DATASETS IN AI AND ML

A Christmas Tale – The Importance of Labels

A label on a gift box can convey a number of things. The obvious reason being who it's for and who has sent the gift. One can also attach heartfelt sentiments giving a guess of what might be inside, but what is for sure is that if there is a gift box without a label, then all kinds of fun can ensue. Imagine for a moment your family sat around the Christmas tree with all the presents laid out in front of them but missing are the all the important labels from their boxes. The chaos that would occur may make entertaining tales in future years, but the process of tackling with who the gift is for and who sent it may even cause a severe headache.

A second scenario could be that the labels are put on the gift boxes, but not necessarily the correct gift box. Again, chaos and confusions would prevail and maybe even offend relations. The process of not labeling, or equally mislabelling a gift box is something we would not encourage. Putting up relevant labels on relevant gift boxes is what all encourage.

Figure1: A basic understanding of the Importance of Labels.

Similarly when we consider Artificial intelligence and Machine learning, datasets with relevant tags play a crucial role. Tagging data for the above applications is accomplished by data labeling or data annotation tools. Labeling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative.

Many organisations do not implement a labeling process for their datasets. Without such a process in place, the value of that data and to whom it applies to can be overlooked and misappropriated. Implementing a data classification solution within your organisation ensures employees to understand the value of the data they handle. This reduces the risk of costly data leak.

Data Labeling Market Outlook – 2025

The data annotation tools market size was valued at USD 316.2 million in 2018 and is projected to register a CAGR of 26.6% from 2019 to 2025. The growth is majorly attributed to increasing adoption of image data annotation tools in the automotive, retail, and healthcare sectors. Data annotation tools enable users to enhance the value of data by adding attribute tags to it or labeling it. The key benefit of using such tools is that the combination of data attributes enables users to manage the data definition at a single location and eliminates the need to rewrite similar rules in multiple places. The rise of big data and surge in number of large datasets are likely to necessitate the use of artificial intelligence technologies in the field of data annotation.

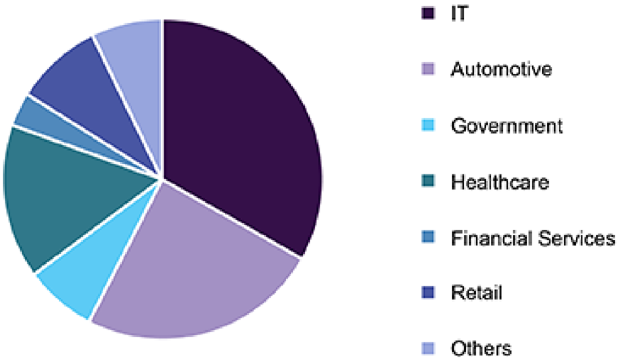

Figure2: Data Labeling Market Outlook.

Fueling the Gold Rush: The Greatest Public Datasets for AI

It has never been easier to build AI or machine learning-based systems than it is today. Open-source tools such as TensorFlow and Torch coupled with the availability of massive amounts of computation power through AWS, Google Cloud, or other cloud providers has made training of cutting-edge models an easy go task. Though not at the forefront of the AI, **the unsung hero of the AI revolution is data** — lots and lots of labeled and annotated data. However, most products involving machine learning or AI rely heavily on proprietary datasets that are often not released. With that said, it can be hard to settle on which public datasets are useful before you collect your own proprietary data. It’s important to remember that good performance on data set doesn’t guarantee a machine learning system will perform well in real product scenarios. Most people in AI forget that the hardest part of building a new AI solution or product is not the AI or algorithms rather it’s the data collection and labeling. Standard datasets can be used as validation or a good starting point for building a more tailored solution.

Figure3: Public Datasets would be a safer choice to start with.

