STREAMLINING ML DEPLOYMENT

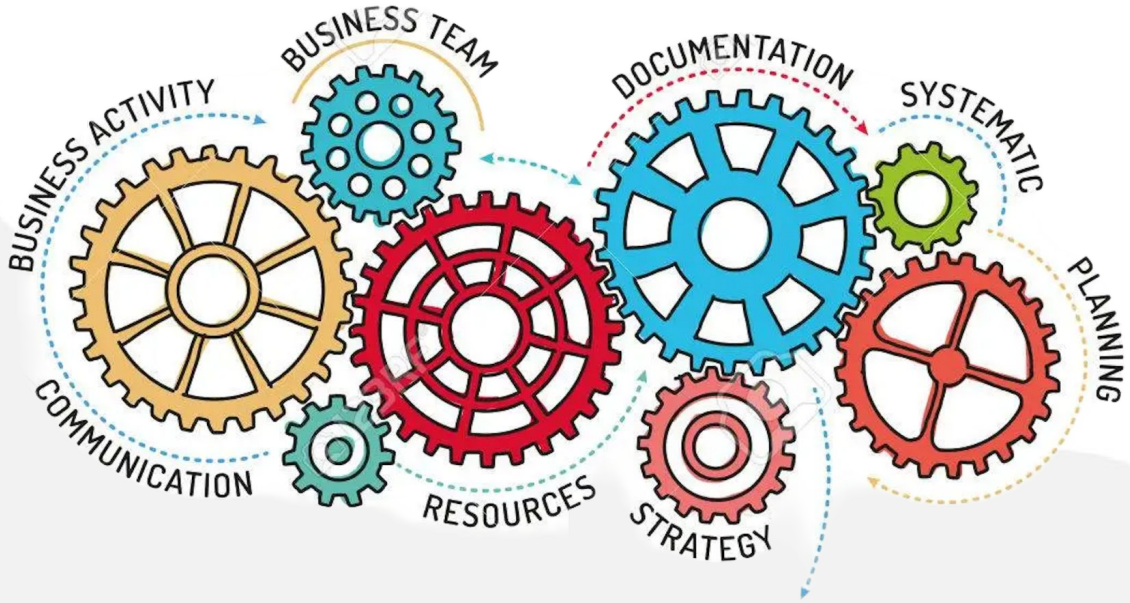
By

Ankita Shetty

Kirtana Kirtivasan

Krishna Sai Tejaswini Kambhampati

Swetha Vijaya Raju (TL)



COMPANY OVERVIEW

About us:

- **Year founded**: 2017
- **Head Quarters**: Palo Alto, CA
- **Employees**: 150-200
- **Type**: GPS Navigation Service

Value Propositions:

- Best real-time route
- Alerts about Road Hazards,accidents,community activities and Traffic

Revenue:

- Location -based brand advertisements
- Carpool commissions

Key activities:

- Development and maintenance of platforms
- Marketing
- Create Data
- Community engagement

Key Resources:

- Software
- Developers
- Algorithms
- IT Infrastructure
- Active User base
- Community

PROBLEM STATEMENT

CONTEXT

Currently Operates in Bay Area, California only

2000 Monthly Active users and counting

500 Volunteers

CHALLENGES

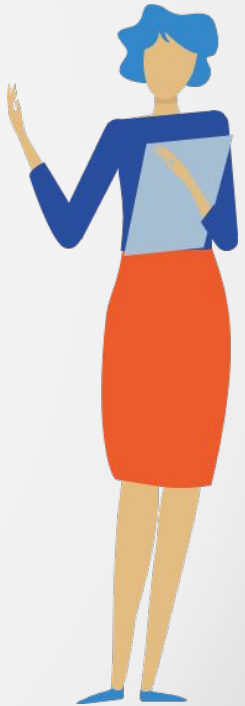
- ➔ Long Deployment cycle due to
  - ◆ Usage of multiple ML frameworks
  - ◆ Semi- manual operations during training and validation of model
  - ◆ Limited monitoring capabilities after model deployment
- ➔ Increasing monthly active users
- ➔ Including business expansion

Project Objectives

- Minimize timelines for deployment
- Reduce manual work along the life cycle of model
- Reduce overall costs with Managed services
- Model Versioning by tracking changes in code and data
- Automate storage of models
- Ensure reusability of models

Success Criteria

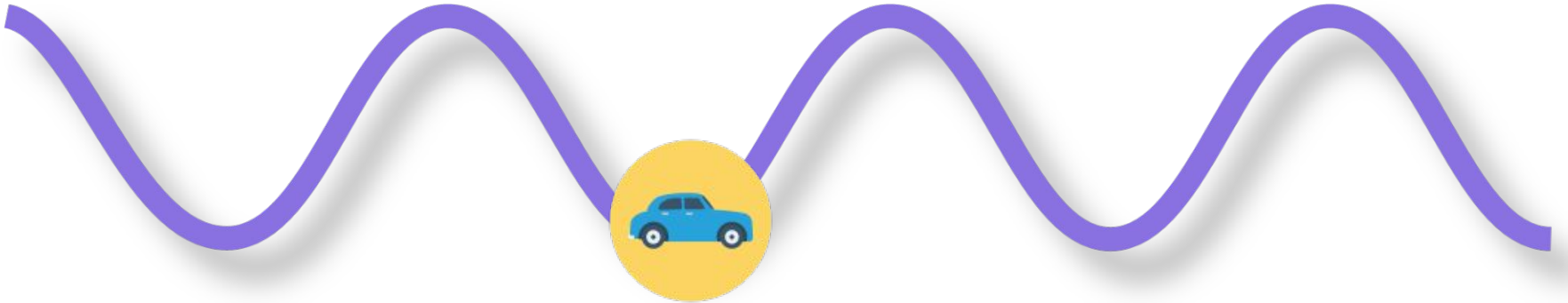
- Bring down the operational cost by ~34% in the next 5 years
- Reduce 40% of the stakeholders along the cycle
- Reduce deployment timeline by ~22%

