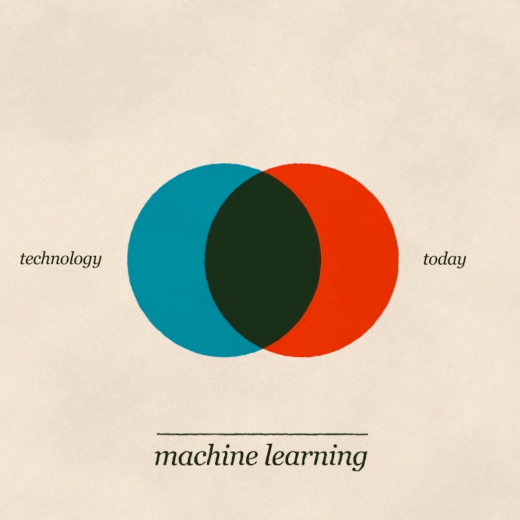
Lesson Overview

Machine Learning is all around us

Machine learning (ML) already powers many products we use every day. But it isn't always apparent to us that ML is behind them. Some examples:

- GPS navigation softwares, such as Google Maps and Waze
- Streaming services, such as Netflix and Spotify
- Search engines, such as Google Search, Baidu and Yahoo
- Social media, such as TikTok, Facebook, and Instagram

Machine learning can be applied to a wide range of fields, from health to retail, and in the development of self-driving vehicles.



For more lessons, visit:

newsinitiative.withgoogle.com/training

You already use Machine Learning

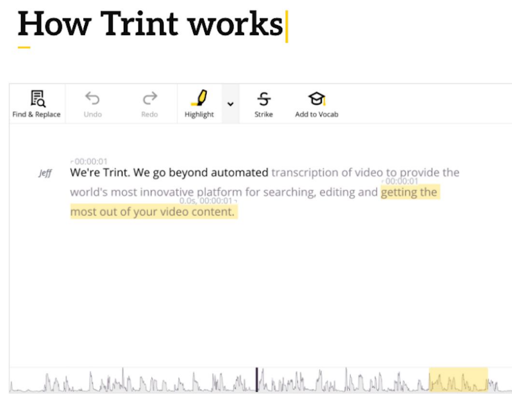
SINGLE STEP

You may not realise it but we have all already come into contact with machine learning in our daily life.

As a journalist, it’s likely that you have already used machine learning technology in your newsroom. Maybe you used Google Translate or another translation tool. Perhaps you used Grammarly or Hemingway to review your writing. Or maybe Trint saved you a lot of time by automatically transcribing your interviews.

Whatever your role in the newsroom or as an independent journalist, it's more likely than not that machine learning has already supported your work at some point.

So what are the main ways in which journalism is using machine learning through the different stages of the reporting process?



Journalism and Machine Learning

SINGLE STEP

Beyond the specific tools we already mentioned, machine learning is slowly but surely making its way into the journalistic process. As the [JournalismAI report] explained, it's mostly doing so through the augmentation of existing processes: freeing up journalists from repetitive tasks and allowing them to work on stories that would be too complex or too time-consuming to report without the help of technology.

So, what exactly can machine learning do for a newsroom? And how can journalists use it to enhance their editorial work?

In the next paragraphs we will look at some practical examples that show how machine learning can be deployed to support news gathering, as well as the production and distribution of news and information.

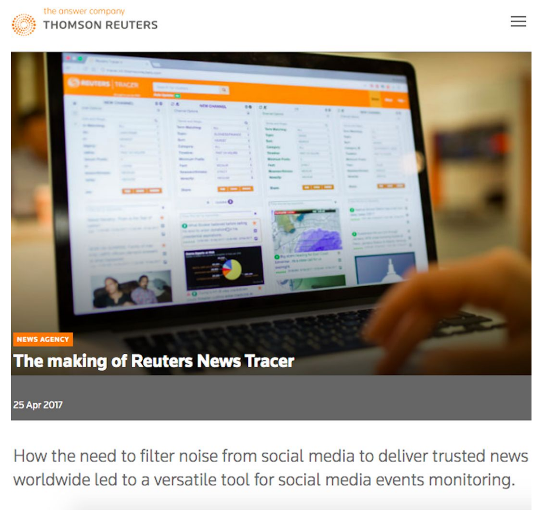
Machine Learning for news gathering

SINGLE STEP

In 2018, Reuters developed News Tracer and Lynx Insight. Both tools use machine learning and artificial intelligence technologies (more on the two terms in the next lesson) to support Reuters journalists in the news gathering process.

News Tracer is designed to help journalists find events that are breaking on Twitter. The tool analyses millions of tweets in real time to flag potential breaking news stories and allow the newsroom to spot the breaking news faster than what would be possible with regular news gathering practices.

Similarly, Lynx Insight is designed to identify trends and key facts in large datasets, suggesting new stories to reporters, while providing additional context and background information.



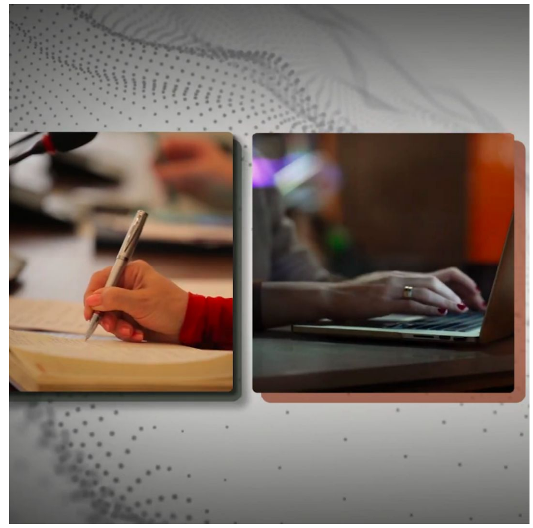
Machine Learning for news production

SINGLE STEP

Tools that use machine learning to automate the time-consuming process of transcribing interviews and translating information are good examples of how the production of news can be augmented by technology. But the use of machine learning in the process of news production goes well beyond that.

A wide array of media organisations – including Bloomberg, The Washington Post, and the Associated Press – have started to deploy different AI and machine learning techniques to automatically produce news stories at scale.

The main goal is to allow journalists to focus on the most creative aspects of their job, leaving repetitive tasks to the machine, but recent case studies show that the benefits could be bigger than we think.



