Yeshiva University - Katz School of Science

Data Analytics and Visualization MS

date: May 2022
from: Jesus Olivera, Independent Study

subject: YU Combating Hate Speech Project MVP Outcomes

Designing and Implementing a Data Labeling Tool Backend for Building a Training Dataset for ML/AI Hate Speech Classification

Agenda | Data labeling Tool MVP Outcomes

- Executive Summary
- Project Recap
- Lessons learned
- Next Steps

Katz School of Science and Health

YU Hate Speech Project | Executive Summary

Context

Antisemitism is on the rise. New technologies and applications are required to combat the negative effects of online hate speech towards the Jewish Community; For this reason, the project will be focused on a complex generic use case for an MVP aimed at designing, building, and deploying the backend of a data labeling pipeline that builds robust labeled datasets that can be used to train Machine Learning and AI algorithms to detect online antisemitic hate speech.

Opportunity

Synthetic data has emerged as a data science tool and its primary purpose is to be versatile and robust enough to be useful for the training of machine learning models. The explored methodology incorporates a controlled statistical process that leverages randomization and a customizable class separation. Yeshiva University Research will be introduced to an alternative to their data collection and data handling practices focused at reducing costs and expediting their data acquisition activities. The data produced in this project can also be useful in gaining training data for edge cases, which are instances that may occur infrequently but are critical for the success of AI.

Outcome

Yeshiva University Research was provided with an end-to-end pipeline that leverages an open-source tool and an automated methodology for the ETL, data validation and data management activities. This project produced a complete set of infrastructure deployment and setup instructions as well as a detailed implementation guideline that will allow the organization to (1) collect, (2) organize, (3) analyze, and (4) infuse data in ML/AI workstreams. The project also provided a generic framework for the development of any synthetic data domain. Next steps are detailed in the presentation.

Katz
Katz School
of Science and Health

MVP | Importance & Broader Impact

Developing techniques and methods for combating online hate speech can provide tech companies with a framework around monitoring content in their platforms. Overtime, these initiatives can enhance regulatory compliance efforts and ultimately provide users a safe space for their regular interactions within online platforms.

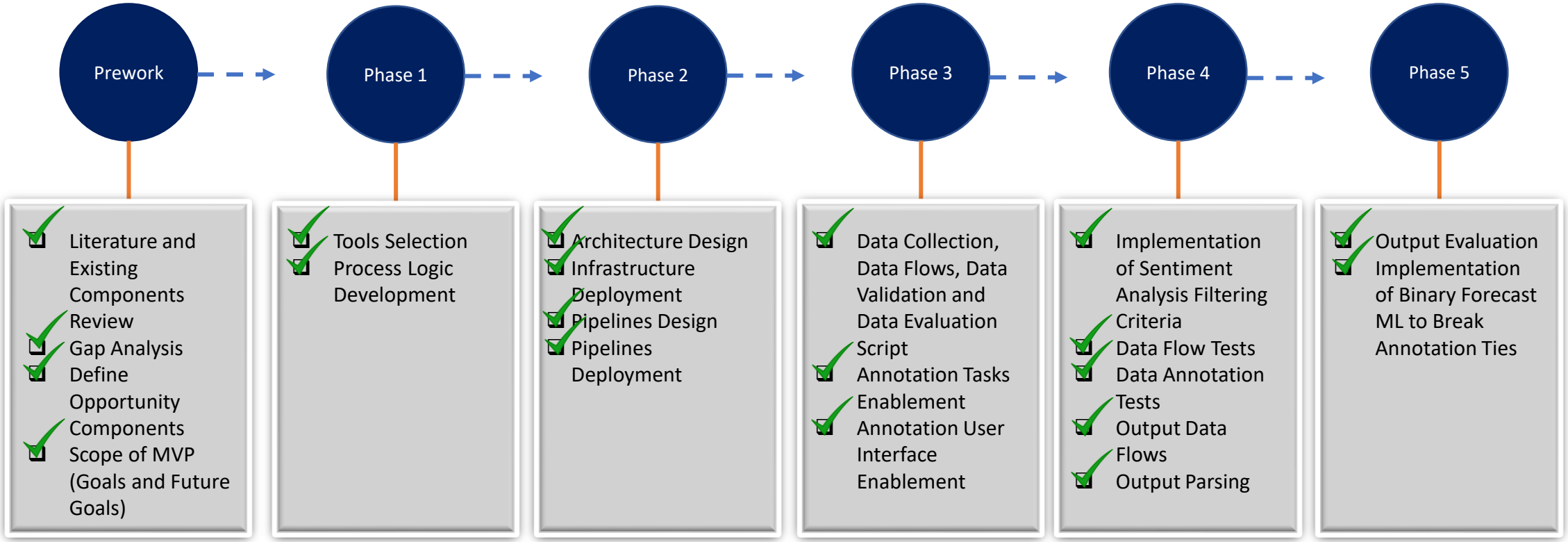
Mental health has declined over the years and some of the variables that have impacted the overall uptake in suicides, depression and other mental health problems can be attributed to the rise in technology and in particular social media.