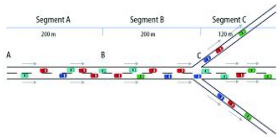


USE CASE

Multiple segments from source to destination in one route

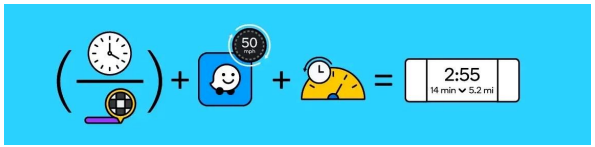
Establish labels:
Ground truth - past eight weeks average speed on specific time



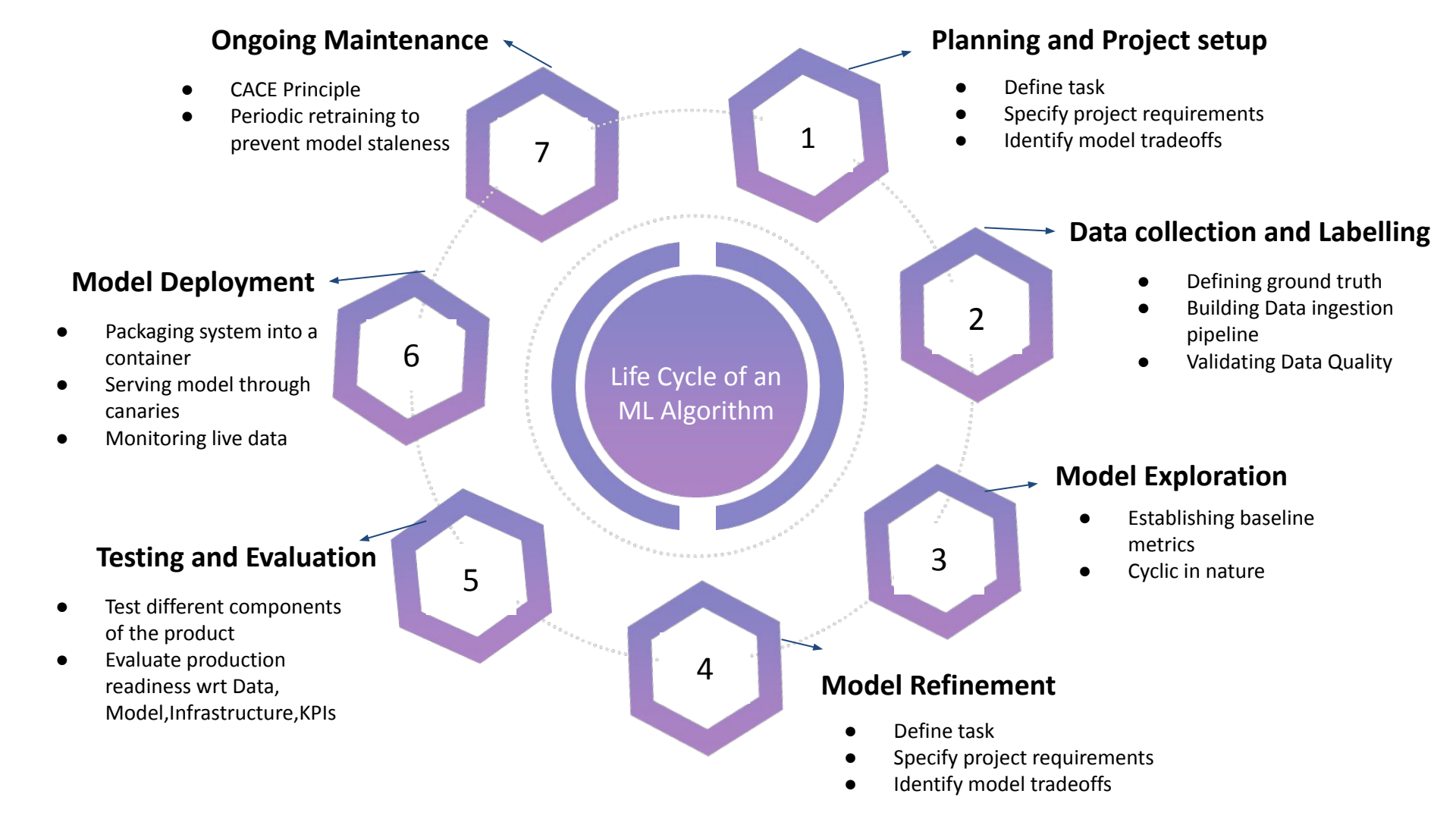


Historical data:
48 samples on each segment every week

ETA Equation:



LIFE CYCLE OF ML MODEL



Planning and Project setup

- Define task
- Specify project requirements
- Identify model tradeoffs

Data collection and Labelling

- Defining ground truth
- Building Data ingestion pipeline
- Validating Data Quality

Model Exploration

- Establishing baseline metrics
- Cyclic in nature

Model Refinement

- Define task
- Specify project requirements
- Identify model tradeoffs

Testing and Evaluation

- Test different components of the product
- Evaluate production readiness wrt Data, Model,Infrastructure,KPIs

Model Deployment

- Packaging system into a container
- Serving model through canaries
- Monitoring live data

