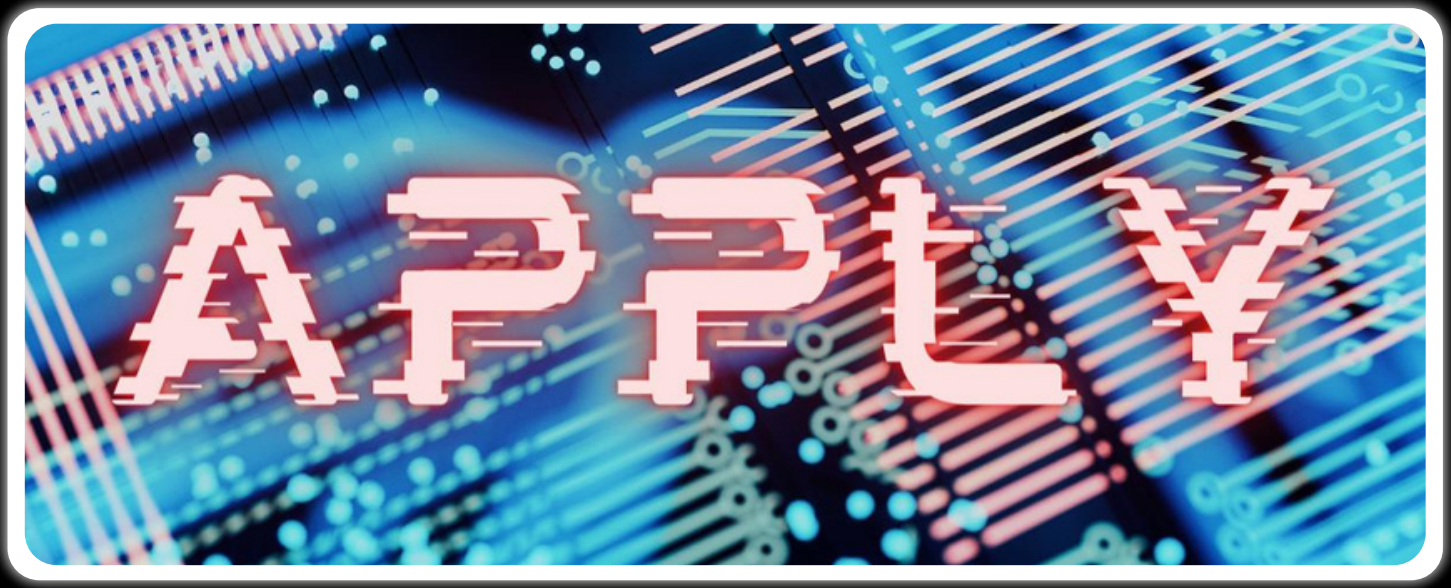
AI Readiness Diagnostic Findings

Step 5: Foundational Recommendations



Step 5 Overview

To find out how you can benefit from examining your institution through a 'data and AI lens', contact our AI & Data Science team at hello@educateventures.com

- The kind of machine learning AI that has been discussed in all the recommendations for the AI Readiness Diagnostic Findings has so far been **supervised** machine learning. This is the type of machine learning used when you want to train the AI to find something **specific** in the data, such as a child's face, or particular grade of exam script
- However, we **do not always know exactly** what we want the AI to find in data, so we need another type of machine learning that can find patterns in the data: **unsupervised machine learning**
- This is the tool we use in a situation where **we do not know what we are looking for** and so we cannot get the **algorithm** to learn what the **target data** we want to find looks like
- With **unsupervised** machine learning, the algorithm looks for **patterns,** searching for **similarities** that might surprise us
- Data that might be fed into an unsupervised machine learning algorithm could be:
- **Log data** from interactions with an online learning platform such as mouse clicks
- **Audio** from user conversations in breakout rooms in Zoom
- **Performance data**
- **Eye-tracking data**
- **Survey responses**
- Preparing this data is key and deciding what machine learning AI technique to apply to it will depend on the **context**
- You may not want to use **all** the data you have with just one AI technique anyway. You may end up applying some machine learning, and with the remaining data, using some more **traditional methods** not based in AI
- Human intelligence will need to be used to help clean – **label** – the data as well, in removing **errors,** and **feature engineering**
- Feature engineering is where humans help **describe** patterns so that the AI isn't scrambling about identifying commonalities with the data that make absolutely **no sense**

