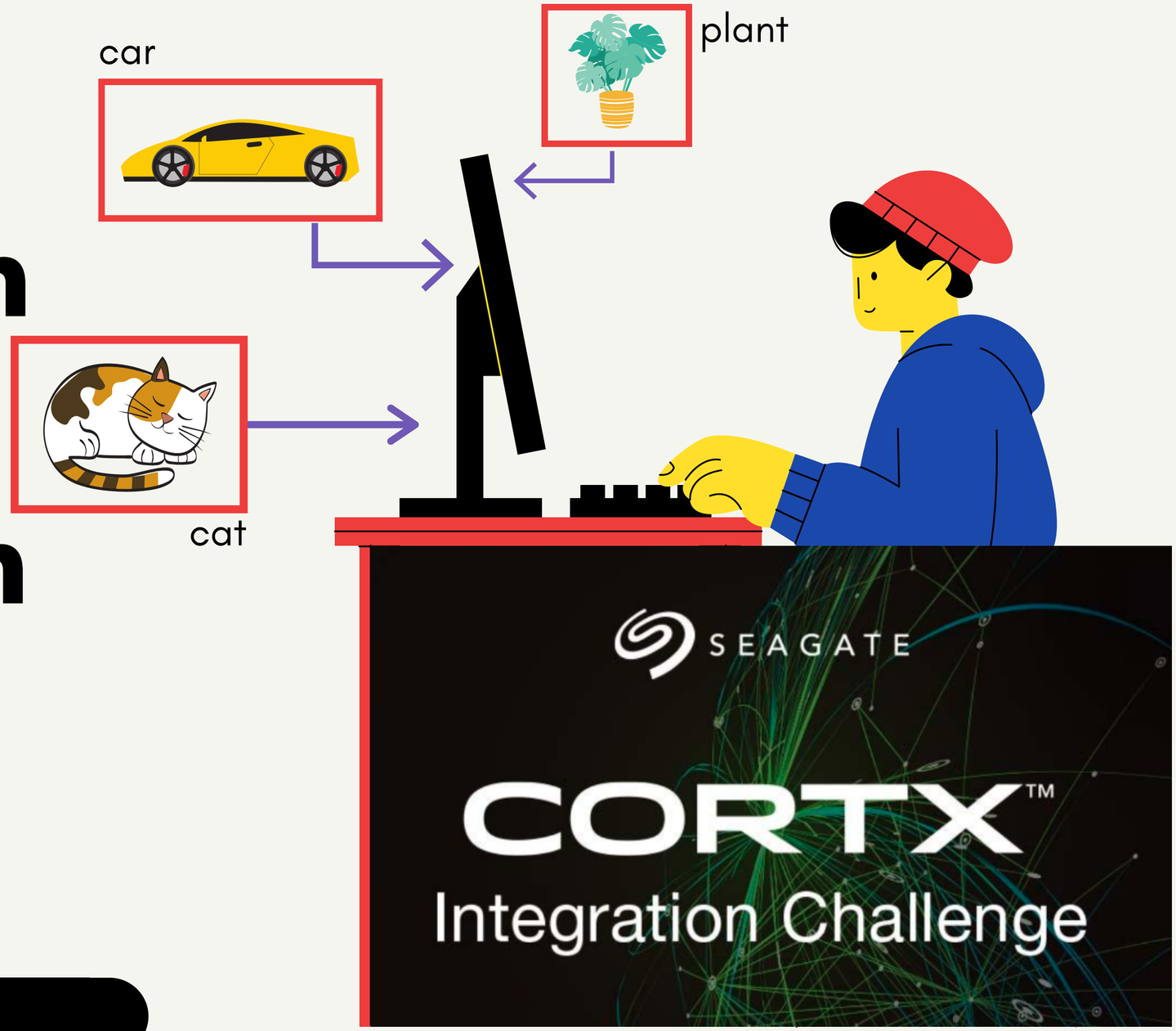
CORTX Integration for Data Annotation Studio,

