

# Hunting Zika Virus using Machine Learning

Subho Majumdar, IBM Social Good program

In collaboration with Cary Institute of Ecosystem Studies

08/25/2016



#Data4Good



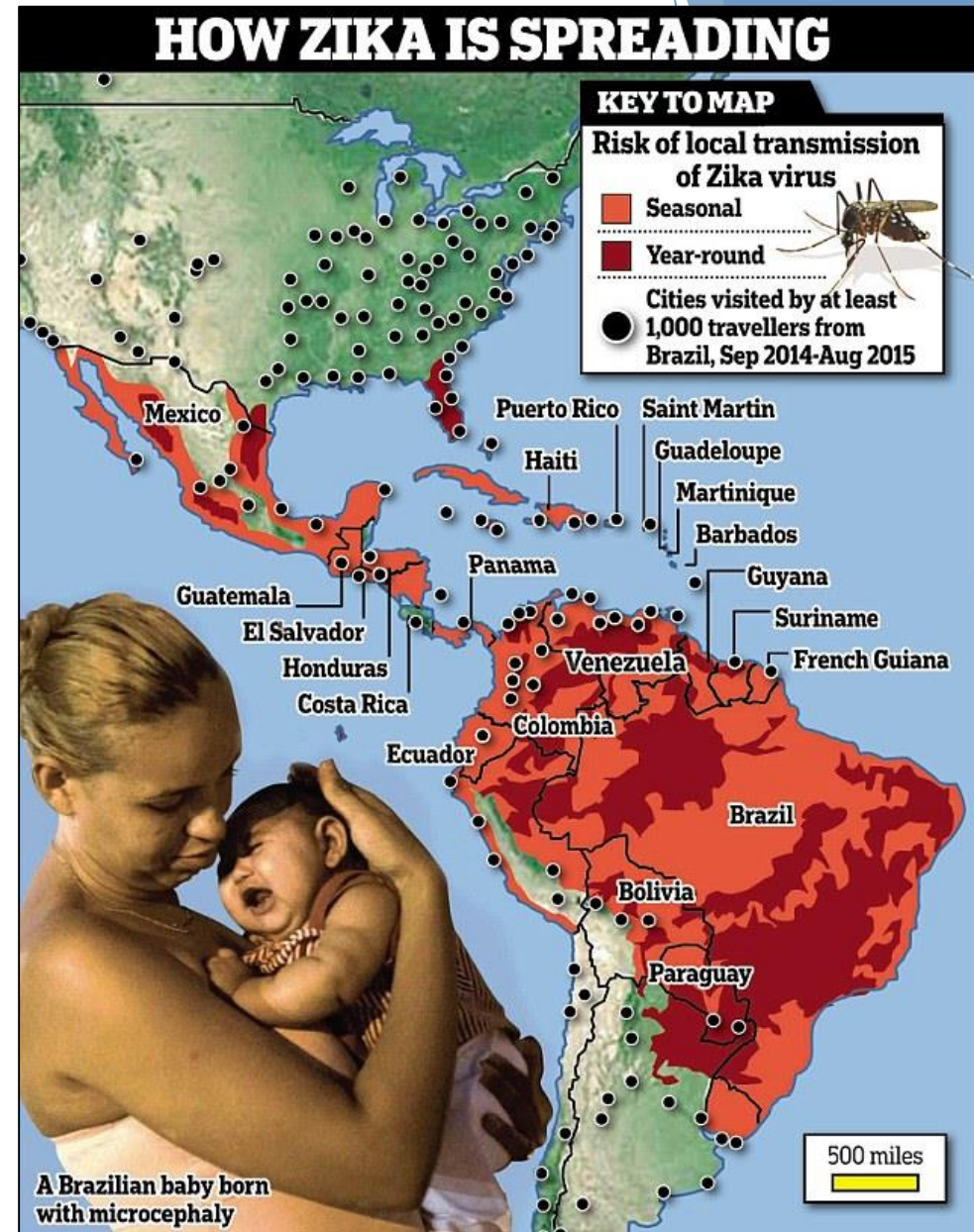
Cary Institute  
of Ecosystem Studies

# The Zika problem

Recently there is a Zika outbreak in Brazil and it is spreading fast.

