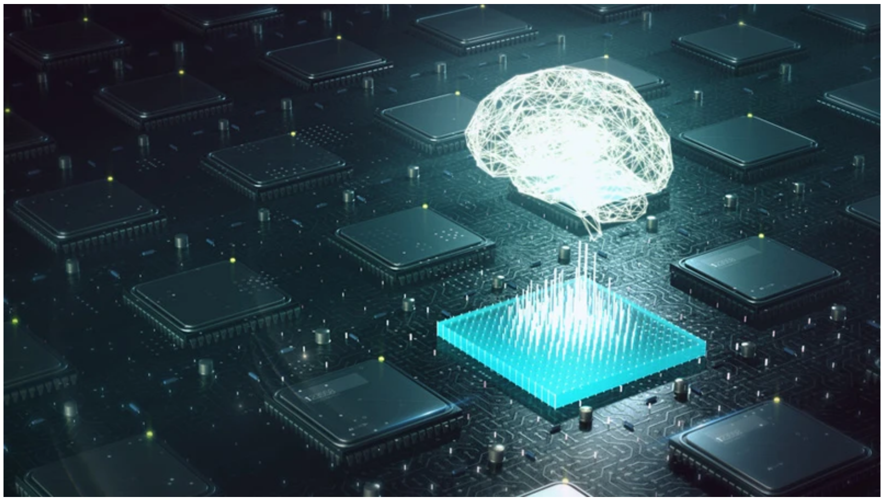
Five approaches to data labeling for machine learning projects

AI DATA POSTED MARCH 2, 2021



The quality of a machine learning project comes down to how you handle three important factors: data collection, data preprocessing and data labeling.

Labeling (also known as data annotation) is often time-consuming and complex. For example, image recognition systems often require bounding boxes drawn around specific objects, while product recommendation and sentiment analysis systems can require complex cultural knowledge. And don't forget that a dataset could contain tens of thousands of samples in need of labeling, if not more.

With this in mind, selecting the right approach for a machine learning project means taking into account the complexity of the task, the size of the project and your project timeline. With these factors in mind, we've listed five common approaches to data labeling along with pros and cons for each.

Data labeling for machine learning

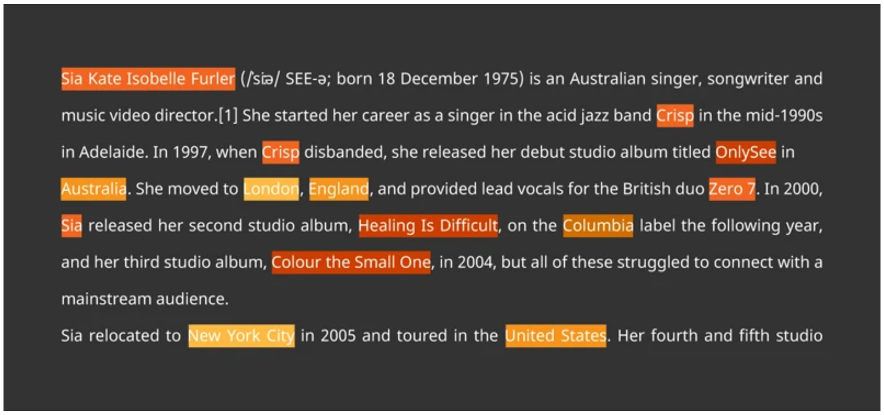
Data labeling for machine learning can be broadly classified into five categories.

In-house: As the name implies, this is when your data labelers are your own team of data scientists. This approach has a number of immediate benefits: tracking progress is simple and accuracy and quality levels are reliable. However, outside of big companies with internal data science teams, in-house data labeling may not be a viable option.

Example of polygon annotation.

Outsourcing: Outsourcing is a good option for creating a team to label a project over a set period of time. By advertising your project through job sites or your company's social media channels, you can create a funnel for potential applicants. From there, an interviewing and testing process will ensure that only those with the appropriate skill set make it onto your labeling team. This is a great way to build a temporary team, but it also requires a certain amount of planning and organization; your new staff will need training to become adept at their new job and complete it to your specifications. Furthermore, if you don't already have one, you might also need to license a data labeling tool for your team to work on.

Crowdsourcing: Crowdsourcing platforms are a way to enlist help from people around the globe to work on particular tasks. Because crowdsourcing jobs can be picked up from anywhere in the world and performed as tasks become available, it is extremely quick and cost effective. However, crowdsourcing platforms can vary wildly in terms of worker quality, quality assurance and tools for project and worker management. Therefore, it's important to be aware of how the platform approaches these factors when looking at crowdsourcing options.