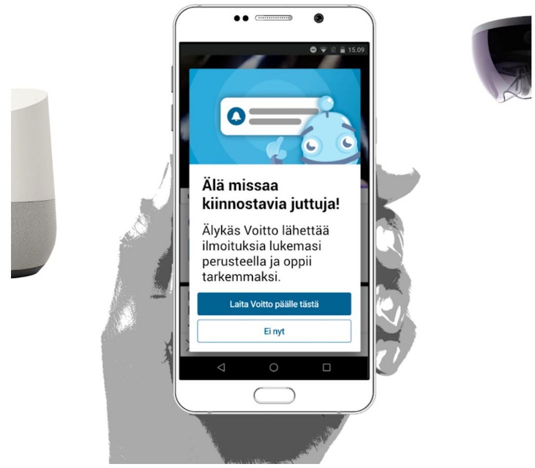
Machine Learning for news distribution

SINGLE STEP

Yle News Lab at the Finnish Public Broadcasting Company used machine learning to create a smart news assistant Voitto for its personalised news app Yle NewsWatch.

Voitto assistant lives on the lock screen of a mobile device and recommends the user interesting news content through alerts, or notifications. Voitto uses machine learning to improve its recommendations by learning from the user's interactions on their lock screen and from the user's reading history. Additionally, the user can teach the assistant by giving it direct feedback through notifications and in the news app itself.

Machine learning can also help news organisations to enhance their business model, for example by fine-tuning a flexible paywall for their subscribers.



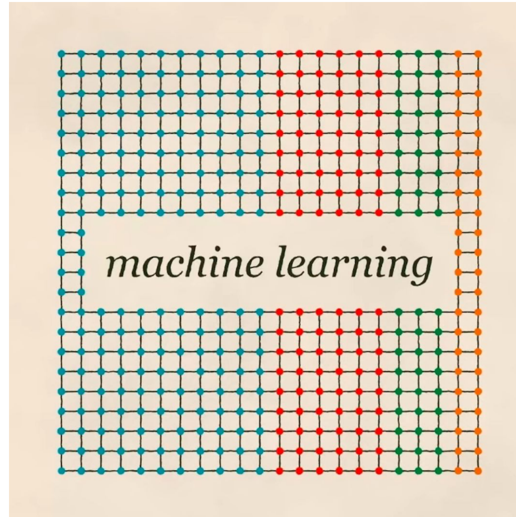
Exploring the potential of Machine Learning

SINGLE STEP

As we have seen, machine learning is already being used to augment journalists' capabilities across the journalistic process. But this does not mean that machine learning is the silver bullet to all journalism problems.

A lot of the potential offered by machine learning is yet to be discovered by journalism, and the new powers come hand in hand with new risks and challenges journalists should be wary of.

In the next lessons of this course, we will learn in depth what machine learning is and how it works. We will explore how it can be used by journalists in innovative ways and what risks must be taken into account for a responsible use of this powerful technology.



LESSON 2

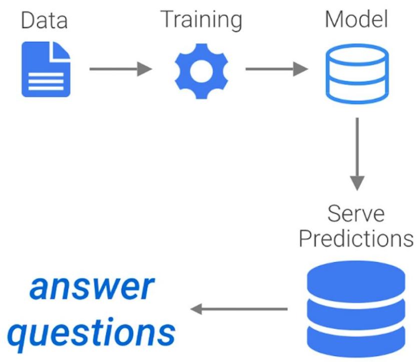
Is Machine Learning the same thing as AI?

Take a bird's eye view of machine learning within the AI landscape.

Lesson Overview

What is Machine Learning?

As with most of the terminology in the field of artificial intelligence, there is no unique definition of machine learning.



In simple terms, what ML does is to use data to answer questions. More formally, it refers to the use of algorithms that learn patterns from data and are able to perform tasks without being explicitly programmed to do so.

Moreover, a defining feature of machine learning systems is that they improve their performance with experience and data. Or in other words: *they learn*.

For more lessons, visit:

newsinitiative.withgoogle.com/training

How does Machine Learning relate to AI?

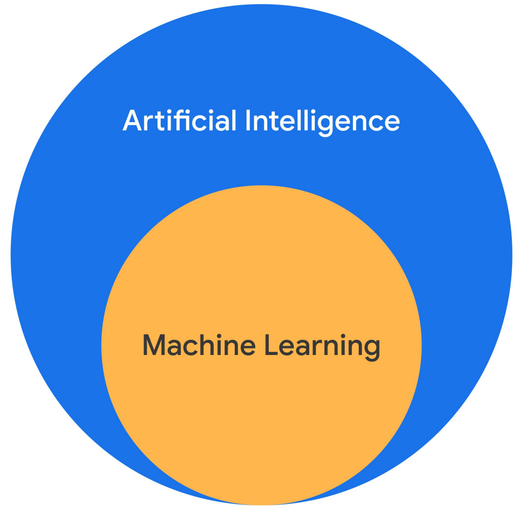
SINGLE STEP

Machine learning is part of a collection of technologies that are grouped under the umbrella term "artificial intelligence" (AI).

The concepts of AI and machine learning often seem to be used interchangeably, but in fact it is more correct to consider machine learning as a subfield of AI – which itself is a subfield of computer science.

AI means different things to different people but we can say that artificial Intelligence refers to the broader concept of machines being able to carry out tasks that normally require human intelligence.

In that context, machine learning refers to specific applications that use data to train a model to perform a given task independently and learn from experience.



AI and Machine Learning: a bit of history

SINGLE STEP

AI and machine learning have become hot buzzwords in recent years. But these topics are not new. Scientists have been working on AI and ML for quite some time.

Artificial intelligence was first discussed in the 1950s. The term was coined by American computer scientist John McCarthy in a workshop at Dartmouth College, New Hampshire, in 1956.

Since then, AI has undergone many evolutions and experienced both golden and darker days. Machine learning entered the scene in the 1980s but it's only in the 2010s that developments in the field started to accelerate exponentially. What explains this change of gear?

Why is everyone talking about ML and AI *now*?

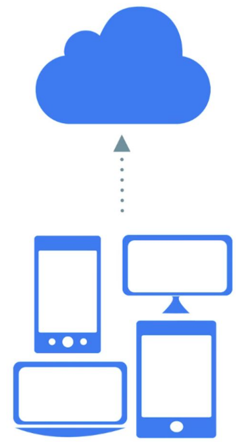
SINGLE STEP

In the last decade, two key factors have contributed to significant developments in the AI field:

First, huge amounts of data are being created every minute. Machines need data to ‘learn’ and the increasing availability means that bigger datasets can be used to improve the training of existing models and also that those models can be tested and applied to new fields.