The Importance of Quantity

The first thing to know about machine learning data is that you need a lot of it. Remember, machine learning helps computers solve problems that are too complex for an algorithm alone. What makes these problems complex? Often, it’s the amount of inherent variation—there are hundreds, thousands or millions of variables. And the resulting system must be able to cope with them all. Think of machine learning data like survey data, the larger and more complete your sample size, the more reliable your results will be. If the data sample isn’t big enough, it won’t capture all the variations or take them into account, and your machine may reach inaccurate results, learn patterns that don’t actually exist, or not recognize patterns that do. Take a speech recognition system, for example. Spoken languages and human voices are extremely complex, with infinite variations among speakers of different genders, ages, and region. You could work with a mathematical model to train a machine on textbook English, but the resulting system would likely struggle to understand anything that strays from the textbook like loose grammar, people with foreign accents or speech disorders, and those who use slang. If you were employing that system for email or text, it would also trip up on the emojis and abbreviations (such as LOL) that appear in typical chat sessions. You would have spent a lot of time and money on something that would utterly fail in the market. The more your machine learning data accounts for all the variation the AI system will encounter in the real world, the better your product will be. Some experts recommend at least 10,000 hours of audio speech data to get a recognizer to begin working at modest levels of accuracy. This same principle applies to new and established products alike.

You need a lot of data to get to market with the best AI solution you can make, as well as to improve and update it. Search engines on retail websites, for example, need constant training to keep up with changing inventory: adding new products, removing discontinued products, adding and removing seasonal items. To ensure customers see relevant results, it’s critical to regularly tune the onsite search algorithm.

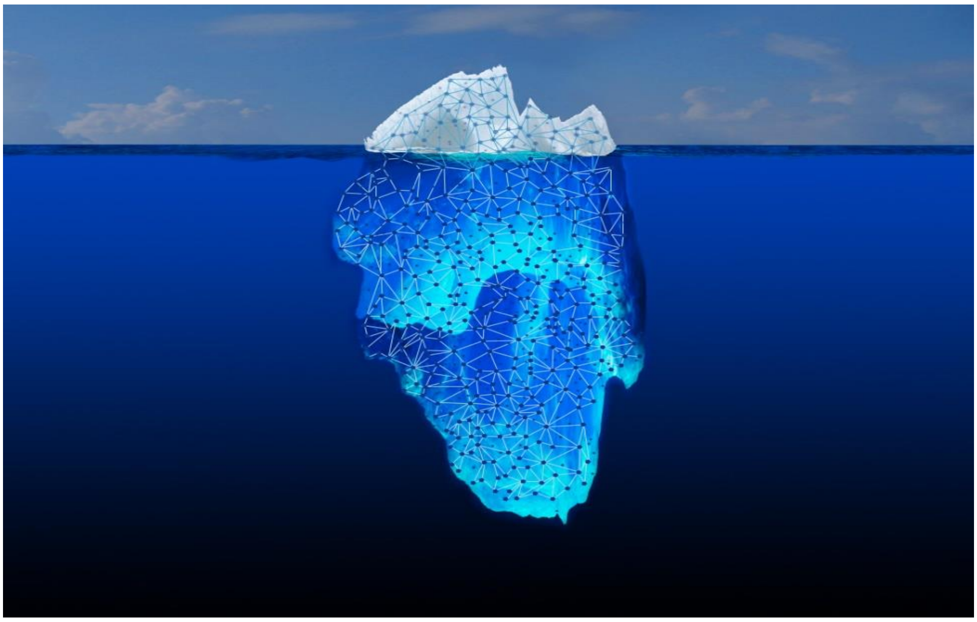

Figure4: Huge Datasets lead to better AI models

The Importance of Quality

Machine learning not only requires a huge volume of data but also the right kind, because ultimately the system will do what it learns from the data. You can have the most appropriate algorithm, but if you train your machine on bad data, then it will learn the wrong lessons, come to the wrong conclusions, and not work as you (or your customers) expect. On the flip side, a basic algorithm won't hold you back if you have good data (and enough of it). Your success, then, is almost entirely reliant on your data.