Ongoing Maintenance

- CACE Principle
- Periodic retraining to prevent model staleness

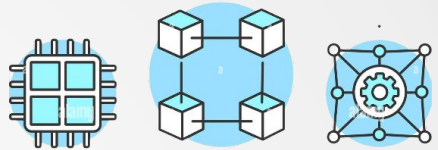
PHASE DETAILS

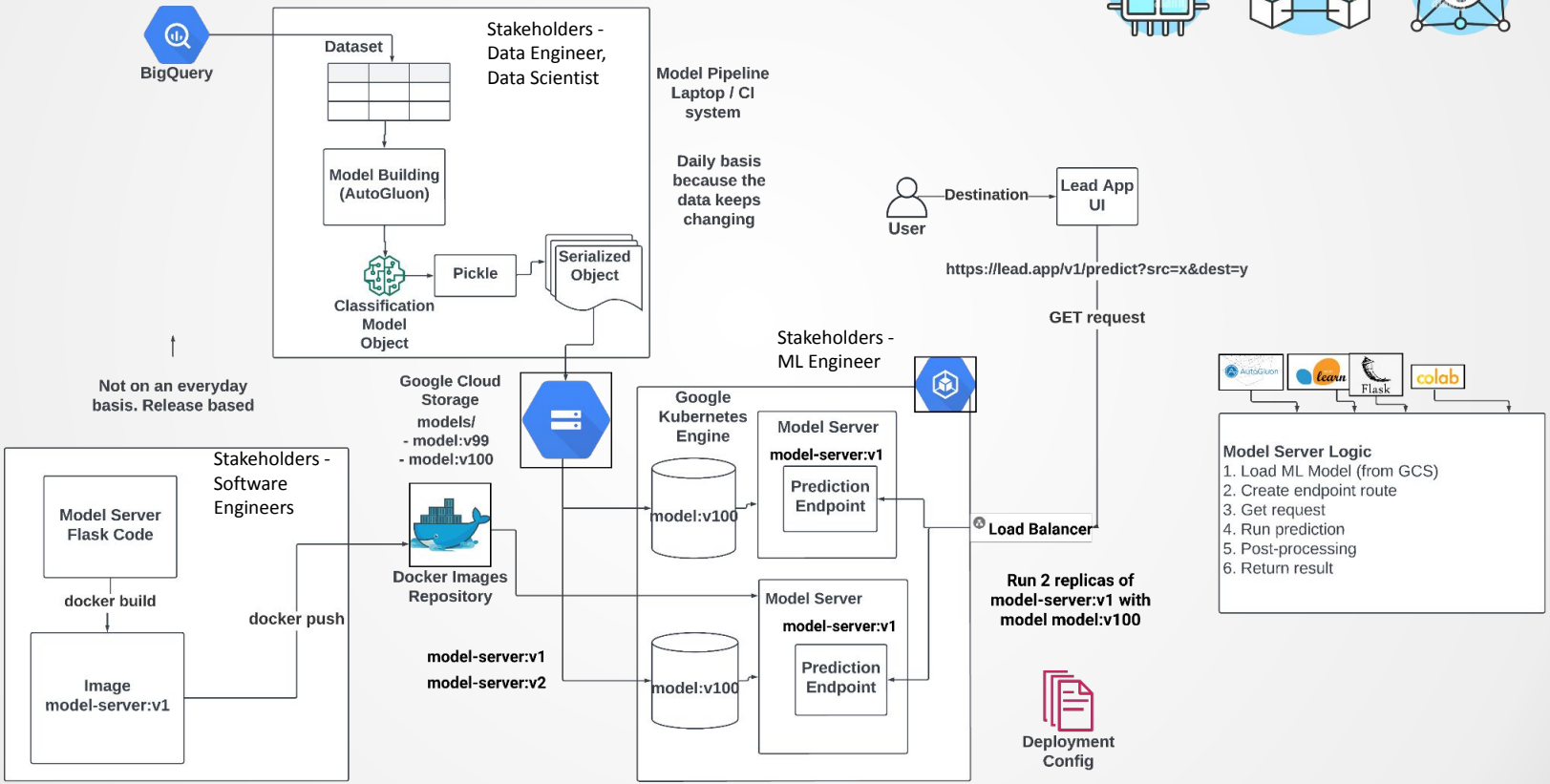
STAGE	DELIVERABLES	TIMELINE ESTIMATE
Discovery	Specification Document	4-6 days
Exploration	Proof of Concept	~ 10 days
Development	Production ready ML Solution	~ 22 days
Improvement	Performance in Production	Continuous (for first 15 days)

CURRENT TEAM

Stakeholders	Team Size	User Story
Data Engineer		ETL / Build data pipelines to load, clean and process the data
Data Scientist		Data Refinement, Data Visualization, Data Analytics
ML Engineer		Build, train and deploy the Machine Learning model
Software Engineer		Integrating the ML model with the rest of the application
Product Manager		Making product decisions and main point of contact for the project

SYSTEM ARCHITECTURE





CURRENT OPERATIONAL COST

Items	Description	Deploy 1 model	Deploy an additional model
Model Infrastructure	Kubernetes clusters with a load balancer	$8,000	+ $3,781
Data Support	Independent data pipeline manager for continuous update of analytic data	$10,000 (labor) + $3,200	+ $2,250 (labor)
Engineering / Deployment	Continuous integration and continuous deployment (CI/CD) system to pull model from registry	$24,500 (labor) + $516	+ $2,900 (labor)
Total Investment / yr		$34,500(labor) + $12,000	$5,150(labor) + $3,781
5 year TCO		$94,500	+ $24,055

1. Cost for deploying 1 model to production = $94,500 over first 5 years
2. Cost for deploying 2 models to production = $119,000 over first 5 years
3. Cost for deploying 100 models to production = $2.5 million over first 5 years

What is the Cost to Deploy and Maintain a Machine Learning Model - https://www.phdata.io/blog/what-is-the-cost-to-deploy-and-maintain-a-machine-learning-model/