The second factor relates to recent advances in processing speeds that allow computers to make sense of all this information much more quickly. This has allowed tech companies and other players in the field to justify bigger and bigger investments in research and development.

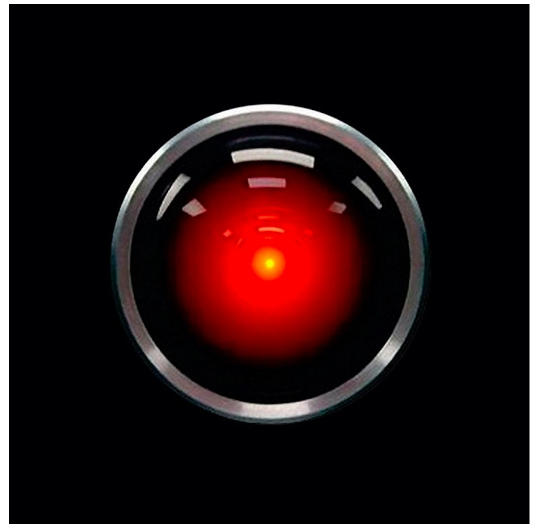
At the current speed, AI will soon become a little less artificial, and a lot more intelligent.



Should you be worried about machines becoming too intelligent?

SINGLE STEP

There is a fundamental misunderstanding about what research on AI tries to achieve. We are nowhere near machines thinking for themselves like the HAL 9000 computer in 2001 Space Odyssey, nor you should be afraid of a robot taking over your job in the foreseeable future.



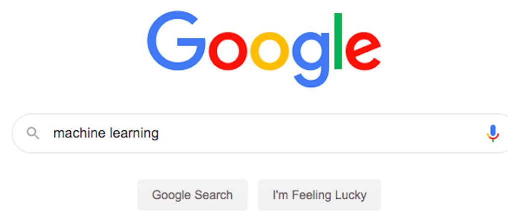
That might happen only if we ever reach Artificial General Intelligence (AGI): hypothetical machines that can handle any intellectual task in a human-like fashion without supervision. But as of today, that's still in the realm of science-fiction.

With the exception of few companies and research labs – DeepMind and OpenAI for example – current AI research focuses on narrow intelligence, with great progress being made in teaching machines to handle specific tasks independently.

Machine Learning: beyond the buzzwords

SINGLE STEP

The popularity of machine learning makes it sometimes difficult to separate what is real from what is just noise. The lack of an officially agreed definition, the legacy of science-fiction, and a general low level of literacy on AI-related topics are all contributing factors.



Hopefully, this lesson gave you a better understanding of what machine learning is and how it relates to artificial intelligence. But even within the field of machine learning there are different types of models and approaches that are important to recognise.

This is the topic of the next lesson.

LESSON 3

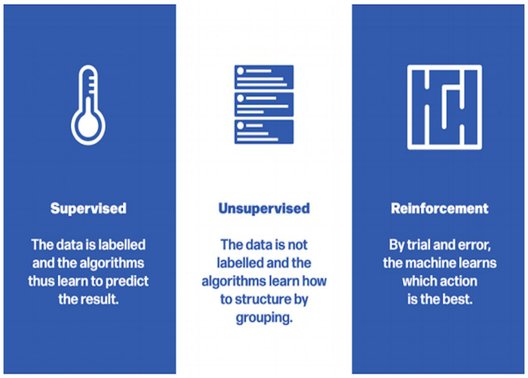
Different approaches to Machine Learning

Learn to recognise what defines different machine learning solutions.

Lesson Overview

There are various ways to learn

There are different ways for a machine to learn. Different approaches to ML are commonly distinguished by the kinds of problems they try to solve, as well as the type and amount of feedback provided by the programmer.



Broadly, we can divide machine learning into three subareas:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Although this might look like a neat categorisation, it's not always easy to place a particular method. Let's see what differentiates these three categories.

1 Supervised Learning

2 Unsupervised Learning

3 Reinforcement Learning

4 And what about Deep Learning?

5 Different learning models...so what?

For more lessons, visit:

newsinitiative.withgoogle.com/training

Supervised Learning

SINGLE STEP

Let's say you want to teach a machine to recognise dogs from cats. You give it as input photographs labelled as "cat" or "dog". Studying the examples, the algorithm will learn to recognise what distinguishes a cat from a dog and to assign the correct label to each new image you ask it to analyse.

In supervised learning, the machine needs labelled examples to learn. Those examples are used to train an algorithm to automatically assign the correct label.

In the journalistic context, supervised learning can, for example, train an algorithm to spot documents that might be interesting for an investigation. On a number of occasions this has already proven useful to investigative journalists having to deal with large volumes of documents.

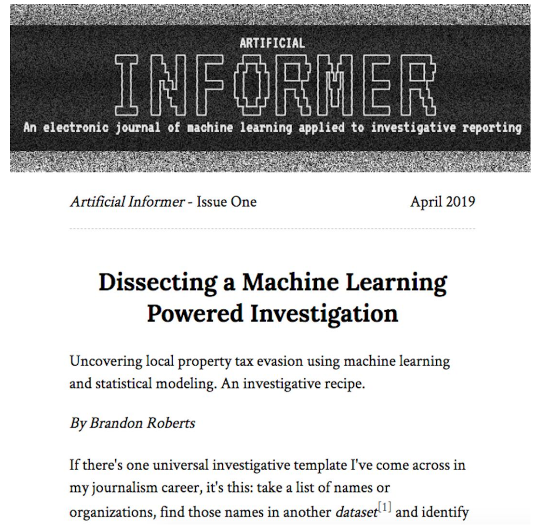
Unsupervised Learning

SINGLE STEP

With unsupervised learning, the examples provided to the machine are not labelled. The algorithm is tasked with learning by itself to recognise patterns in the data, for example with the goal of clustering together records that share similar characteristics.

In other words, the algorithm is trained to discover some structure in the unlabelled data that you ask it to analyse. This might be used by a business to better understand its customers, for example by grouping them into categories that show similar shopping behaviours.

In journalism, these kinds of techniques have been deployed by investigative journalists to [uncover tax evasion]() and to help campaign finance reporters link multiple donation records to the same donor.


