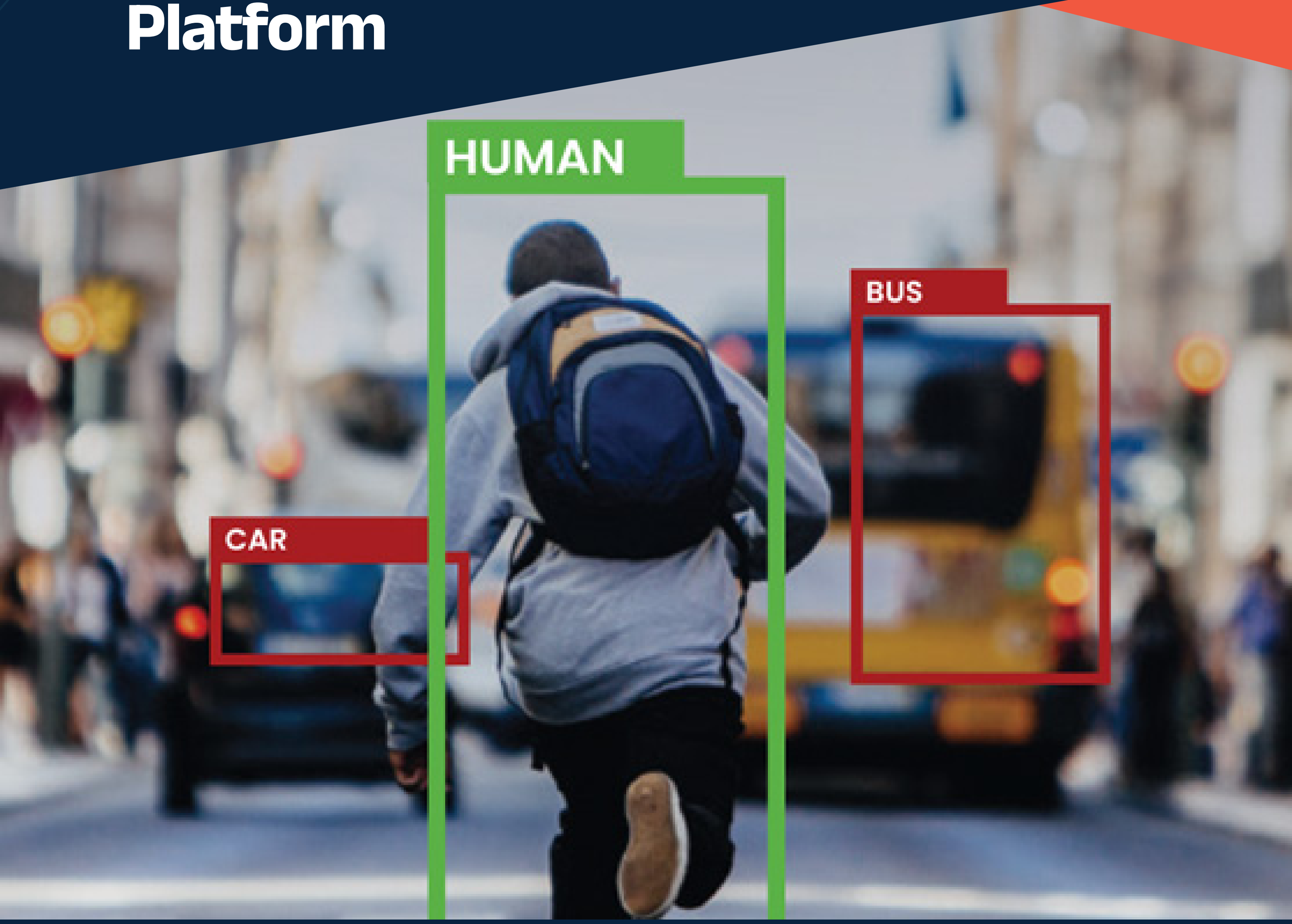
A White Paper

Build a Scalable Data Annotation Pipeline with Zastra™, an Active Learning-based, End-to-End Data Curation and Annotation Platform



Abstract

Data is the backbone of artificial intelligence (AI) and machine learning (ML) algorithms. However, having raw data is not sufficient in and of itself. For the ML algorithm to correctly recognize the items in each image, comprehend human speech, and perform a variety of other functions, you need to have this data annotated.

