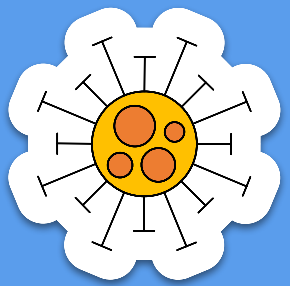
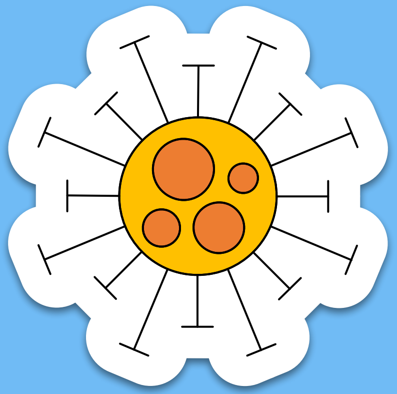
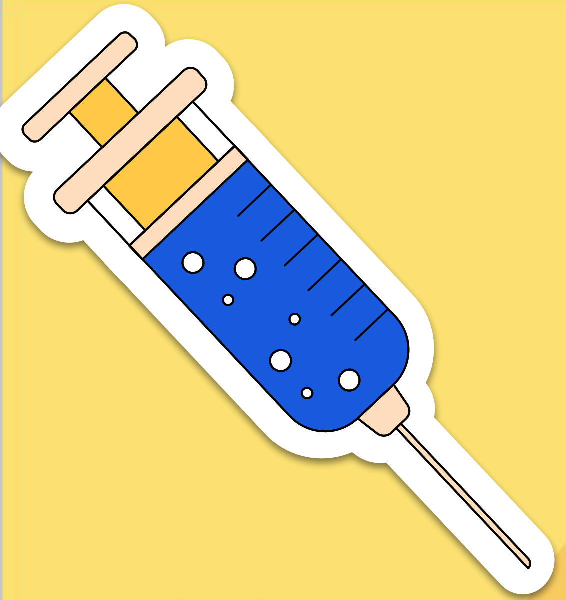
Predictive Modeling of Immune Responses to Pertussis Vaccination

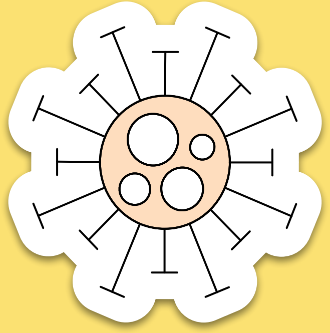


Group 4:
Javier Garcia (Business Manager & Data Engineer)
Weikang Guan (Data Analyst & Visualization Developer)
Brian Qian (Data & Dashboard Developer)
Peng Cheng (Solution Architect & Budget Controller)

Advisors:
Barry Grant (Professor)
Jason Hsiao (Biological Sciences PhD Student)

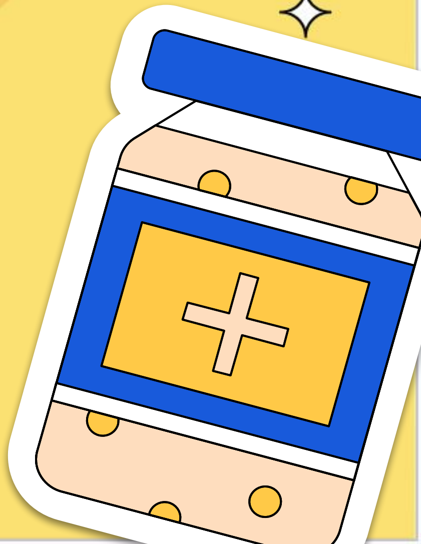






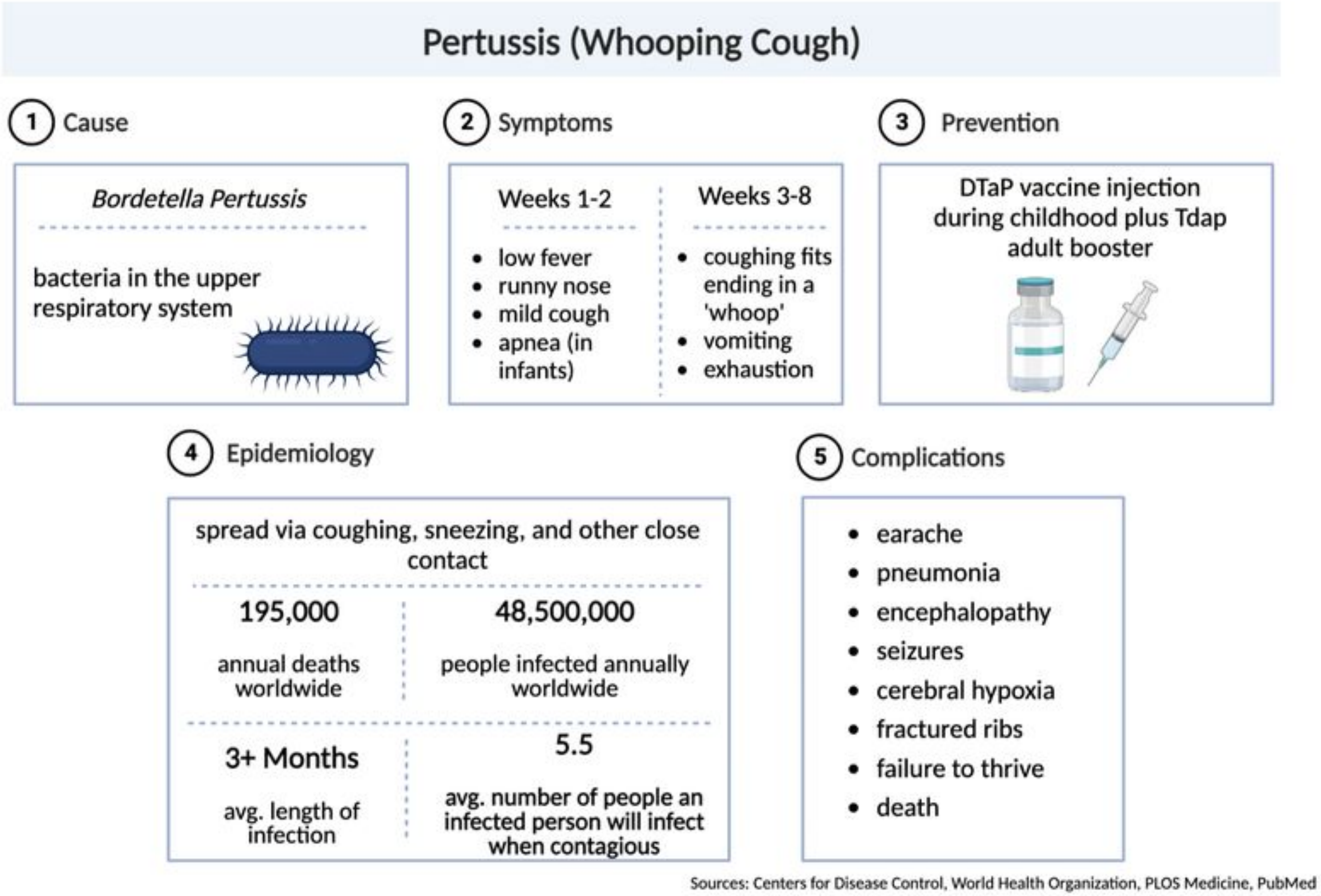
01

Project Background and Definition



1. Background and Definition

- Pertussis, or Whooping cough, is a highly contagious lung infection
- Two vaccines: whole-cellular (wP) and acellular (aP)
- Challenges of balancing vaccine safety and efficacy
- Importance of ongoing research and monitoring of vaccine effectiveness over time
- How does vaccine induced immunity change over time per person?



Pertussis (Whooping Cough)

1 Cause

Bordetella Pertussis

bacteria in the upper respiratory system

2 Symptoms

Weeks 1-2

- low fever
- runny nose
- mild cough
- apnea (in infants)

Weeks 3-8

- coughing fits ending in a 'whoop'
- vomiting
- exhaustion

3 Prevention

DTaP vaccine injection during childhood plus Tdap adult booster

4 Epidemiology

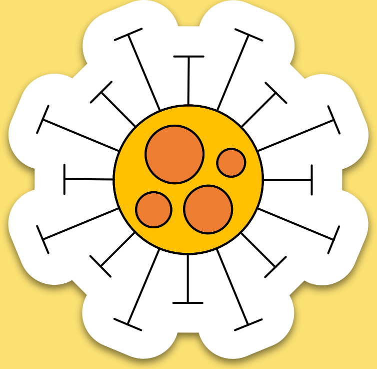
spread via coughing, sneezing, and other close contact

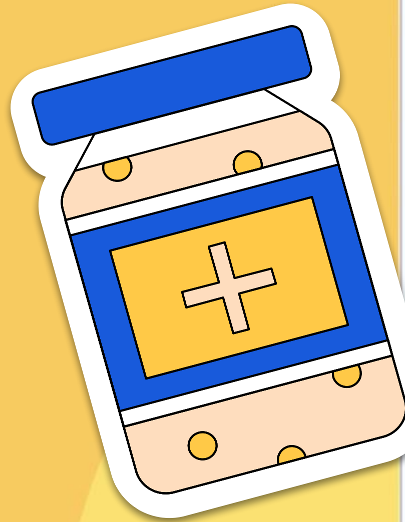
195,000 annual deaths worldwide	48,500,000 people infected annually worldwide
3+ Months avg. length of infection	5.5 avg. number of people an infected person will infect when contagious

5 Complications

- earache
- pneumonia
- encephalopathy
- seizures
- cerebral hypoxia
- fractured ribs
- failure to thrive
- death

Sources: Centers for Disease Control, World Health Organization, PLOS Medicine, PubMed





16 Million

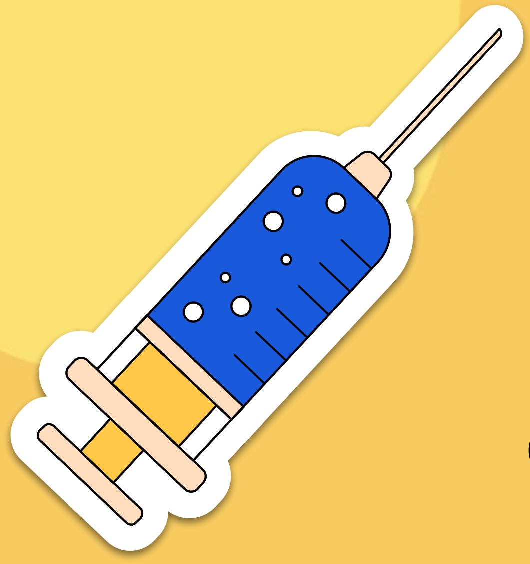
Estimated Annual Pertussis Cases Worldwide