Label Studio

An open-source project



Open source: Better solutions and a more inclusive society

Introduction

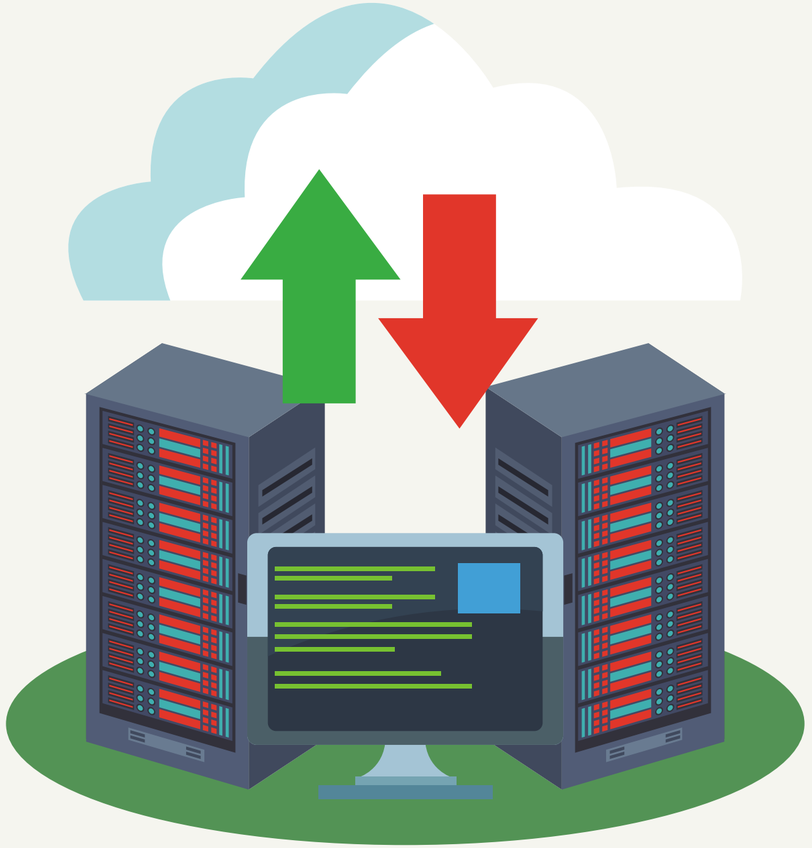
Background

Overview of the project

A need for scalable, efficient storage service for ever growing unstructured data due to massive adoption of AI.

Building Seagate Cortex S3 storage solutions INTEGRATIONS to store and automate, one of the most complex task in AI, data annotations.

Building user interface to make annotation exports and project tracking simpler.



The Problem

What we want to solve

The problem statement is from my personal story, I was working on a wildlife conservation project to save animals from poaching and using biometric sensors to monitor body vitals, with tremendous hardwork I collected enough amount of data, not an easy task- I had to daily meet park rangers to get data but my negligence ruined all my hardwork after fatal HDD crash. Such huge loss in terms of data and time, the project could not be completed yet, due to it. But this time I will ensure that the data generated by the data revolution can be harnessed to the full extent of its potential using CORTX S3 data services.

Hypothesis

Integrating Label Studio, an open source data annotation tool widely used by global AI brands. Integrating, S3 services for result storing and data retrieval for data analytics and other data related tasks. Label Studio is compatible with cloud services in it's latest version, so it is good and compatible to integrate Cortx S3 with it.

Objectives

What we want to achieve

Solve problems in handling unstructured data which are very useful in supervised machine learning with annotated data.

Make data annotation workflow easier, less cumbersome, simplify data uploading, management and retrieval tasks.

Not restricted to region and data volume, hence we need a highly scalable data storage system with multi node, hybrid resource operations features.