For enterprises, achieving this large data set of labeled images to create an annotation pipeline is challenging and expensive. According to Google, annotating a full dataset can easily take 15,000 hours of labor[1]. Open-source tools may accomplish this target, however, when trying to scale, such tools become obsolete.

This white paper elucidates enterprise decision-makers to devise data annotation strategies, for effective data labeling, highlights the challenges, and underscores the need to select the right framework. It also introduces Zastra™ – a data annotation framework – and looks at how it can help businesses become more competitive.

[1] https://arxiv.org/pdf/1806.07527.pdf

"By 2030, AI has the potential to generate an additional $13 trillion in global economic activity. Given all of the potential advantages of AI, it is crucial that all of the data be properly labeled in order to maximize its usefulness. "

- McKinsey

Why does Data Annotation Matter?

Today, smart AI has several real-world applications – autonomous driving, medical diagnostics, weather forecasting, smart assistants, web search optimization, and so on. Given these scenarios, humans make decisions based on the input they receive. This input could be images, text, or audio files.

> *85% of AI projects fail to make it into production largely due to data.*
>
> \- Gartner

Computers are unable to analyze visual information in the same way as the human brain, therefore, we must instruct them on what to interpret and provide them with the context in order to make decisions. This capability of algorithms to deliver on these promises depends on data annotation – the act of accurately categorizing information to educate AI to conclude.

According to Research And Markets' findings, globally, the data annotation market was valued at $695.5 million in 2019 and is projected to reach $6.45 billion by 2027 and is expected to grow at a CAGR of 32.54% from 2020 to 2027.

Factors Contributing to the Growth of Data Annotation Needs

This increase in demand has been influenced by AI-based services being used across industries.

- To create content assets and enhance customer experience, industries like healthcare, transportation, telecommunication, and e-commerce are gathering datasets from many sources and classifying them according to their context, need, kind, and feature.
- The other driving factor includes the combination of mobile computing platforms and digital image processing. Utilizing this technology, banking, financial, and insurance institutions can connect with customers in real-time and assist with document verification.
- The agricultural industry is using this technology more and more for things like soil testing and crop monitoring.
- Through a variety of digital platforms, such as social media, websites, and apps, among others, users and organizations have created, shared, and interacted with a tremendous amount of digital information, including photographs, videos, and text. Data annotation services are helpful for these companies as they use online material to add value and attract clients.

Data Annotation & Labeling Market Trends

Manual vs. Automated Annotation	Despite conventional annotation, automatic options are gaining traction as research in IoT, and ML products advance. By 2030 it is expected to grow at a CAGR of 18%. In AI, auto-labeling is the upcoming big thing.
Data Annotation Tools	With the surge in big data developments and large datasets, the demand for data annotation tools is spiking up. As per Grand View Research, the global market is estimated at $494 million in 2020 and is projected to rise at a CAGR of 27.1% from 2021 to 2028. The use of picture annotation tools would be more in the automotive, retail, and healthcare sectors.

Image and Video Annotation	The fastest-growing categories of data annotation are photos and videos, with a CAGR of around 17% between 2020 and 2030. Both types of annotation are used in manufacturing, energy & utilities, and the automotive industry. Most image annotation uses are in the field of medicine as medical imaging.
Text and Audio Annotation	The use of audio labeling is low, and text annotation is growing with the surge in e-commerce and medical research. Text annotation is dominating the global labeling market to fine-tune AI's capacity to recognize patterns in text, voice, and semantic connection of annotated data.

Source: https://labelyourdata.com/articles/trends-in-data-annotation-market-forecast-2022

Challenges of Data Annotation

The process of data labeling has its downsides, making it one of the most complex, time-consuming, and expensive processes if not managed properly. An estimated 80% of the time spent on AI projects is spent on data-related tasks.