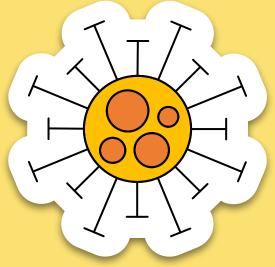
195,000

Annual Deaths Due to Pertussis

85%

Global Vaccination Coverage for DTP3 in 2021





1. Background and Definition

Goals

- Help save lives from preventable pertussis cases
- Share academic research with growing community
- Advance the understanding of immunology and use models in the real world to improve vaccine effectiveness

UC San Diego

JACOBS SCHOOL OF ENGINEERING

MAS Data Science and Engineering

CMI-PB

COMPUTATIONAL MODELS OF IMMUNITY

PERTUSSIS BOOST

02

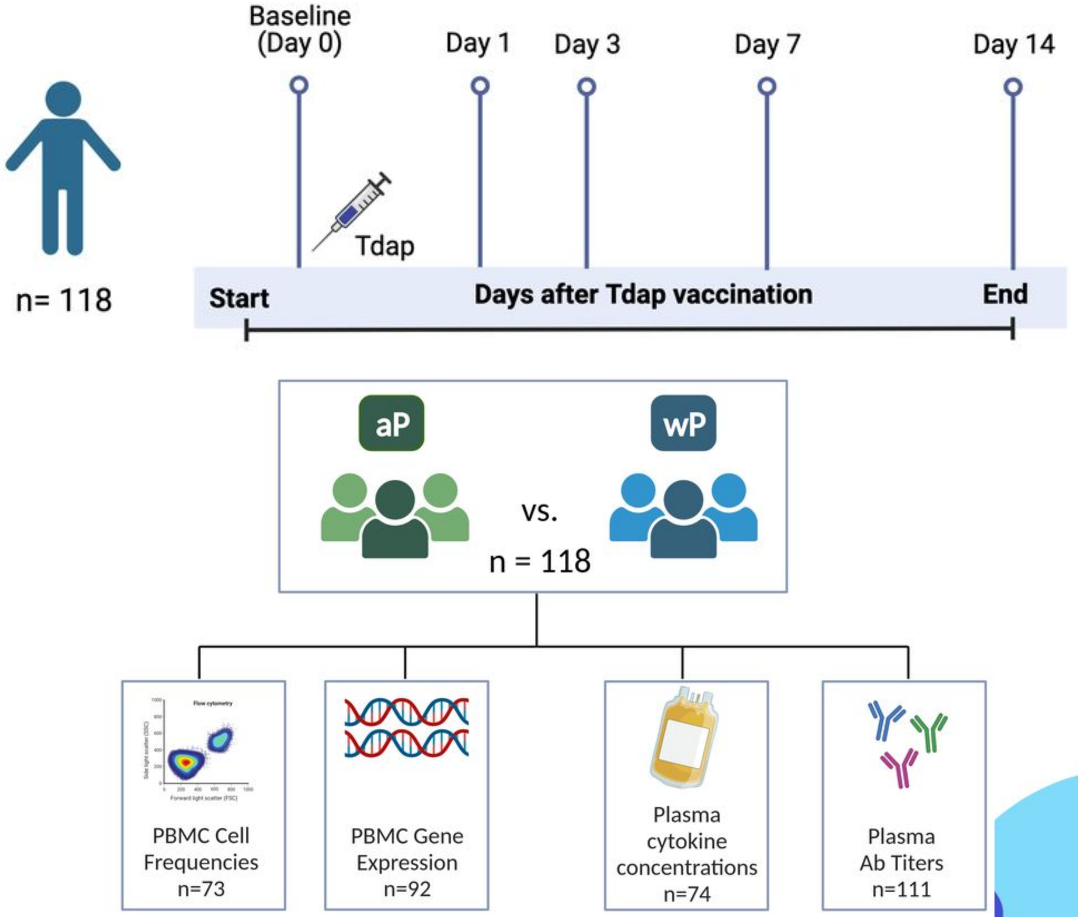
Datasets Overview

2. Overview of Data Sets

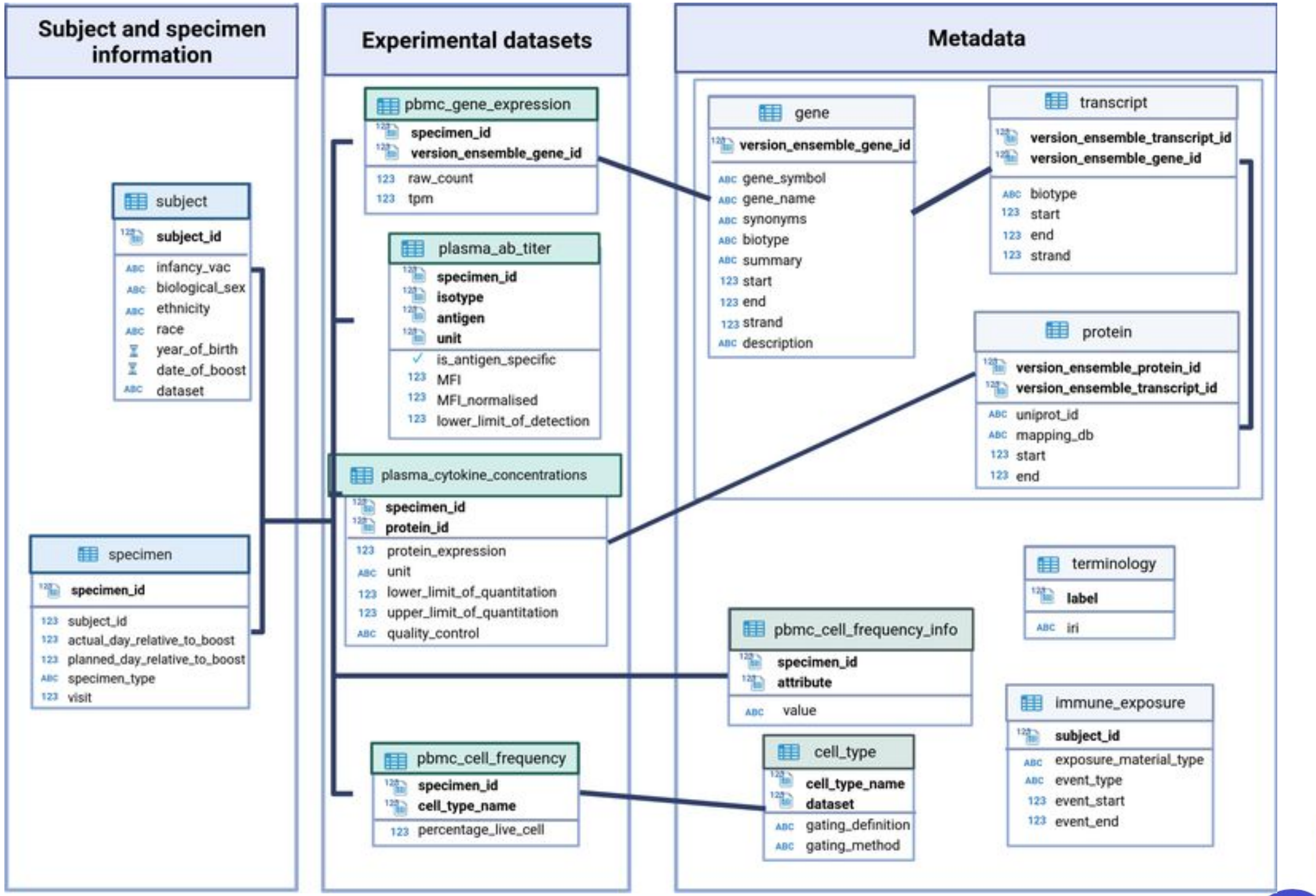
Objective: Analyze immune responses post-Tdap booster vaccination.

Subjects: 118 individuals contributing 500+ blood specimens.

Timeframe: Pre- and post-vaccination (days 1, 3, 7, and 14).


Baseline (Day 0)
Day 1
Day 3
Day 7
Day 14
Tdap
n= 118
Start
Days after Tdap vaccination
End
aP
wP
vs.
n = 118
Flow cytometry
PBMC Cell Frequencies n=73
PBMC Gene Expression n=92
Plasma cytokine concentrations n=74
Plasma Ab Titers n=111

2. Overview of Data Sets



Subject and specimen information

subject
- **subject_id**
- infancy_vac
- biological_sex
- ethnicity
- race
- year_of_birth
- date_of_boost
- dataset

specimen
- **specimen_id**
- subject_id
- actual_day_relative_to_boost
- planned_day_relative_to_boost
- specimen_type
- visit

Experimental datasets

pbmc_gene_expression
- **specimen_id**
- **version_ensemble_gene_id**
- raw_count
- tpm

plasma_ab_titer
- **specimen_id**
- **isotype**
- **antigen**
- **unit**
- is_antigen_specific
- MFI
- MFI_normalised
- lower_limit_of_detection

plasma_cytokine_concentrations
- **specimen_id**
- **protein_id**
- protein_expression
- unit
- lower_limit_of_quantitation
- upper_limit_of_quantitation
- quality_control

pbmc_cell_frequency
- **specimen_id**
- **cell_type_name**
- percentage_live_cell