- **Key Takeaway:**

- Unpacking what AI can do with the data that you've got will let you make greater sense of both the **data** you've collected, and **the challenge itself.** It may even reveal something in the data you had no idea was there. But it takes a lot of time to **prepare** the data, and if it isn't **clean,** you can get a lot of nonsense information out the other end. With an **increased understanding** of your challenge, you will be in a much better position to select the **AI tools and products** you need to make your life easier in your educational setting or business

Recommendation: Approaches to applied AI, part 1

Before we begin: it is not expected that you are able to apply the following AI techniques to your data, rather, that by familiarising yourself with them, you understand more about AI and how it might help you approach your challenges

SUMMARY: help contextualise the process of applying AI by likening it to the process of cooking. Cooking methods, ingredients, and washing and chopping all map on to the steps needed to prepare data for AI

• It's time to prepare to apply some of the **AI techniques** at our fingertips to the **data** that we've collected on our selected **challenge.** For more on **how to pick your challenge and why,** visit **Step 2** in the AI Readiness Framework. An easy way to **conceptualise** our situation then, is to view it a bit like **cooking.** There's lots of ways to make a **recipe**

• The point of discussing cooking is to show that the **fundamental principles** are not so dissimilar when it comes to data and the application of AI, and particularly the use of machine learning. Think of the data we have as being **ingredients**

• In terms of AI techniques - **cooking methods** - we have many that are available to us. Once the options have been refined to one (in this case the discussion will be about **unsupervised machine learning** - arbitrarily equivalent in our analogy to the cooking method of **baking**) we will still have choices to make about how precisely we wish to **combine** the data - our ingredients

• There are also lots of different AI techniques that we can use if we add in **Good Old-Fashioned AI** (GOFAI) as well as machine learning and then **deep learning.**

It's really about trying to pick what's best for the **ingredients** - for the data - and for the purpose of exploring our data to **learn more** from it

• Our goal in the analogy is now to produce 12 identical **desserts.** The question becomes **how,** and **what type** of dessert. Meringue, trifle, cake or souffle?

• Let's make a souffle:

• We have to go through a set of **preparations** in order to be able to bake. Think of the baking as being a machine learning **algorithm,** and we've got to do a lot of preparation before we start

• First, we have to **wash** our raspberries. Then we have to **crack** the eggs and **whisk** them. Then we have to **add** the sugar into the eggs, and then **whip** up the cream

• There is a bit of **muscle action** needed for the whipping up of the cream. We have to **add** that cream to the beaten eggs to which we've **added** the sugar. And then we need to **scoop** it all together in a bowl and **pour** the mixture into little souffle dishes, and **add** a raspberry on the top as well

• Then we need to **repeat** until we have our 12 beautiful raspberry souffles, ready to go into the oven to be **baked**

• **It probably takes longer to do all that preparation than it does to bake them,** because a souffle is really quick to bake

Recommendation: Approaches to applied AI, part 2

SUMMARY: help contextualise the process of applying AI by likening it to the process of cooking. Cooking methods, ingredients, and washing and chopping all map on to the steps needed to prepare data for AI

• AI is not the only way to **analyse** data and many of the traditional ways of analysing data, such as **statistical analysis,** are well-tested and perfectly valid analytical techniques. We use AI, however, to discover something **extra,** and we use it to learn more about both it, and our challenge

• Recall that we have data - our **ingredients** - around our fictional challenge (that of maintaining the quality of teaching and learning online during the pandemic) and this data consists of perhaps **recordings** of students taking online courses, or **log data** from interacting with a voice-activated personal assitant, or maybe some graded **assignments**

• We need to think what **broader** set of ingredients we can include: the extra kinds of things that are going to make a difference

