Big Data for Reproductive Health (bd4rh)

bd4rh
Big Data *for* Reproductive Health

Data+

Students: Dennis Harrsch, Elizabeth Loschiavo, Zhixue (Mary) Wang
Mentors: Amy Finnegan, PhD[1, 2, 3]; Megan Huchko, MD, MPH[1, 2, 4]; Kelly Hunter[1, 2, 5]
Affiliation: [1]Duke Global Health Institute, [2]Duke Center for Global Reproductive Health, [3]Evidence Lab; [4]Duke OB/GYN, [5]Sanford School of Public Policy

1 in 3 women in low-income countries **quit using** modern methods of **contraception** within the first year of use. Our team scaled existing visualization tools and used machine learning to change that.

Scalability

START

Originally **only three countries,** each with one survey, were available per visualization. **Access times** were upwards of 30 seconds, potentially inhibiting use for stakeholders.

To improve latency, we **cleaned the data,** removing unnecessary variables and regrouping, then organized clean files by country and information available.

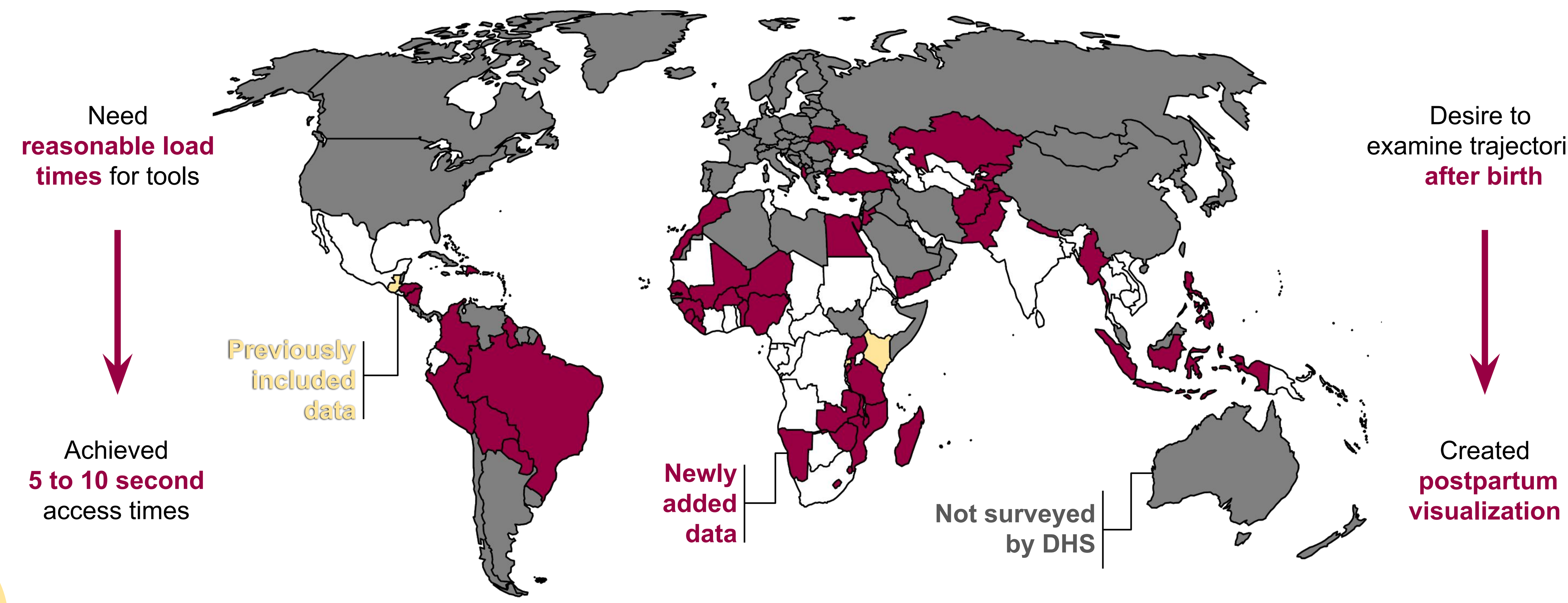
Next, we **rewrote existing code** to ensure that only the necessary data would be downloaded. We also added new functionalities to better visualize specific patterns.

Finally, we **shifted hosting** from an R Shiny-owned server to a temporary, local Duke-hosted virtual machine, further reducing latency and dependencies.