Metadata

gene
- **version_ensemble_gene_id**
- gene_symbol
- gene_name
- synonyms
- biotype
- summary
- start
- end
- strand
- description

transcript
- **version_ensemble_transcript_id**
- **version_ensemble_gene_id**
- biotype
- start
- end
- strand

protein
- **version_ensemble_protein_id**
- **version_ensemble_transcript_id**
- uniprot_id
- mapping_db
- start
- end

terminology
- **label**
- iri

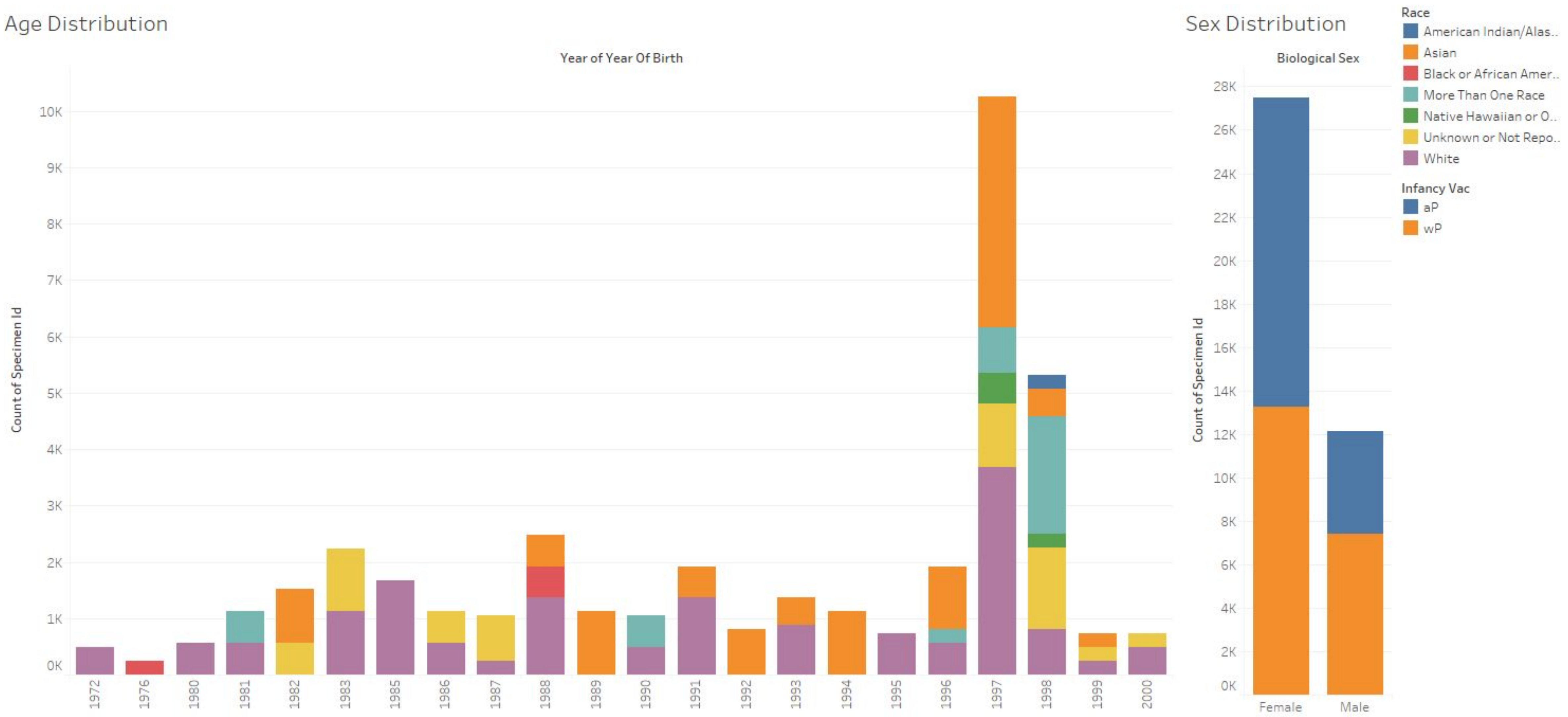
pbmc_cell_frequency_info
- **specimen_id**
- **attribute**
- value

immune_exposure
- **subject_id**
- exposure_material_type
- event_type
- event_start
- event_end

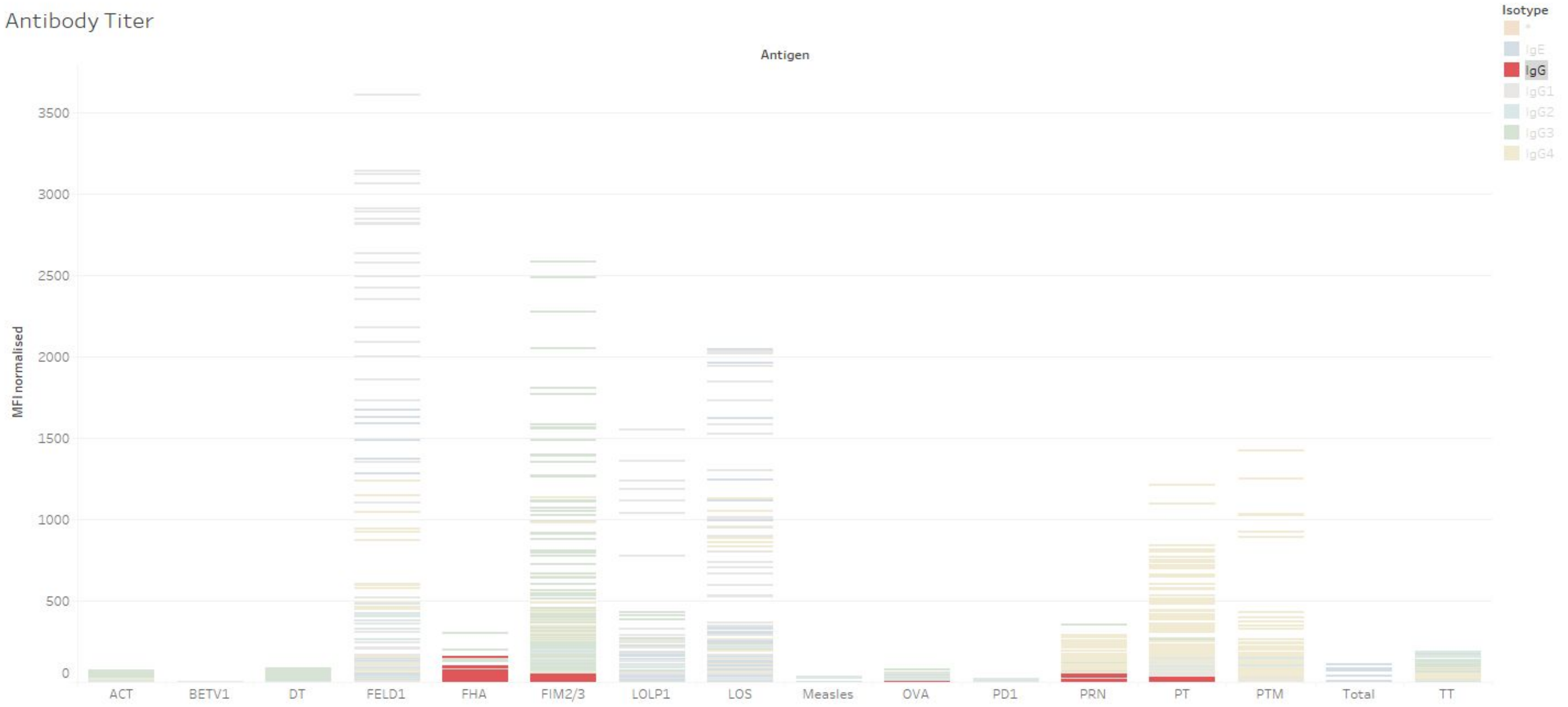
cell_type
- **cell_type_name**
- **dataset**
- gating_definition
- gating_method