Symptoms include rash, fever, muscle and joint pain, headache;

Zika is non-lethal but causes Microcephaly (= small head size) and eye problems in babies of infected mothers.





# How Zika spreads

- ▶ The virus naturally resides in the blood of host animals;
- ▶ The animals don't get infected because they have antibodies, or have subclinical infections;
- ▶ Virus survives by passing from generation to generation of animals.



*Known primate hosts of Zika*

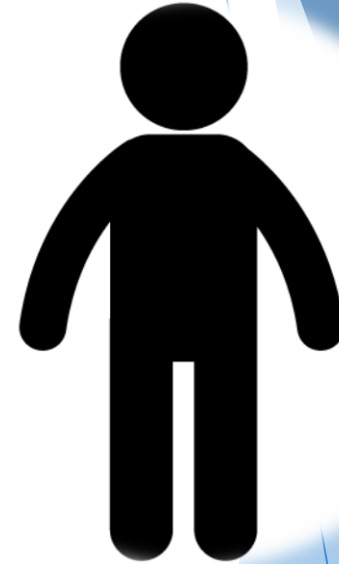
# How Zika spreads



Mosquito  
bites  
monkey



Mosquito  
bites  
human



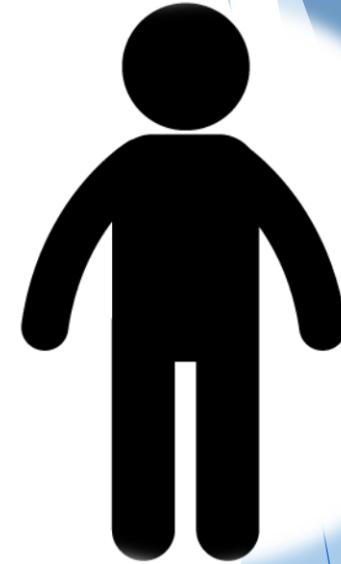
# Present: REACTIVE approach to contain outbreak



