WHITE PAPER

HOW CAN MACHINE LEARNING HELP EDUCATIONAL PUBLISHERS WITH THEIR CONTENT?

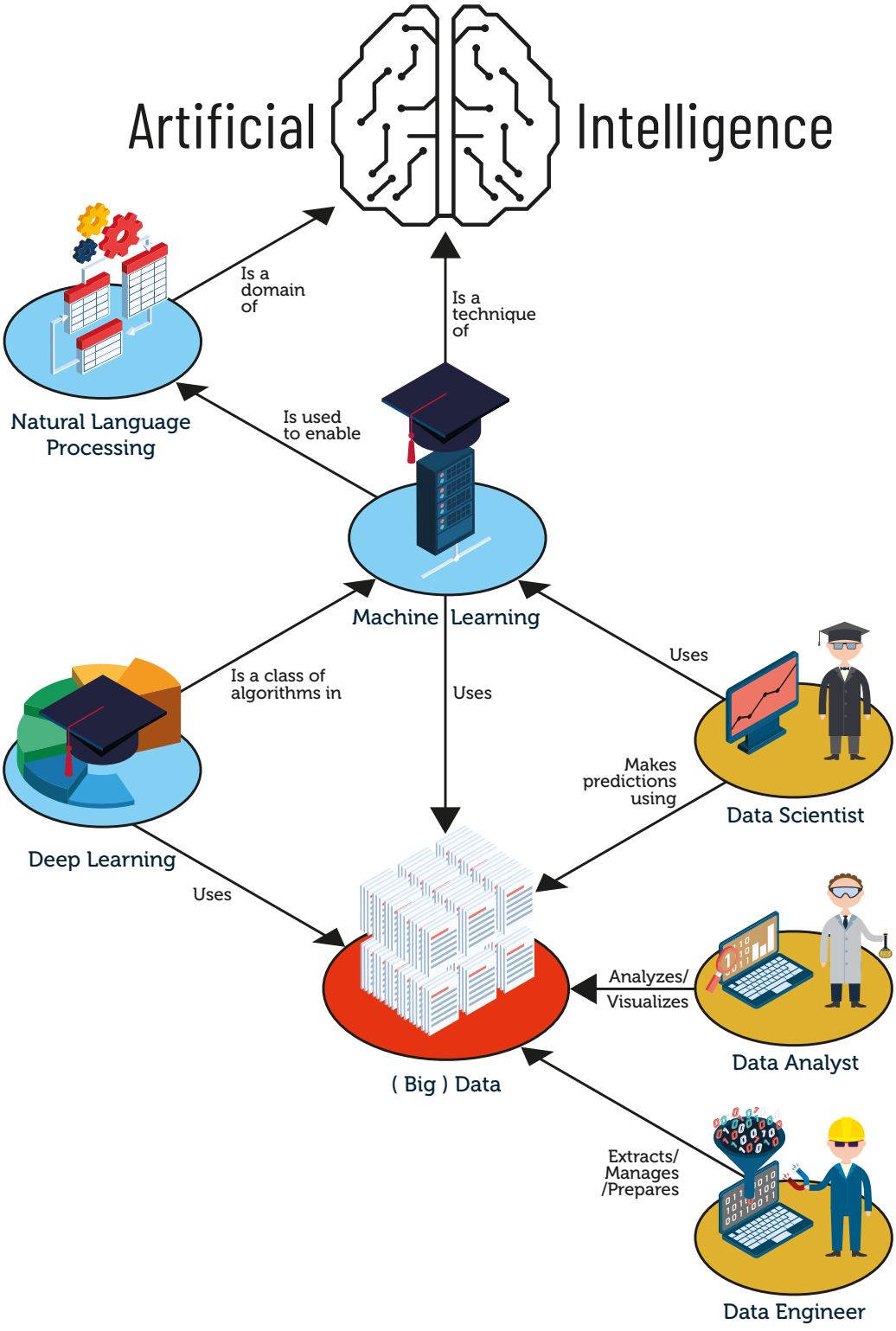
EDIA

You're a Director of Product, Content Manager, or Publishing Innovation Lead and under pressure to make the most out of your content resources.

That's why you want to know how you can improve content digitization and process automation. It's important that you make a well-informed decision about how to manage technology in your business strategy. After all, it's about keeping pace with your customers' happiness.