Source: CMI-PB

2. Overview of Data Sets

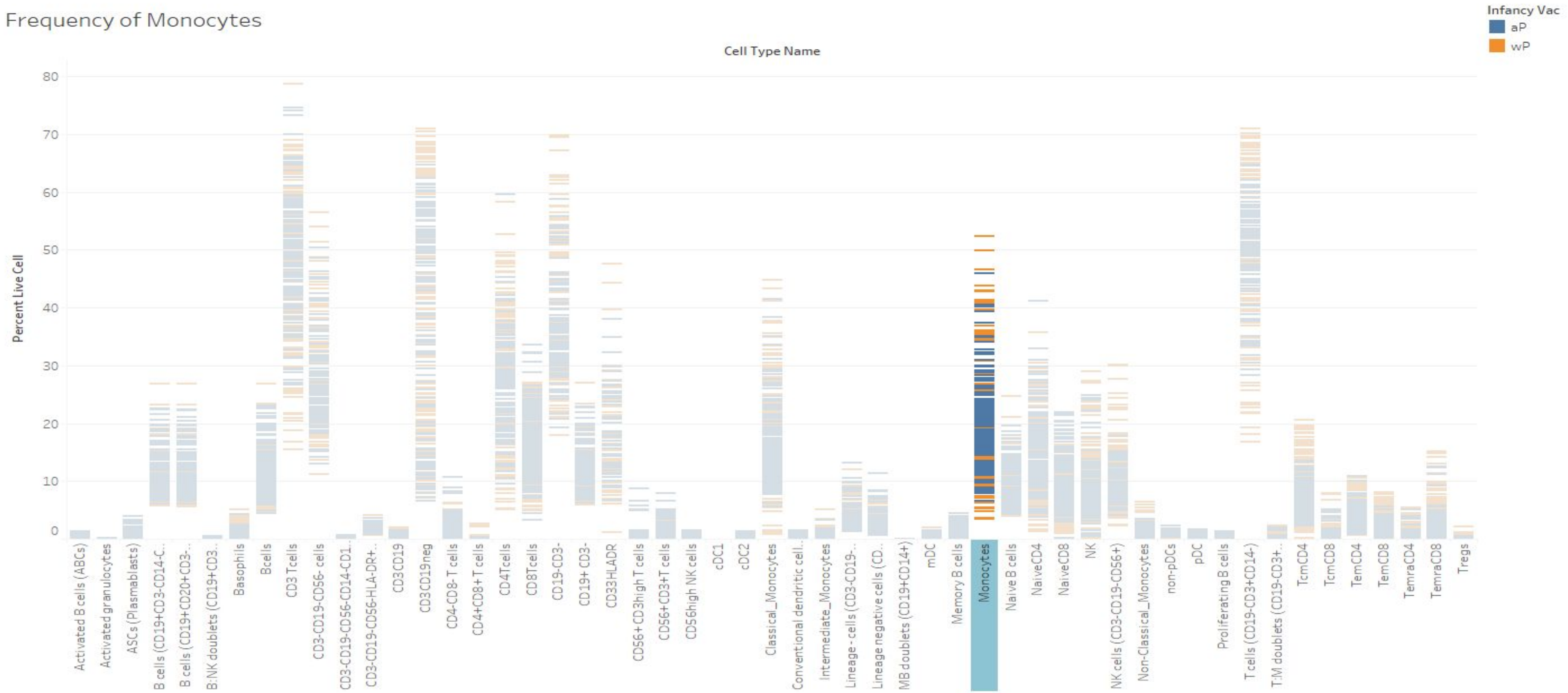


2. Overview of Data Sets



Immunoglobulin

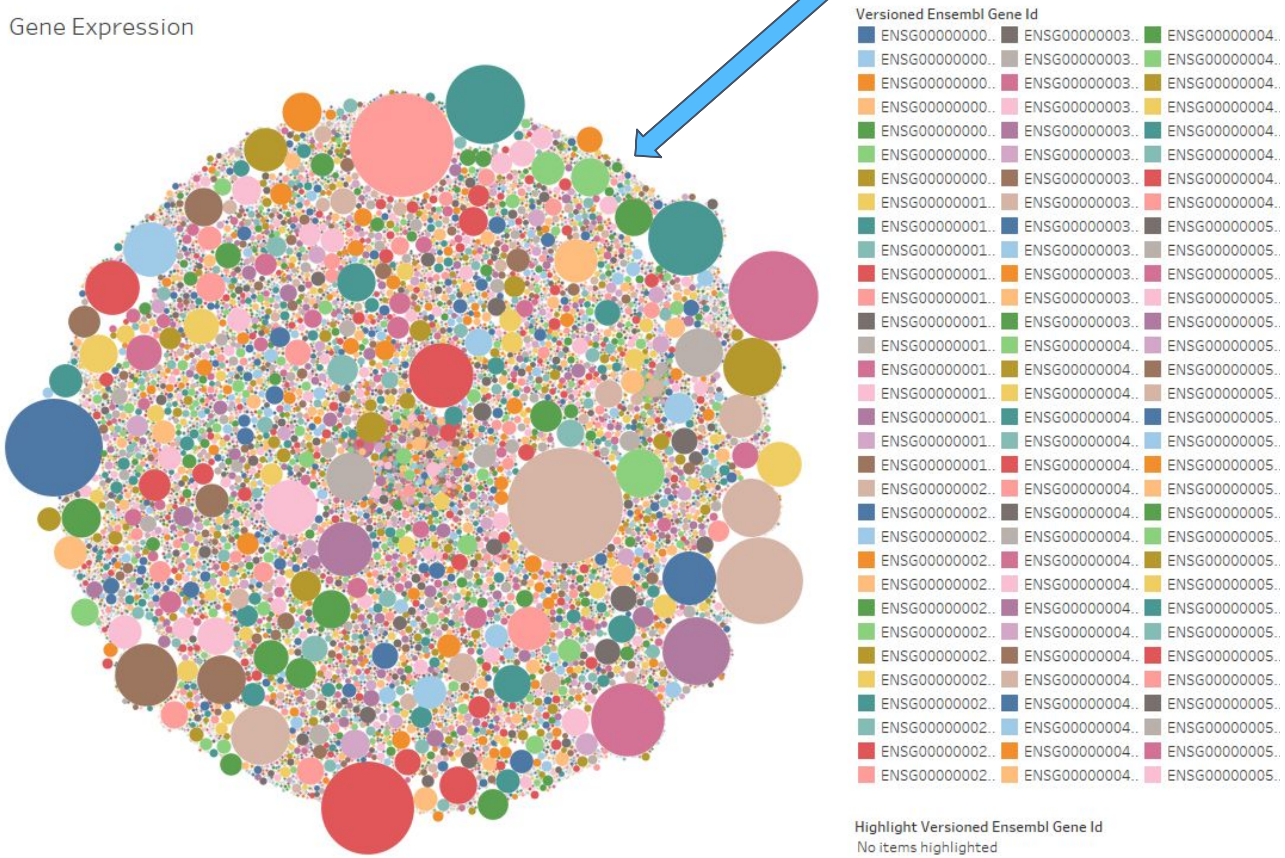
2. Overview of Data Sets

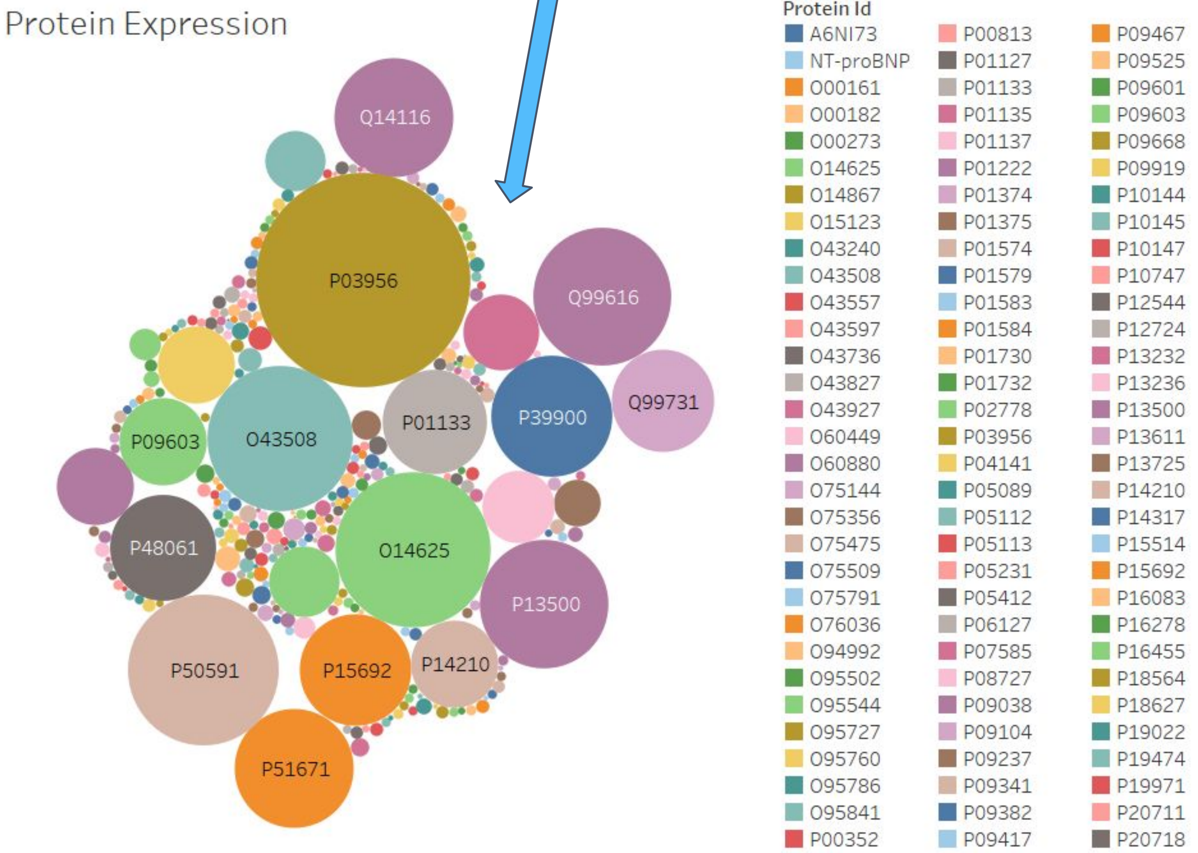


Frequency of Monocytes

2. Overview of Data Sets

DNA → mRNA(Gene Expression) → Protein Expression → CCL3





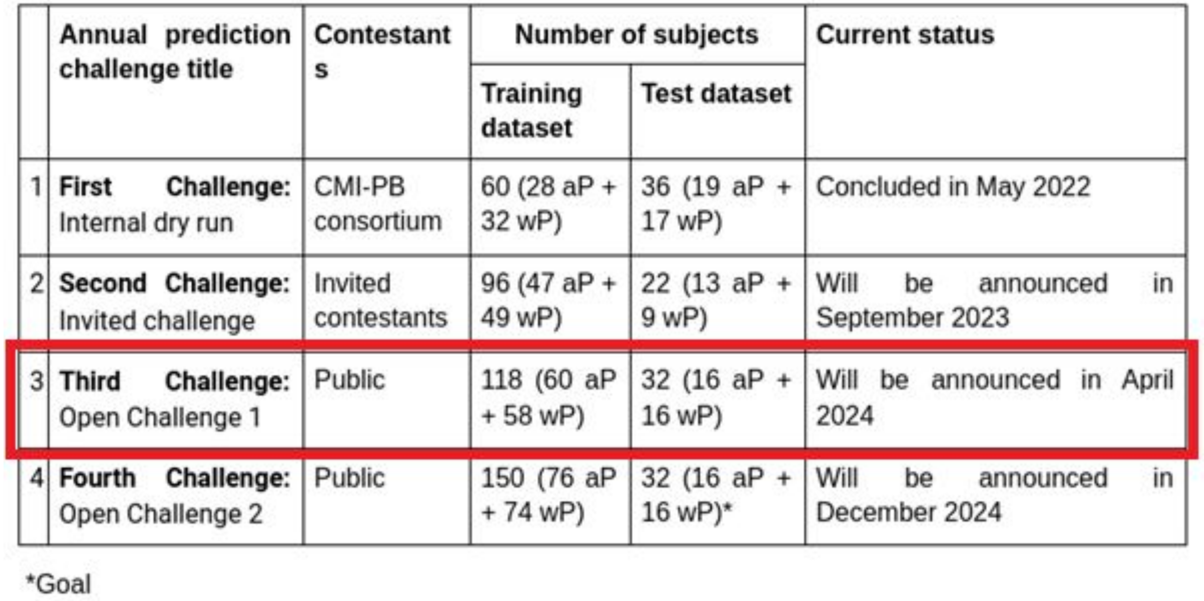
03

Project Objectives

3. Project Objectives

- **Building Computational Models:** To predict vaccination outcomes for newly tested individuals.

- **Participation in the Prediction Challenge:** Showcase intuition and analytical skills by engaging in the community prediction challenge

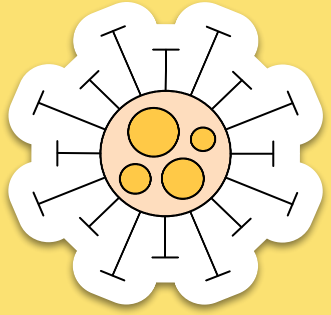


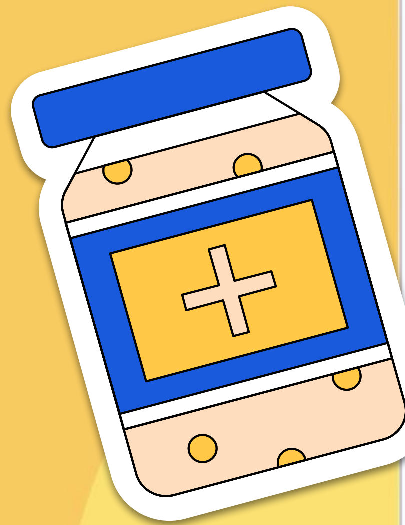
	Annual prediction challenge title	Contestant s	Number of subjects		Current status
			Training dataset	Test dataset	
1	**First Challenge:** Internal dry run	CMI-PB consortium	60 (28 aP + 32 wP)	36 (19 aP + 17 wP)	Concluded in May 2022
2	**Second Challenge:** Invited challenge	Invited contestants	96 (47 aP + 49 wP)	22 (13 aP + 9 wP)	Will be announced in September 2023
3	**Third Challenge:** Open Challenge 1	Public	118 (60 aP + 58 wP)	32 (16 aP + 16 wP)	Will be announced in April 2024
4	**Fourth Challenge:** Open Challenge 2	Public	150 (76 aP + 74 wP)	32 (16 aP + 16 wP)*	Will be announced in December 2024

*Goal

3. Project Objectives

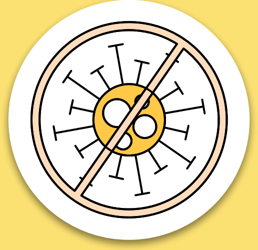
- Accurate prediction of immune responses
- Identifying what variables induce a strong response
- Real-time access to dynamic immune response data
- Efficient handling of diverse features