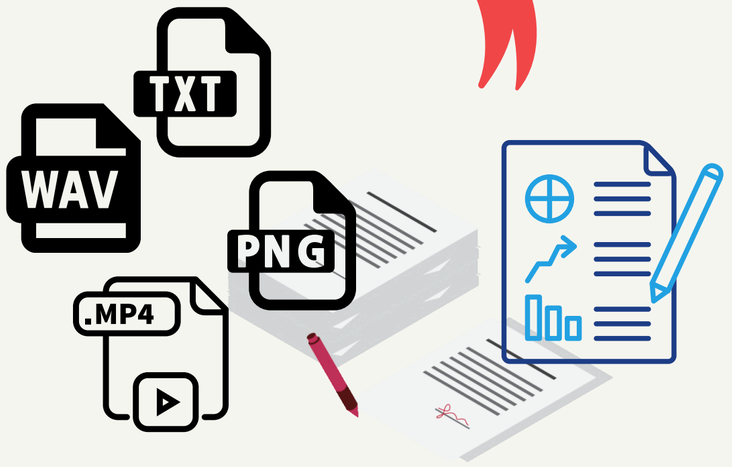
CORTX™

AI/ML Application Developer

Now any developer can download the data and annotations from Cortx S3 for any model training and app.

Receiving all data in S3 bucket in one sync function calling, label data over variety of methods and upload the labeled results in JSON, CSV, COCO and PASCAL VOC format to S3

Data Annotator



Data

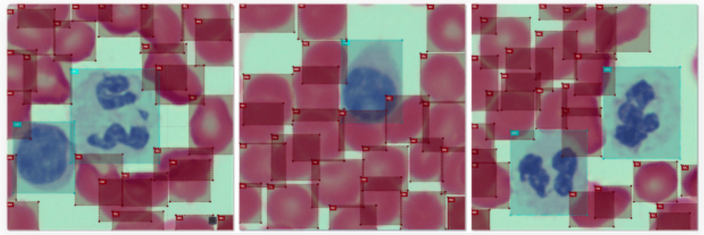
Preciously collected dataset by engineers, scientists and subject experts.

Uploading all the data to Cortx S3 bucket for annotating in Label Studio

Label Every Data Type

Images | Audios | Texts | Time Series | Multi-Domain

Computer Vision

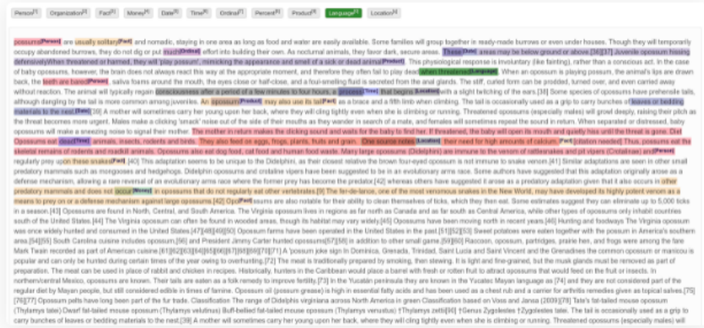


- **Image Classification**
  Put images into categories
- **Object Detection**
  Detect objects on image, bboxes, polygons, circular, and keypoints supported
- **Semantic Segmentation**
  Partition image into multiple segments. Use ML models to pre-label and optimize the process

Label Every Data Type

Images | Audios | Texts | Time Series | Multi-Domain

NLP, Documents, Chatbots, Transcripts



- **Classification**
  Classify document into one or multiple categories. Use taxonomies of up-to 10000 classes
- **Named Entity**
  Extract and put into pre-defined categories relevant bits of information
- **Question Answering**
  Answer questions based on context
- **Sentiment Analysis**
  Determine whether document is positive, negative or neutral

Label Studio

Label Every Data Type

Images | Audios | Texts | Time Series | Multi-Domain

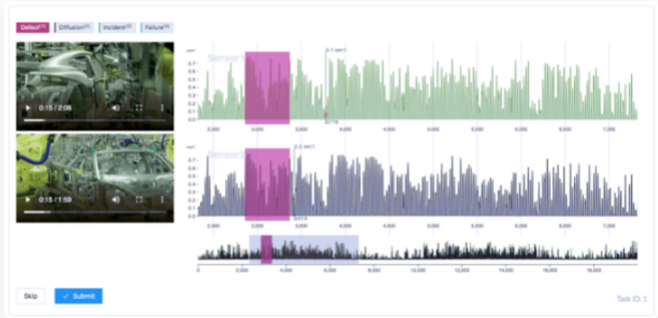
Audio & Speech Applications

- **Classification**
  Put audio into categories
- **Speaker Diarization**
  Partition an input audio stream into homogeneous segments according to the speaker identity
- **Emotion Recognition**
  Tag and identify emotion from the audio.
- **Audio Transcription**
  Write down verbal communication in text

