Predictive Model for Dengue

Zheng Ting, Yu Sheng, Sunil, Sherena

Data Science Process

01

Define the Problem

02

Obtain the Data

03

Explore the Data

04

Model the Data

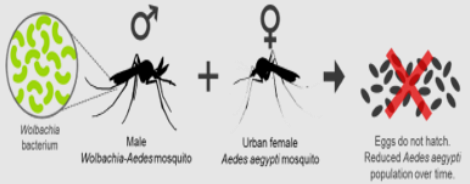
05

Evaluate the Model

06

Answer the Problem

Background



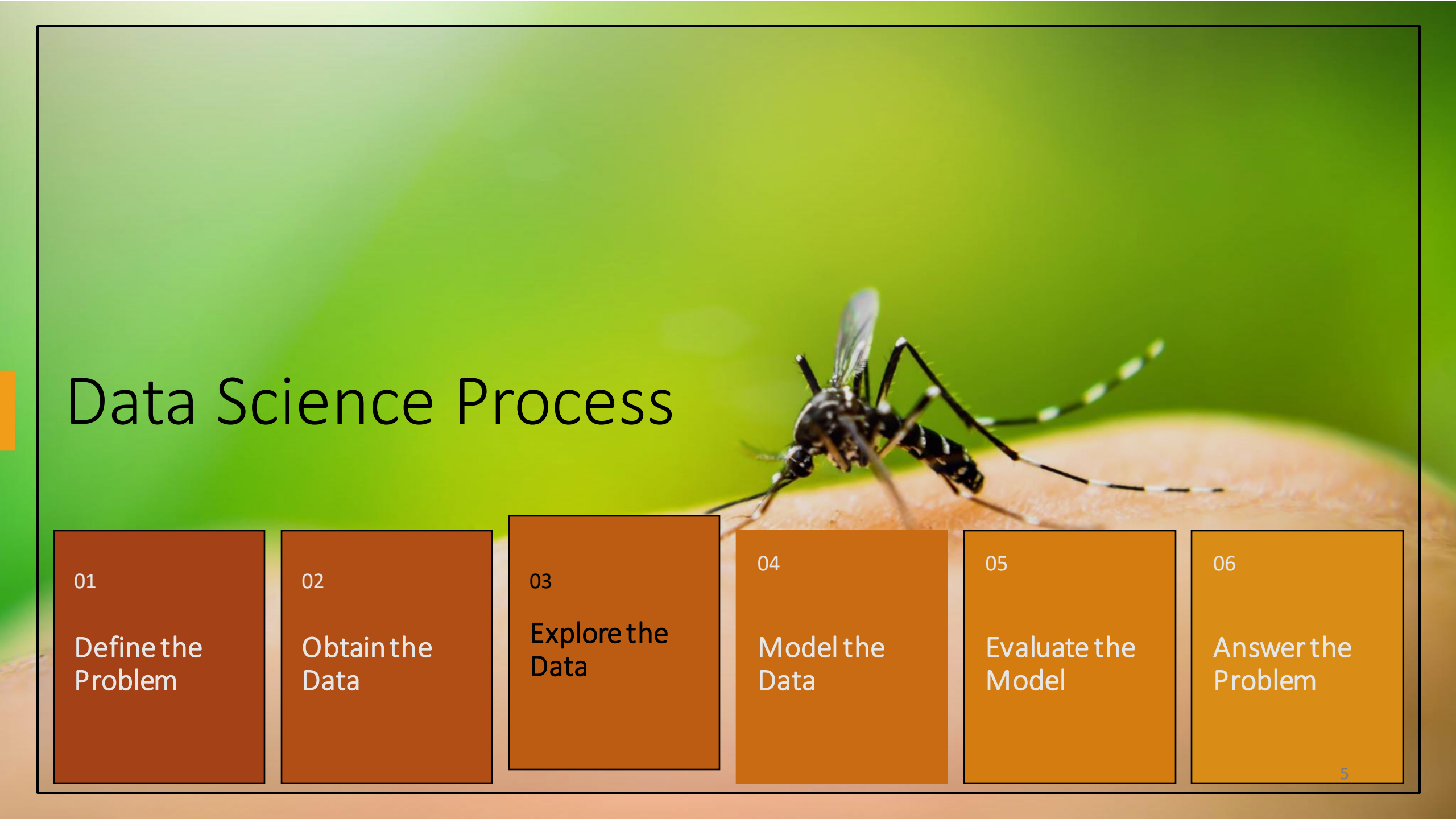
- Estimated **1 in 4** dengue virus infections are symptomatic. **1 in 20** patients with dengue virus disease progress to develop severe, life-threatening disease called severe dengue.
- Dengue virus (DENV) has **4 different serotypes**: DENV-1, DENV-2, DENV-3, DENV-4.
- Since the 1990s, periodic spikes in dengue cases have been occurring in five- to six-year cycles. Several existing theories have been proposed to explain this cyclical pattern:
  - Switch in predominant **virus serotype**
  - **Weather variables** on mosquito activity
  - Low herd immunity due to successful implementation of **Aedes Control Program** in the 1970s
- In 2016, Singapore embarked on a multi-phased field study named Project Wolbachia.

Problem Statement

When and where should the National Environment Agency (NEA) and Ministry of Health (MOH) allocate resources for dengue control more effectively?

Objective: Provide a **16-week ahead forecast of weekly dengue cases** to enable authorities to have sufficient lead time to plan control measures if cases are expected to spike