• With the raspberry souffle analogy, we decide how we mix all those ingredients, and use different techniques on them to prepare them. We **wash**, we **clean**, we **mix** two different types of data to produce an **outcome** - our desserts

• From our example challenge, we could examine: Data from **interactions** that students have, as they're learning online - log data interactions with technology

• That could include **button clicks** for example – how quickly or slowly students are clicking the **mouse** or pressing the **keys**

• **Conversations**, perhaps recorded through the online platform, or in breakout rooms or actually conversations that happen face to face

• **Performance data**

• We could have a whole host of different types of data, and some of it could be from the last few weeks, and some of it could be more historical. We could have lots of **live recordings** of the sessions, and we might be able from that data to have included something about **eye tracking** so we can see the user looking at the camera, at the other people. Are they looking at the task, are they looking at something else? We're talking about situations where we either have:

• Some of the users in a physical location, in the lecture hall, in the seminar room, and some **online,** or

• We have **everybody** online

• Some back present in a **physical location** and a few people **online**

• It's a **mix.** So think about data that comes not just from online but also for students who are present physically for at least part of the time that we're interested in analysing and discovering more about

• And what has all this got to do with **machine learning?** Read on to find out!

Recommendation: Types of AI that can be applied, part 1

SUMMARY: a quick introduction to supervised and unsupervised machine learning, and how the data for either should be prepared

- Machine Learning is a huge set of different sorts of **techniques** and there isn't space here to explore all of them. Increasingly, people become **specialists** in a particular technique or a particular **category** of techniques. The point of covering them here is to **familiarise** you with the kinds of things AI can do, to help you learn more about your **challenge** and how AI might assist with it

- For our cooking analogy, which we have developed for our challenge - that of maintaining the quality of **teaching and learning online** during the pandemic - **supervised machine learning** and **unsupervised machine learning** will be the AI techniques we study

• In a **supervised** situation, we **know** what we're looking for. The algorithm is set off to explore the data, and **identify** what we're looking for

• In an **unsupervised** learning situation, however, we are in a situation where we **don't know** what we're looking for. The algorithm explores the data to look for **patterns** or **similarities** in the data that might tell us something we **don't** already know

• They are two different types of machine learning for two different sorts of **activities.** What we want to achieve is either **identifying** something within the data, or exploring the data to look for **patterns** in the data

• With **supervised** learning, **classification** can be used to classify the particular types of things within a data set. With **unsupervised** learning, **feature reduction** and **clustering** will be used. Revisit the steps above to remember what data we might have at our **disposal**

• In terms of the workflow of our example, there are four different sorts of data: **logs, performance data, recordings, and survey results**

• Now we prepare that data. In the cooking analogy, we **washed, cleaned, mixed, added, broke some eggs etc.**

• In our example challenge, the baking is like applying the **algorithm,** and there are all those other things we have to do before we open up the oven. So that's precisely what we need to do now: we need to bring all these different ingredients **together**

• **Clean the data**

• **Organise the data**

• **Transform the data into a dataset**

• Cleaning and organising: we are going to make the data **uniform;** try to make it look **neat** and **nice** and **easy to work with,** like you might in a spreadsheet

• It's really important that the data we use in order to conduct our exploration is **very high quality.** It should be **accurate, complete** as can be, it should be **consistent** and **uniform**

• We have to try not to change the **values** of the data that we have, only the way that it's **organised,** in order to ensure that we can use it with our **unsupervised machine learning algorithm**

• Don't think of anything as ever being **raw data.** Raw data doesn't exist, because all data is collected for a **reason;** somebody has made a decision to collect that data, or for that data to be collected whilst other activities are going on. There's a **context** for that data, none of it is just raw, there's always added **contextual information** that's important

• Remember that as soon as we're trying to get hold of that contextual data, we're also starting to run the risk of being **intrusive** and we have to get the balance right. For example, if we're recording everything that's going on when users are online, we're capturing a lot of the **context** of their home background, context of the noises that are happening in their home background, for example. We have to make sure we're **ethical** in the way that we treat that data, if it contains something concerning. Perhaps it invades **privacy,** then we would need to make sure that this was **removed**