ARTIFICIAL
INFORMER
An electronic journal of machine learning applied to investigative reporting

Artificial Informer - Issue One April 2019

Dissecting a Machine Learning Powered Investigation

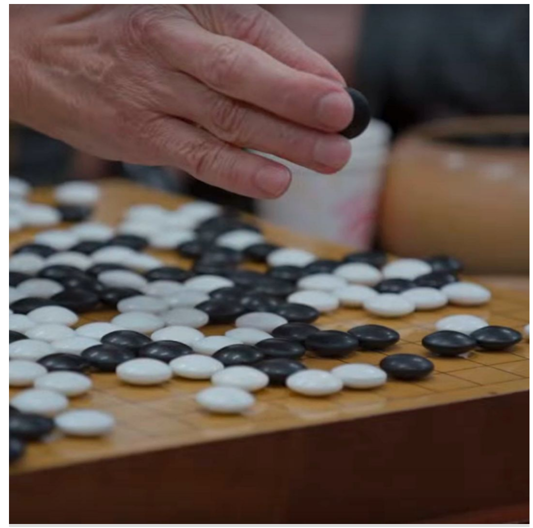
Uncovering local property tax evasion using machine learning and statistical modeling. An investigative recipe.

By Brandon Roberts

If there's one universal investigative template I've come across in my journalism career, it's this: take a list of names or organizations, find those names in another *dataset*[1] and identify

Reinforcement Learning

SINGLE STEP



The third type is reinforcement learning. Similarly to unsupervised learning, it doesn't need labelled data. It is instead based on the idea of learning what actions to take through trial and error, or in other words: by making mistakes. Initially the algorithm acts randomly, exploring the environment, but it learns with time by being rewarded when it makes the right choices.

Reinforcement learning is commonly used to teach machines to play games, with the most famous example being AlphaGo, the computer program developed by DeepMind that in 2016 managed to beat world's top player Lee Sedol at the Chinese board game Go.

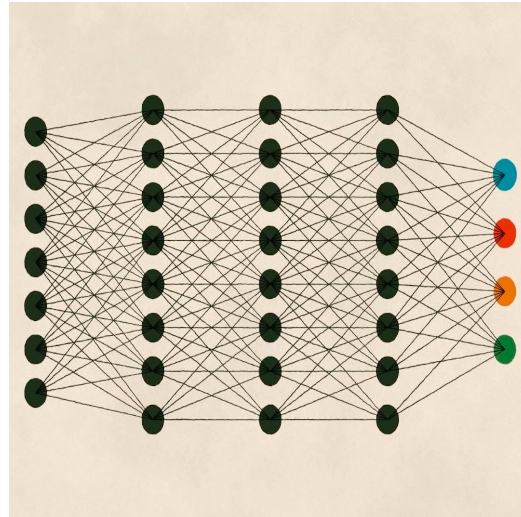
Journalistic applications are still rare, but reinforcement learning is used, for example, for headline testing.

And what about Deep Learning?

SINGLE STEP

Deep learning is another type of learning that has made a name for itself in recent years thanks to the increased computing powers we already discussed. It's in itself a subfield of machine learning, but differently from the approaches we just studied, deep learning is defined by the complexity and depth (hence the name) of the mathematical model involved.

The depth of the model refers to the use of multiple layers of analysis that allow the algorithm to learn progressively more complex structures. Deep learning is based on [artificial neural networks](), whose architecture is inspired by human biological systems, for example by how visual information is processed by our brain through our eyes.



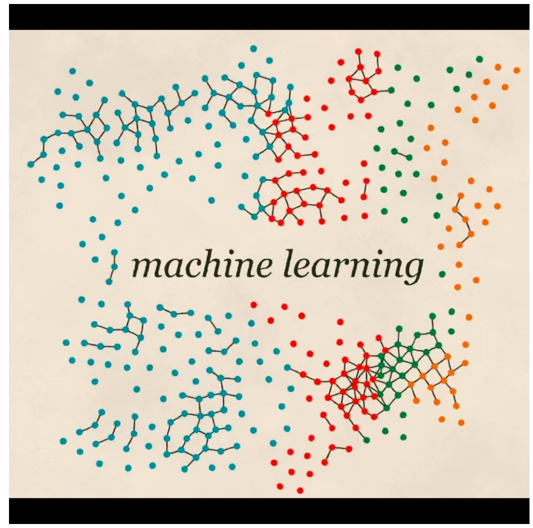
Different learning models... so what?

SINGLE STEP

Supervised, unsupervised, reinforcement, neural networks... your head must be spinning.

This lesson was not designed to put you off. It's important to understand the complexity of the field of machine learning and meet its subfields, but unless you want to dive deeper (pun intended) into the data science rabbit hole, what you should retain from this lesson is fairly simple: different problems require different solutions and different ML approaches to be tackled successfully.

In the next lesson, we will look at what situations in your work might welcome a machine learning solution. After that, we will explore the process that allows a machine to learn and introduce the concept of bias, with a few tips on how to deal with it.



LESSON 4

How you can use Machine Learning

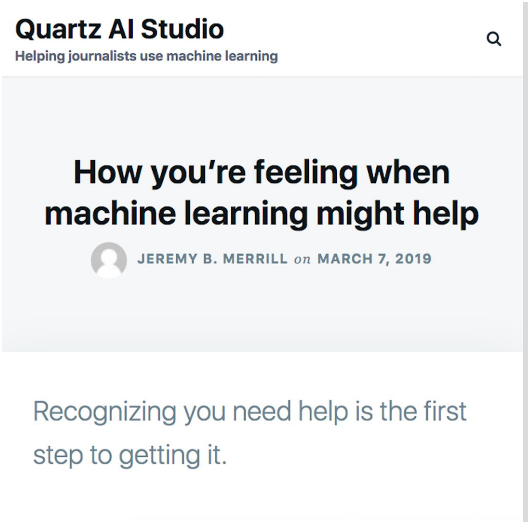
Understand in what cases ML might be the solution to your problem.

Lesson Overview

How you are feeling when Machine Learning might help

Now that you have a better sense of what machine learning is and what are the different approaches to training a model, you are probably wondering how machine learning can help in your daily work. This lesson will address just that.

No-one framed this conversation in a more effective way than the Quartz AI Studio. In the following paragraphs we will borrow their model (with permission) to help you understand some of the situations and feelings you might have when machine learning could help.