Mosquito  
bites  
monkey



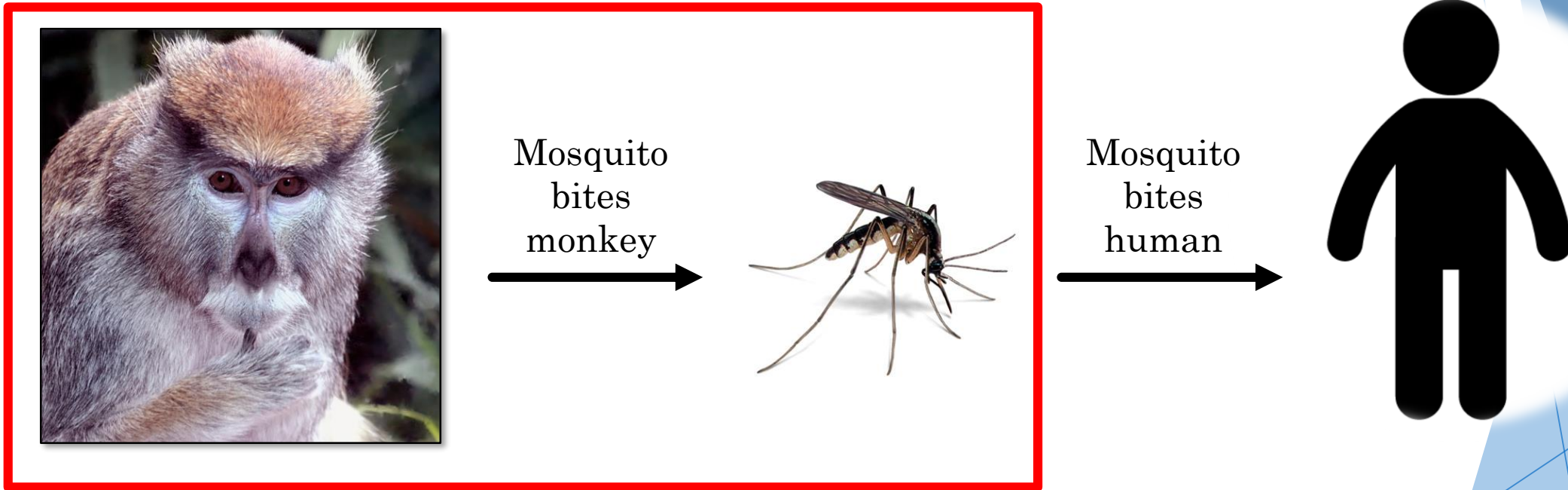
Mosquito  
bites  
human



Eradicate mosquitos in spillover areas  
**Not Good!**



# Objective : PROACTIVE approach to stop outbreak



Find out the source animals to prioritize mosquito eradication efforts

**Kills the problem at source**

# Data

Dataset 1: Reservoir status

Animal	Carries Zika?	Carries Dengue?	Carries Yellow fever?	...
Monkey 1	No	Yes	No	...
Monkey 2	Yes	No	No	...
Monkey 3	No	Yes	No	...
Monkey 4	No	No	No	...
Monkey 5	Yes	No	Yes	...
Monkey 6	No	Yes	No	...
...	...	...	...	...

376 monkeys, 8 diseases

Dataset 2: Animal characteristics

Animal	Body mass	Litter size	Maximum longevity	...
Monkey 1				...
Monkey 2				...
Monkey 3				...
Monkey 4				...
Monkey 5				...
Monkey 6				...
...	...	...	...	...

50 characteristics

# Challenges

Dataset 1: Reservoir status

Animal	Carries Zika?	Carries Dengue?	Carries Yellow fever?	...
Monkey 1	No	Yes	No	...
Monkey 2	Yes	No	No	...
Monkey 3	No	Yes	No	...
Monkey 4	No	No	No	...
Monkey 5	Yes	No	Yes	...
Monkey 6	No	Yes	No	...
...	...	...	...	...

376 monkeys, 8 diseases

- ▶ Reservoirs are extremely rare;
- ▶ Only 4 known reservoirs for Zika;
- ▶ There are only 26 positive entries in this matrix;
- ▶ Need specialized methods to deal with the situation.



# Challenges

- ▶ Data on animal characteristics are not complete: lot of entries are missing in many animals;
- ▶ Some characteristics are almost completely missing for all animals;
- ▶ Some animals have almost all variables missing;

Dataset 2: Animal characteristics

Animal	Body mass	Litter size	Maximum longevity	...
Monkey 1				...
Monkey 2				...
Monkey 3				...
Monkey 4				...
Monkey 5				...
Monkey 6				...
...	...	...	...	...