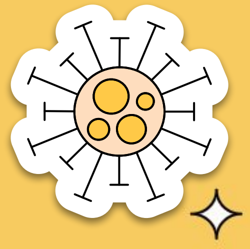




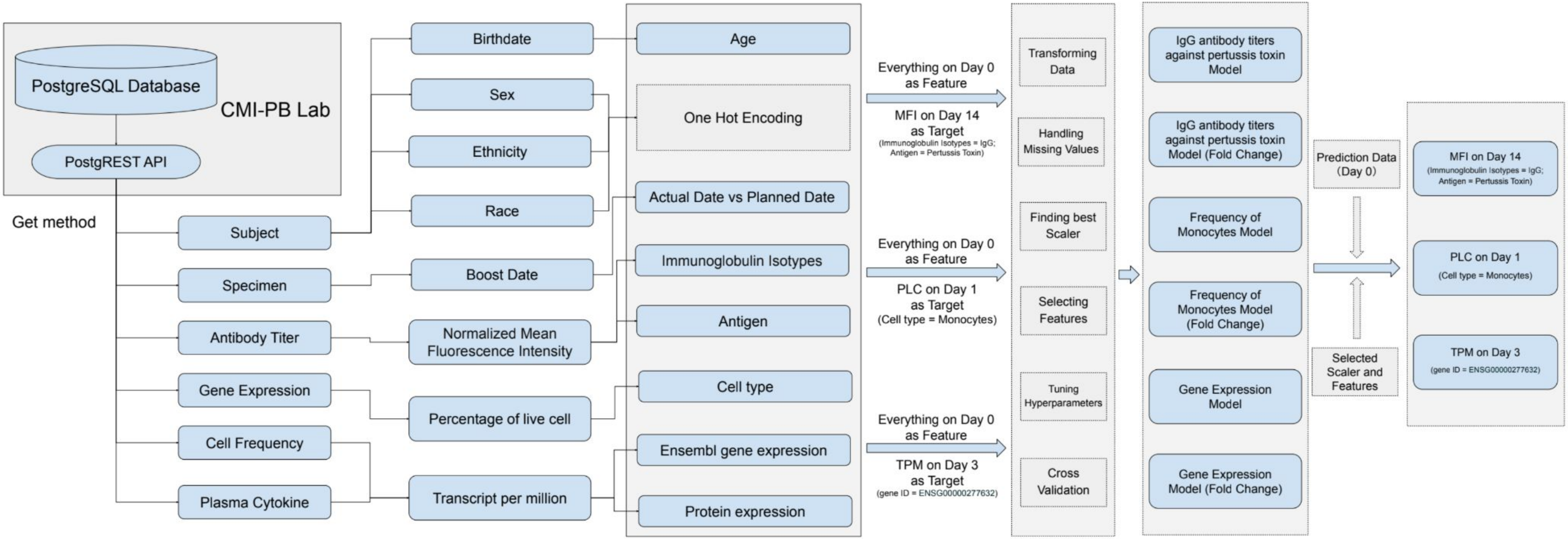
04

Methods and Techniques





4. Methods and Techniques



PostgreSQL Database

CMI-PB Lab

PostgREST API

Get method

Subject

Specimen

Antibody Titer

Gene Expression

Cell Frequency

Plasma Cytokine

Birthdate

Sex

Ethnicity

Race

Boost Date

Normalized Mean Fluorescence Intensity

Percentage of live cell

Transcript per million

Age

One Hot Encoding

Actual Date vs Planned Date

Immunoglobulin Isotypes

Antigen

Cell type

Ensembl gene expression

Protein expression

Everything on Day 0 as Feature

MFI on Day 14 as Target (Immunoglobulin Isotypes = IgG; Antigen = Pertussis Toxin)

Everything on Day 0 as Feature

PLC on Day 1 as Target (Cell type = Monocytes)

Everything on Day 0 as Feature

TPM on Day 3 as Target (gene ID = ENSG00000277632)

Transforming Data

Handling Missing Values

Finding best Scaler

Selecting Features

Tuning Hyperparameters

Cross Validation

IgG antibody titers against pertussis toxin Model

IgG antibody titers against pertussis toxin Model (Fold Change)

Frequency of Monocytes Model

Frequency of Monocytes Model (Fold Change)

Gene Expression Model

Gene Expression Model (Fold Change)

Prediction Data (Day 0)

Selected Scaler and Features

MFI on Day 14 (Immunoglobulin Isotypes = IgG; Antigen = Pertussis Toxin)

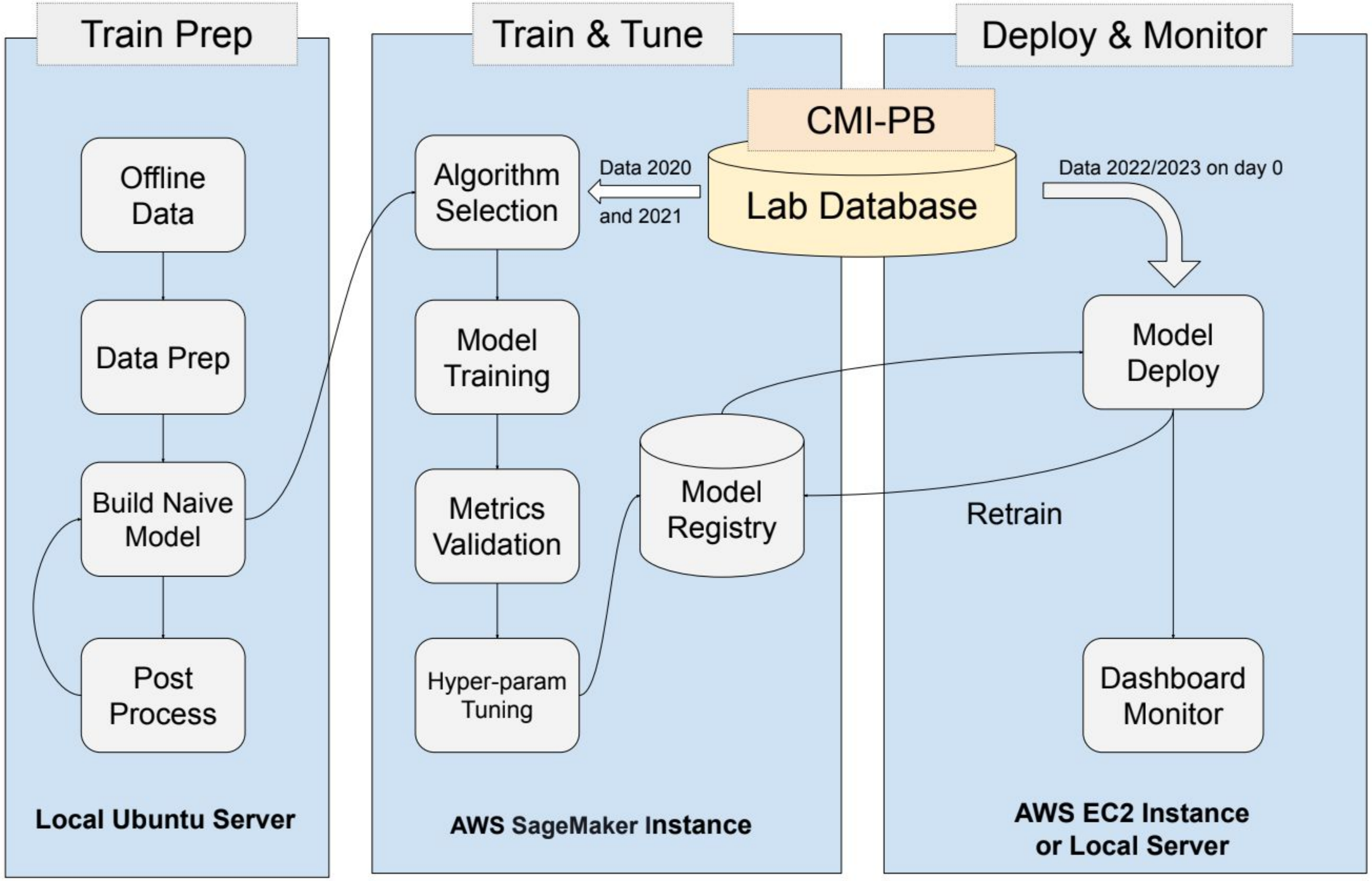
PLC on Day 1 (Cell type = Monocytes)

TPM on Day 3 (gene ID = ENSG00000277632)