1 How will I be able to read all these documents?

2 How can I find out what's unique about this text?

3 How will I be able to analyse so many images?

4 How can I find more records like these?

5 What problems can Machine Learning help you with?

For more lessons, visit:

newsinitiative.withgoogle.com/training

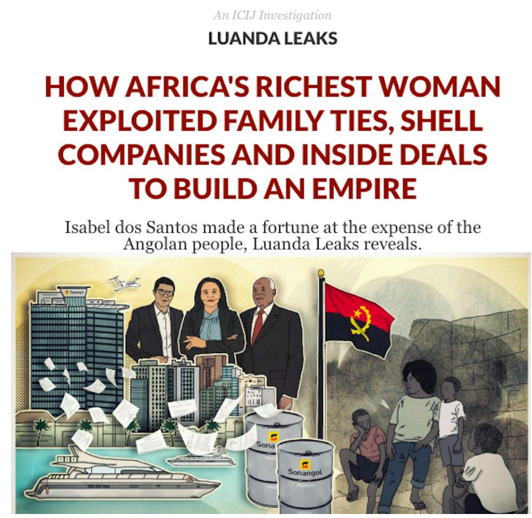
How will I be able to read all these documents?

SINGLE STEP

Movies about journalism tend to glorify investigations where reporters spend months reading boxes of documents in a windowless room to uncover big stories of corruption. What if we could achieve the same results in a fraction of the time?

Machine learning can help you do exactly that and for this reason is already being used by investigative journalists all across the world.

In 2019, the International Consortium of Investigative Journalists (ICIJ) received more than 700,000 leaked documents, collectively known as the Luanda Leaks. In order to analyse all those files, the ICIJ partnered with Quartz, whose investigation team built a machine learning model to help journalists find the kinds of documents they expected in the cache of leaks.



How can I find out what's unique about this text?

SINGLE STEP

Another kind of challenge a reporter might face when working on a story is the ability to compare a set of documents with a corpus of similar nature. For example, a political reporter might want to compare one president's State of the Union speeches to all those pronounced by other president's decade after decade.

It happens that this is another challenge machine learning is pretty good at dealing with.

Back in 2017, ProPublica [used a computer model](#) to analyse press releases from individual members of the US Congress in comparison with all Congressional press releases published during the same time. This allowed reporters to learn what topics members of Congress cared the most about, or at least talked about more than their peers.

Chamber of Secrets: Teaching a Machine What Congress Cares About

Want to know what distinctive topics your members of Congress are concerned about? Represent's got you covered.

by Jeremy B. Merrill, Oct. 4, 2017, 4:32 p.m. EDT

How will I be able to analyse so many images?

SINGLE STEP

Our world is photographed zillions of times a day. And this translates into an unprecedented amount of images reporters might find stories in. If only there was a way to teach computers to find specific details in a database of visual information... You know where this is going: enter machine learning.

The Ukrainian data journalism agency Texty used machine learning to detect illegal amber mines across Ukraine. Combining different algorithms, they were able to train the ML system on existing examples of amber mining, so that it could find new examples in a set of satellite images.

The resulting story included an online map in which a viewer can zoom into pictures of amber mines across the country.

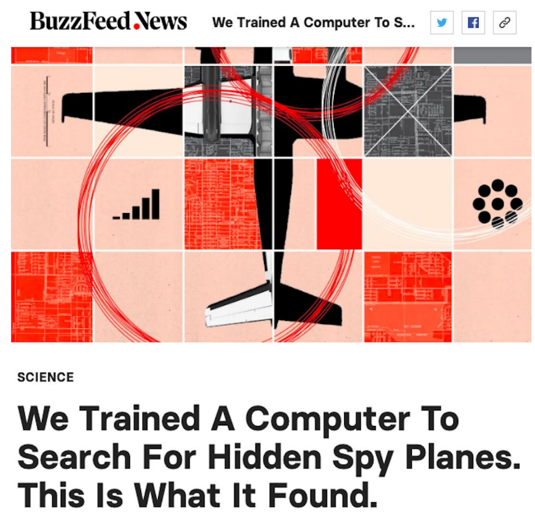
How can I find more records like these?

SINGLE STEP

Words, images, and now numbers. Among the many things computers can do better than humans, there is processing numeric data at scale. If you have thousands of numeric records to analyse, especially if you want to spot patterns and similarities, you are dealing with another case when machine learning can help.

That's what BuzzFeed News did in 2017 for their story on [hidden spy planes](), which made quite some noise as one of the early high-level examples of journalism applying machine learning for reporting purposes.

They trained a computer to find surveillance aircraft by letting a "random forest" algorithm sift for planes with flight patterns that resembled those operated by the FBI and the Department of Homeland Security.



What problems can Machine Learning help you with?

SINGLE STEP

Amber mines, corruption scandals, spy planes and State of the Union speeches. As you can see, machine learning can be quite handy in supporting your work by augmenting your ability to find and tell important stories with data.

By now, though, it should also be clear that machine learning is not magic. You might even say that it can't do anything you couldn't do – if you just had a thousand tireless interns working for you.

It's still entirely up to you to consider whether machine learning is the right tool to aid the story you want to report. After that assessment is made, you can count on machine learning to help you sift through an unmanageable amount of information and empower your journalism with the findings.

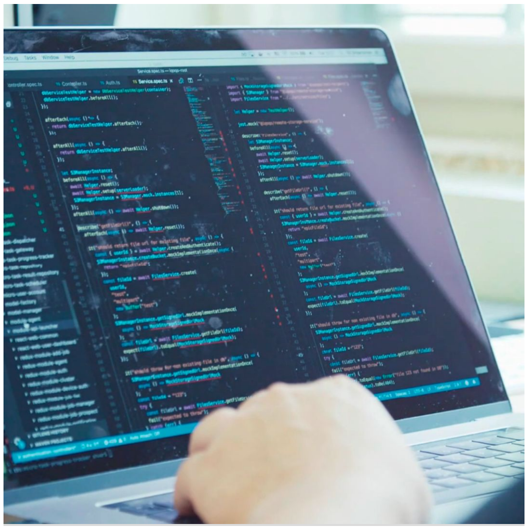
LESSON 5

How does a machine learn?

A step-by-step overview of the ML training process.

Lesson Overview

Training your Machine Learning model



So far, we have hinted at the fact that a ML model needs to be 'trained' in order to produce the expected outcome. In this lesson you will learn what steps are involved in the training process, through the lens of a specific case-study.

The goal is to help you understand how machines learn, not yet to be able to replicate the process on your own.

Before you decide to use machine learning, ask yourself: What question am I trying to find answers for? And do I need machine learning to get there?