The most common challenges faced by organizations related to data annotation and labeling can be clubbed under the following classifications:

- Labor Intensive
- Lack of understanding of Technology and Access
- Huge Maintenance Costs
- Data Security & Compliance

Open-source tools may be reliable, and flexible and come with added benefits to accomplish these tasks, but there are many instances when they just aren't enough.

Cigniti's flagship data annotation tool Zastra^{TM} is an Active Learning-based, data curation and annotation platform for image and video data that reduces the time required to label and annotate data to power your business applications.

Zastra^{TM} provides a unified, collaborative environment for managing the entire AI model development, training, and deployment process.

Its key capabilities include Active-Learning based object classification, object detection, localization, and segmentation (upcoming feature). Mentioned here is a list of use cases that Zastra^{TM} had used for successful outcomes.

Data Annotation – Industry Use Cases

Autonomous Vehicles – annotation of images

The object in an image or video is labeled using the bounding boxes and defining other attributes to help the ML model understand and recognize the object detected by the sensors of the vehicle.

But the number of images generated by a single autonomous vehicle is very large. For example, a self-driving car travels 10,000 miles at an average speed of 20 miles per hour. This comes to around 500 hours of travel time. Assume the cameras on the car take a bare minimum of 10 frames per second. The total number of images taken during the travel time of the car comes down to 18 million images.

Imagine the tedious task of annotating these 18 million images using manual work. This is humanely impossible.

Here comes the annotation capability of Zastra™. A minimum number of images can be annotated using manual work and the model uses an active learning method to annotate all the available images to build a data set. This data set can be used in the ML model to distinguish different objects by the autonomous vehicle, thereby providing a safe, smooth, and worry-free driving experience to the end customer.

Satellite Imagery – detection of defects

One of the research projects done at Cigniti is the detection of defects present on windmills. Usually, there is a contract period between power generation companies and windmill manufacturers to provide replacements for the defective parts if the defects are reported within a certain duration.

So, power generation companies are in a hurry to find these defects before the expiration date. But the windmills are enormous and numerous units of windmills are present across a gigantic coastal or desert area. Covering a single windmill and finding defects in it by manual work may take up to 2 days. Whereas, covering 1000 windmills across the area takes nearly 2000 days of manual work. By this time, the contract period would end.

With the advancement of technology now, this time-consuming and expensive process can be reduced to a few weeks of work. Drone cameras are used to capture images and the inbuilt ML model, and the annotated data helps to detect the defects instantaneously on the ground and send the report to the team in a matter of seconds.

Healthcare – diagnosis of diseases

Medical image annotation refers to the labeling of medical images such as X-Rays, CT-Scans, MRIs, and PET scans.

Radiologists can detect life-threatening diseases with objectivity using data annotation and computer vision technology. Use cases such as breast cancer detection, tumor classification, finding hidden fractures, detecting neurological abnormalities, etc. One such capability we developed is the diagnosis of Chronic venous insufficiency (CVI).

Chronic venous insufficiency occurs when your leg veins don't allow blood to flow back up to your heart. Normally, the valves in your veins make sure that blood flows toward your heart. But when these valves don't work as they should in normal conditions, But when these valves don't work as they should in normal conditions, blood can also flow backward. This can cause blood to collect (pool) in your legs.

Left untreated, these damaged veins can cause serious complications, such as ulcers, bleeding, and a potentially fatal condition called deep vein thrombosis.

In the process of finding a solution for this health condition, we at Cigniti have used MRVenograms of adult patients. A total of 3640 individual leg images were extracted from the MRVenograms of 26 patients. We developed a neural network model that can diagnose the disease with 97% accuracy. This achievement helps in the diagnosis of CVI in an objective way, which is not possible at present.

Minimize Data Annotation Efforts with Zastra™

Zastra™ is an end-to-end, enterprise-grade annotation workflow platform that minimizes data annotation efforts and maximizes collaboration.

It uses state-of-the-art Active Learning methods to reduce annotation efforts by up to 70% and delivers high-quality detection, classification, and segmentation of image and video datasets.