05

Solution Architecture

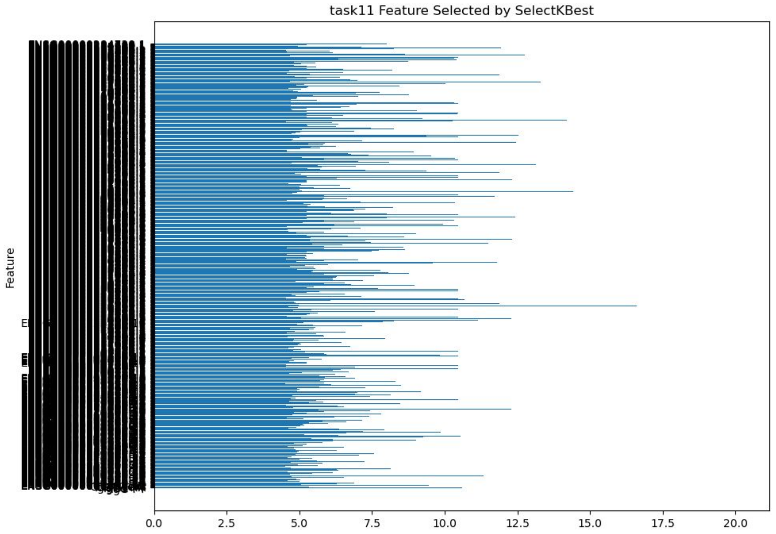
5. Solution Architecture

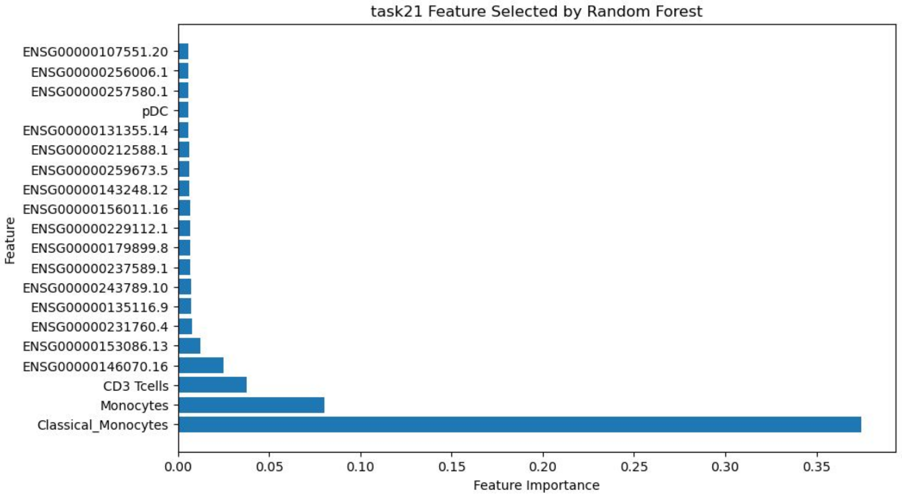


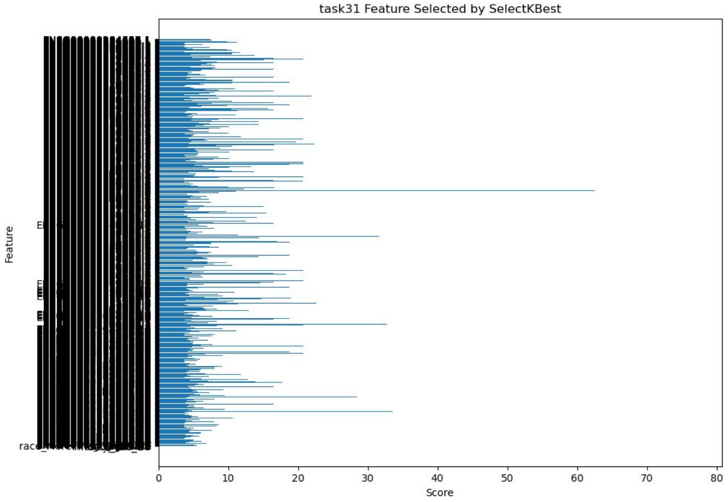
06

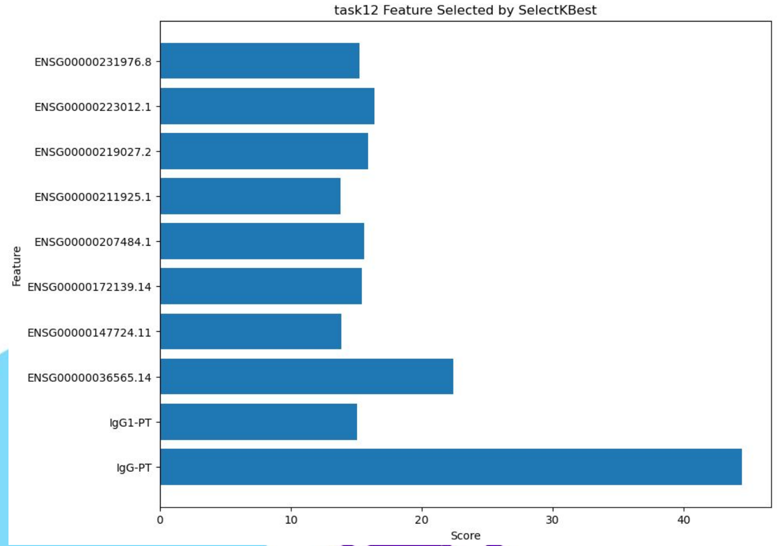
Analysis of Results

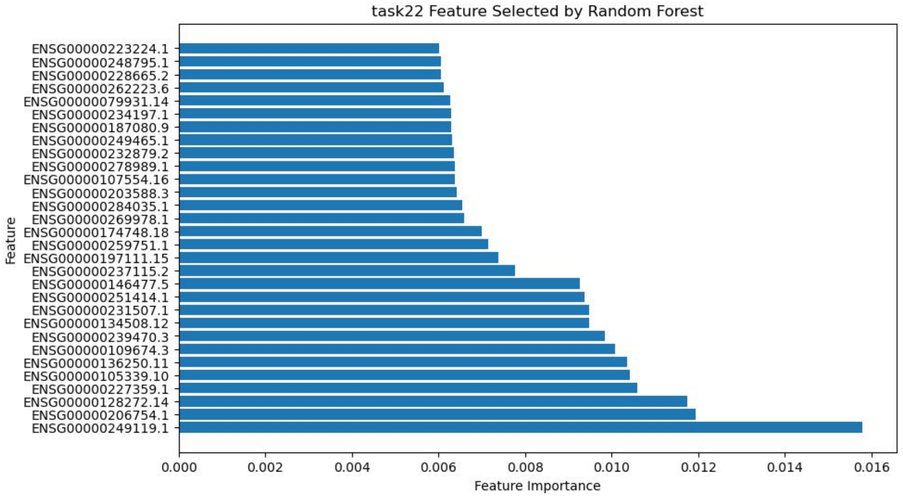
6. Analysis of Results

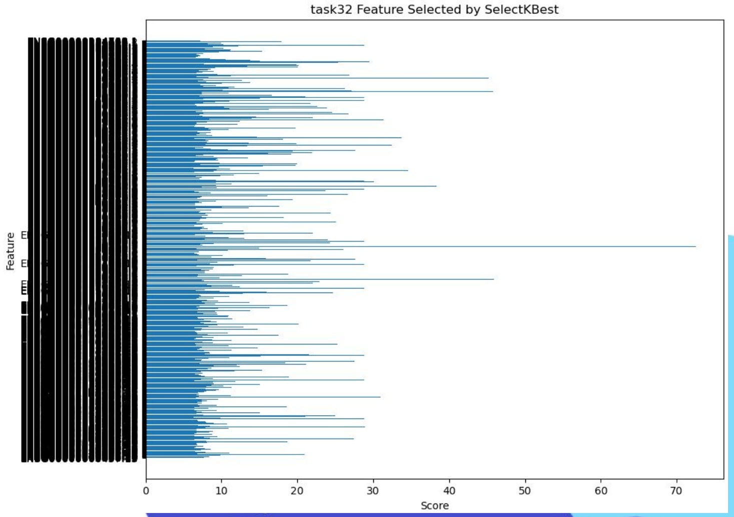




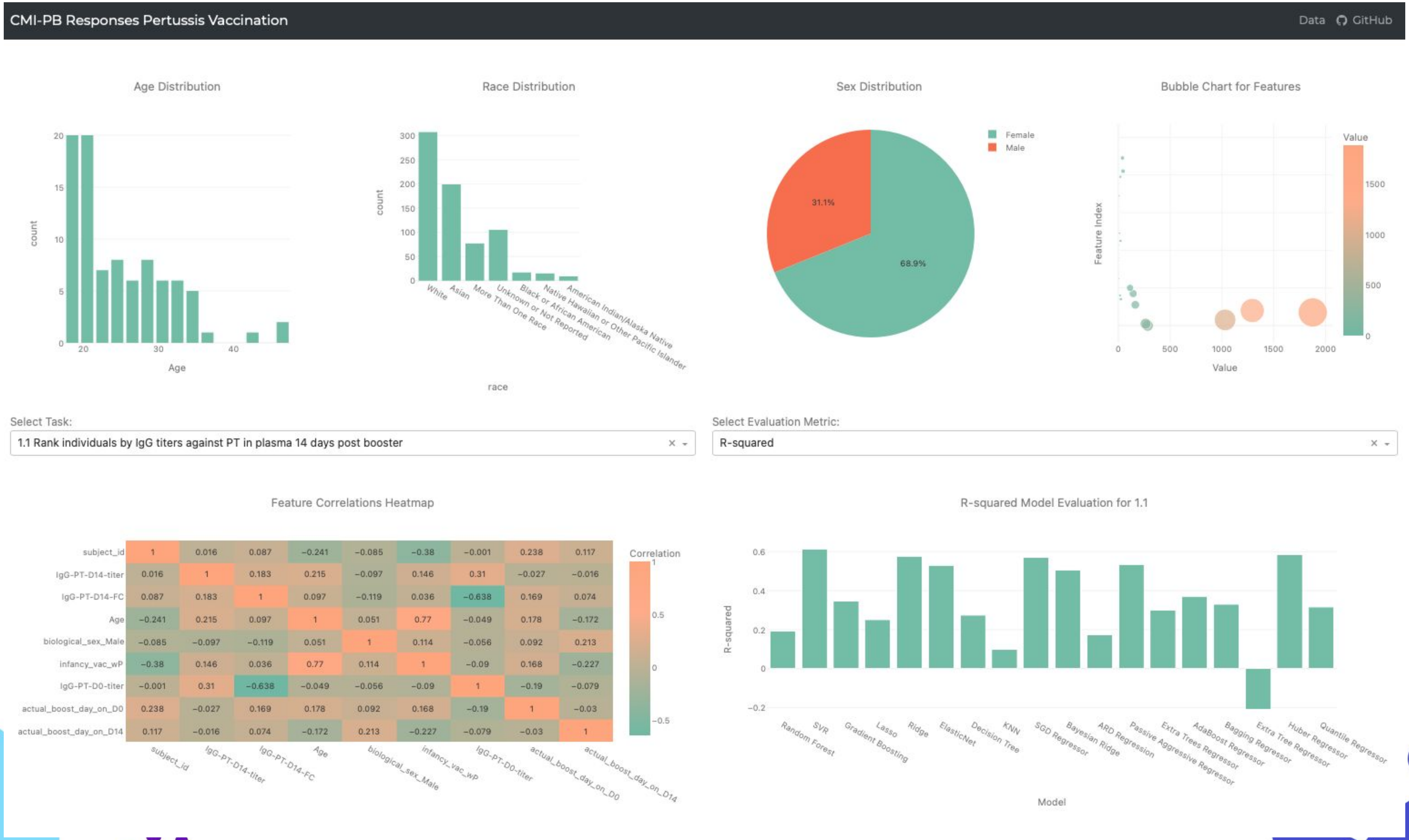








6. Dashboard



CMI-PB Responses Pertussis Vaccination

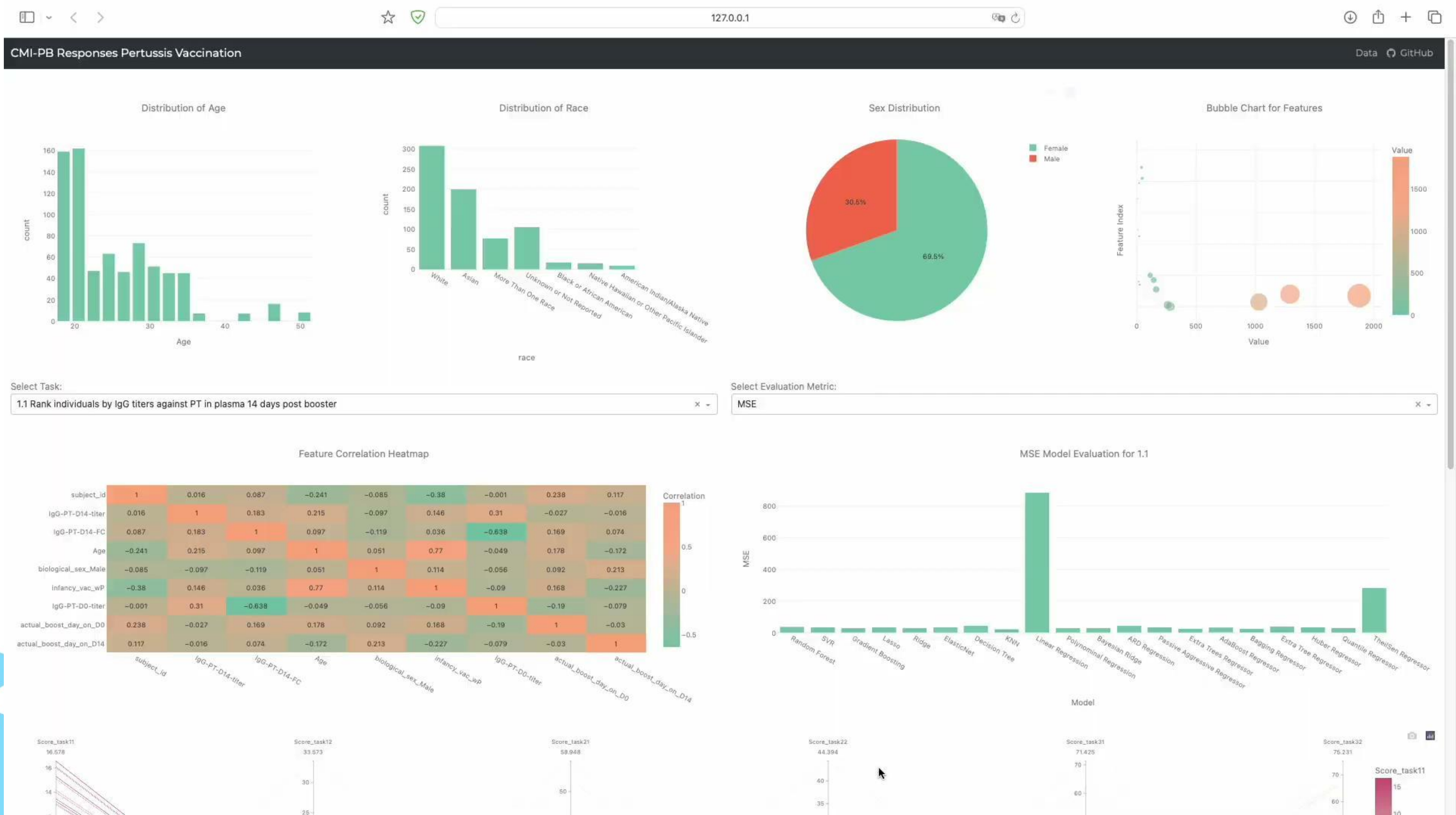
Select Task: 1.1 Rank individuals by IgG titers against PT in plasma 14 days post booster

Select Evaluation Metric: R-squared

Feature Correlations Heatmap

	subject_id	IgG-PT-D14-titer	IgG-PT-D14-FC	Age	biological_sex_Male	infancy_vac_wP	IgG-PT-D0-titer	actual_boost_day_on_D0	actual_boost_day_on_D14
subject_id	1	0.016	0.087	−0.241	−0.085	−0.38	−0.001	0.238	0.117
IgG-PT-D14-titer	0.016	1	0.183	0.215	−0.097	0.146	0.31	−0.027	−0.016
IgG-PT-D14-FC	0.087	0.183	1	0.097	−0.119	0.036	−0.638	0.169	0.074
Age	−0.241	0.215	0.097	1	0.051	0.77	−0.049	0.178	−0.172
biological_sex_Male	−0.085	−0.097	−0.119	0.051	1	0.114	−0.056	0.092	0.213
infancy_vac_wP	−0.38	0.146	0.036	0.77	0.114	1	−0.09	0.168	−0.227