Data Science Process



01 Define the Problem

02 Obtain the Data

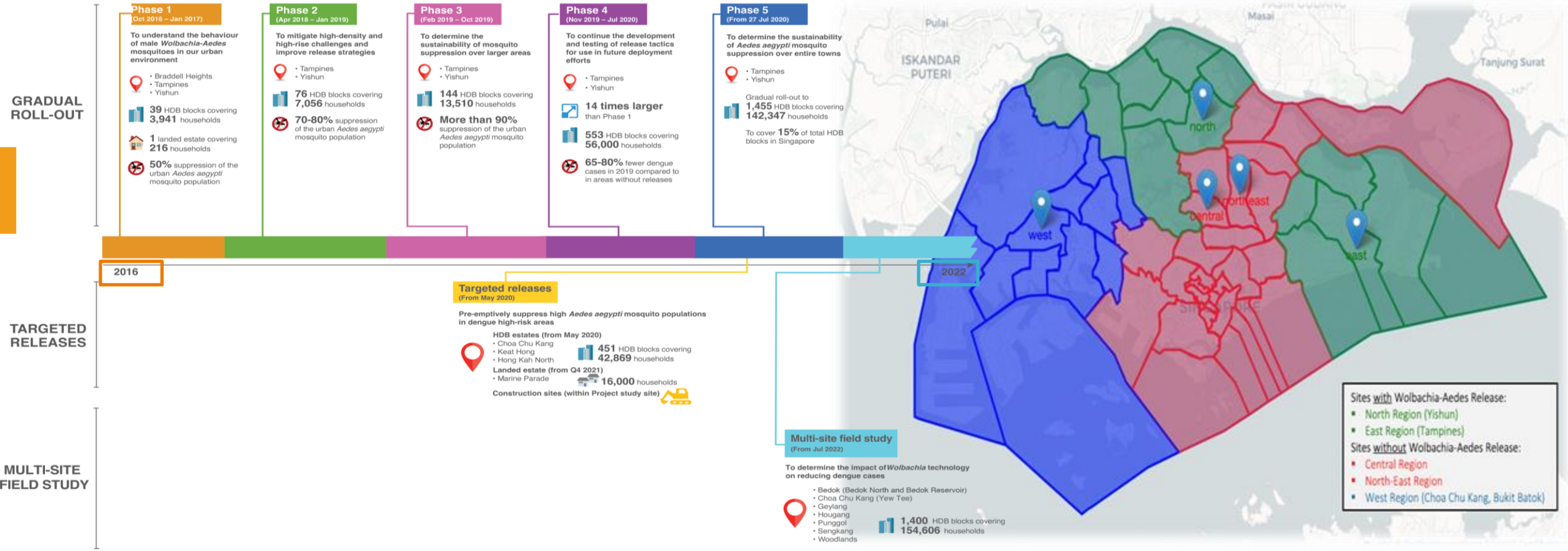
03 Explore the Data

04 Model the Data

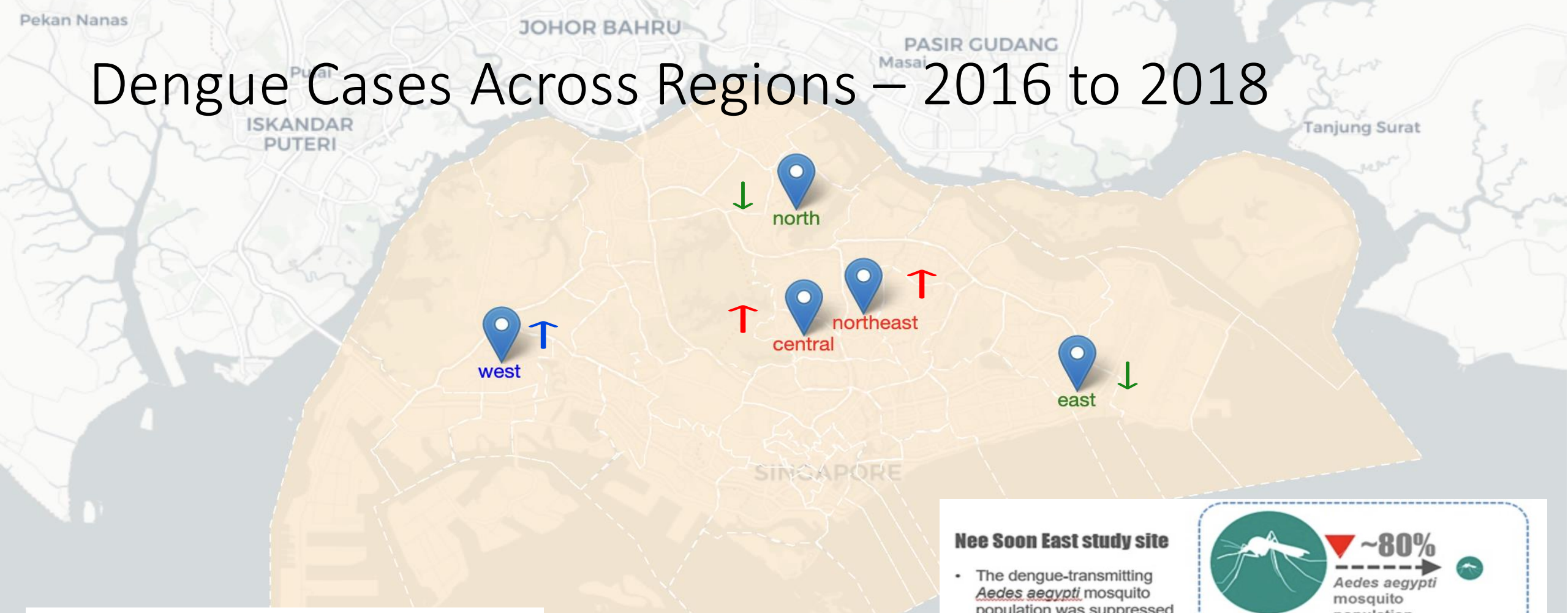
05 Evaluate the Model

06 Answer the Problem

Project Wolbachia Timeline – 2016 to 2020



Dengue Cases Across Regions – 2016 to 2018



Legend

Project *Wolbachia* study site

Low dengue risk — High dengue risk

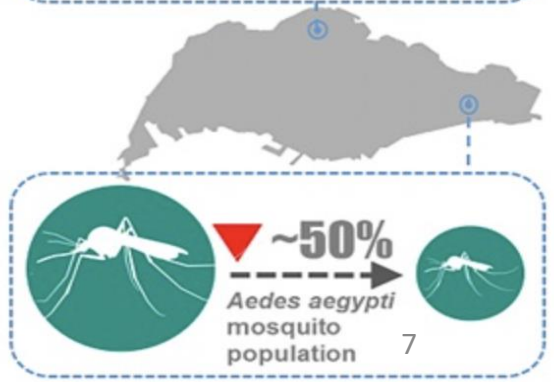
Increasing number of dengue mosquitoes

Nee Soon East study site

- The dengue-transmitting *Aedes aegypti* mosquito population was suppressed by about **80%**

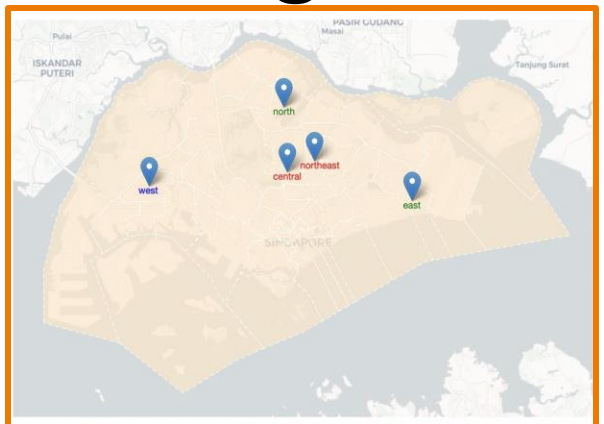
Tampines West study site

- A suppression of about 50% was observed
  - The reduced effectiveness could be due to the shorter duration of the release period, multiple mosquito breeding habitats found and other environmental factors.