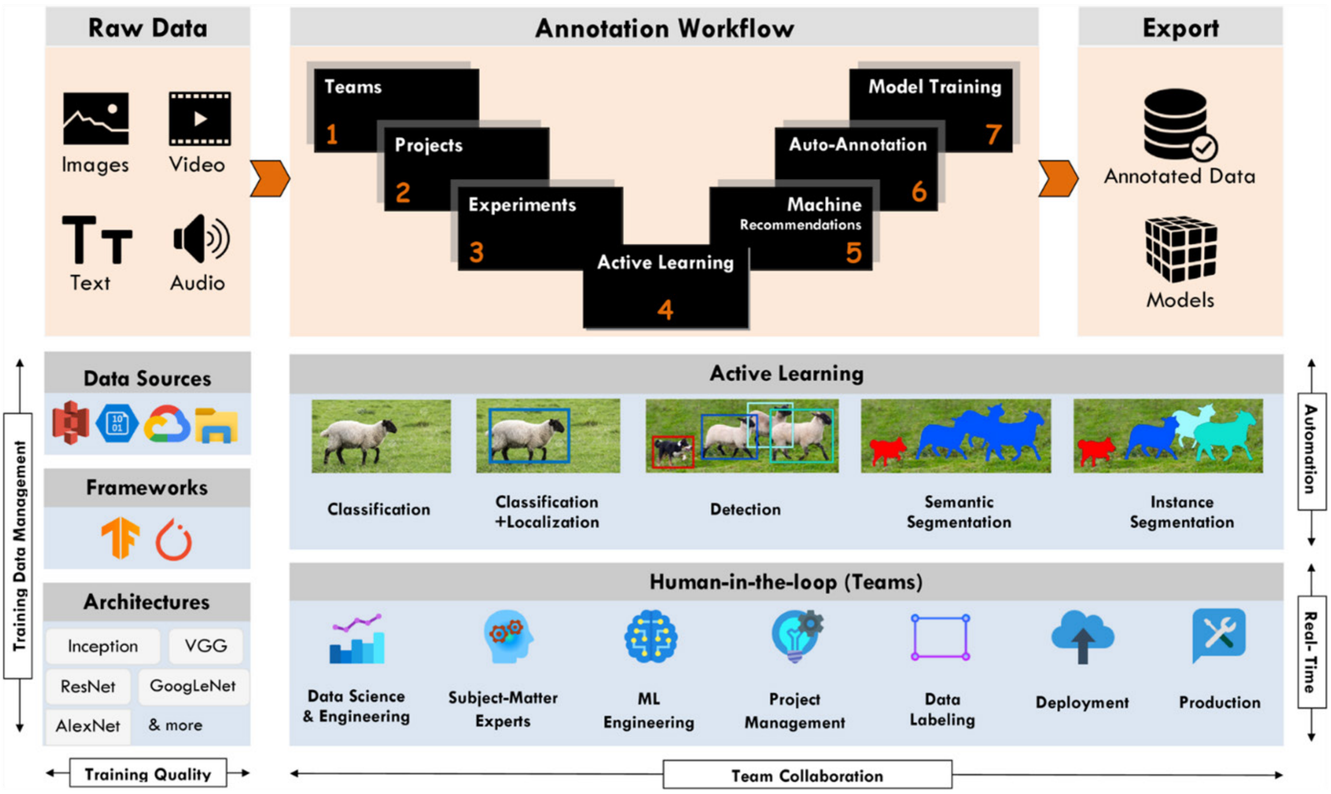
Why was Zastra™ Designed?

The following important points motivated us to design Zastra™ and address the problems faced by data scientists.

Constructing the datasets today by providing labels and annotations is extremely time-consuming and inefficient. Data annotation and labeling is a multi-dimensional activity and its speed, accuracy, and completeness have a direct impact on the effective mainstreaming of AI applications in an enterprise.

Current collaborative environments too often 'lock out' different stakeholders depending on project stage and maturity, do not allow for intelligent collaboration and handover between teams, and introduce inefficiency and risk by not providing adequate visibility into project status and updates.

