

## Motivation

- **Dengue** is a mosquito-borne virus that causes 390 million infections per year [1], straining public health and the economy.
- Early detection can reduce these costs, but **predictive systems** for environmentally-driven disease have been underdeveloped.
- To address this need, we develop a **scalable workflow** that integrates large, diverse data sources to predict dengue in real-time at various spatiotemporal resolutions of Brazil.
- We fit **thousands of time series models in parallel** to help inform public health policy and mitigate impact of future outbreaks.

## Dengue in Brazil

We consider predictors of dengue from the following data sources:

### Predictors of Dengue

Google Searches Satellite/Weather Demographic



- Statistical modeling of spatially distributed phenomena relies on careful selection of the spatial resolution (scale).
- From a public health perspective, allocation of resources depends on the spatial scale.
- Hence, we model the spread of dengue at various scales, e.g.,

### Spatial Scales of Brazil

Region State Municipality



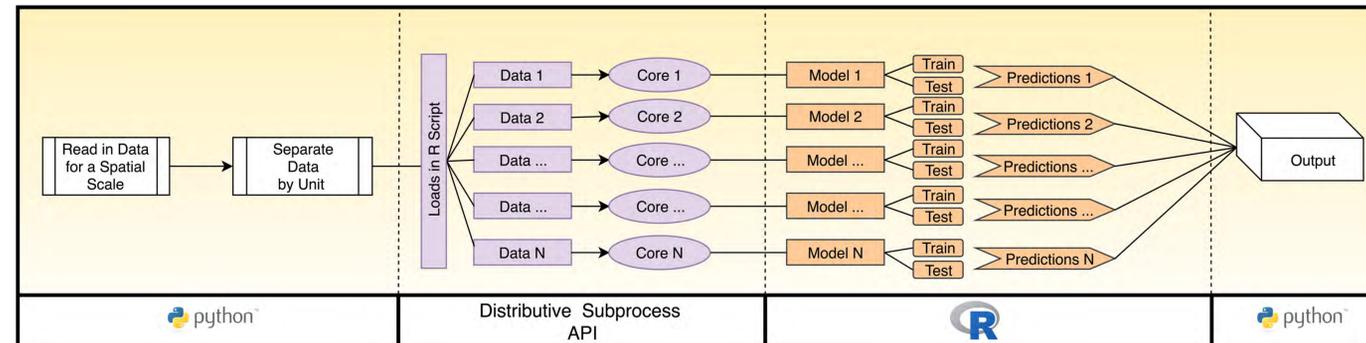
The large number of independent models and need for real-time forecasting motivate the use of parallelism. For example, **27 states × 344 weeks × 1097 predictors = 10,188,936 elements**

We develop a **multi-language parallel workflow** that combines

- the sophisticated statistical methods of R
- the parallel capabilities of Python

## Multi-Language Parallel Workflow

We adapt the algorithmic structure of a program to optimize runtime compilation. We create a self-contained API that automates the transition of **Single Instruction Multiple Data (SIMD)** programs to **Multiple Instruction Multiple Data (MIMD)** modeling, improving architecture utilization.



## Statistical Models

For each unit of space, time series variables are imputed with Kalman smoothing (to estimate missing values) [2], then lagged exhaustively. The training set and testing set are defined for years 2010-2014 and 2015-2016, respectively. Assume  $Y \in \mathbb{R}$  to be the dengue count,  $X \in \mathbb{R}^p$  the predictor vector,  $(x_i, y_i)$  the observation pairs in time, and  $E(Y|X = x) = \beta_0 + x^T \beta$ .

### Negative Binomial GLM

$Y \sim \text{NegBinom}(\mu, \mu + \frac{\mu^2}{\theta})$ , so that  $Y$  has the probability density function

$$f(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(\theta) y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{y+\theta}} \quad (1)$$

If  $\theta$  is known, this represents a **generalized linear model (GLM)** [5].