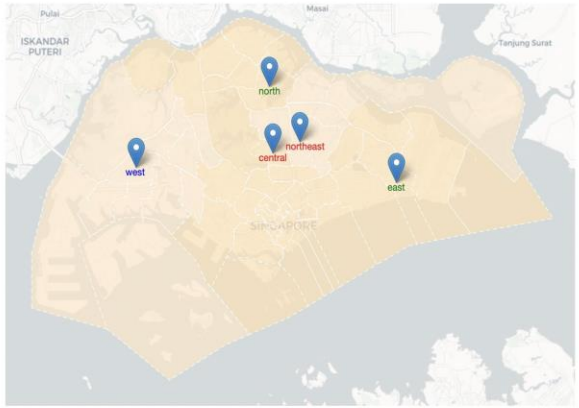


Dengue Cases Across Regions – 2019 to 2020

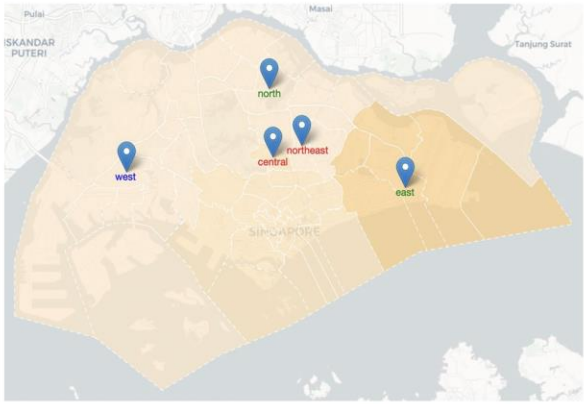
Phase 3 (Feb 2019 – Oct 2019)



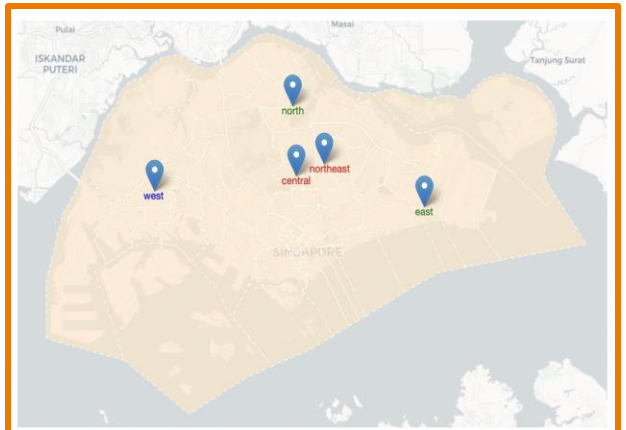
Feb 08, 2020 – Release of Wolbachia Aedes Mosquitoes in the North (Yishun) and East (Tampines) Regions



Jun 09, 2019 – Predicted Spikes in Field Study Sites (Yishun/North Region and Tampines/East Region)



Jul 07, 2019 – Situation in the North (Yishun) under control; Cases in the East (Tampines) remain elevated as anticipated



Oct 06, 2019 – Situation in the North (Yishun) and East (Tampines) Regions under control

Start of Phase 3 (Feb 2019)
Week 1

Middle of Phase 3 (Jun/Jul 2019)
Weeks 16 – 20

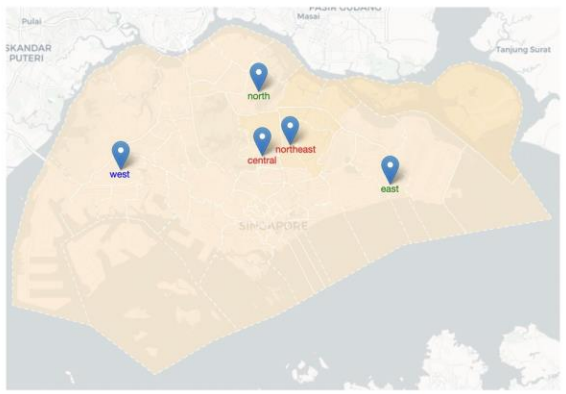
End of Phase 3 (Oct 2019)
Week 32

Week 1
Start of Phase 4 (Nov 2019)

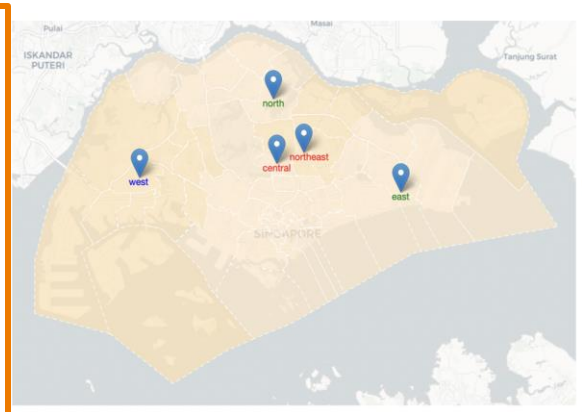
Weeks 16 – 20
Middle of Phase 4 (Mar/Apr 2020)

Week 32
End of Phase 4 (Jul 2020)

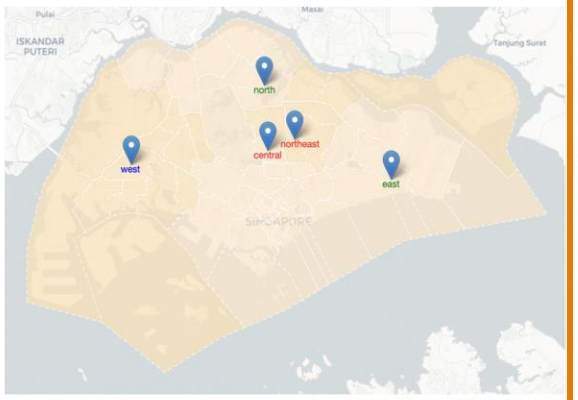
Phase 4 (Nov 2019 – Jul 2019)



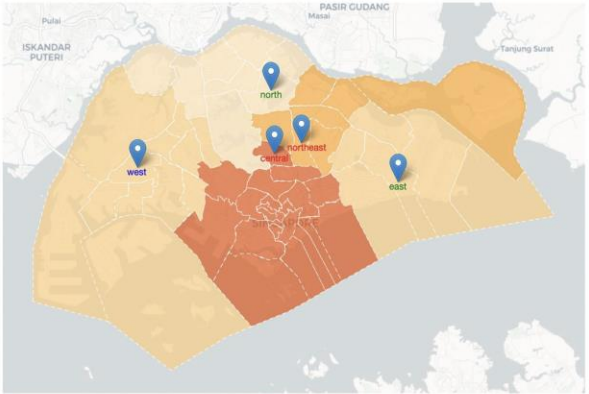
Nov 03, 2019 – Release of more Wolbachia Aedes mosquitoes in the North (Yishun) and East (Tampines) Regions