A future state for this project can include providing access to external organizations, via a data marketplace or a deployment space, to high-quality antisemitic labeled datasets for their models and/or access to trained models focused on detecting antisemitism online.

This project has the potential of contributing to the overall benefit of society by mitigating some of the negative effects of online hate speech towards Jews and the other communities.

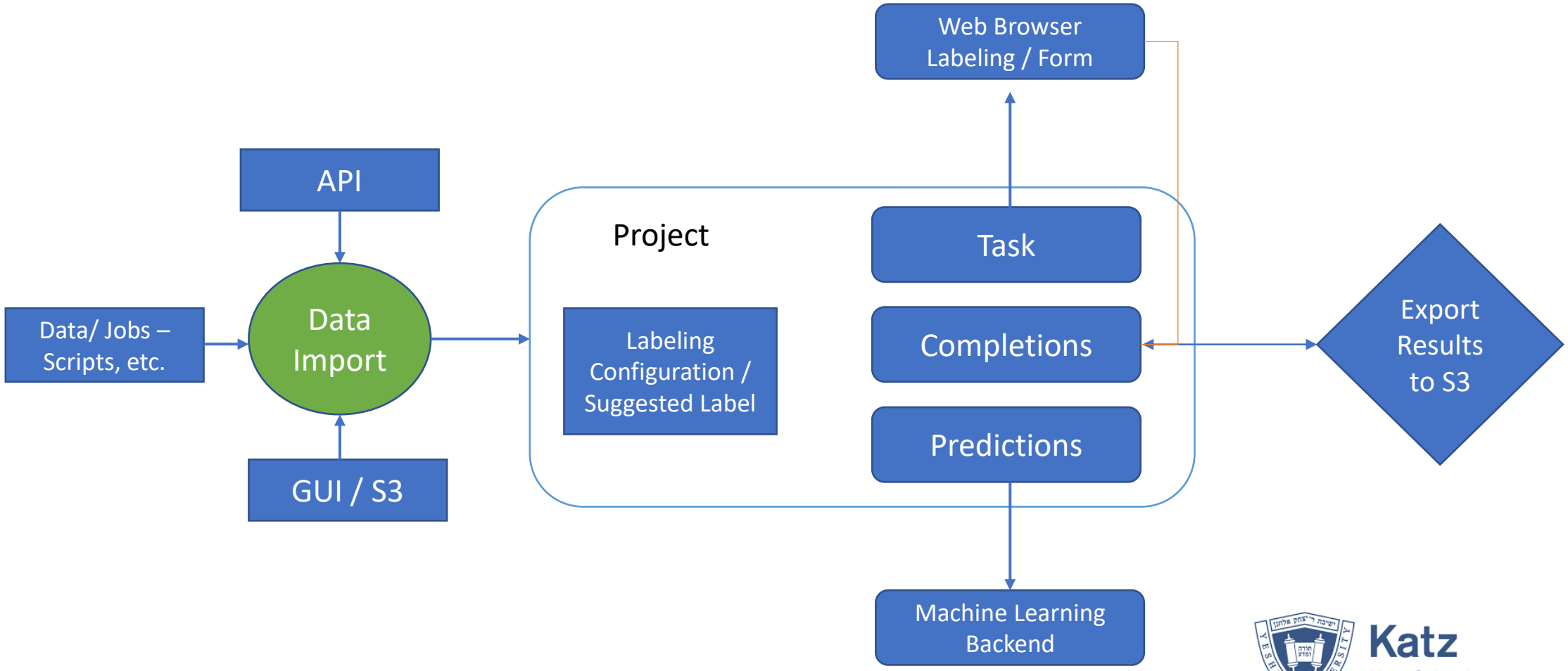
Katz
Katz School
of Science and Health

MVP | Project Recap



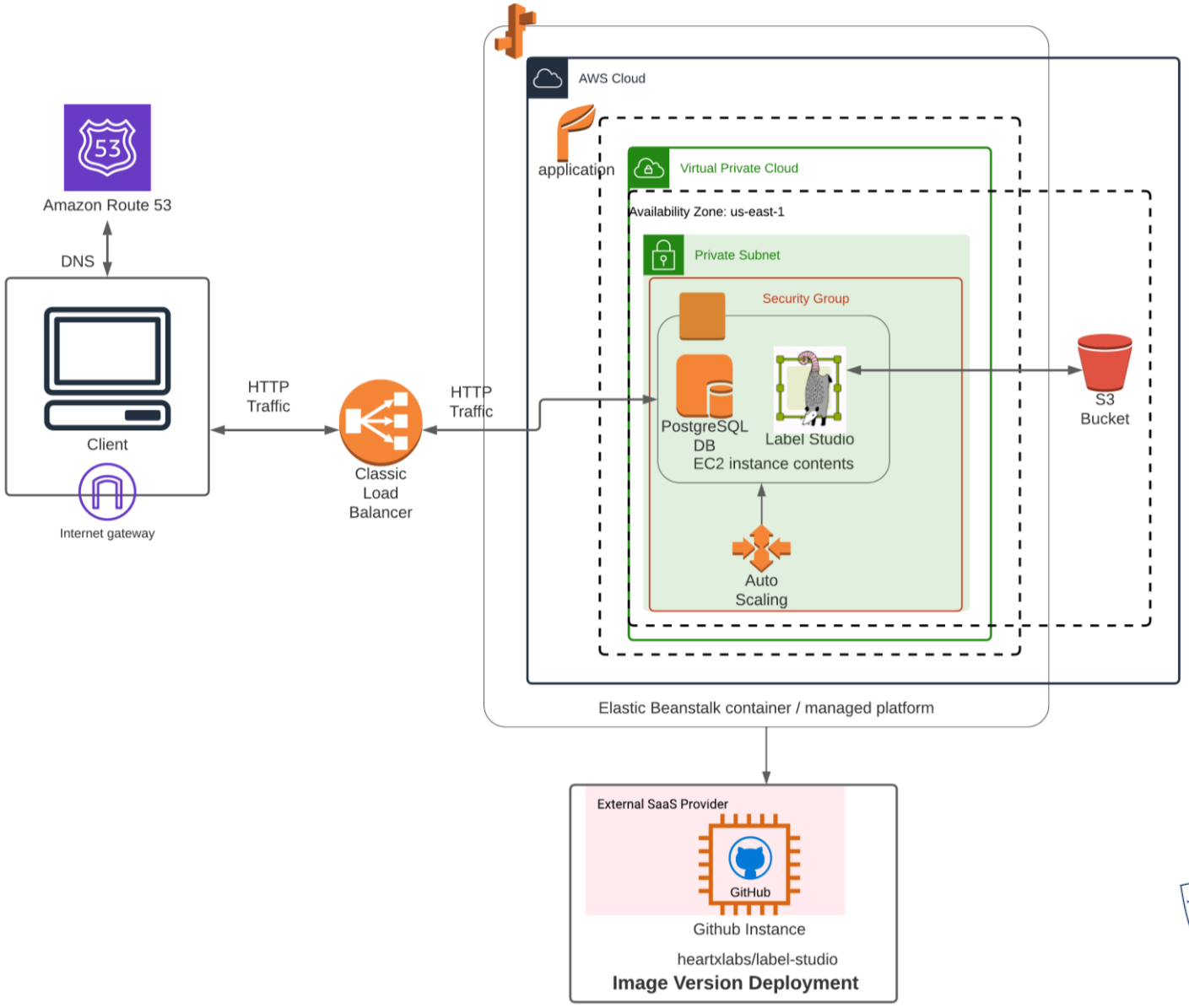
Katz
Katz School of Science and Health

MVP | Application High-Level Logic



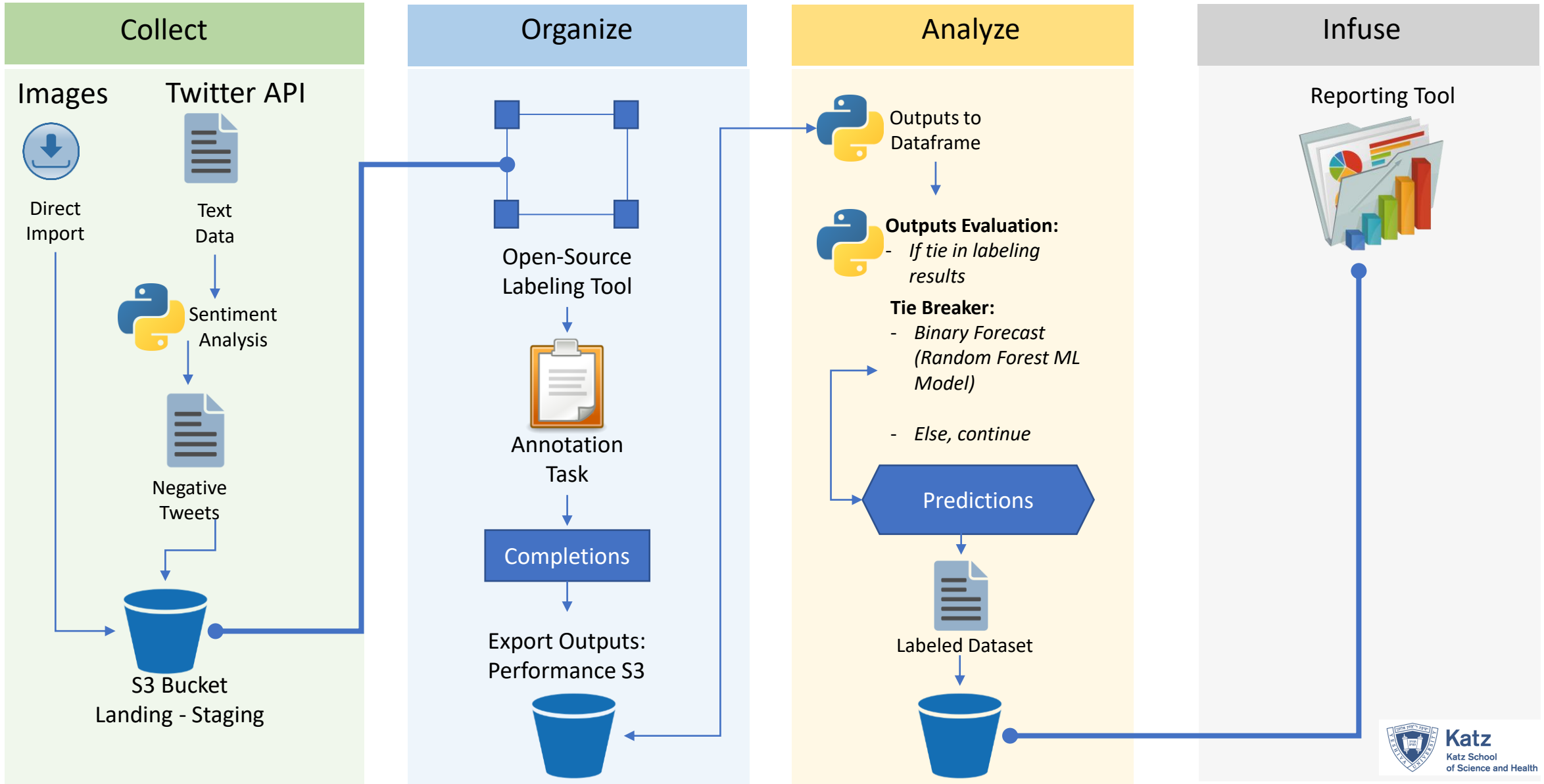
Katz
Katz School
of Science and Health