IgG-PT-D0-titer	−0.001	0.31	−0.638	−0.049	−0.056	−0.09	1	−0.19	−0.079
actual_boost_day_on_D0	0.238	−0.027	0.169	0.178	0.092	0.168	−0.19	1	−0.03
actual_boost_day_on_D14	0.117	−0.016	0.074	−0.172	0.213	−0.227	−0.079	−0.03	1

6. Dashboard - Demo



127.0.0.1

CMI-PB Responses Pertussis Vaccination

Data GitHub

Select Task:
1.1 Rank individuals by IgG titers against PT in plasma 14 days post booster

Select Evaluation Metric:
MSE

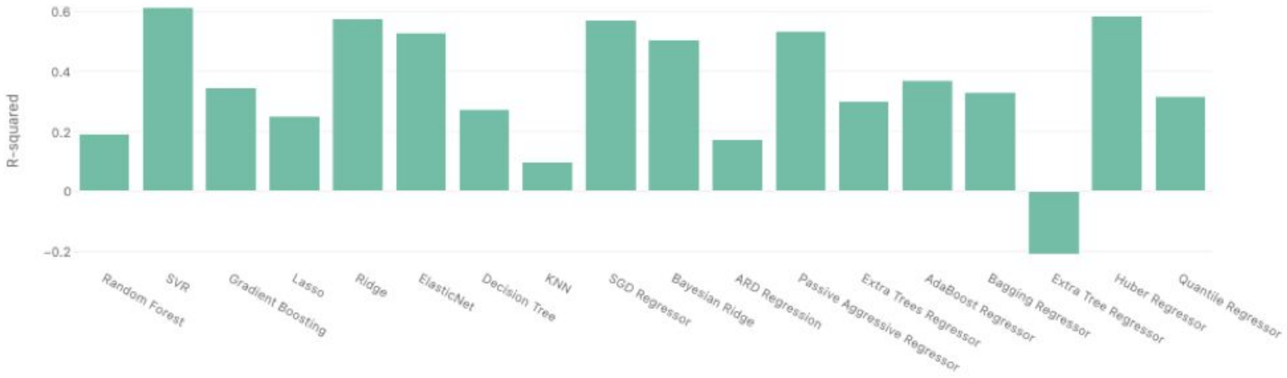
Feature Correlation Heatmap

	subject_id	IgG-PT-D14-titer	IgG-PT-D14-FC	Age	biological_sex_Male	infancy_vac_wP	IgG-PT-D0-titer	actual_boost_day_on_D0	actual_boost_day_on_D14
subject_id	1	0.016	0.087	−0.241	−0.085	−0.38	−0.001	0.238	0.117
IgG-PT-D14-titer	0.016	1	0.183	0.215	−0.097	0.146	0.31	−0.027	−0.016
IgG-PT-D14-FC	0.087	0.183	1	0.097	−0.119	0.036	−0.638	0.169	0.074
Age	−0.241	0.215	0.097	1	0.051	0.77	−0.049	0.178	−0.172
biological_sex_Male	−0.085	−0.097	−0.119	0.051	1	0.114	−0.056	0.092	0.213
infancy_vac_wP	−0.38	0.146	0.036	0.77	0.114	1	−0.09	0.168	−0.227
IgG-PT-D0-titer	−0.001	0.31	−0.638	−0.049	−0.056	−0.09	1	−0.19	−0.079
actual_boost_day_on_D0	0.238	−0.027	0.169	0.178	0.092	0.168	−0.19	1	−0.03
actual_boost_day_on_D14	0.117	−0.016	0.074	−0.172	0.213	−0.227	−0.079	−0.03	1