Recommendation: Types of AI that can be applied, part 2

SUMMARY: a quick introduction to supervised and unsupervised machine learning, and how the data for either should be prepared

• We have to ensure the **reliability** and **credibility** of the data that we're going to process when we apply our machine learning algorithm. We need to remove **errors,** and look for **impossible values and incorrect values.** There are always **mistakes** in data, so we have to go through it carefully to check that they've been removed. We need to look for **duplicate entries,** for example, **irrelevant data.** Perhaps, in the context of our example challenge, there is data about student age but it isn't relevant to our investigation because it was in an **existing dataset** that we wanted to use, but we didn't want all of the data in that dataset

• Look for **missing data** and look for **outliers.** Outliers are values that are **radically different** to most of the values in the dataset. We have to be **careful** though without those in the main dataset. We can **remove** them, but we have to be careful not to remove too many outliers otherwise we just have **averages.** A **mix** of data is required, but we can decide what action to take about outliers and then make sure that we apply that approach **consistently** to all our data

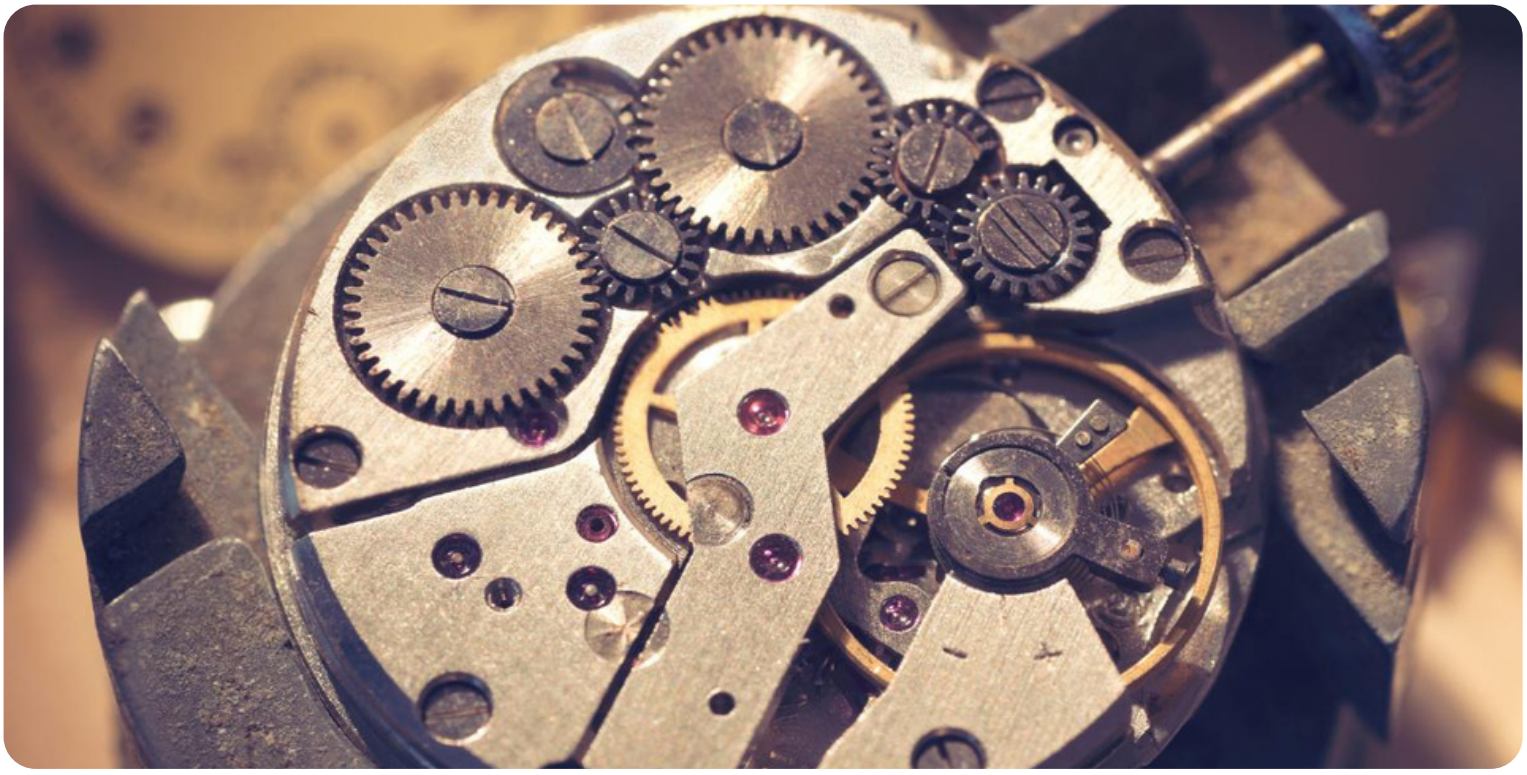
• Different datasets have different **labelling** methods, and so we might have user gender, for example, which could be described in one data set as female or male, and another as girl or boy. There are **typos, syntax,** or **conventions,** and we've got to make sure that all the labels use the **same conventions** so that they are recognised by the algorithm to be the same thing if they are indeed the same. The algorithm won't know that female could be **classified** the same as girl. So we have to make sure that we only use "female" or only use "girl" in the labelling of the data