Mar 08, 2020 – No predicted spikes in Field Study sites (Yishun/North and Tampines/East Regions) due to timely release of Wolbachia Aedes mosquitos



Apr 05, 2020 – Dengue cases in the North (Yishun) and East (Tampines) stay low due to preventive measure

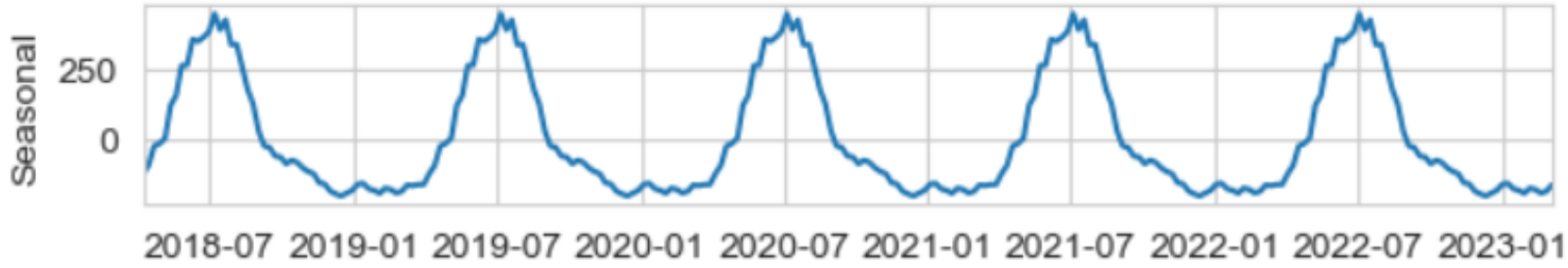


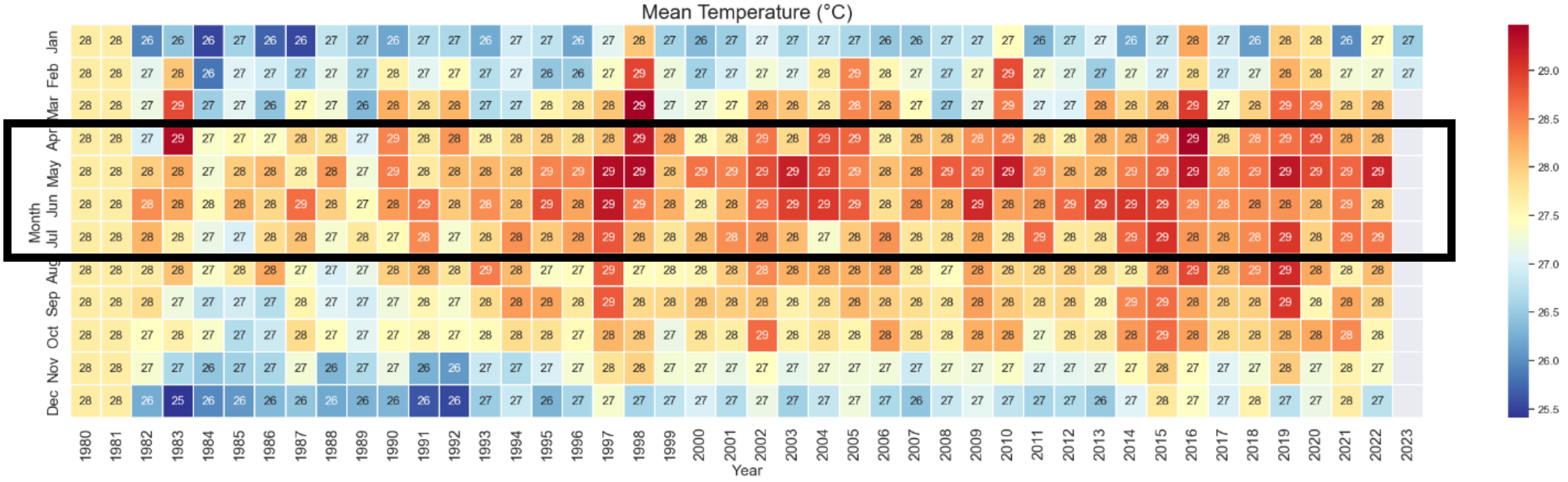
Jul 05, 2020 – Dengue cases in the Central, Northeast, and West (to a lesser extent due to targeted releases from May 2020) spikes due to absence of preventive measures