1 What question do you want to answer?

2 Assessing your use case

3 Getting the data

4 Getting the data in shape

5 Choosing an algorithm

6 Training, validating and testing the model

7 Evaluating the results

8 Journalistic evaluation

For more lessons, visit:

newsinitiative.withgoogle.com/training

What question do you want to answer?

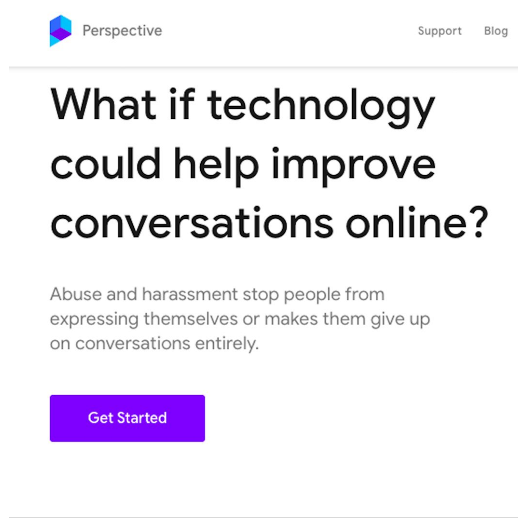
SINGLE STEP

Imagine your website gives readers the opportunity to comment on articles. Every day thousands of comments are posted and, as it happens, sometimes the conversation gets a little nasty.

It would be great if an automated system could categorise all comments posted on your platform, identify those that might be ‘toxic’ and flag them to the human moderators, who could review them to improve the quality of the conversation.

That's a type of problem machine learning can help you with. And in fact, it already does. Check Jigsaw's Perspective API to find out more.

This is the example we are going to use to learn how a machine learning model is trained.But keep in mind that the same process can be extended to any number of different case-studies.



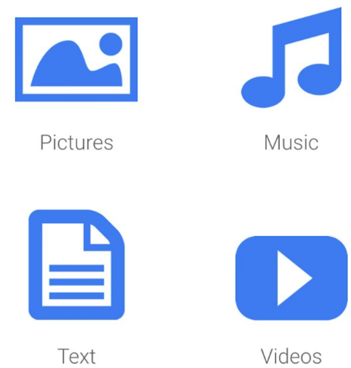
Assessing your use case

SINGLE STEP

To train a model to recognise toxic comments, you need data. Which in this case means examples of comments you receive on your website. But before you prepare your dataset, it’s important to reflect on what is the outcome you are trying to achieve.

Even for humans, it's not always easy to evaluate whether a comment is toxic and should therefore not be published online. Two moderators might have different views on the ‘toxicity’ of a comment. So you shouldn't expect the algorithm to magically "get it right" all the time.

Machine learning can handle a huge number of comments in minutes, but it's important to keep in mind that it's just ‘guessing’ based on what it learns. It will sometimes give wrong answers and generally, make mistakes.



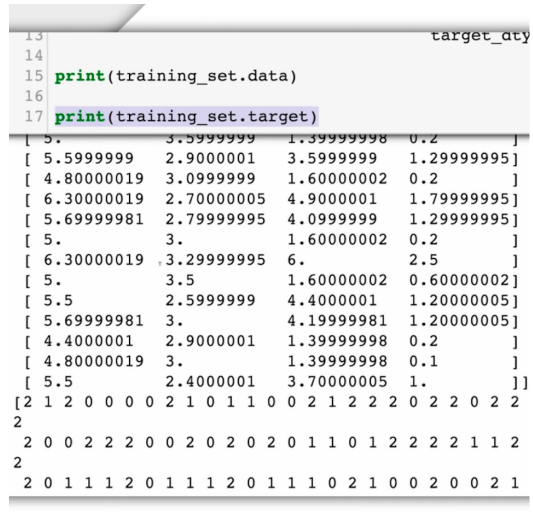
Getting the data

SINGLE STEP

It's now time to prepare your dataset. For our case-study, we already know what kind of data we need and where to find it: comments posted on your website.

Since you are asking the machine learning model to recognise comments' toxicity, you need to supply labelled examples of the kinds of text items you want to classify (comments), and the categories or labels you want the ML system to predict ("toxic" or "non toxic").

For other use cases you might not have the data so easily available, though. You will need to source it from what your organisation collects or from third-parties. In both cases, make sure to review regulations about data protection in both your region and the locations your application will serve.



Getting your data in shape

SINGLE STEP

Once you have collected the data and before you feed it to the machine, you need to analyse the data in depth. The output of your machine learning model will be only as good and fair as your data is (more on the concept of 'fairness' in the next lesson). You must reflect on how your use case might negatively impact the people that will be affected by the actions suggested by the model.

Among other things, in order to successfully train the model you will need to make sure to include enough labelled examples and to distribute them equally across categories. You must also provide a broad set of examples, considering the context and the language used, so that the model can capture the variation in your problem space.

```
INFO:tensorflow:Using default config.
INFO:tensorflow:Using config: {'_save_checkpoints_
onfig': None, '_keep_checkpoint_max': 5, '_tf_rand
ckpoint_every_n_hours': 10000, '_log_step_count_st
points_steps': None, '_model_dir': '/tmp/iris_mode
s': 100}
```

```
def input_fn(dataset):
    def _fn():
        features = {feature_name: tf.constant(
        label = tf.constant(dataset.target)
        return features, label
    return _fn

print(input_fn(training_set)())

# raw data -> input function -> feature colum
```

```
# Fit model.
classifier.train(input_fn=input_fn(training_se
```

Choosing an algorithm

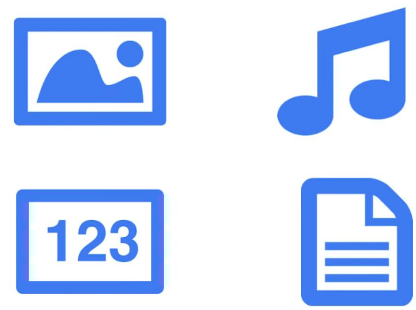
SINGLE STEP

After you are done preparing the dataset, you have to choose a machine learning algorithm to train. Every algorithm has its own purpose. Consequently, you must pick the right kind of algorithm based on the outcome you want to achieve.

In previous lessons we have learned about different approaches to machine learning. Since our case-study requires labelled data in order to be able to classify our comments as "toxic" or "non toxic", what we are trying to do is supervised learning.