### Regularized Regression

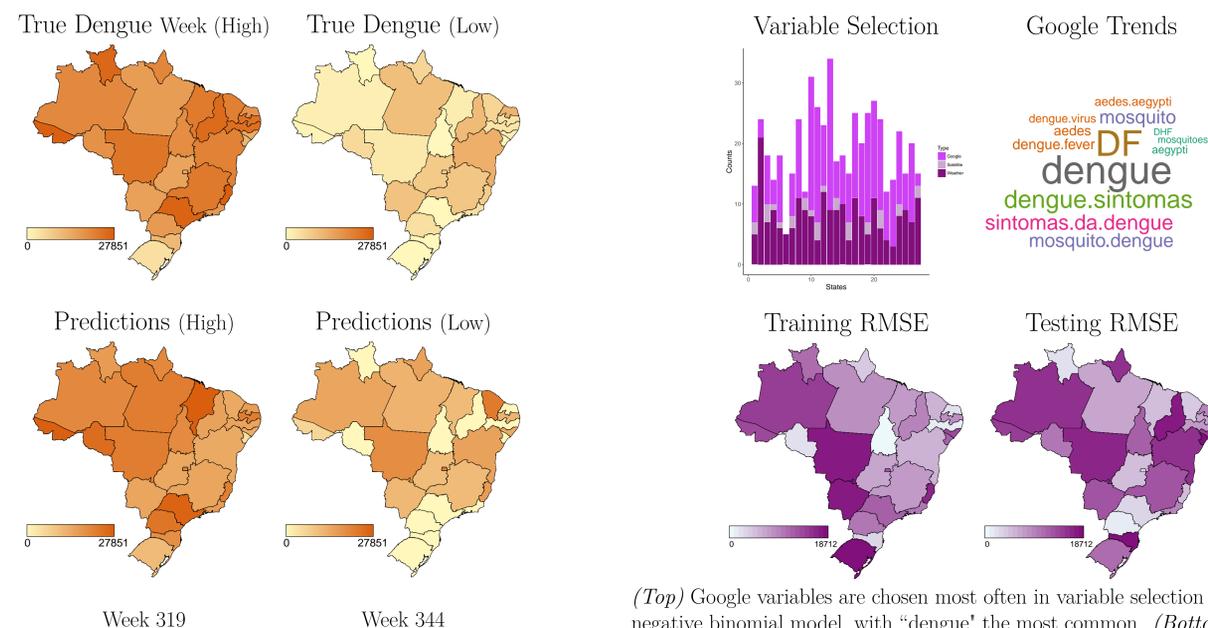
For  $\alpha \in [0, 1]$  and hyperparameter  $\lambda$ , **lasso**, **ridge**, and **elastic net** regression solve the following problem [3], [4]:

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta), \quad (2)$$

where

$$P_\alpha(\beta) = (1 - \alpha) \frac{1}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1. \quad (3)$$

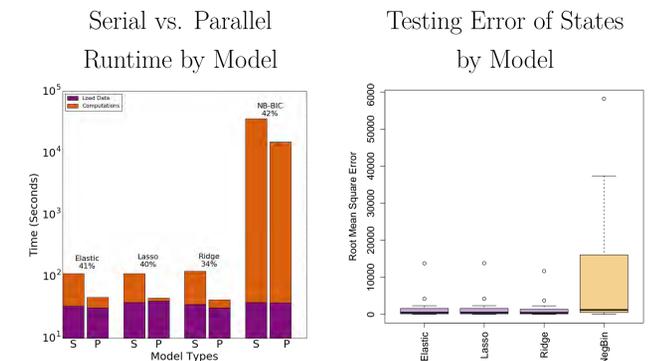
## Results at State Level



Heatmaps of actual versus predicted dengue indicate goodness of fit for the elastic net model with  $\alpha = 0.5$ .

(Top) Google variables are chosen most often in variable selection for the negative binomial model, with "dengue" the most common. (Bottom) Training and testing root mean square error (RMSE) by state for the elastic net model. **Total dengue cases per state for this period range from 9,814 to 1,734,669.**

## Model Comparison



(Left) Parallelism decreases runtime up to 42%. Data loading represents a nontrivial portion of runtime. (Right) The regularized regression models outperform the negative binomial model. We omit three outliers with high RMSE.

## Discussion

We achieve **high predictive accuracy** for the states, given the severe underreporting of dengue [1] and magnitude of incidence.

Error **varies between states**, reinforcing the

- heterogeneity of data in space and time
- need for multiple spatial resolutions
- importance of independent models for each spatial unit

We **reduce the runtime by up to 42%**, verifying the power of **parallelism** for this problem.

In the future, we may

- increase speedup by parallelizing variable selection
- repeat the analysis at all spatial levels of Brazil and compare levels

## Acknowledgments

- We would like to thank our mentors Carrie Manore and Geoffrey Fairchild and the Parallel Computing co-leads Hai Ah Nam, Bob Robey, Kris Garrett, and Joseph Schoonover.
- Computations conducted on Darwin cluster at Los Alamos National Laboratory. Support for this work was provided by U.S. Department of Energy at Los Alamos National Laboratory supported by Contract No. DE-AC52-06NA25396

## References

- [1] Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., ... & Myers, M. F. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504.
- [2] Moritz, S., & Bartz-Bielstein, T. (2017). imputeTS: time series missing value imputation in R. *The R Journal*, 9(1), 207-218.
- [3] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1.
- [4] Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).
- [5] Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R package version*, 1, 29.