Seasonality and Weather Features

Aedes Mosquito has Life-span: 42-56 days

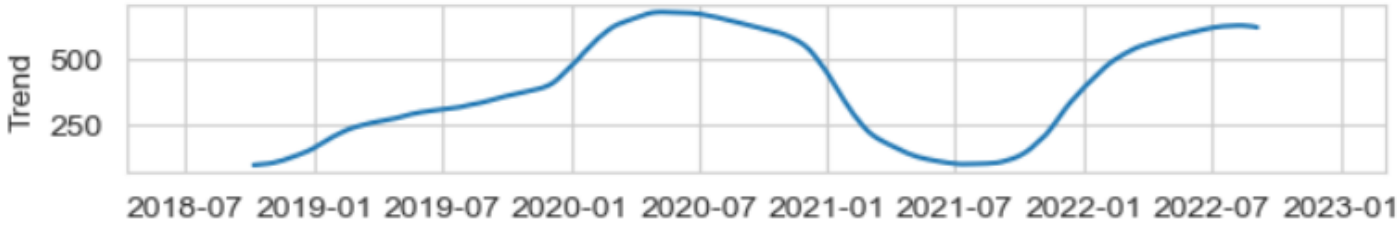
Hottest Months: April to July

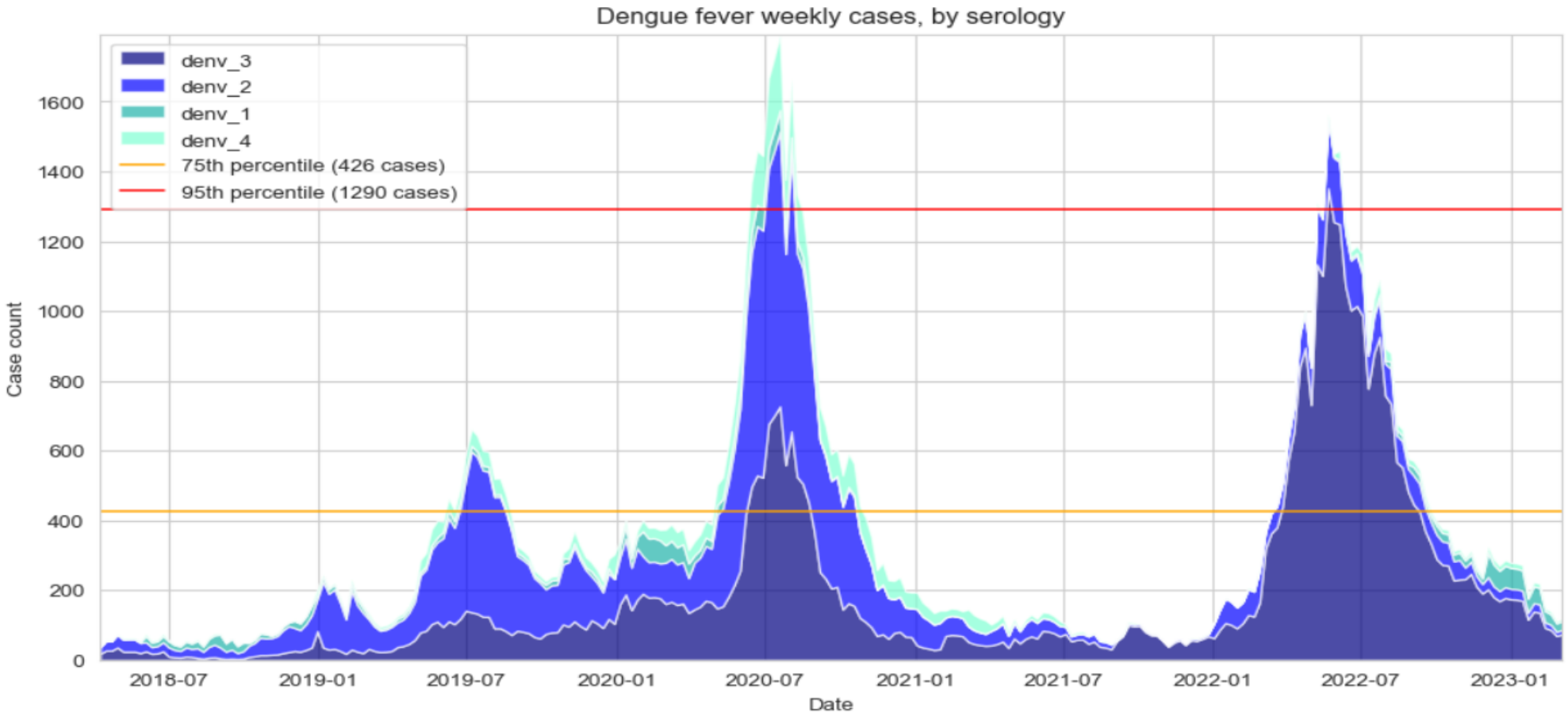




Trends and Serology

DENV 3 is the dominant strain from 2020 onwards

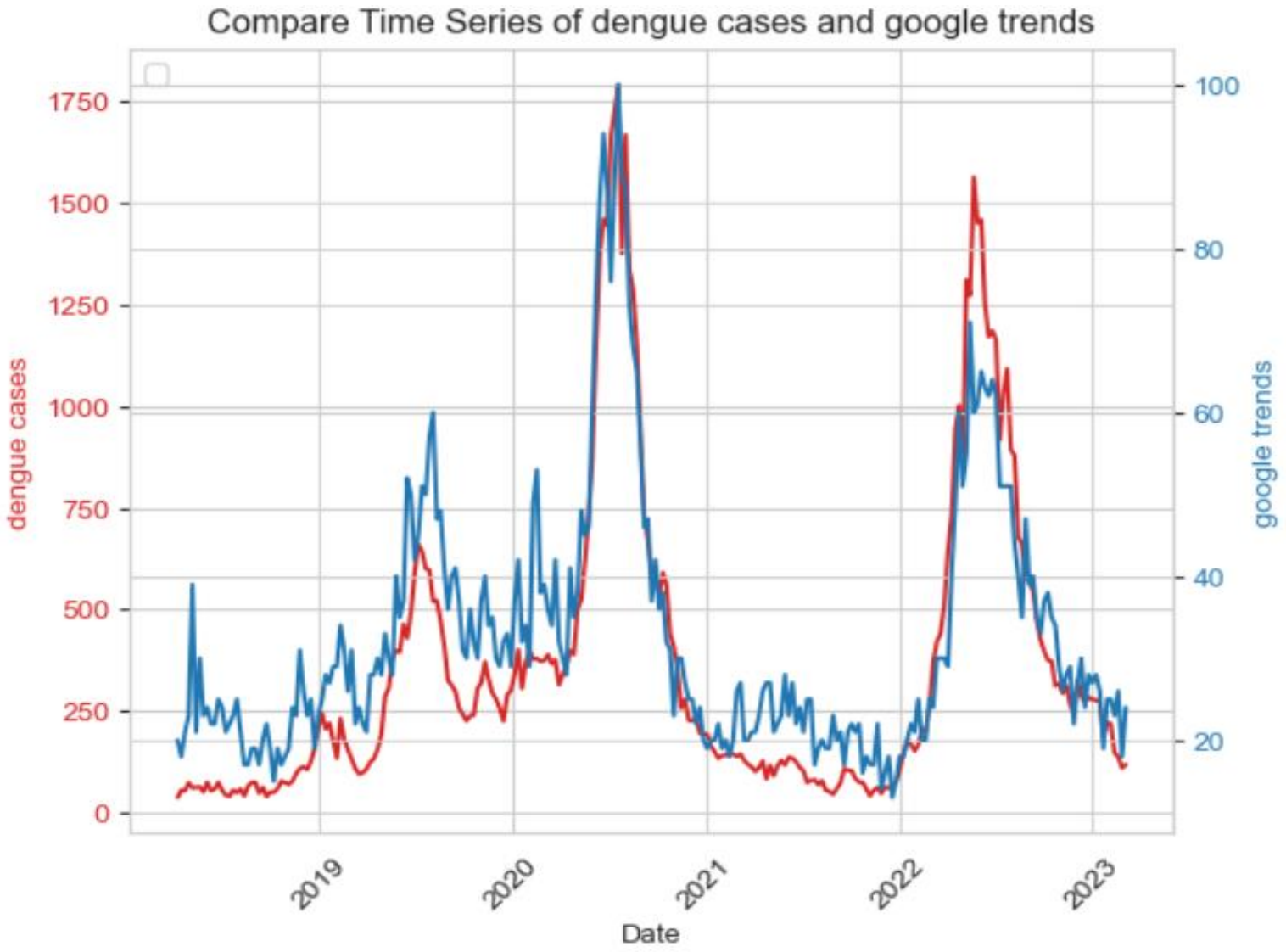




Google Searches

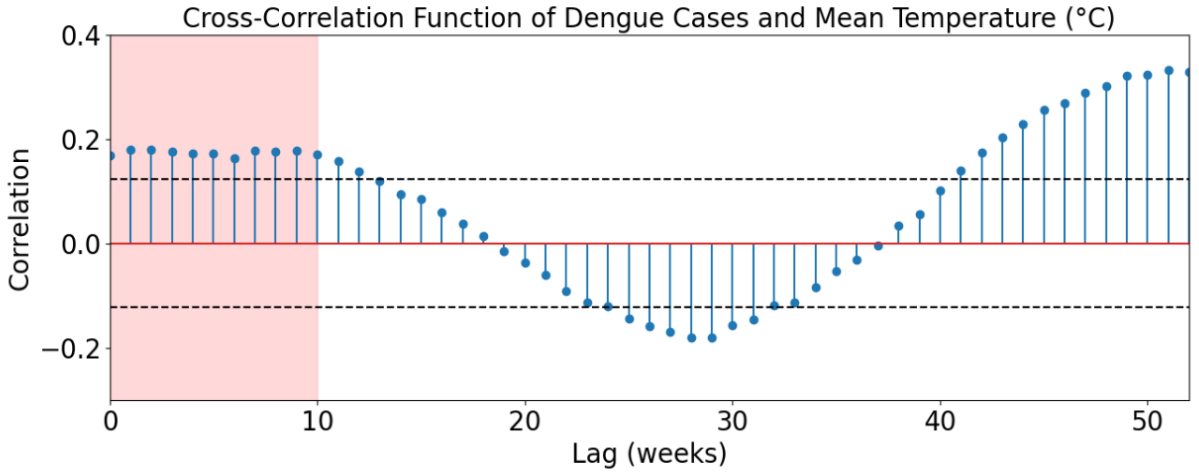