[Google Cloud AutoML Natural Language] is one of many algorithms that allow you to achieve our desired outcome. But whatever algorithm you choose, make sure to follow the specific instructions on how it requires the training dataset to be formatted.

Choosing a Model



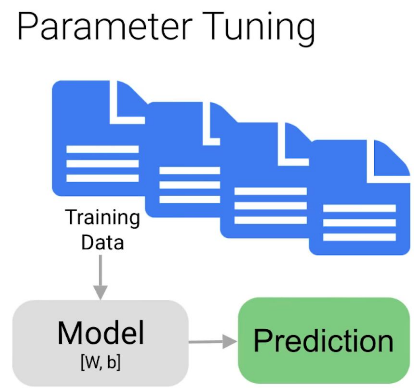
Training, validating and testing the model

SINGLE STEP

Now we move onto what is the proper training phase, in which we use the data to incrementally improve our model's ability to predict if a given comment is toxic or not. We feed most of our data to the algorithm, perhaps wait a few minutes, and voilà, our model is trained.

But why only "most" of the data? To make sure the model learns properly, you must divide your data in three:

- The training set is what your model "sees" and initially learns from.
- The validation set is also part of the training process but it's kept separate to tune the model's hyperparameters, variables that specify the model's structure.
- The test set enters the stage only after the training process. We use it to test the performance of our model on data it has not yet seen.



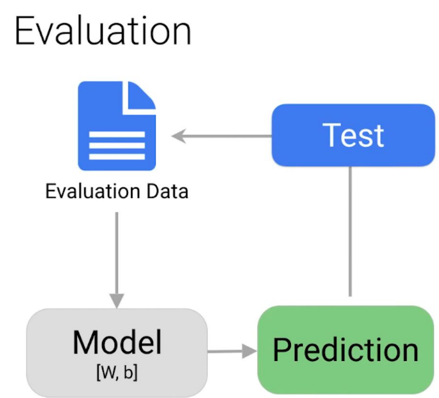
Evaluating the results

SINGLE STEP

How do you know if the model has correctly learned to spot potentially toxic comments?

When the training is complete, the algorithm provides you with an overview of the model performance. As we already discussed, you can't expect the model to get it right 100% of the time. It's up to you to decide what is 'good enough' depending on the situation.

The main things you want to consider to evaluate your model are false positives and false negatives. In our case, a false positive would be a comment that is not toxic, but gets marked as such. You can quickly dismiss it and move on. A false negative would be a comment that is toxic, but the system fails to flag it as such. It's easy to understand which mistake you should want your model to avoid.



Journalistic evaluation

SINGLE STEP

Evaluating the results of the training process doesn't end with the technical analysis. At this point, your journalistic values and guidelines should help you decide if and how to use the information the algorithm is providing.

Start by thinking whether you now have information that was not available before, and about the newsworthiness of that information. Does it validate your existing hypothesis or is it shedding light on new perspectives and story angles you were not considering before?

You should now have a better understanding of how machine learning works and you might be even more curious to try out its potential. But we are not ready yet. The next lesson will introduce the number one concern machine learning brings with it: Bias.

LESSON 6

Bias in Machine Learning

Understand what ML bias is and how to avoid it.

Lesson Overview

Fairness in Machine Learning

So far, this course showed how machine learning can enhance your work, from saving precious time on existing tasks to opening up new opportunities. ML can do a lot for you, but it comes with challenges you shouldn't overlook.

To address those challenges, a growing number of researchers and practitioners focus on the topic of "fairness" in machine learning. Its guiding principle is that ML should equally benefit everyone, regardless of the societal categories that structure and impact our lives.

1	What is bias?
2	Three types of bias
3	Asking the right questions to avoid bias
4	Considering the main sources of bias
5	Preventing bias: it starts with awareness

For more lessons, visit:

newsinitiative.withgoogle.com/training

What is bias?

SINGLE STEP

What are the negative consequences that might derive from the use of machine learning? The short answer is: Bias.

As humans, we all have our biases. They are tools our brain uses to deal with the information that is thrown at it every day.

Take this example: close your eyes and picture a shoe. Most likely you pictured a sneaker. Maybe a leather men's shoe. It's less likely that you thought of a high-heeled women's shoe. We may not even know why but each of us is biased toward one shoe over the others.

Now imagine that you want to teach a computer to recognise a shoe. You may end up exposing it to your own bias. That's how bias happens in machine learning. Even with good intentions, it's impossible to separate ourselves from our own biases.

Three types of bias

SINGLE STEP

There are different ways in which our own biases risk to become part of the technology we create:

Interaction bias
Take the example before: if we train a model to recognise shoes with a dataset that includes mostly pictures of sneakers, the system won't learn to recognise high heels as shoes.

Latent bias
If you train a ML system on what a scientist looks like using pictures of famous scientists from the past, your algorithm will probably learn to associate scientists with men only.

Selection bias
Say you're training a model to recognise faces. If the data you use to train it over-represents one population, it will operate better for them at the expense of others, with potentially racist consequences.

So what can we do to avoid these biases?

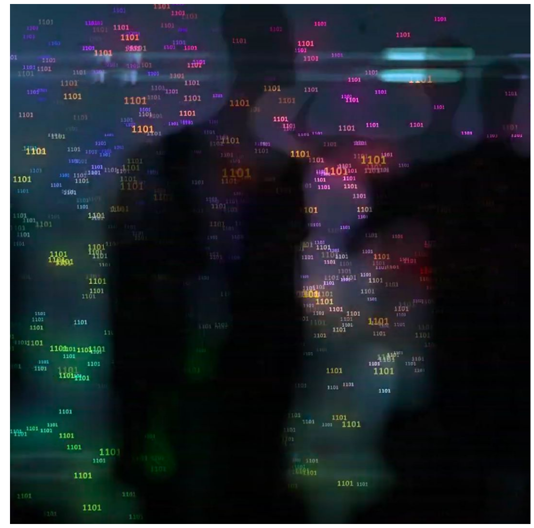
Asking the right questions to avoid bias

SINGLE STEP

As a journalist, a first line of defence against bias is firmly within your reach: the same values and ethical principles you apply every day in your profession should extend to assessing the fairness of any new technology that is added to your toolbox. Machine learning is no exception.

Furthermore, in all cases you should start by considering whether the consequences might negatively impact individuals' economic or other important life opportunities. This is critical especially if the data you use includes sensible personal information.

Often, the unfair impact isn't immediately obvious, but requires asking nuanced social, political and ethical questions about how your machine learning system might allow bias to creep in.