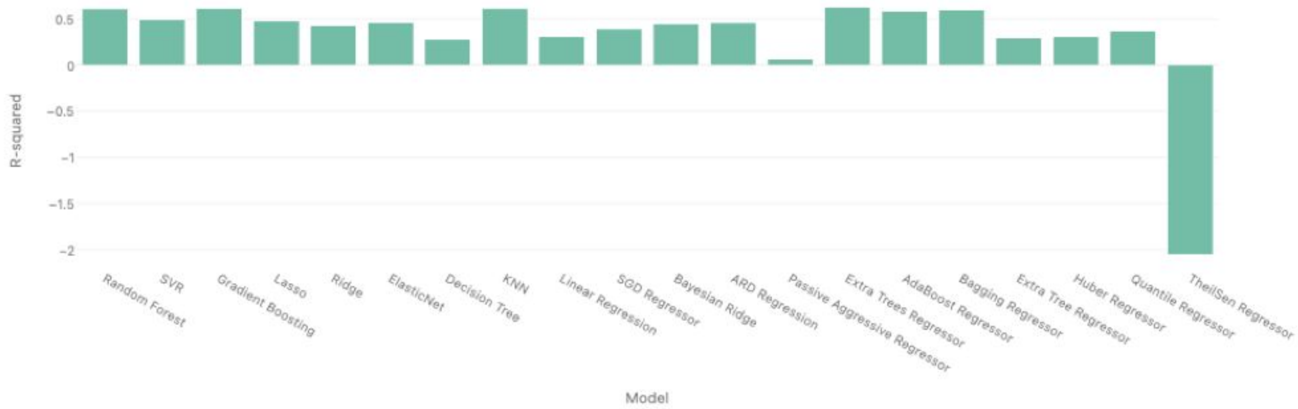
6. Analysis of Results

– Predict IgG antibody titers against pertussis toxin on day 14

R-squared Model Evaluation for 1.1



R-squared Model Evaluation for 1.2



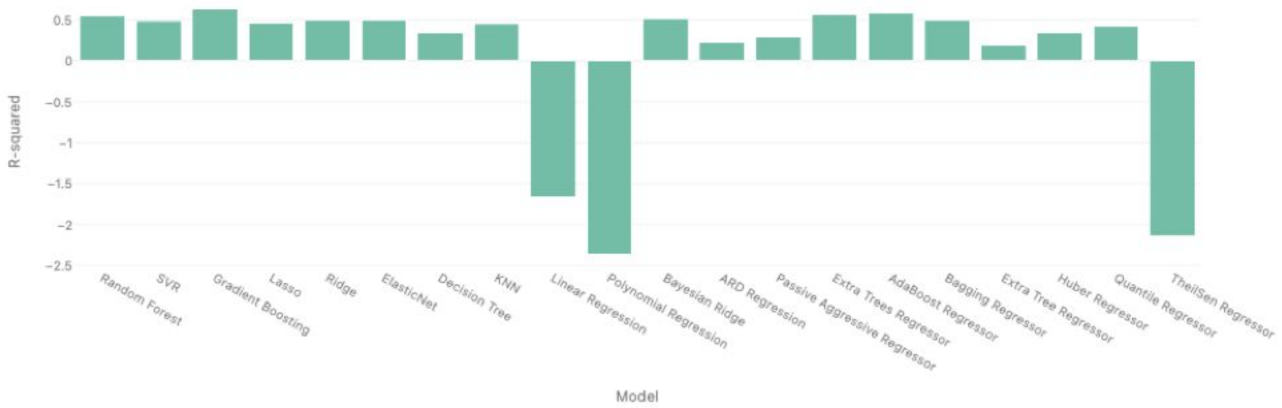
Take Log2 Fold Change
the Target

$$\log_2\left(\frac{MFI\ on\ day\ 14}{MFI\ on\ day\ 0} - 1\right)$$

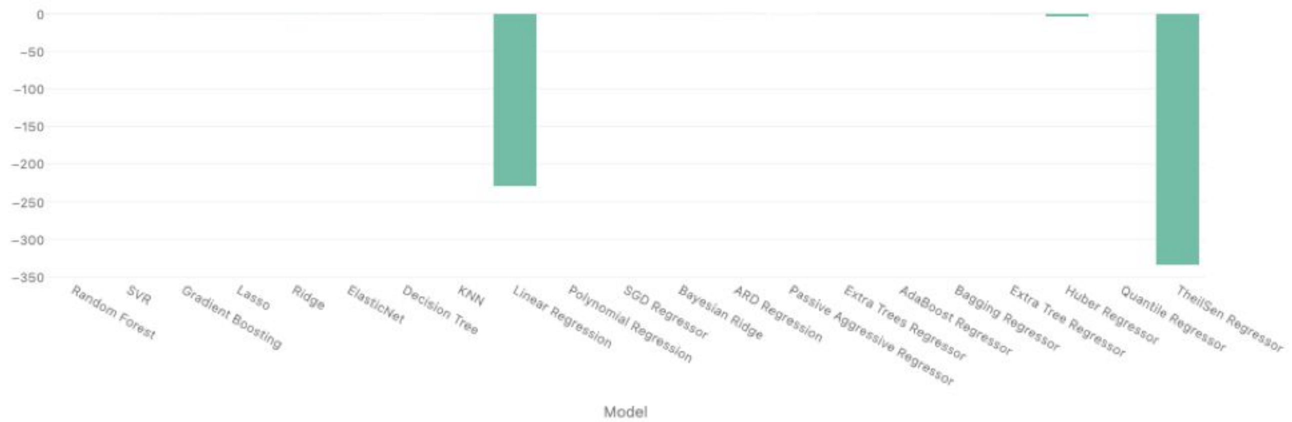
6. Analysis of Results

– Predict frequency of Monocytes on day 1

R-squared Model Evaluation for 2.1



R-squared Model Evaluation for 2.2

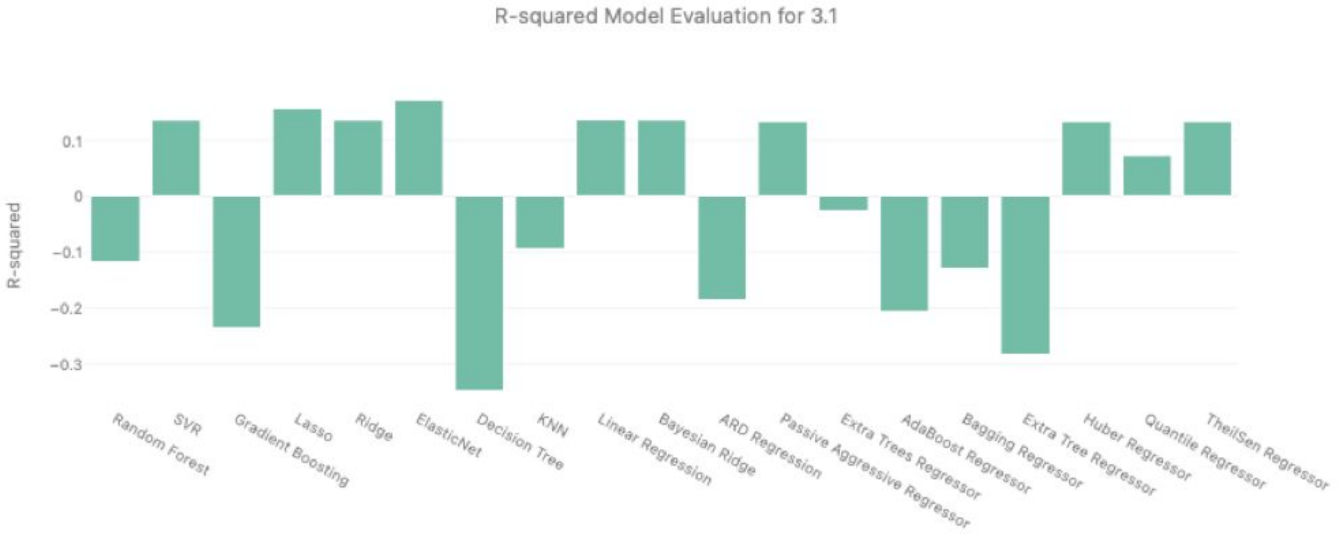


Take Log2 Fold Change
the Target

$$\log_2\left(\frac{PLC\ on\ day\ 1}{PLC\ on\ day\ 0}-1\right)$$

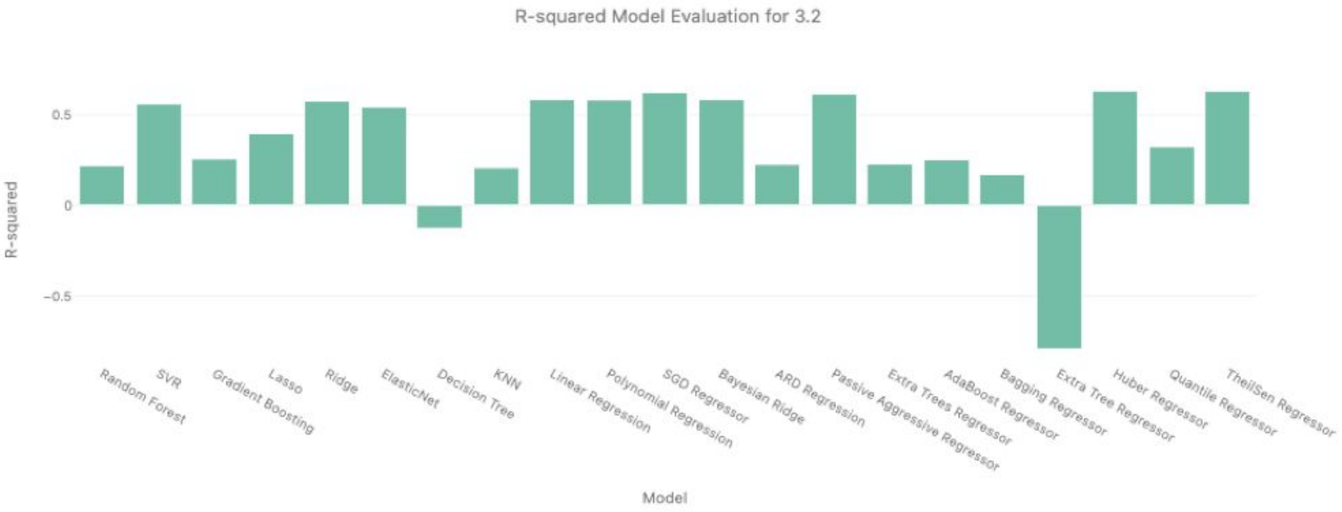
6. Analysis of Results

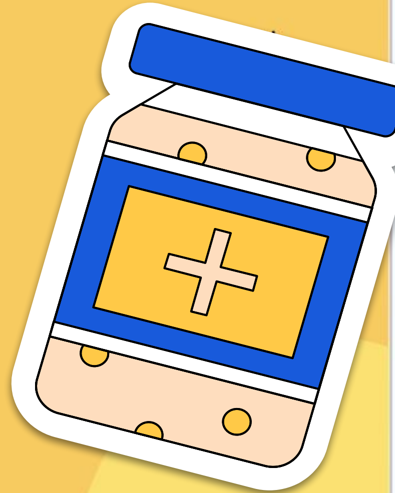
– Predict gene expression of CCL3 on day 3


R-squared Model Evaluation for 3.1

Take Log2 Fold Change
the Target

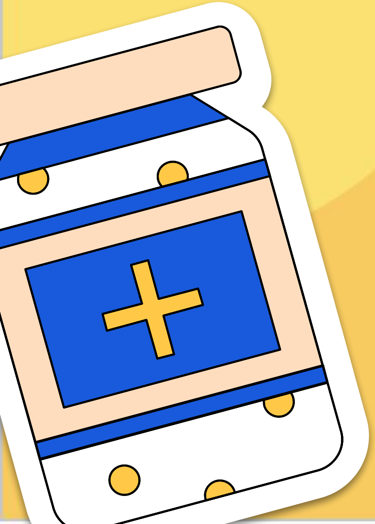
$$\log_2\left(\frac{TPM\ on\ day\ 3}{TPM\ on\ day\ 0} - 1\right)$$


R-squared Model Evaluation for 3.2



07

Insights



7. Insights

- *Support Vector Regression (SVR)*: The model performed well in predicting antibody titer levels. SVR is effective in high-dimensional spaces and can handle nonlinear relationships, making it suitable for handling our complex biological data.
- *Extra Tree Regressor and Gradient Boosting*: These ensemble methods outperformed other methods in predicting the Log2 fold change of antibody levels at day 14 relative to day 0. They were able to reduce overfitting through bagging and boosting techniques, making them a good fit for our dataset.

7. Insights

- *Gradient Boosting for Monocyte Frequency*: Gradient Boosting showed the best performance in predicting monocyte frequency at day 1. The approach of this model helps minimize the prediction error, making it applicable to a variety of tasks.

- *ElasticNet for Gene Expression Levels*: Initially, ElasticNet performed best in predicting CCL3 gene expression levels at day 3. However, simpler models such as Stochastic Gradient Descent(SGD) Regressor and TheilSen Regressor outperformed ElasticNet when predicting the Log2 fold change of the target variable. This shift highlights the importance of model simplicity and regularization in dealing with large amounts of gene expression data.

08

Next Steps Plan

8. Next Steps Plan

Model performance metrics:

- *Regular re-evaluation*: Continuously monitor performance metrics (MSE, MAE, R^2) to ensure the model remains accurate as new data becomes available.
- *Model retraining*: Regularly retrain the model with updated datasets to maintain prediction accuracy and adapt to any changes in data patterns.

Operational metrics:

- *Data quality:* This involves regular data cleansing and validation processes. High-quality data is critical to maintaining the integrity of the model and ensuring reliable predictions.
- *System availability*: Monitor user feedback on dashboard usability for iterative improvements. Usability metrics include user satisfaction scores, system response time, and frequency of use, which can provide insight into the effectiveness of the dashboard and areas for improvement.

Thank you!