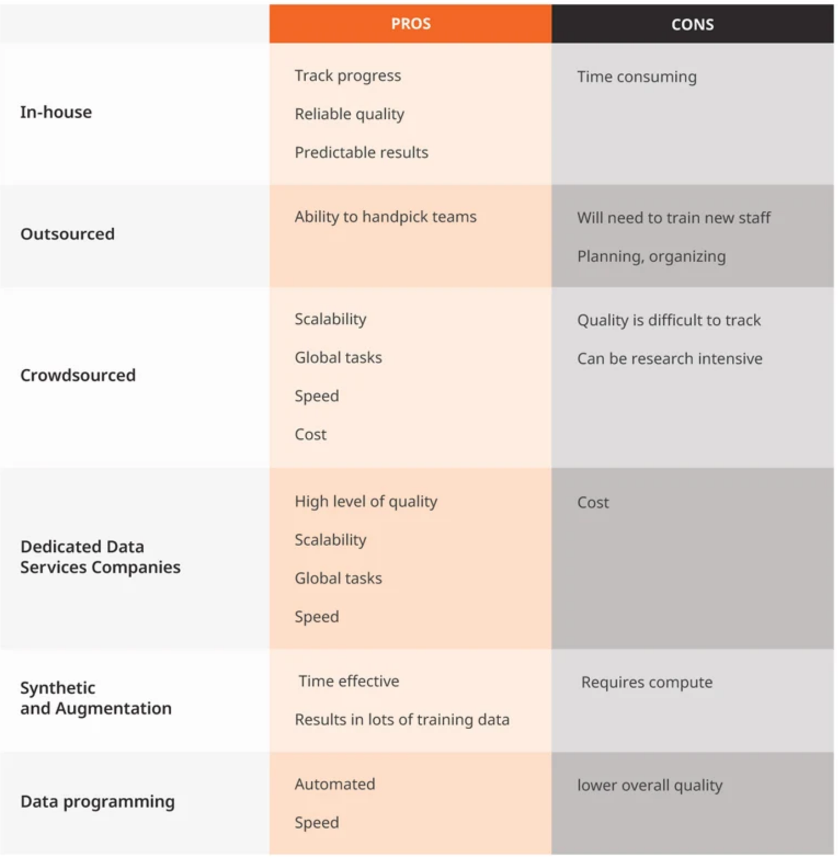
Sia Kate Isobelle Furler (/ˈsiːə/ SEE-ə; born 18 December 1975) is an Australian singer, songwriter and music video director.[1] She started her career as a singer in the acid jazz band Crisp in the mid-1990s in Adelaide. In 1997, when Crisp disbanded, she released her debut studio album titled OnlySee in Australia. She moved to London, England, and provided lead vocals for the British duo Zero 7. In 2000, Sia released her second studio album, Healing Is Difficult, on the Columbia label the following year, and her third studio album, Colour the Small One, in 2004, but all of these struggled to connect with a mainstream audience.

Sia relocated to New York City in 2005 and toured in the United States. Her fourth and fifth studio

Example of entity annotation.

Synthetic: Synthetic labeling is the creation or generation of new data that contains the attributes necessary for your project. One way to perform synthetic labeling is through generative adversarial networks (GANs). A GAN utilizes two neural networks (a generator and a discriminator) that compete to create fake data and distinguish between real and fake data respectively. This results in highly realistic new data. GANs and other synthetic labeling methods allow you to create all-new data from pre-existing datasets. This makes them time effective and excellent at producing high quality data. However, at present, synthetic labeling methods require large amounts of computing power, which can make them very expensive.

Programmed: Programmatic data labeling is the process of using scripts to automatically label data. This process can automate tasks including image and text annotation, which eliminates the need for large numbers of human labelers. A computer program also does not need rest, so you can expect results much faster than when working with humans. However, these procceses are still far from perfect. Programmatic data labeling is therefore often combined with a dedicated quality assurance team. This team reviews the dataset as it is being labeled.

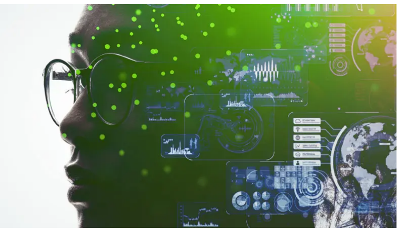


	PROS	CONS
In-house	Track progress Reliable quality Predictable results	Time consuming
Outsourced	Ability to handpick teams	Will need to train new staff Planning, organizing
Crowdsourced	Scalability Global tasks Speed Cost	Quality is difficult to track Can be research intensive
Dedicated Data Services Companies	High level of quality Scalability Global tasks Speed	Cost
Synthetic and Augmentation	Time effective Results in lots of training data	Requires compute
Data programming	Automated Speed	lower overall quality

A chart summarizing approaches to data annotation for easy reference.

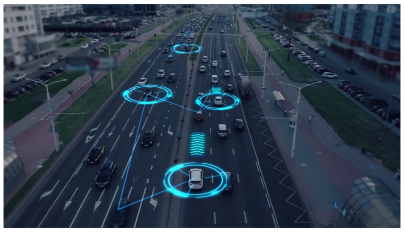
Each different approach to data labeling has its own strengths and weaknesses. Knowing which approach is best for you depends on a number of factors. These can include the complexity of your use case, the training data, the size of your company and data science team, your finances and your deadline. Be sure to keep these in mind when considering data labeling solutions.

Related insights



AI BEST PRACTICES

The essential guide to AI training data



DATA ANNOTATION

AUTOMOTIVE

Enhancing an autonomous vehicle bot’s scenario understanding with tailored fine-tuning datasets

DATA ANNOTATION

Everest Group Data Annotation and Labeling (DAL) Solutions for AI/ML PEAK Matrix® Assessment 2024

Digital Services

Trust, Safety & Security

Fintech & Financial Services

Travel & Hospitality

Games

Retail & Ecommerce

Healthcare

Automotive

Our Awards

Newsroom

Technology Partners

Subscribe to Newsletter

WillowTree, a TELUS Digital Company

Cookie Preferences Do Not Sell my Personal Information

Website User Terms | Privacy Policy | TELUS Digital is a subsidiary of: TELUS Communications Inc.