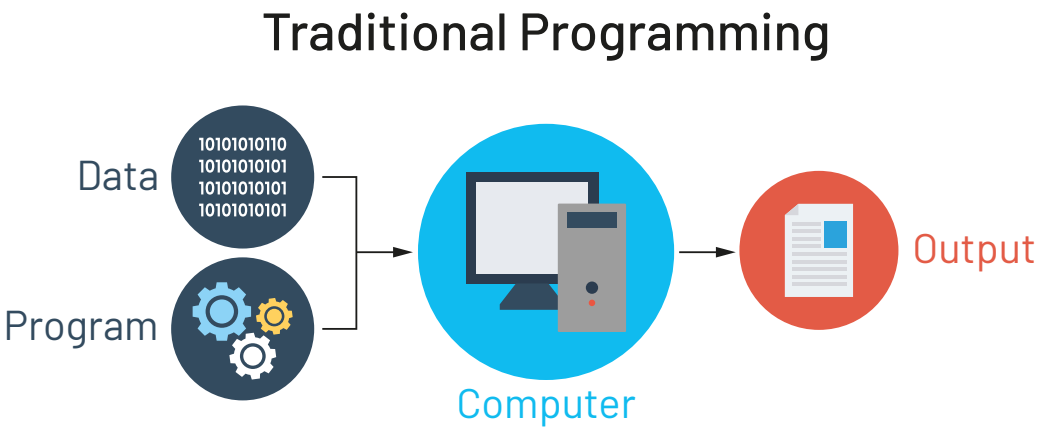
This White Paper gives a clear picture of what machine learning is all about and where to begin with it. After reading this paper you have the information you need to make a choice.

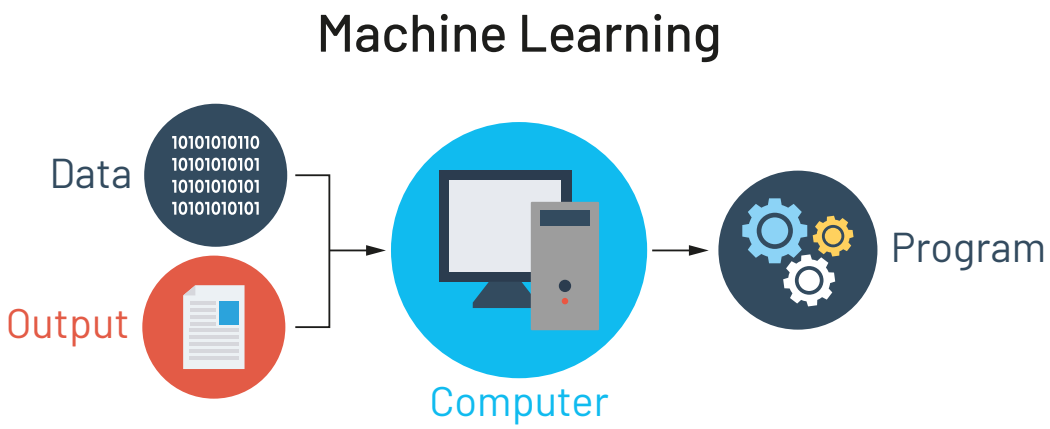
Artificial Intelligence and its subfield, Machine Learning, are frequently cited as the technology of the future. A science fiction dream realized, they allow us to outsource many boring and repetitive tasks to a computer, leaving more time for human employees to work on decisions requiring a human mind. However, while AI generates a lot of buzz, most organizations find it difficult to comprehend how it and machine learning could fit into their workflows.



What is machine learning?