MARKET RESEARCH

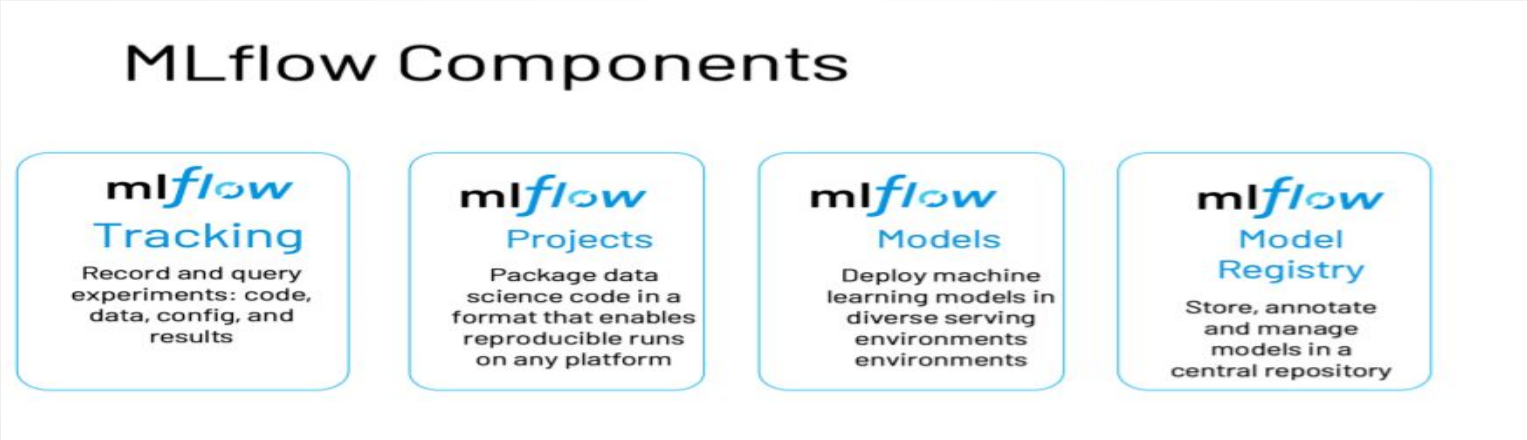
MLFlow - By Databricks

ML Workflow challenges:

- Difficult to Tracking Experiments
- Difficulty in Reproducing Code
- No Standard way to Package and Deploy Models
- No Central Storage to manage Model Versions and Stage Transitions

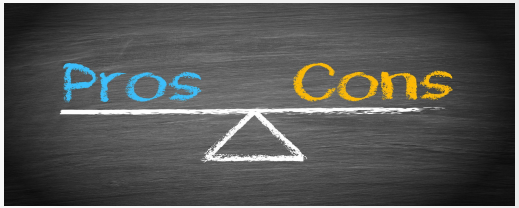
What is MLFlow?

- **Open source platform** for managing the end-to-end ML lifecycle
- Provides 4 components to help manage ML workflow:



PROS

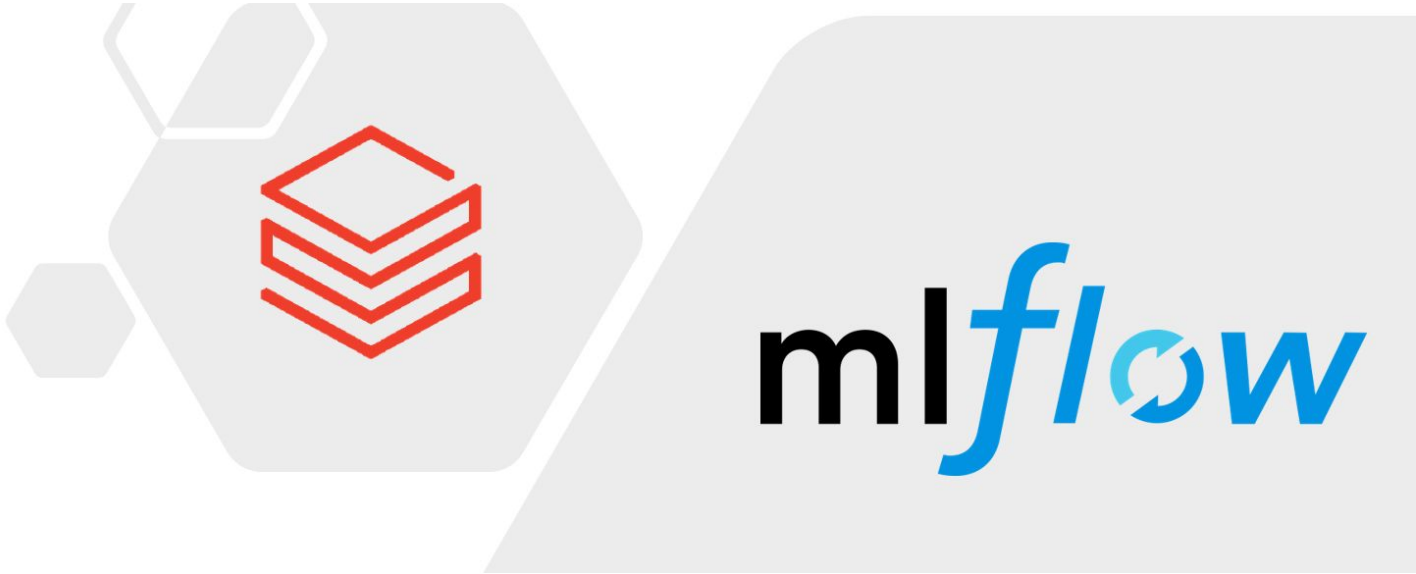
- Open Source Tool (All features offered are free)
- Track and manage model, package ML codes, centralize lifecycle stage transitions
- Python package: simple import in code
- Model Deployment via Model Registry
- Registry - model versioning, model lineage, annotation and stage transitions
- Can promote models on cloud environments like Amazon Sagemaker, Microsoft Azure, GCP
- Automatically visualizes your model and able to compare metrics between two/more Runs
- Parameters can be supplied to pipeline as arguments saving time as compared to re-running notebook



CONS

- Data cleaning and preprocessing to be done on our own (not automated unlike Google's Vertex AI)
- AutoML not supported
- Model Development takes several days to months
- MLFlow requires current architecture; MLFlow Recipes still in experimental phase

DEMO - MLFLOW



MLFlow and Azure Databricks https://cprosenjit.medium.com/mlflow-azure-databricks-7e7e666b7327

What is Vertex AI?