What defines "bad" data? Many things. The data may be irrelevant to your problem, inaccurately annotated, misleading, or incomplete.
Consider search engine evaluation. To improve a search engine's performance, working with human judges to rate how good each result is for a particular query. For example, if you were searching to find the hotels to stay in a particular geographical area and got two results, one from the trivago and one from a local hotel site, both of which had the answer, which would you trust more? Most people would trust the trivago one, because the site itself looks more trustworthy. This comparison illustrates why preference and relevance are important. A computer can find the data— a hotel to stay in a particular geographical area—but doesn't know which source is better unless it is told. If the evaluators don't interpret the original intention correctly and they train the search engine with the bad data, the resulting model would ultimately fail in the market. In this example, qualifying your evaluators is a key to ensuring you're creating high-quality data.

Figure5: Quality is the signature of trust a dataset needs to carry in the field of AI

Where does the Data Comes From?

Where will your data come from? Broadly speaking, there are four main sources: real-world usage data, survey data, public data sets, and simulated data.

Real-World Usage Data

When your AI products are already in-market, real-world data from actual users is a great resource. With a search engine or search feature, for example, you can look at queries, total results, which results people click on, and what they look at and purchase. Social media sites can gather data about what users post, like, share, and comment on. Speech recognition solutions from smartphones, in car systems, or home assistants can collect spoken queries and the machines' responses. There's also broadcast data from music services and sites like YouTube that may track what people look at.

The benefits of using real data are that you know it accurately reflects how people use your system, and you don't have to pay to create it. However, there are legal questions associated with collecting it, as well as privacy concerns. Some companies have had trouble collecting this data and faced lawsuits when they overstepped.

Survey Data

The second source for machine learning data is surveys. You go directly to your users, or prospective users, and ask what they like or don't like, and what you can improve about the product. This approach gives you data from actual users, and gets around privacy concerns and legal issues as, by taking the survey, people are opting to participate. Surveys provide context and the opportunity to follow up on anything that's unclear. You also have some control over what people say and do in that you can direct them to the specific topics you want to address. On the other hand, survey data is somewhat unreliable, because what people say they do and want on your survey might be quite different than what they actually do and want. Additionally, survey data is often skewed toward dissatisfied users, as people who get what they want are less motivated to provide feedback.

Public Data Sets

There are a number of different types of public data sets available from search engines, social media, Amazon Web Services, Wikipedia, universities, data science communities, and other data repositories. There's also an enormous amount of public data from academic efforts in speech and language processing from the last 40 years, licensable from various organizations. For most commercial purposes, affordability is the real advantage of these data set. This kind of data is often used for applications like basic language recognition or machine translation.

Engineered or Collected Data

The fourth main way to collect quality data is to make it yourself. This is often the only way to proceed with a new solution, when there aren't any users or usage data yet. You can simulate the user experience by hiring speakers and professionals, gathering and annotating the data your project specifically needs. You can mimic the conditions where people will use your product like driving in a car on a city street, etc.

On one hand, you can get exactly what you need faster this way because you're in control. You always know the context. You can follow-up with your professionals and speakers if there's a question. And, since you're not using real data, there are no legal or privacy concerns. Most important, your model will produce a better end result.

On the other hand, this type of data collection will require a larger investment. To do it well, be sure you work with an experienced data collection vendor. There's a lot of management involved to ensure you’re getting the right kind of data. This annotation, like the same done for real world data, also requires qualifying crowds of people to ensure they can label and categorize the data, allowing machines to know what to do with it. If you spend the time and money to build a custom database solution, it would be a waste to end up with messy data from an inexperienced vendor. Simulated data is also not something most companies should attempt themselves. He could hire a data collection vendor to create a set of data.

Figure6 : Looking out for Datasets and choosing the best.

Why is data annotation important in some machine learning projects?

Data annotation is important in machine learning because in many cases, it makes the work of the machine learning program much easy. This has to do with the difference between supervised and unsupervised machine learning. With supervised machine learning, the training data is already labeled so the machine can understand more about the desired results. For example, if the purpose of the program is to identify cats in images, the system already has a large number of photos tagged as cat or not. It then uses those examples to contrast new data to make its results. With unsupervised machine learning, there are no labels, and so the system has to use attributes and other techniques to identify the cats. Engineers can train the program on recognizing visual features of cats like whiskers or tails, but the process is hardly ever as straightforward as it would be in supervised machine learning where those labels play a very important role. Data annotation is the process of affixing labels to the training data sets. These can be applied in many different ways. For example, in the medical field, data annotation may involve tagging specific biological images with tags identifying pathology or disease markers for other medical properties. Data annotation takes work and is often done by team of people, but it is a fundamental part of what makes many machine learning projects function accurately. It provides that initial setup for teaching a program what it needs to learn and how to discriminate against various inputs to come up with accurate outputs.