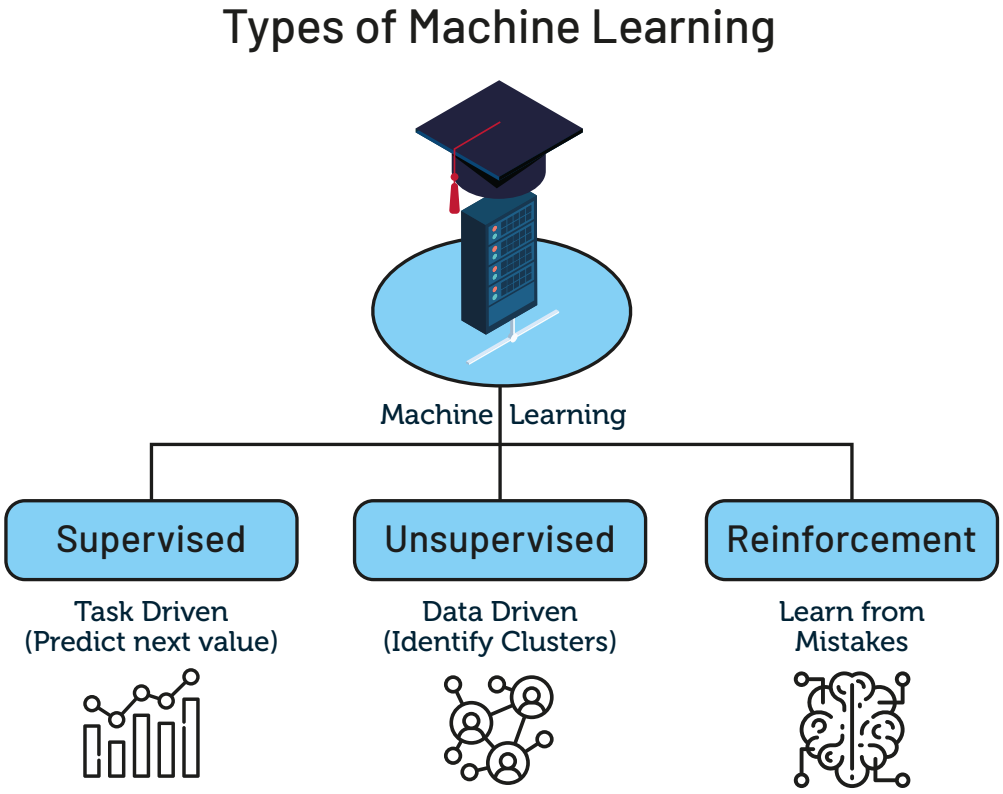
"The science of getting computers to act without being explicitly programmed" - Andrew Ng





Machine learning is a sub-branch of Artificial Intelligence – a technique that allows computers to mimic human thinking in ways which include learning from past experiences, adapting to new information, self-correction and synthesizing new hypotheses.

Machine learning draws specifically on the ability of computer algorithms to learn. Traditional computing requires detailed instructions to perform a function, with each action corresponding to a line of code. For machine learning the computer is given some initial data and instructions which allow it to learn from examples and experience rather than using predetermined instructions from the outset. It's the difference between giving a computer a fish and teaching the computer to fish for itself. Machine learning is divided into various subsets, including unsupervised and supervised learning.