36 common terms, filtered to 7 keywords:

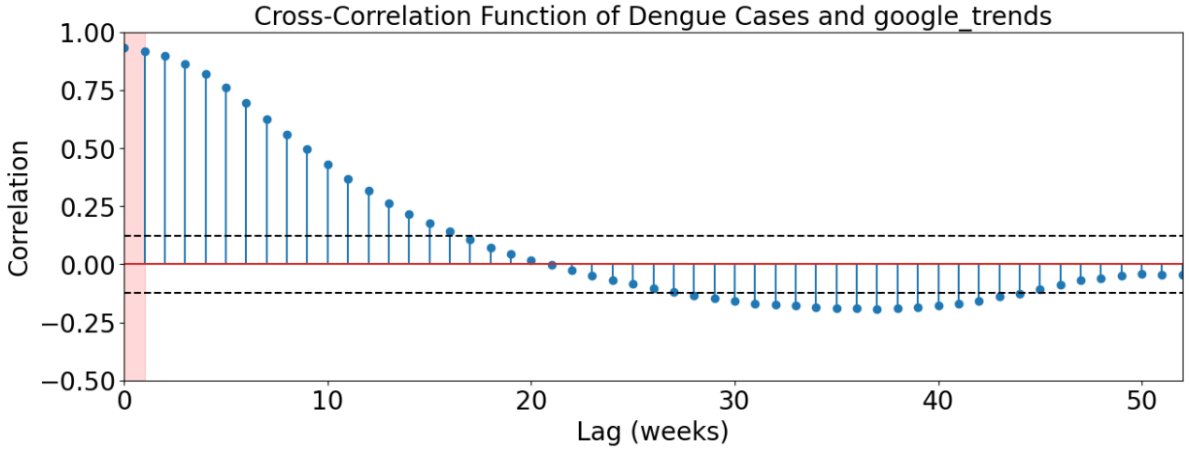
- Repellent
- Dengue Fever
- Dengue Cluster
- Mosquito
- Mosquito Repellent
- Dengue
- Aedes

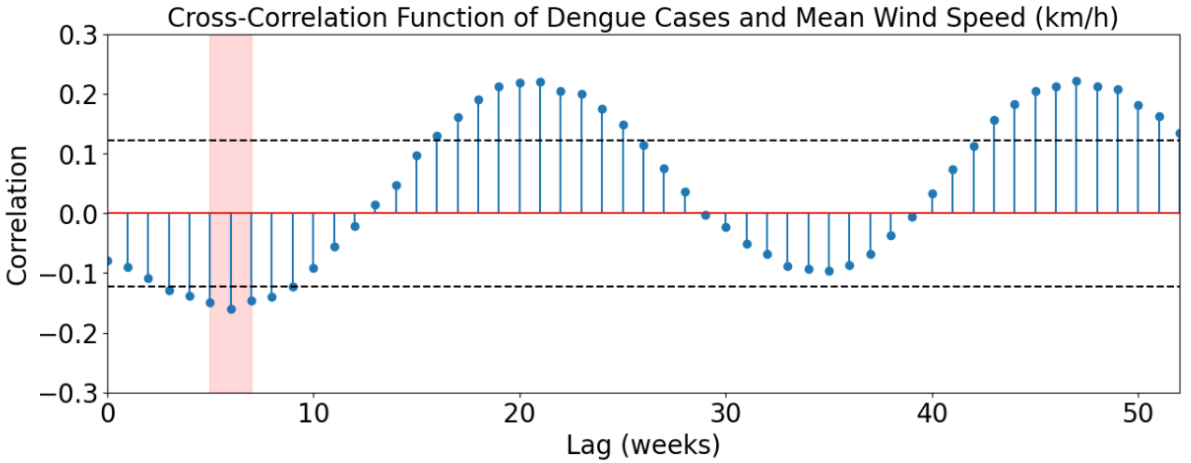


Features of Importance

- **0-1** week lags for google trends
- **0-10** weeks lag for mean temperature
- **5-7** weeks lag for mean wind speed