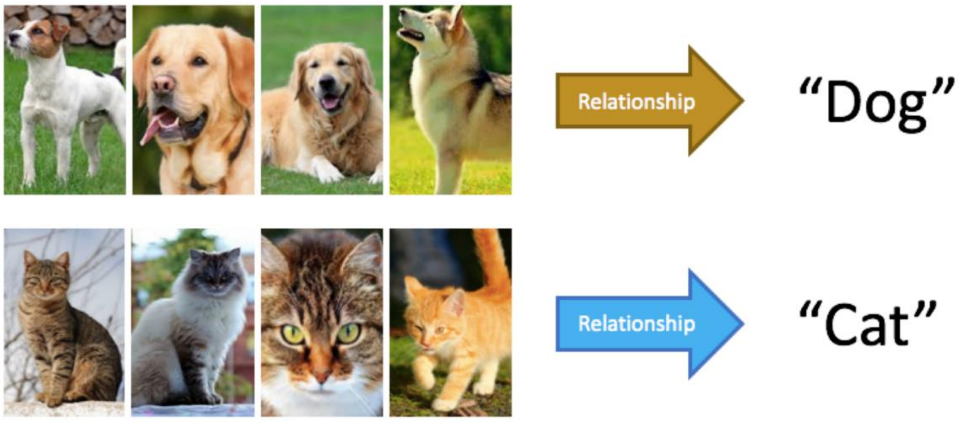

Figure 7 and 8: Labeled Data is the fuel of Supervised Learning

Importance of Data Labeling in Customer Experience

Companies don’t face a lack of data; they have an overabundance of data that isn’t labeled. Data labeling helps improve insights to create Intelligent machines. Data labeling allows companies to attach meaning to data gathered from customer interactions and analyze what is happening in each. By understanding the nature and emotion of each experience, the overall machine performance can be enhanced. Through data labeling and applied analytics, new patterns emerge that were unrecognizable before.

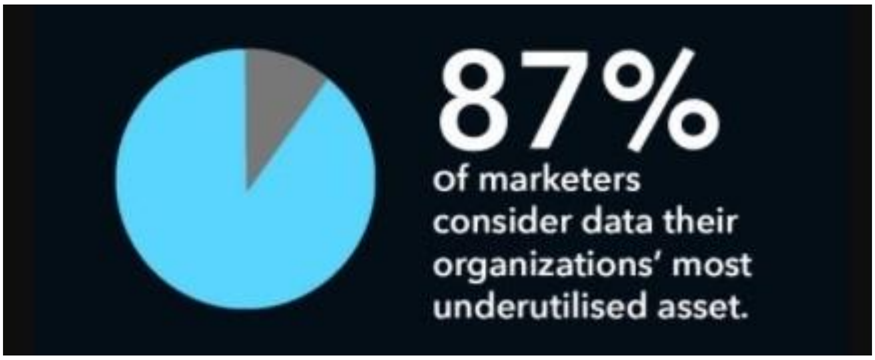

Figure9: Over-abundance of Unlabeled Data

CONCLUSION

Adopting AI and ML is a journey, not a bullet that will solve problems in an instant. It begins with gathering data into simple visualizations and statistical processes that allow you to better understand your data and get your processes under control. From there, you’ll progress through increasingly advanced analytical capabilities, until you achieve that goal of perfect production, where you have AI helping you make products as efficiently and safely as possible. Companies like Google, Amazon and Facebook dominated their industries because they were the first to begin building data sets. Their data sets have become so large, and their data collection and analysis so sophisticated that they are able to grow their competitive advantage. Datasets coming with valid features is a key solution for the success of AI models.