AI / Machine Learning - December 2, 2021

# The Importance of Data in Artificial Intelligence (AI)

Mikaela Pisani

AI effectively mimics the reasoning and thought processes of the human brain to replicate in our everyday applications. This is seen a lot in Cybersecurity with task

But just like a car, at the heart of any AI system is the fuel that it is being fed. But rather than gasoline, it's data and lots of it. So, the focus of this article is to help you understand the key role data plays in AI.

### Come join us

*Do you want to work with data and AI? We're hiring!*

## Why is Data in AI important?

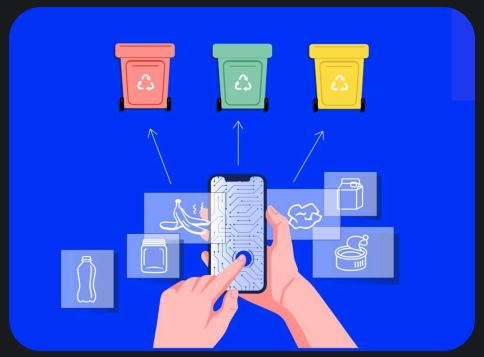
Good question, here are some key reasons why AI needs good data:

### 1. It's garbage in and garbage out

The answer you seek from an AI system is known as the “output”, and the only way you

any of these are off-scale in any way, your output will get skewed, and your results will send you in the wrong direction.

A prime example of this (and no pun intended on the garbage reference) is when we used machine learning to build a waste classifier app. Data was absolutely key to the success of this project.



Waste classifier app

## 2. What are the characteristics of a good data set?

application of which the AI system is serving. But, in general, the following are features you should look out for when parsing through datasets:

- **It is complete:** By this, there are no empty spots or cells in your datasets. Every slot has a piece of data in it, and there are no visible holes in them.
- **It is comprehensive:** The datasets are as complete as they can get. For example, with Cybersecurity if your goal is to model a threat vector, then all of the signature profiles from which it emerged must have all of the necessary information.
- **It is consistent:** All of the datasets must fit under the variables that it is has been assigned to. For instance, if you are modeling gasoline prices, your selected

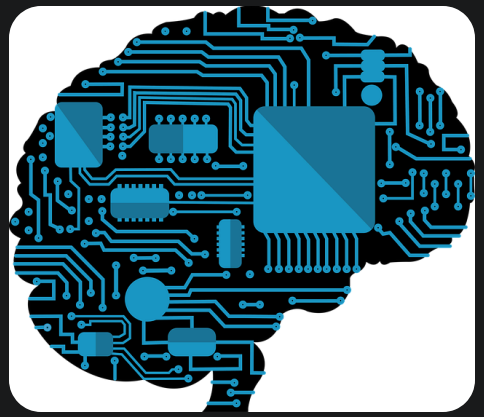
data to fall into those categories.

- **It is accurate:** This is key. As you will be selecting various feeds for your AI system, you must trust these data sources. If there are chunks that are not accurate, your output will be skewed, and you will not get a correct answer.
- **It must be valid:** This is crucial with time series datasets. You don't want old data that could interfere with the learning process of the AI system when analyzing recent datasets. So, let it learn from recent data. How far back depends on your application. With Cybersecurity, for example, going back a year is typically enough.
- **It is unique:** Similar to consistency, each piece of data must be unique to the variables it is serving. For instance, you

## 3. Not all AI systems are built equally

With actual datasets, we often think of a long series of numbers i.e. quantitative data. But, there are also datasets in qualitative data i.e. videos, pictures, etc.

With AI systems, these datasets are known as “Structured” and “Unstructured”, respectively. It's important to note that not all AI systems can handle both of these sets.



Machine learning

important to select the right dataset for your system, or your output could yield a different answer than what you imagined.

## 4. The issue of quality versus quantity

For an AI system to learn and produce the desired outputs; it must first ingest and learn from a lot of data. It doesn't take a long time to process this, so the question now arises: quality over quantity? The latter is always preferred.

Although it will take the AI system longer if datasets are shorter in nature, you will have some guarantee that your output will be robust and relevant. It's not productive to feed an AI system lots of data just for the hope that it will learn something from it.

# What to take away

sources, you still need to do your due diligence in making sure that the datasets conform to your requirements.

This requires targeted testing and sampling, and possibly running smaller training exercises to ensure they are being fully optimized. This hard work will pay off in the long run.

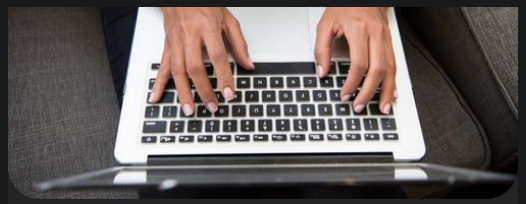
← Back to blog

## Featured articles


Ruby

## How to Use Counter Caches in Rail...

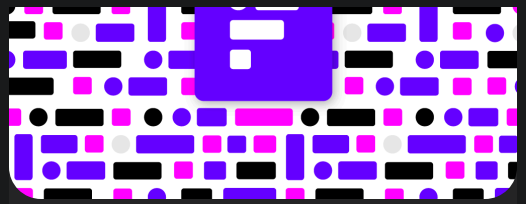
Read →



Hiring

## The Fastest & Cheapest Ways to Hire ...

Read →



yaaf

## Introducing yaaf

Read →

## Never Miss an Update!

Join our community of insiders and never miss out on exciting news, product launches, and more.