Unsupervised machine learning

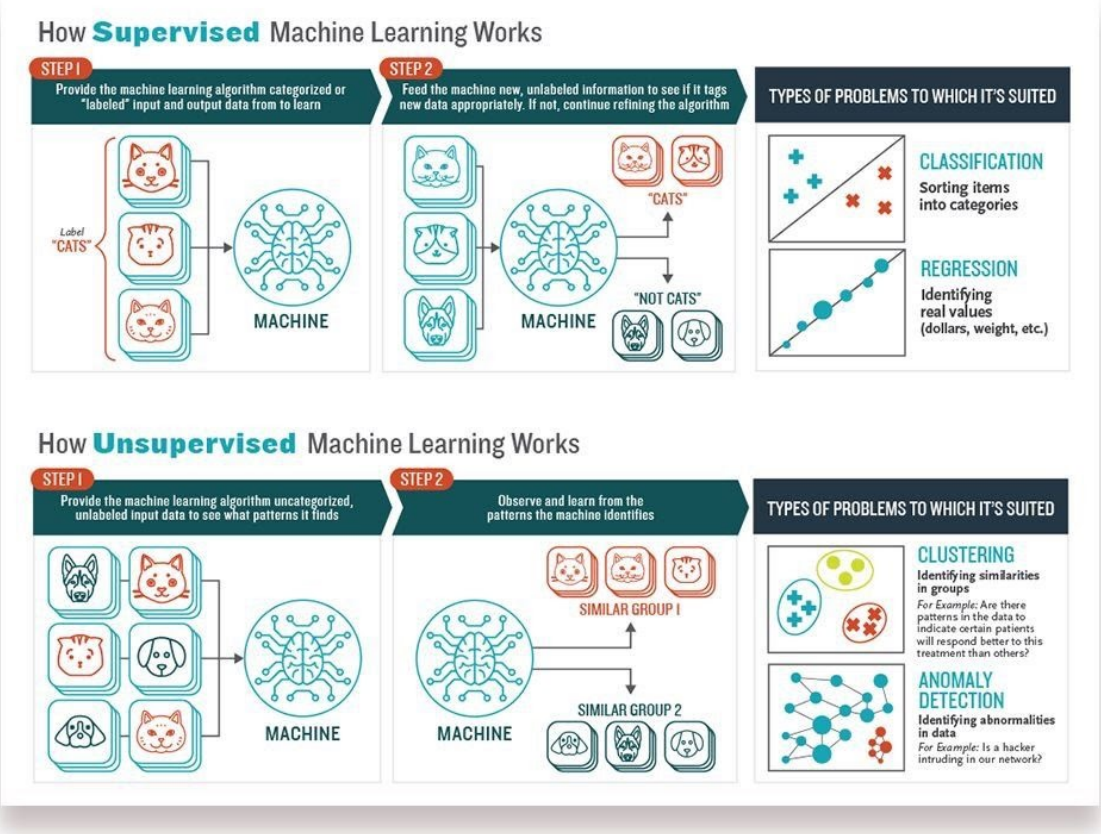
Unsupervised machine learning is simple. It asks a computer to recognize patterns from a set of data without specific directives. This is useful for clustering data or finding anomalies. It is often used for large data sets and is exploratory – such algorithms are rarely helpful for specific or detailed insights but look at broad patterns in data.

For example, unsupervised machine learning is commonly used for fraud detection. Algorithms run through huge amounts of data - in this case financial transactions - and identify outliers - financial transactions which do not match the normal pattern of transactions.

Unsupervised machine learning can also be used to cluster data, grouping data together to bring some organization. This is a particularly useful first step in an analysis where the contents of the dataset are unknown.

Supervised machine learning

Supervised machine learning provides a more specific and detailed analysis. It involves 'training' the algorithm with an example set of data to teach it how to behave. Supervised learning involves feeding input into the algorithm along with output. Eventually, the algorithm learns the connection and can repeat it with a broader set of data.



For example, an algorithm is shown five images of a dog, with the label 'dog'. It is then shown five other images, labeled 'not dog'. The computer learns from this and can determine whether a new image is a 'dog' or 'not dog'.

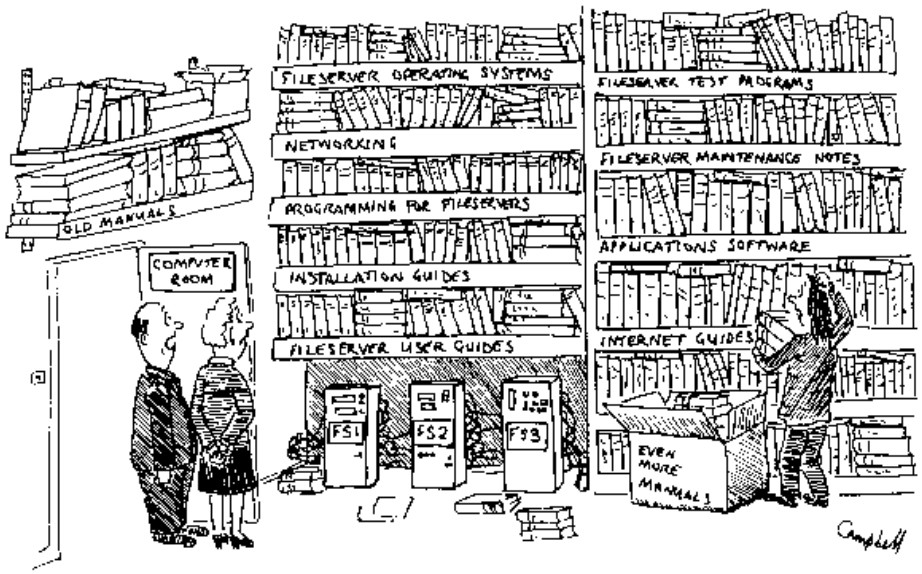
Supervised learning provides more detailed outcomes than unsupervised learning because the algorithm is able to label the data it is working with. For this reason, supervised learning is often used for classifying material – sorting it into different groups with a high level of accuracy.