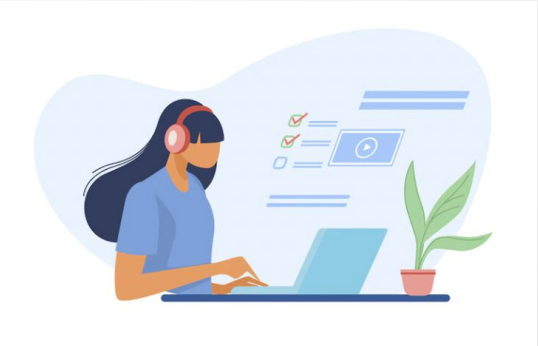
vertex.ai

It is a managed machine learning platform offered by Google.It helps ML workflow by providing services like: dataset creation and upload data, Train an ML model on your data, Upload and store your model and Deploy your trained model to an endpoint for serving predictions.

KEY FEATURES

DATA

MODEL DEPLOYMENT



DEPLOY AND USE

PROS

1. Easy to access, takes as low as 30 seconds to login and access Vertex AI.
2. Accessing data, data source can be local, big query or google cloud. Uploading 1M rows of data from local to google cloud took few seconds.
3. AutoML model predict/classify the data with no code. This can also be used to create an initial proof of concept
4. Monitors various tasks performed on the tool (ex:training completion) and alerts the users via email.
5. In depth information like documents and videos about vertex AI is made freely available by Google.

CONS

1. Auto ML is expensive.It charges $21 for training tabular data for a node hours.
2. Vertex AI natively uses infrastructure from Google. The team members should be trained to use Vertex AI from data readiness to deployment.
3. The data scientists who will be using Vertex AI should have some familiarity about the cloud technologies like containerising, cloud buckets, etc

COMPETITIVE ANALYSIS

S.no	Feature	Vertex AI	MLflow
1	Python programming language support	Yes	Yes
2	R programming language support	Yes	Yes
3	Read any data source using SQL syntax	Metadata only	Uses Hive metastore
4	Ability to save, load, list, tag and version multiple models	Yes (training models)	Yes (training models)
5	Store metrics & details of ML model training	Yes	Yes (mlflow experiments)
6	Feature store - Save pre calculated tabular data to be used by other team members	Yes	No
7	Scheduling - Run notebooks, jobs or pipelines on regular intervals	Requires cloud scheduler	No