MVP | Proposed Application Architecture



Katz
Katz School of Science and Health

MVP | Project Methodology Diagram



MVP | Lessons Learned

- The proposed solution addresses the business requirements, shortens time-to-value, reduce manual efforts, and overall project costs
- The proposed solution streamlines user activities and provides a methodology for creating synthetic datasets to train ML/AI models
- The proposed solution can be leveraged as content to create an agnostic methodology for an automated data collection, organization, analysis and infusion into ML/AI model training activities
- Label Studio support integration with all the components in the system
- Label Studio is suited for the needs of Yeshiva University Research
- The proposed solution allow users to create datasets properties and export to the appropriate formats
- The project demonstrated the ability to quickly ensemble the infrastructure of the system leveraging open-source technologies
- Integration of ML models made possible the streamline of data validation and organization by infusing Natural Language Processing and Predictive Analytics into the workstream
- The system allow to easily identify and modify subject areas and concepts related to the dataset domain

Katz
Katz School
of Science and Health

MVP | Next Steps

Infrastructure

- Select a cloud provider to host the infrastructure
- Deploy infrastructure and create an infrastructure instance image
- Deploy all pipelines
- Revise design and enable user interface
- Enable secure public access via user interface
- Refine and update infrastructure requirements and create a private GitHub repo with the deployment configuration files, step-by-step instructions and develop a CI strategy (continuous integration).

Label Studio Configuration

- Create a single project per data type and synthetic dataset domain
- Develop a logical application for data syncing
- Customize user interfaces per data type and dataset domains
- Create user accounts and provide access to annotation users to the projects via user interface
- Review out of the box application features and evaluate their usability in each project

System

- Parameterize scripts
- Leverage Functions
- Setup job schedules and triggers
- Enable a monitoring infrastructure system that capture logs and alerts admin of non-compliant events and bugs