Considering the main sources of bias

SINGLE STEP

While no training data will ever be perfectly ‘unbiased’, you can greatly improve your chances of building a fair model if you carefully consider potential sources of bias in your data, and take steps to address them.

The most common reason for bias creeping in is when your training data isn't truly representative of the population that your model is making predictions on. You must make sure to have enough data for each relevant group.

A different kind of bias manifests itself when some groups are represented less positively than others in the training data. You should consider reviewing your data before using it to train a model, in order to verify whether it carries any prejudices that might be learned and reproduced by the algorithm.

Preventing bias: it starts with awareness

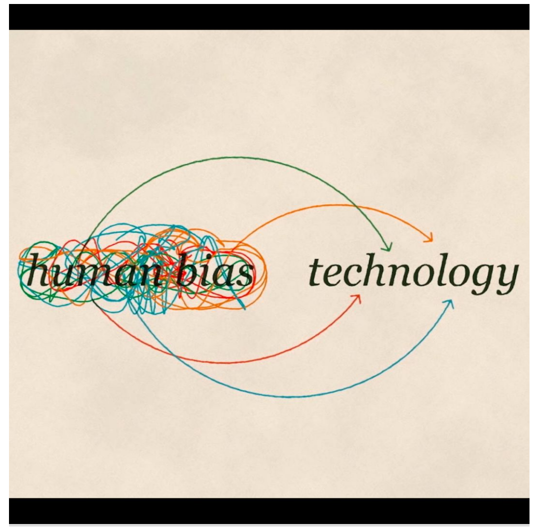
SINGLE STEP

Bias can emerge in many ways: from training datasets, because of decisions made during the development of a machine learning system, and through complex feedback loops that arise when a ML system is deployed in the real world.

Some concrete questions you might want to ask in order to recognise potential bias include:

- For what purpose was the data collected?
- How was the data collected?
- What is the goal of using this set of data and this particular algorithm?
- How was the source of data assessed?
- How was the process of data analysis defined before the analysis itself?

Bias is a complex issue and there is no silver bullet. The solution starts with awareness and with all of us being mindful of the risks and taking the right steps to minimise them.



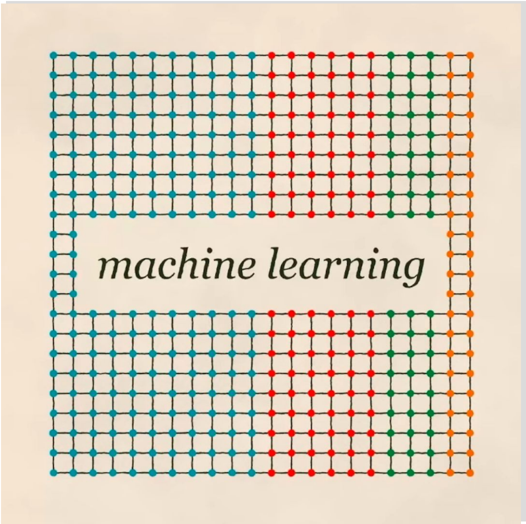
LESSON 7

Looking ahead to ML-powered journalism

Key learnings and recommended resources to deepen your ML knowledge.

Lesson Overview

Moving the next steps



You have reached the end of this introduction to machine learning. You should now be better equipped to report on developments in the field of AI and ML, or find it easier to communicate with data scientists in your organisation with a shared language. This course might have triggered your curiosity to explore the topic more in depth or maybe you can't wait to try out ML in the reporting of your next story.

Whatever your next steps will be, there are some key learnings you should retain.

1 Key concepts and learnings

2 Where to learn more

3 Credits

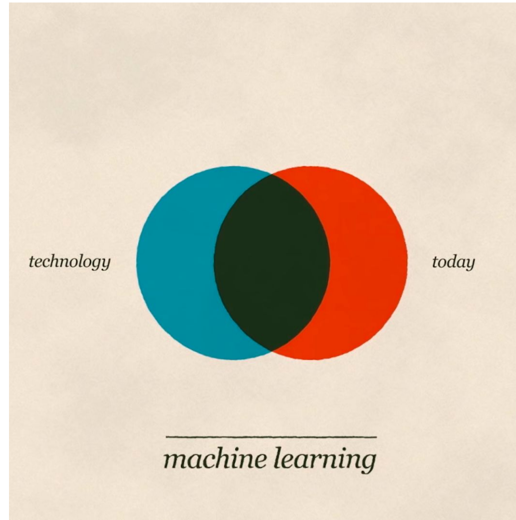
For more lessons, visit:

newsinitiative.withgoogle.com/training

Key concepts and learnings

SINGLE STEP

- Machine learning is not a synonym of AI but rather a subfield of it.
- Machine learning is already used to augment your work across the journalistic process.
- ML refers to algorithms that perform tasks without being explicitly programmed to do so.
- Machines have different ways to learn, each one best suited to achieve a specific outcome.
- Machine learning can augment your ability to find and tell important stories with data.
- But your journalistic question might not have or need a machine learning solution.
- To find the answers you are looking for, ML models need to be trained on existing data.
- The sourcing and preparation of the training dataset is crucial.
- Reproducing human biases is the biggest risk and you can and should take steps to avoid it.



Where to learn more

SINGLE STEP

If you want to learn more, we recommend you check out the following resources, which have been used to design this course:

- The JournalismAI report "New powers, new responsibilities. A global survey of journalism and artificial intelligence" that will give you a detailed overview on what other journalists are thinking and doing with ML and AI
- The EDUbox on artificial intelligence developed by VRT NWS
- The work of Professor Nick Diakopoulos, especially the free online course "News Algorithms: The Impact of Automation and AI on Journalism"
- The online course Elements of AI, developed by Reaktor and the University of Helsinki
- All the materials published by the Quartz AI Studio
- The Google Cloud guide on Inclusive ML

Credits

SINGLE STEP

This course was developed by JournalismAI in collaboration with Tom Van de Weghe and the team at VRT NWS, and with the contribution of Jarno Koponen, Head of AI at Yle News Lab.

VRT NWS is the news service of the Flemish Radio and Television Broadcasting Organisation.

JournalismAI is a project of Polis – the journalism think-tank at the London School of Economics and Political Science – and it's funded by the Google News Initiative.

A step-by-step guide to training a machine learning model will be added to this platform by Autumn 2020.

Sign up to the JournalismAI newsletter to stay informed about project activities and be the first one to know about the release of the new training course.