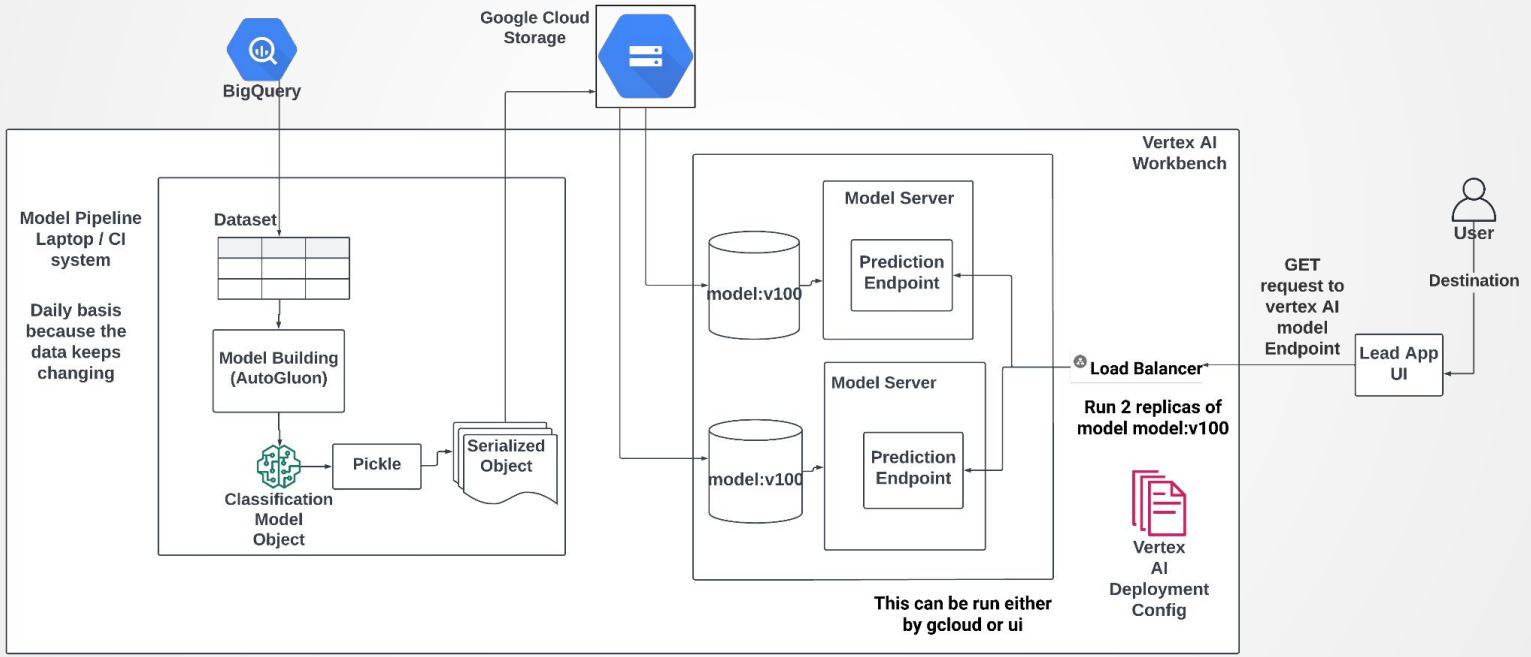
COMPETITIVE ANALYSIS

S.no	Feature	Vertex AI	MLflow
8	Integration	Natively integrated with BigQuery, Dataproc and Spark	Plugins allow to integrate with third-party storage solutions for experiments data (metrics and hyperparameters), artifacts and models
9	API Abstraction with simple one line code or User interface - Unified UI and API	Yes	Yes
10	AutoML - Try to automatically find a model that gives the best results	Yes	No
11	User Interface support for ML workflow	Supports end to end - development, serving, monitoring	Supports only for tracking and has its own registry

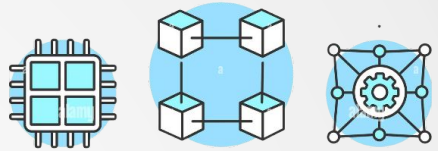
COMPETITIVE ANALYSIS

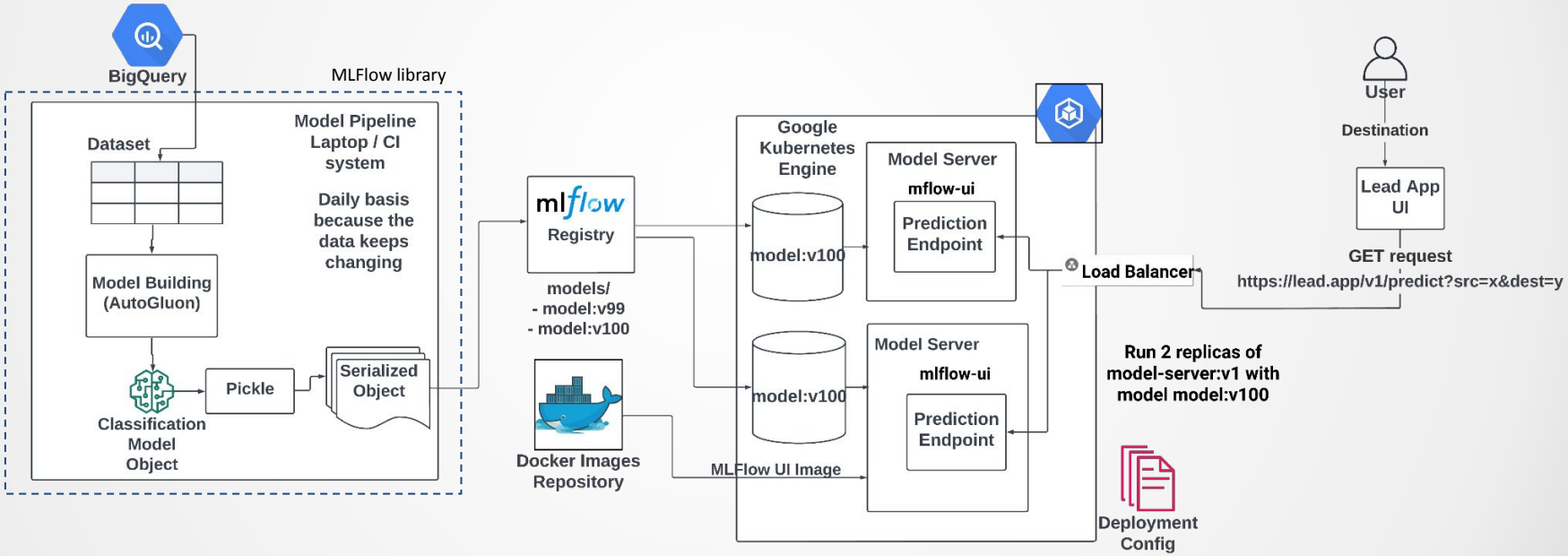
S.no	Feature	Vertex AI	MLflow
12	Support for open source frameworks - tensorflow, pytorch, scikit-learn	Yes	Yes
13	Use cases	● Data readiness ● Feature Engineering ● Training & hyperparameter tuning ● Model serving ● Model tuning and understanding ● Model monitoring ● Model management	● Tracking ● Projects ● Models ● Registry
14	Time reduction through MLOps automation	~ 15 days of manual labor	~ 6-8 days of manual work

PROPOSED SYSTEM - VERTEX AI



PROPOSED SYSTEM - MLFLOW





COST COMPARISON

	Vertex AI		MLFlow	
Items	**Deploy 1 model**	**Deploy additional model**	**Deploy 1 model**	**Deploy additional model**
Model Infrastructure	$6,000	+ $2,561	$8,000	+ $3,781
Data Support	$10,000 (labor) + $3,000	+ $2,250 (labor)	$7,000 (labor) + $3,200	+ $1,575 (labor)
Engineering / Deployment	$6,000 (labor) + $200	+ $750 (labor)	$18,000 (labor) + $300	+ $2,250 (labor)
Total Investment / yr	$16,000(labor) + $9200	$3,000(labor) + $2,561	$25,000(labor) + $11,500	$3,825(labor) + $3,781
5 year TCO	$62,000	+ $15,805	$82,000	+ $22,730