Label Every Data Type

Images | Audios | Texts | Time Series | Multi-Domain

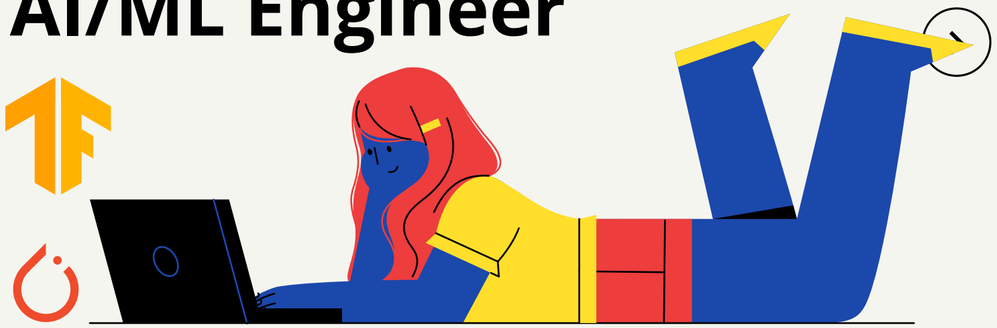
Multi-Domain Applications



- **Dialogues Processing**
  Call center recording can be simultaneously transcribed and processed as text
- **Optical Character Recognition**
  Put an image and the text right next to each other
- **Time Series with Reference**
  Use video or audio streams to easier segment time series data

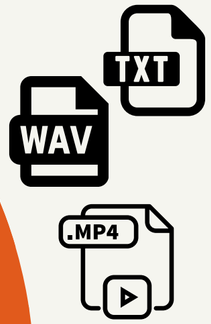
AI App User

AI/ML Engineer



Training ML models based upon annotations

CORTX™



COCO
Common Objects in Context

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Data Annotator

Significance of the Project

In terms of scope

In the world of unstructured data, data annotation is the key component in it. Not only those data annotation has helped to build sentiment analysis chatbot, retail monitoring solutions and self-driving cars too.

In terms of resources

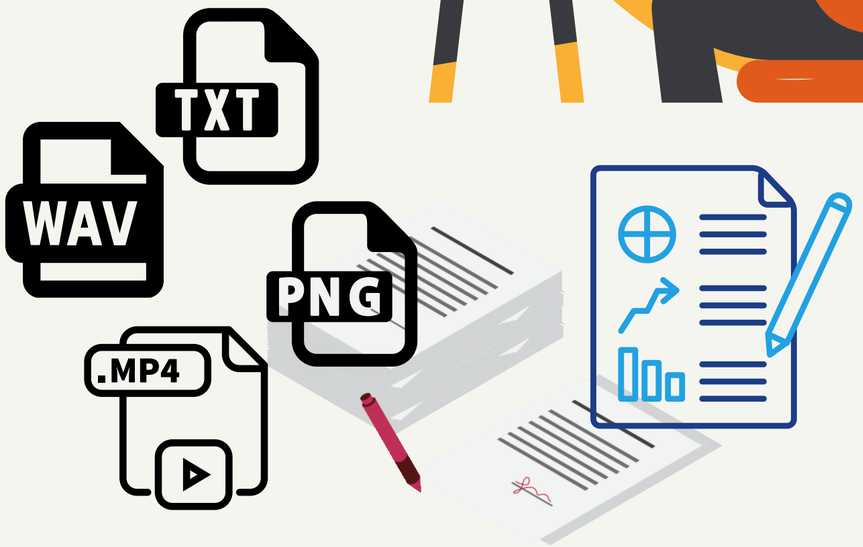
For all this unstructured data, data storage system plays a huge role. There is always a fear of losing data and data competitiveness, using highly scalable, efficient and secure data system is the key.

In terms of time

Label Studio is a great tool to start integrating S3 storage solutions to make bulk file data annotation and exports happen in no time.