Data Science Process

01 Define the Problem

02 Obtain the Data

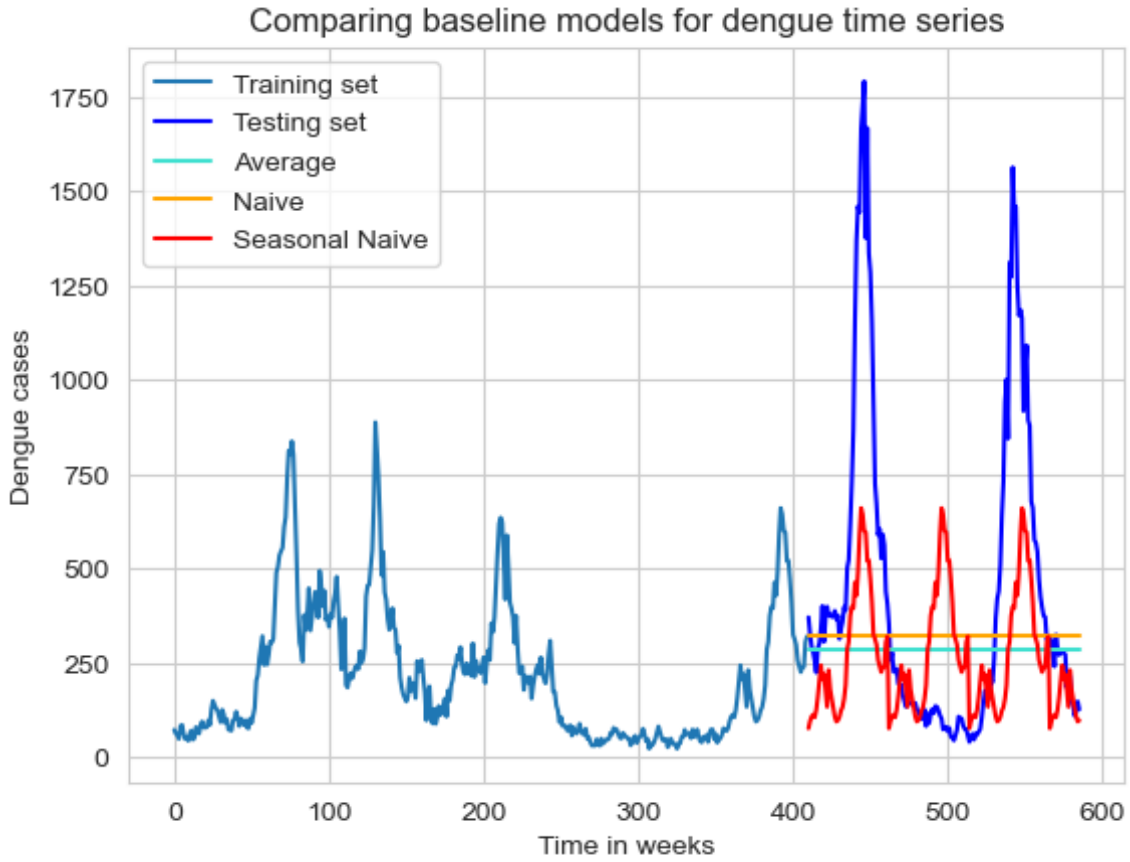
03 Explore the Data

04 Model the Data

05 Evaluate the Model

06 Answer the Problem

Baseline Model



	Baseline Model	Description	Test RMSE	Test MAPE
1	Average	Overall historical mean	449	0.99
2	Naive	Naive forecast based on last observation	438	1.13
3	Seasonal Naive	Replicate last seasonal cycle	397	1.01

Pipeline Steps

1

Step 1

Data Transformation and Feature Engineering

2

Step 2

Model Selection

1-week forecasting model

3

Step 3

Model Selection & Tuning

16-week forecasting model

4

Step 4

Model Evaluation

Model Selection – 1-week forecasting

- Performance metric: **Root Mean Squared Error (RMSE)**
  - i.e. Typical difference observed between predicted and actual number of dengue cases

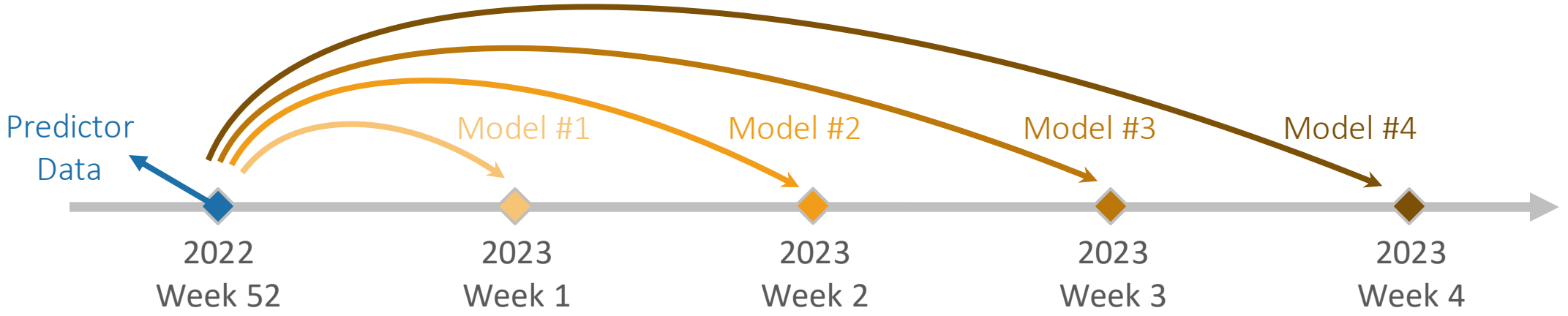
- Start off with 1-week ahead forecast to understand performance of model families[1]

Model Family	Model	RMSE
Time Series	ARIMA (3,1,0)	617
	ARIMAX (1,1,2)	262
	SARIMA (1,1,2)(1,1,0,51)	359
	SARIMAX (1,1,2)(1,1,0,48)	102
Boosting	Gradient Boosting	176
Decision Trees	Decision Tree and Bagging	207
	Random Forest	192
Support Vector	Support Vector Machine	141

Note 1: Models trained on 2018~2021 data and validated on 2022 data, 2023 data is used as test set at the end for verification

Model Selection – 16-week forecasting

- To forecast 16 weeks ahead from the current time, we use current data and fit one dedicated model for each week's forecast