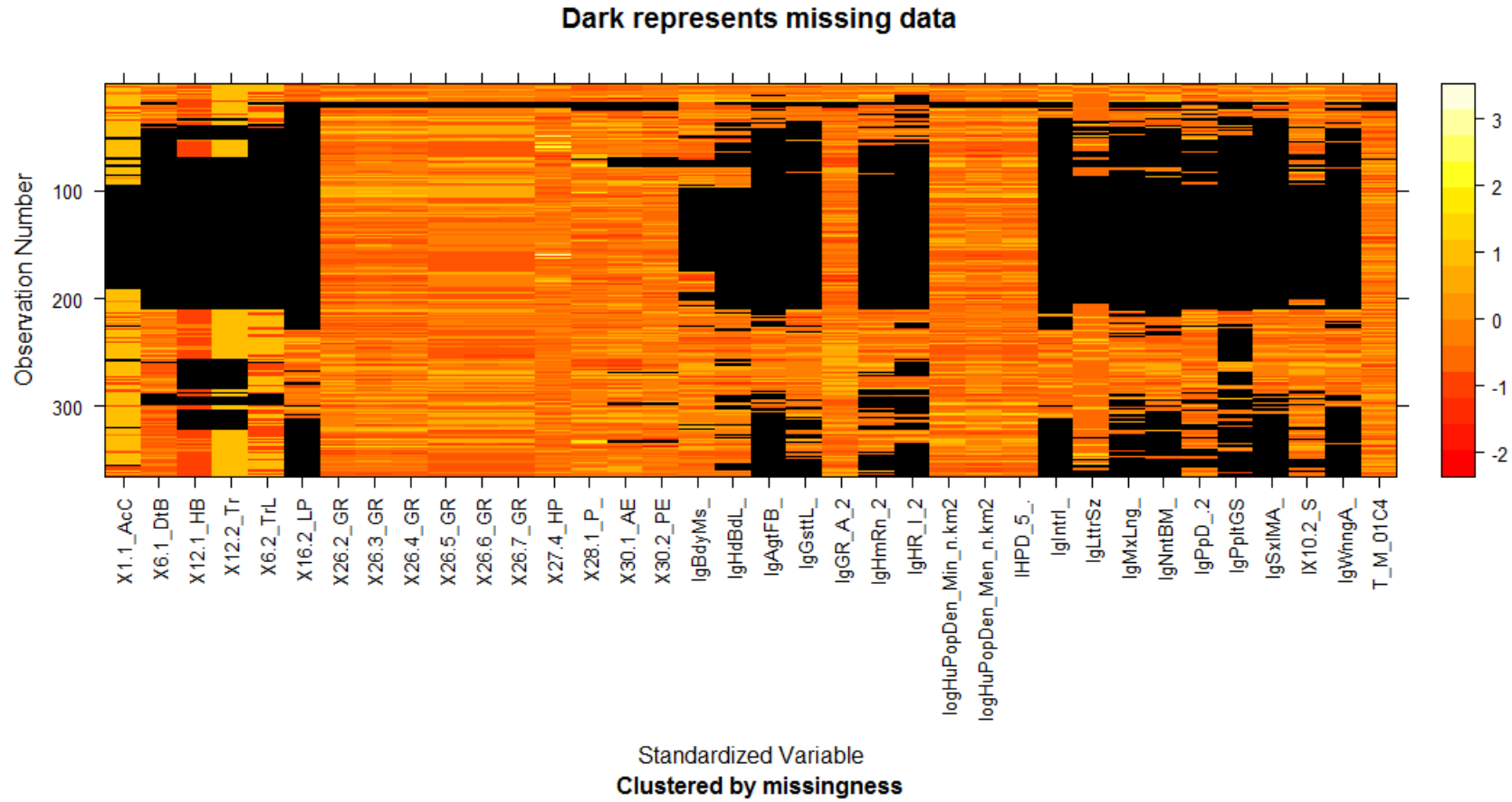
# Challenges

- ▶ Data on animal characteristics are not complete: lot of entries are missing in many animals;
- ▶ Some characteristics are almost completely missing for all animals;
- ▶ Some animals have almost all variables missing;

Dataset 2: Animal characteristics

Animal	Body mass	Litter size	Maximum longevity	...
Monkey 1				...
Monkey 2				...
Monkey 3				...
Monkey 4				...
Monkey 5				...
Monkey 6				...
...	...	...	...	...

# The missing data problem



# Modelling approach

## 1. **Missing data imputation:**

use a multiple imputation procedure called Multiply Imputed Chained Equation (MICE: [Raghunathan \*et al\*, 2001](#))