Teams working at various stages of the AI project lifecycle have no meaningful coordination. For example, upstream input teams such as labeling and annotation, Midstream teams such as model training and tuning, and downstream, output teams, such as model deployment and production.



Key Capabilities of Zastra™

Real-Time Collaboration: Brings together disparate teams (such as SMEs, data scientists, labeling teams, project management, ML engineering, deployment, and production)so they are collaborating effectively and reducing time-to-market.

Active Learning Driven: Helps in active-learning-based object classification, object detection, localization, and segmentation (upcoming feature). Zastra™ can do this for images, video, and point cloud data.

Topological Data Analysis: Understands the 'bias' in the annotated data.

No Data Redundancy: Uses data across projects without duplication and is compatible with (Blob, S3).

Pre-Built Algorithms: Integrates within the platform for detection and classification.

Popular Frameworks: Supports PyTorch & TensorFlow, including the ability to change hyper-parameters.

Re-Use / Re-Purpose: Adds convenience in using one or more datasets; Multiple Projects – Multiple Experiments.

Easy Export: Helps in moving not just the labeled datasets but also the models to an external location.

Benefits of Zastra™

- Reduces time in the data annotation process and thereby reduces the time taken for the ML project
- Reduces annotation time and effort by up to 70%
- Delivers high-quality detection, classification, and segmentation of image and video datasets
- Increases collaboration and effectiveness of the entire AI development and deployment process by providing a unified platform

Conclusion

Data labeling may be quite difficult and complex, especially when done on a big scale, which is necessary nowadays for many use cases. Without resolving these issues, the data may be of low quality or may include additional complexity layers and costs. It is ideal to outsource this process to a partner you can rely on, as well as those who deliver quality and speed while properly addressing all these challenges.

Zastra™ from Cigniti is a powerful multi-source (picture, and video) customizable data annotation platform that is dedicated to providing you with the highest quality, most accurate training data at scale, in the manner of your choice, when you need it, and at the right price.

In order to pre-annotate the data before human labeling, ML teams may use Zastra™ to incorporate their models into the annotation platform. Labeling teams can reduce the amount of time spent on each batch by about 70% by converting the human annotation process into a straightforward auditing operation.

See how Zastra™ can make your data annotation process fast?
Book a demo with one of our experts to find out.

Schedule a Demo

Success Stories

Senior IT leaders share how our services helped them win in the platform age.

CTO Speak

Data pipeline buildout, Software development, Salesforce development, AWS System admin/DevOps, BI/Dashboard. The execution has been very good.

- Satyadeep “Bobby” Patnaik, CTO

Lafayette Square

CEO Speak

They understood that product development was iterative and they patiently worked through our requirements even as they rapidly evolved.

- Dr. Ganesh Naidoo, CEO

medmate

CTO Speak

Proven technical ability in both web and mobile development; strong project/product management expertise; the ability to become part of the extended BA365 team.

- Graeme Dollar, CTO

BUSINESS ACCELERATOR 365

COO Speak

I have worked with hundreds of service providers and consultants, RoundSqr (Part of Cigniti) is absolutely one of the best. The people are highly skilled, very hard-working, and have a “can do” attitude.

- Mark Mortimer, COO

Adelphoi

We Are Recognized by Global Analysts & Advisors

Gartner

ISG

FORRESTER

IDC

HFS

NelsonHall

Everest Group
From insight to action.

zinnov

About Coforge

Coforge is a global digital services and solutions provider, that leverages emerging technologies and deep domain expertise to deliver real-world business impact for its clients. A focus on very select industries, a detailed understanding of the underlying processes of those industries and partnerships with leading platforms provides us a distinct perspective. Coforge leads with its product engineering approach and leverages Cloud, Data, Integration and Automation technologies to transform client businesses into intelligent, high growth enterprises. Coforge's proprietary platforms power critical business processes across its core verticals. The firm has a presence in 21 countries with 26 delivery centers across nine countries.

Learn more at www.coforge.com.

Website LinkedIn YouTube Blog Facebook Twitter