Katz
Katz School
of Science and Health

MVP | Next Steps

Synthetic Dataset Domain Extraction Logic Development

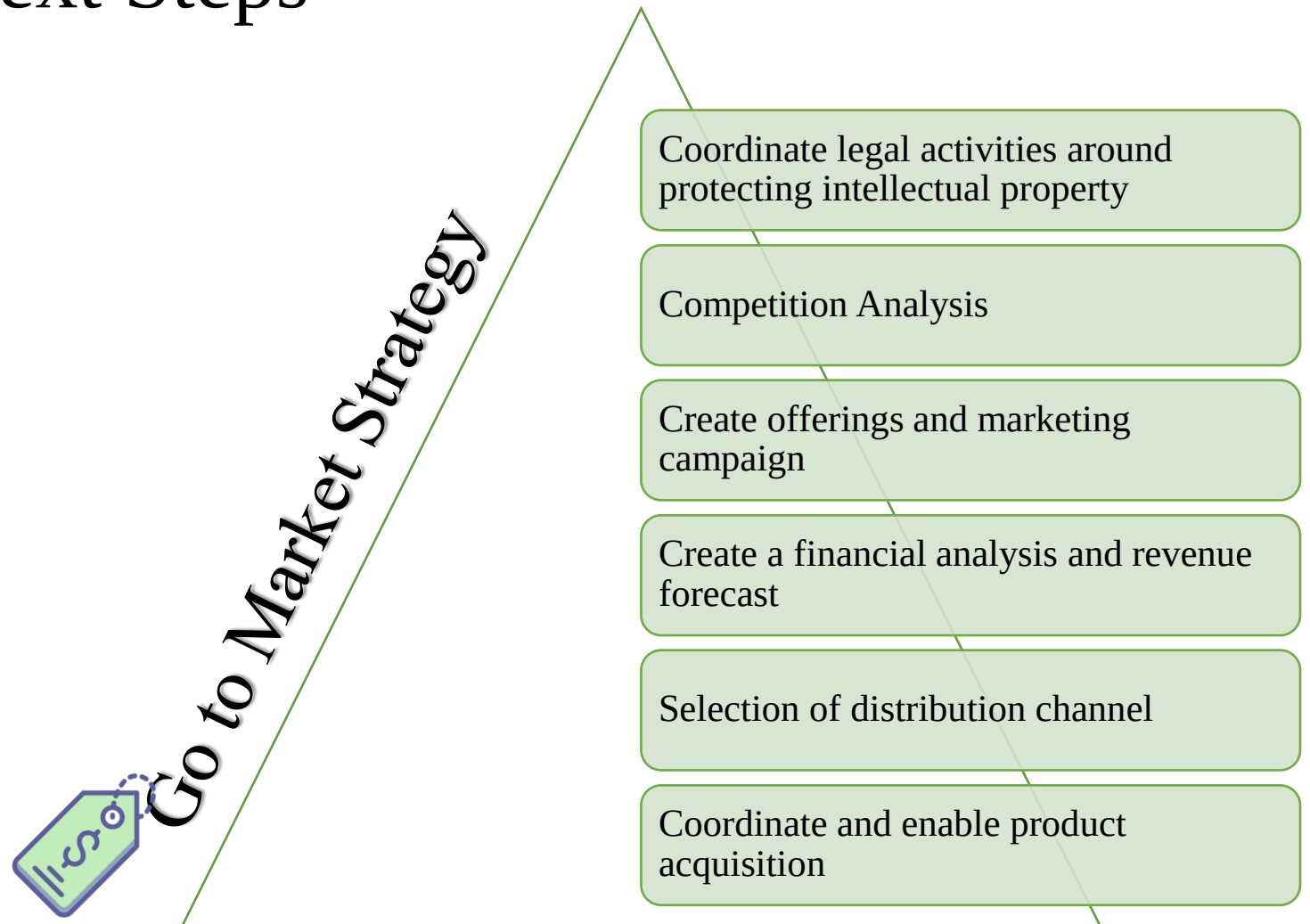
- Develop a robust antisemitism glossary
- Enable the antisemitism glossary to be accessed by the system via S3 bucket
- Continue developing a robust data cleansing job post sentiment analysis

Annotation Outputs Analysis

- Continue development of initial outputs wrangling logic
- Complement initial outputs with other data like sentiment analysis results
- Dummify all variables on the initial output
- Create a co-relation analysis to better understand importance of variables
- Enhance the binary forecast model by including relevant variables as inputs
- Continue tunning model to increase its accuracy
- Develop a model monitoring logic, setup thresholds and enable a messaging system that alerts administrator about thresholds being reached (bias, drift, accuracy. Etc.)
- Develop final synthetic dataset data model in a repository that will be leveraged by the end-user

Katz
Katz School of Science and Health

MVP | Next Steps



Katz
Katz School of Science and Health

Katz
Katz School
of Science and Health

Thanks

YU Combating Hate Speech Project MVP Outcomes

Jesus Olivera
Independent Study

May 2022