- Always aim for **quality,** and always **document** the approach adopted, so that if for example the analysis doesn't behave in the **expected way,** we can go back and look at that documentation, and see if there was anything about the approach used that perhaps could have been done **better**

- The precise actions we take depends on the dataset in the same way that, in the baking analogy, we wash the **raspberries** but we don't wash the **eggs** – we **crack** them. We weren't using the **shells,** we certainly didn't wash the **sugar** or the **cream.** It sounds trivial but it's really important that different types of data be treated differently. Doing so ensures reliability and credibility

- **This is a lot of work, and will take at least 80% of the time involved in applying the AI to your data**

- Once the data has been cleaned, and pulled all together, it needs to be **integrated** and any **transformations** performed. **Transformation** means that there may be aspects of the **structure** or **format** of the data that need to be changed in order for it to be analysed more **accurately.** For example, perhaps the format of data that describes actions taken through eye tracking is different for different types of **capture.** Perhaps we thought that we were using the same video capture device throughout the recordings, but actually a **change** was made by our provider without us **realising.** That makes a difference to the **structural format** of the data. So this is not about transforming **values.** This is about **organising** the structure and the format, so that there is **consistency**



Who can help me?

We are specialists in ethical AI solutions for schools and education and training businesses - contact our team for help

The EDUCATE AI and Data Science team was formed to consult on and co-design ethical AI solutions to complex problems in data-driven technology ventures and schools. Our team of computer scientists, educationalists, and world-renowned experts can take you from zero AI to a comprehensive evidence- led strategy and beyond, with effective, scalable AI-powered teaching and learning solutions.

To find out how you can benefit from examining your institution through a **'data and AI lens'**, and leveraging the transformational power of AI to tackle your challenges, contact the **AI and Data Science Team** at EDUCATE Ventures Research at **hello@educateventures.com**.

Thanks for reading!

- The EDUCATE Ventures Research Team Summer 2023

Further Reading

Below you can find a selection of resources, books, podcasts, webinars, and research papers appropriate to your stage of AI Readiness. Good luck!

- AI for School Teachers, Byte-Sized Edition
  An easy-to-read 10-page byte-sized summary of the book of the same name, written by Professors Rose Luckin, Mutlu Cukurova, and Headteacher Karine George, members of the senior team actively developing and using the AI Readiness Framework from which these recommendations derive

- Alan Turing Institute: Three Question
  The Turing Lecture mini-series is designed to reflect on the use of AI and data science in a post-lockdown world. Professor Luckin's lecture centred on the use of AI and tech in education - particularly in a virtual setting due to the pandemic. In addition, she gives her personal perspective on the use of data and tech to decide exam results across the UK

- China's Grand Experiment in Education
  An MIT Technology Review article on the country's intelligent education revolution