CORTX

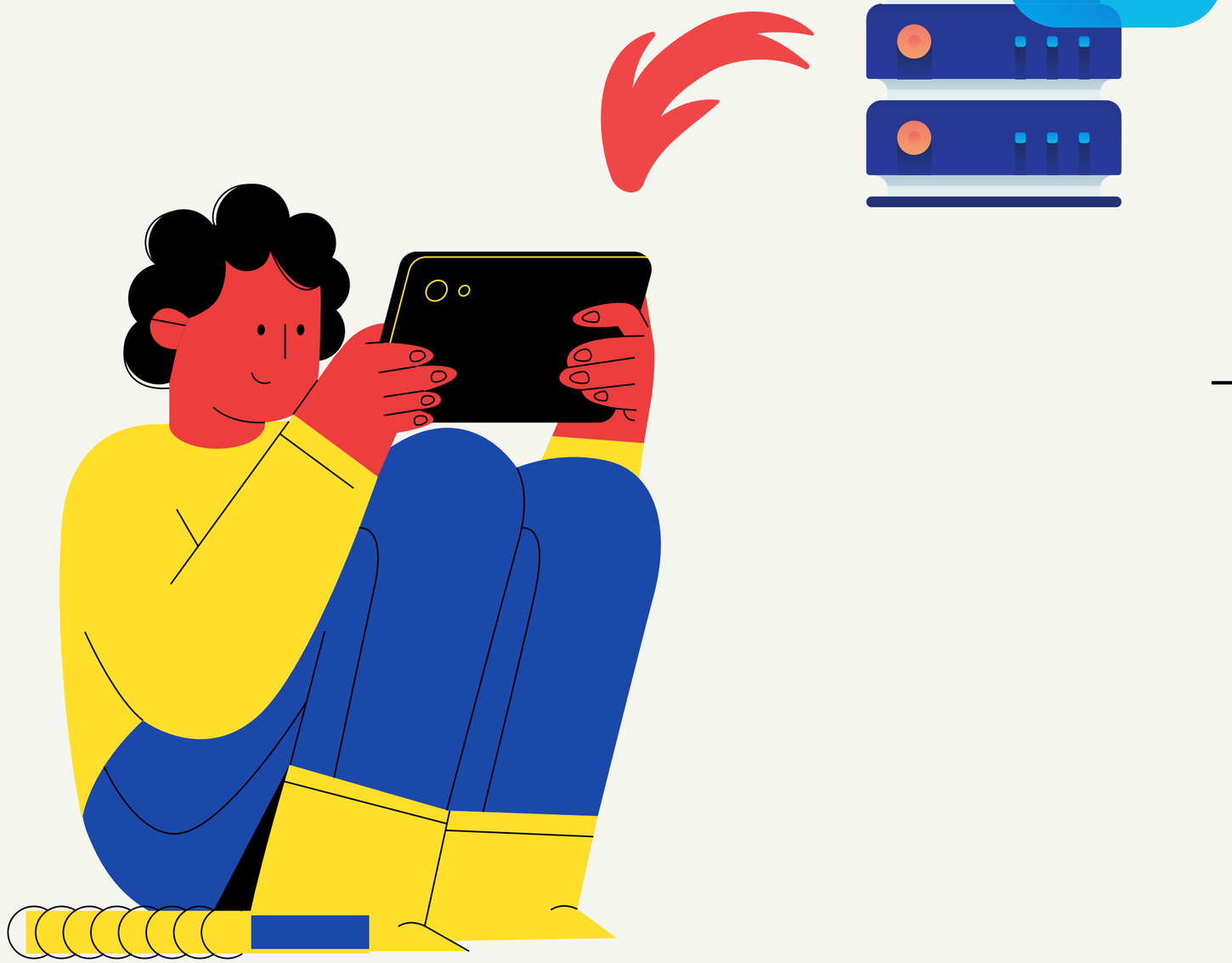
UPLOAD DATASET TO S3



CORTX™

Objective

Download the desired file from Cortx S3 and use it with any platform or toolchain



Examples

Download annotation or training files in local jupyter notebook environment for AI/ML or save file on your local disk. There are variety of things you can put your data to use.

Contributions of the Project

Suggestions for future research

Sumit Kumar

17 y/o high schooler who loves to solve problems through technology. Also Founder and CEO of steptostem.com