1. Vertex AI - Cost for deploying 100 models to production = $1.63 Million over first 5 years
2. MLFlow - Cost for deploying 100 models to production = $2.3 Million over first 5 years

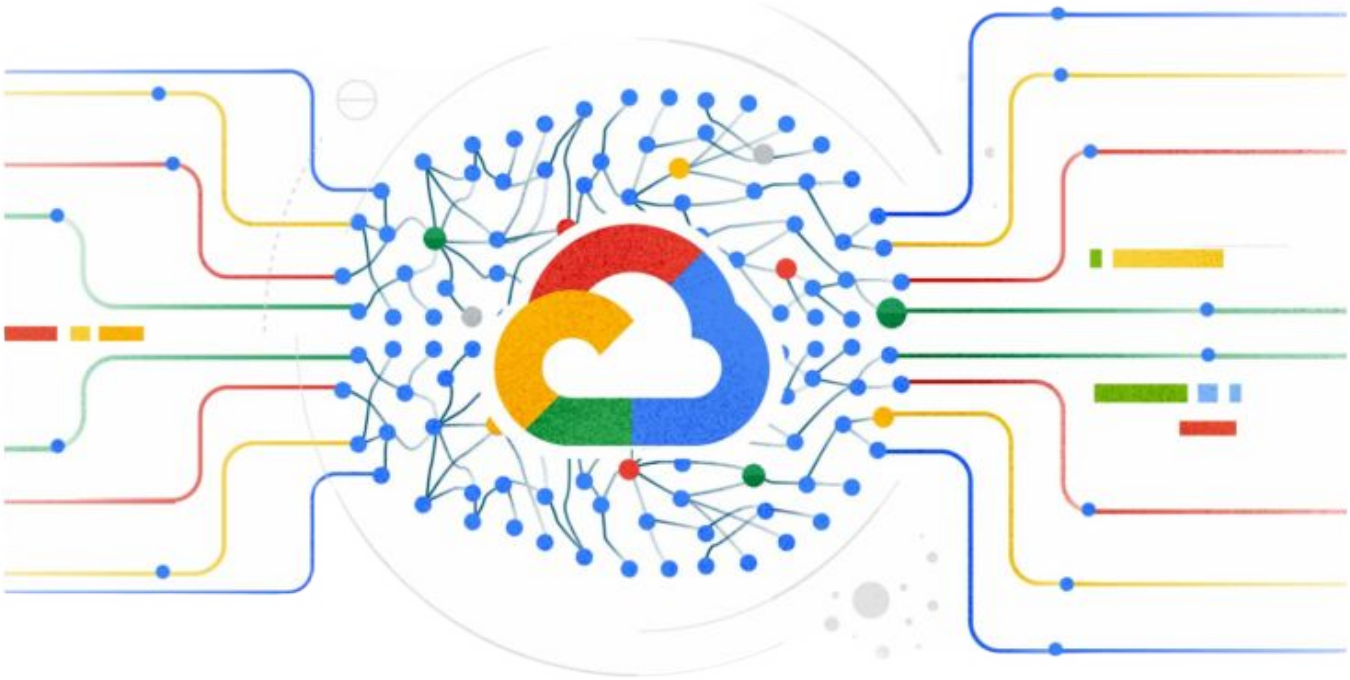
RECOMMENDATION - VERTEX AI

Vertex AI provides a **unified platform** that can accommodate data readiness to ML model deployment

A Data Scientist could close the product cycle from research to production, which **saves labor cost and time** for the entire lifecycle

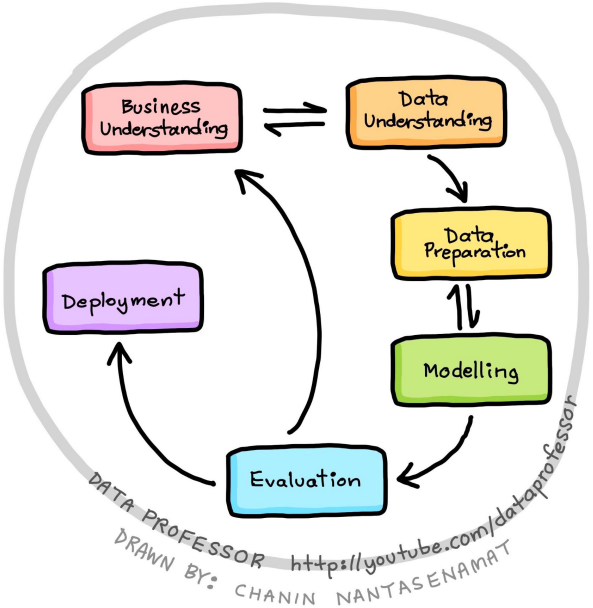
Adoption of Vertex AI to the current system is more **compatible**

DEMO - VERTEX AI



AlgoWorks - https://www.algoworks.com/blog/google-vertex-ai/

PROCESS MODEL - CRISP DM



Generalizable among **Data Science** projects

Research phase follows a **horizontal** slicing & **implementation** phase follows a **vertical** slicing

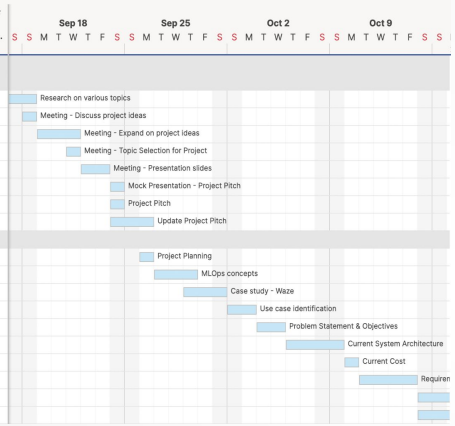
'Mix & Match' model which accommodates waterfall & agile models

Suitable for **small team size** (4 members)