- For each model, best performing model is chosen from the 3 model families

Data Science Process

01 Define the Problem

02 Obtain the Data

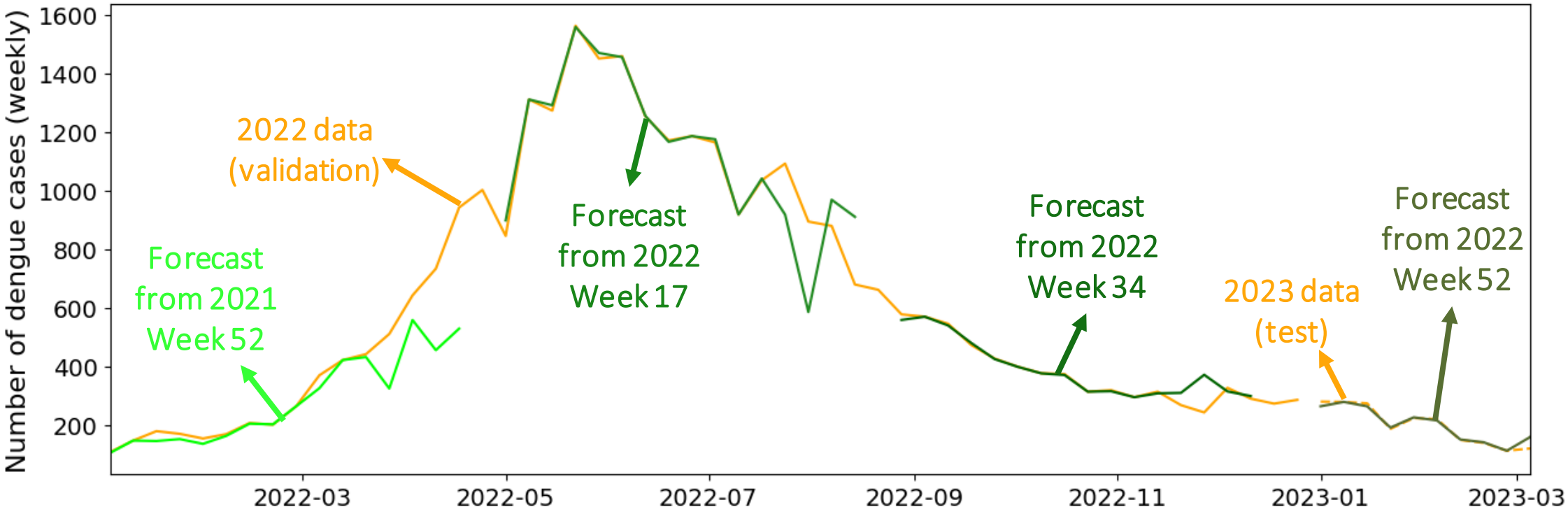
03 Explore the Data

04 Model the Data

05 Evaluate the Model

06 Answer the Problem

Model Evaluation



Dataset	RMSE
2022 (validation)	103
2023 (test)	13[1]

Note 1: Test dataset only has first 10 weeks of 2023 data, hence this RMSE is based on the first 10 weeks of 2023 data only

Data Science Process

01 Define the Problem

02 Obtain the Data

03 Explore the Data

04 Model the Data

05 Evaluate the Model

06 Answer the Problem

Existing Measures

Reactive Measures	
Distribute mosquito repellent	
Fogging	

Preventive Measures	
Home Inspections	
Anti-Dengue Campaign	
Gravi-traps	
Project Wolbachia	PROJECT WOLBACHIA SINGAPORE

Cost-Benefit Analysis

Economic Impact of Dengue	Project Wolbachia
Economic impact of over $1 Bil per Annum between 2010-2020[1]	$108 Mil per Annum for Nation-Wide Deployment[1]
	$0.40 per Mosquito[2]
Higher impact expected in 2020 and 2022 due to huge spike in cases	Up to 88% reduction in dengue cases[2]
	3-4 months to suppress mosquito population[3]

Expected Savings from Nation-Wide Deployment of Project Wolbachia

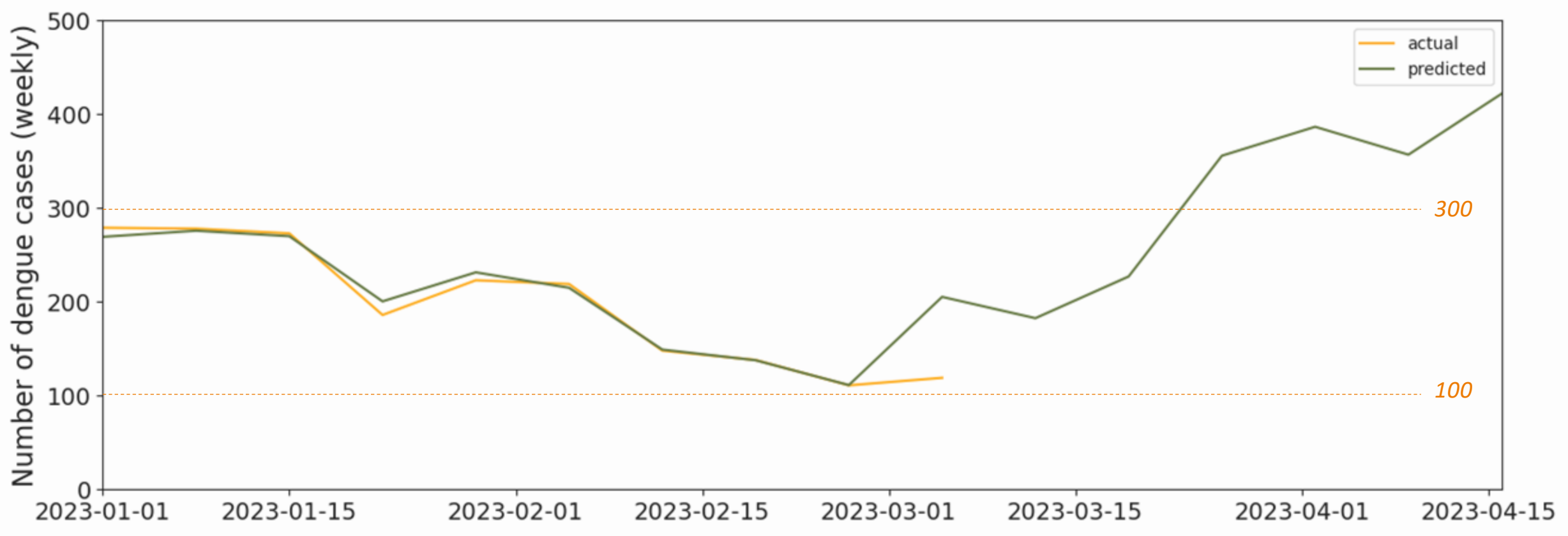
Approx. 770 Mil

[1] https://journals.plos.org/globalpublichealth/article?id=10.1371/journal.pgph.0000024

[2] https://www.straitstimes.com/singapore/health/about-200m-wolbachia-aedes-mosquitoes-released-from-mosquito-factory-nea

[3] https://www.nea.gov.sg/corporate-functions/resources/research/wolbachia-aedes-mosquito-suppression-strategy/frequently-asked-questions

Cost-Benefit Analysis



Conclusion & Recommendations

Dengue Prediction	Project Wolbachia
Model predicts dengue cases with a RMSE of 91 and identifies seasonality and trends well	Optimal window to release the Wolbachia-Aedes mosquitoes • 16-weeks before predicted spike
Serves as an early detection tool to engage town councils • Minor spikes: use existing measures • **Major spikes: deploy Wolbachia mosquitos**	To adopt at national-level, achieving savings of over $700 Mil
	Keeps dengue cases throughout the nation low

Limitations

Increase Data Collection	Domain Expert
Town-level dengue cases	Greater expertise in feature-selection
Age-group of populations in town	Greater expertise in feature-engineering
More historical data	Provide an edge by utilizing in-depth knowledge of mosquito life-cycle, habitats and breeding habits


Q&A

Feature Importance of Different Forecast Models

- Different week's model focuses on different predictors
  - E.g., near future

- Start off with 1-week ahead forecast to understand performance of model families[1]

Note 1: Models trained on 2018~2021 data and validated on 2022 data, 2023 data is used as test set at the end for verification