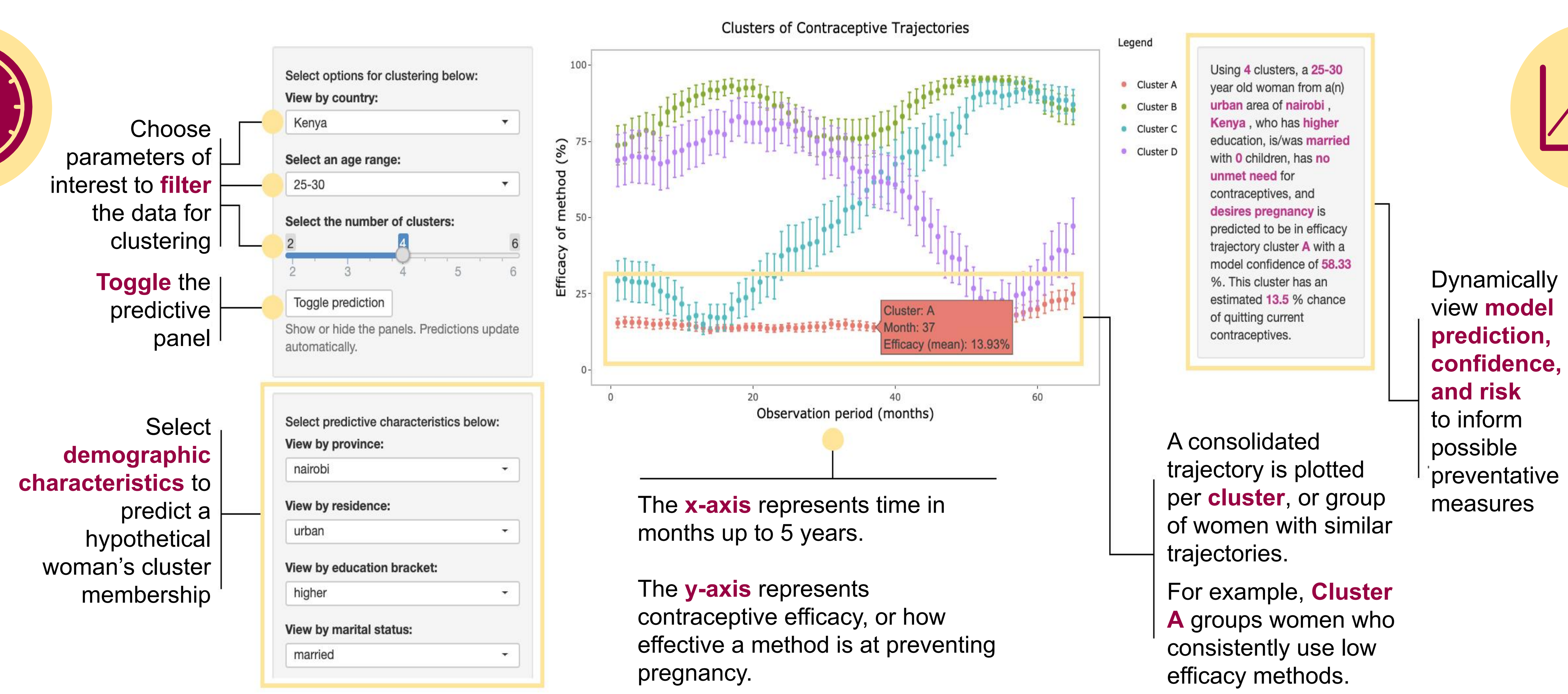
Moving Forward

We want to **continue scaling** our apps by adding more countries and surveys and further reducing access times. We will incorporate **user feedback** to ensure our apps are useful for stakeholders.

SCALABILITY: We incorporated data from **55** countries, **135** surveys, and over **1,800,000** women, while improving latency and adding new functionalities.



MACHINE LEARNING: We created a tool for users to dynamically apply predictive modeling to women's contraceptive efficacy trajectories.



Machine Learning

START

Starting with **static** clusters on around 600 thirty-year-old women in Kenya, the results showed intuitive high, high-low, low-high, or low **efficacy trajectory** clusters.

To make this accessible, we created a public app to let users **interactively cluster** data of interest. We added visualizations of **sequences** and **cluster demographics** to aid user data exploration.

Next, we ran supervised machine learning, using k-nearest neighbors and logistic regression, to **predict dynamic cluster membership** based on demographic indicators.

We found class imbalance with **deteriorating accuracy** as the number of classes increased, e.g. **82%** accurate on 2 clusters and **50%** accurate on 4, even with oversampling.

Moving Forward

We want to **improve the model accuracy** via preprocessing techniques and other supervised learning algorithms. We also want to run usability studies to determine utility for stakeholders.

Duke | the center for GLOBAL REPRODUCTIVE HEALTH | DUKE GLOBAL HEALTH INSTITUTE | Evidence Lab at the Duke Global Health Institute | Duke UNIVERSITY | BASS CONNECTIONS

Acknowledgements: We would like to extend our deepest thanks to our team leads (Amy Finnegan, Megan Huchko and Kelly Hunter) for their mentorship this summer. We would also like to thank our colleagues at RTI International, FHI 360, DHS, Avenir Health, and Track20 for their ongoing feedback and support of this project. We thank the Duke Data+ and Bass Connections programs, the Duke Global Health Institute, the Duke Evidence Lab, and the Duke Center for Global Reproductive Health.