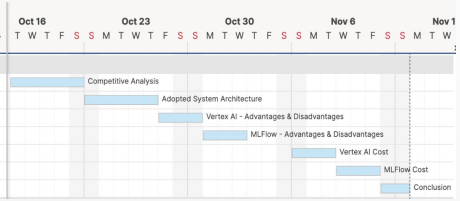
The Data Science Process - https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

PROJECT PLANNING

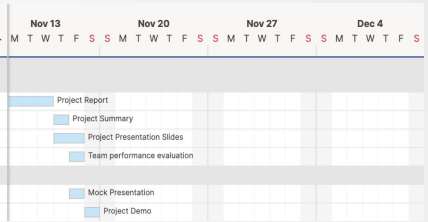
Task Title	Task Owner	Start Date	Due Date	Du...	PCT OF TASK COMP...
RESEARCH & REPORT					☑
Project Conception and Initiation					☑
Research on various topics	Team	09/17/22	09/18/22	2	☑
Meeting - Discuss project ideas	Team	09/18/22	09/18/22	1	☑
Meeting - Expand on project ideas	Team	09/19/22	09/21/22	3	☑
Meeting - Topic Selection for Project	Team	09/21/22	09/21/22	1	☑
Meeting - Presentation slides	Team	09/22/22	09/23/22	2	☑
Mock Presentation - Project Pitch	Team	09/24/22	09/24/22	1	☑
Project Pitch	Team	09/24/22	09/24/22	1	☑
Update Project Pitch	Team	09/24/22	09/26/22	3	☑
Business Understanding					☑
Project Planning	SVR	09/26/22	09/26/22	1	☑
MLOps concepts	Team	09/27/22	09/29/22	3	☑
Case study - Waze	TJ	09/29/22	10/01/22	3	☑
Use case identification	Team	10/02/22	10/03/22	2	☑
Problem Statement & Objectives	TJ	10/04/22	10/05/22	2	☑
Current System Architecture	SVR	10/06/22	10/09/22	4	☑
Current Cost	SVR	10/10/22	10/10/22	1	☑
Requirements Gathering	SVR	10/11/22	10/14/22	4	☑
Research on MLFlow	KK, TJ	10/15/22	10/17/22	3	☑
Research on Vertext AI	AS, SVR	10/15/22	10/17/22	3	☑



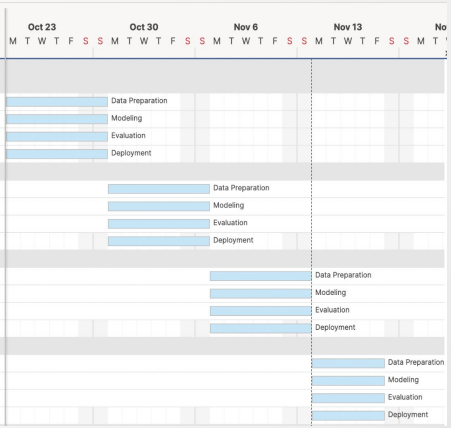
Task Title	Task Owner	Start Date	Due Date	Du...	PCT OF TASK COMP...
Platform Comparison					☑
Competitive Analysis	Team	10/18/22	10/22/22	5	☑
Adopted System Architecture	SVR	10/23/22	10/27/22	5	☑
Vertex AI - Advantages & Disadvanta	AS	10/28/22	10/30/22	3	☑
MLFlow - Advantages & Disadvantag	KK	10/31/22	11/02/22	3	☑
Vertex AI Cost	SVR, AS	11/06/22	11/08/22	3	☑
MLFlow Cost	TJ, KK	11/09/22	11/11/22	3	☑
Conclusion	Team	11/12/22	11/13/22	2	☑



Task Title	Task Owner	Start Date	Due Date	Du...	PCT OF TASK COMP...
PROJECT DOCUMENTS					☑
Deliverables					☑
Project Report	SVR, TJ	11/14/22	11/16/22	3	☑
Project Summary	SVR, TJ	11/17/22	11/17/22	1	☑
Project Presentation Slides	AS, KK	11/17/22	11/18/22	2	☑
Team performance evaluation	SVR	11/18/22	11/18/22	1	☑
Project Presentation					☑
Mock Presentation	Team	11/18/22	11/18/22	1	☑
Project Demo	Team	11/19/22	11/19/22	1	☑



Task Title	Task Owner	Start Date	Due Date	Du...	PCT OF TASK COMP...
IMPLEMENTATION					☑
Sprint 1					☑
Data Preparation	AS, KK	10/24/22	10/30/22	7	☑
Modeling	AS, KK	10/24/22	10/30/22	7	☑
Evaluation	AS, KK	10/24/22	10/30/22	7	☑
Deployment	AS, KK	10/24/22	10/30/22	7	☑
Sprint 2					☑
Data Preparation	AS, KK	10/31/22	11/06/22	7	☑
Modeling	AS, KK	10/31/22	11/06/22	7	☑
Evaluation	AS, KK	10/31/22	11/06/22	7	☑
Deployment	AS, KK	10/31/22	11/06/22	7	☑
Sprint 3					☑
Data Preparation	AS, KK	11/07/22	11/13/22	7	☑
Modeling	AS, KK	11/07/22	11/13/22	7	☑
Evaluation	AS, KK	11/07/22	11/13/22	7	☑
Deployment	AS, KK	11/07/22	11/13/22	7	☑
Sprint 4					☑
Data Preparation	AS, KK	11/14/22	11/18/22	5	☑
Modeling	AS, KK	11/14/22	11/18/22	5	☑
Evaluation	AS, KK	11/14/22	11/18/22	5	☑
Deployment	AS, KK	11/14/22	11/18/22	5	☑



THANK YOU!


THANK
you