When EDIA formed it set out with a mission – to create an AI ecosystem connecting publishers, teachers and students, making sure that educational content is helpful and beneficial for all parties. These three parties make up the key pillars of education yet often communication and cooperation between them can be complicated.

Many publishers have extensive libraries of educational content, which has tagged labels such as subject or age level. But beyond that, tagging is limited, making it difficult for a publisher to determine exactly which content they have on fractions or Shakespearean sonnets. Manual tagging is a laborious and expensive task which few publishers are able to invest in, making it the perfect task to be automated through AI.

The use of AI for educational publishing is set to become increasingly popular because it can drastically improve the accuracy and scope of content tagging, allowing publishers improved access to their own repositories of content.

However, it quickly became clear that a barrier to this vision was the educational content itself. Publishers are in possession of extensive repositories of educational content, making it difficult to determine what content is available and where it is located. In a digital era, content is difficult to work with if it doesn't also include meta tagging, which digitally identifies it. This allows content to be used in a variety of different ways. Well-labelled content is easier for users to find or identify quickly. Such content can be more effectively used to assess what a learner knows and consider where they should progress to, using principles of adaptive learning.



Furthermore, content is still being produced according to the mindset of traditional publishing. Content frequently conforms to the style of a textbook used in a classroom. In contrast, adaptive learning requires granular and smaller content, to allow for different pathways through content.

In order to allow innovation within the field of digital education, educational publishers need help in how they interact with their content. Content creation, management and distribution processes need to be reconsidered. Luckily, machine learning is a tool that can help publishers make data-informed decisions and automate common tasks, improving efficiency and accuracy.

Issues to consider within machine learning

In an ideal world, each individual publisher would have an individualized learning algorithm to classify content according to their standards. However, this is problematic to achieve because often publishers don't have enough content to 'teach' an algorithm. Attempting to teach an algorithm with a very small sample of data doesn't work effectively, as it overfits the data, creating false correlations which are unhelpful when extrapolated out into broader data.

Multi-task learning combines two directives into one algorithm. The algorithm takes into account the terminology used by a specific company but combines it with broader learning

data, so the learning sample is not too small. The two directives work together to ensure that data is classified in an appropriate manner, without resulting in flawed data or lacking the personalization necessary for each individual system. This provides a tailored approach to classifying content.

Using complex algorithms and sophisticated machine learning, EDIA is able to significantly increase the speed and scope of content meta tagging and classification. Publishers looking to modernize can employ AI in a way that fits into their current business model while improving the usefulness of their own content.

What machine learning techniques does EDIA use?

EDIA uses a combination of unsupervised and supervised machine learning to find and tag content. Algorithms use cluster analysis to create basic groupings of content, helping to explore and determine what content is available. But EDIA technology can also fit content into a detailed taxonomy of content – for example, sorting it into maths, then within that into a subset of geometry than a subset of the Pythagorean theorem. This taxonomy makes navigating content clear and easy for publishers, as well as grouping similar content together.

EDIA also integrates algorithms which can determine the specific reading level of a piece of content, linked to the Common European Framework of Reference for Language (CEFR). This is a particularly complicated algorithm to perform, as the CEFR doesn't rely on formulae or more specific algorithms but written classifications which are subjective and open to interpretation. This is an example of supervised machine learning, in which an algorithm is 'taught' to operate with a sample set of data. Based on this, an algorithm can classify content without specific and constant directions. Data can be classified in a fraction of the time it would take a human.

Machine Learning Awareness

We think it is important that you have sufficient knowledge to get your organization closer to personalized and adaptive learning. That is why we share our machine learning expertise in understandable language through different channels, and help you gain insight about various options for working on automation of your content libraries. In addition, we regularly have phone calls, presentations and workshops on this subject and we have made some useful tools.

Do you have questions or is something unclear? Don't hesitate to contact us.

EDIA
+31 20 716 3612
info@edia.nl
www.edia.nl