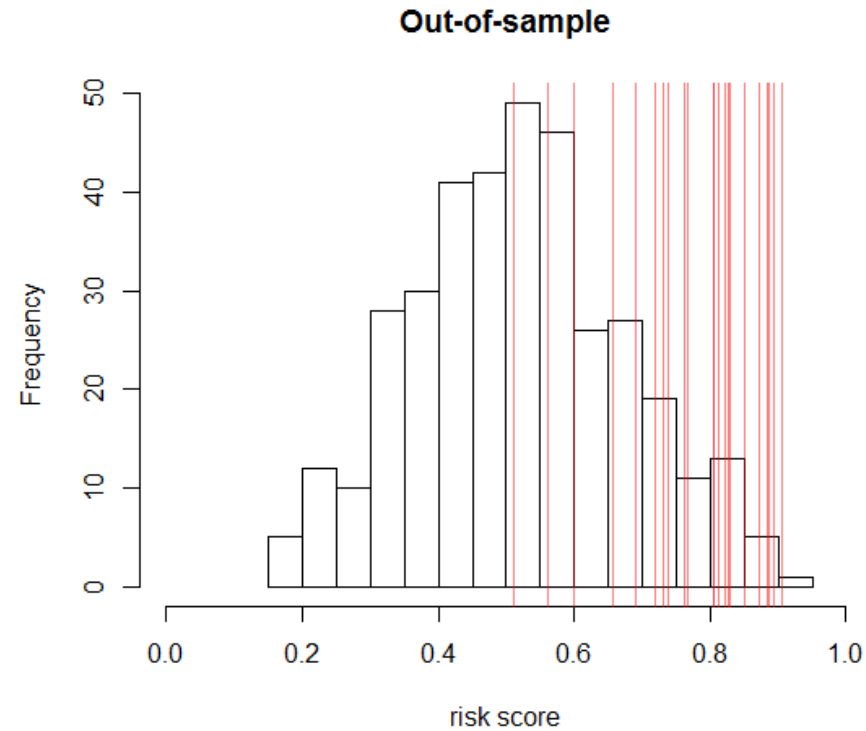
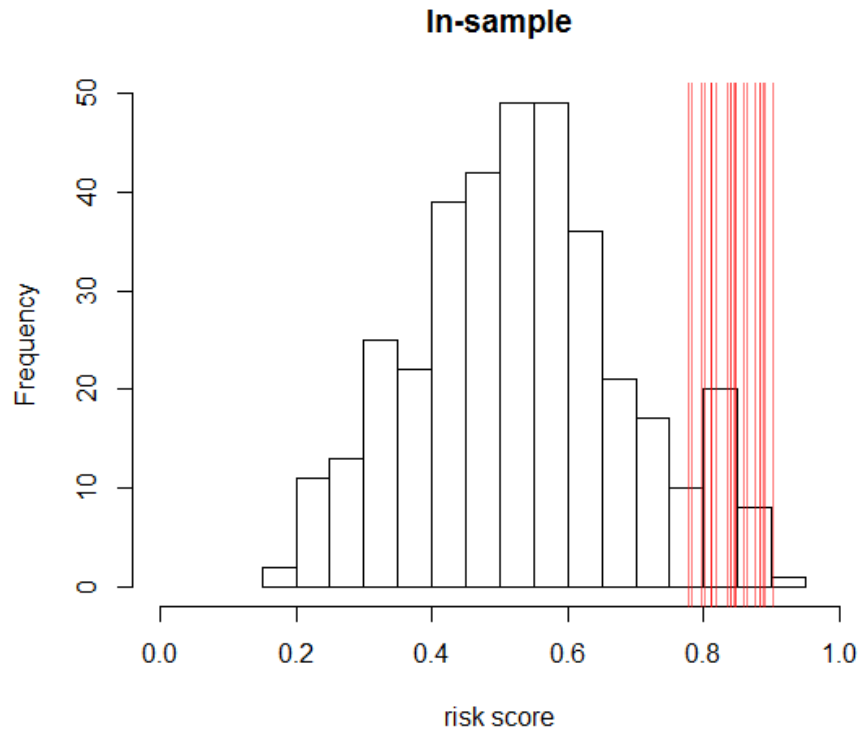
## 2. **Predictive model:**

- ▶ Model reservoir status for all primates simultaneously;
- ▶ Use a Bayesian model that assumes the response variable is generated through a hierarchical process, taking into account covariate information in a nonlinear fashion ([Rai \*et al\*, 2015](#))



# Validation

Delete a positive entry in dataset 1, model on that data and use that model to get risk score for that sample.



# Results

- ▶ We are interested in high-risk animals in South America that haven't been detected yet.
- ▶ Top 5 high-risk species:



Black-tufted  
Marmoset  
(*Callithrix penicillata*)



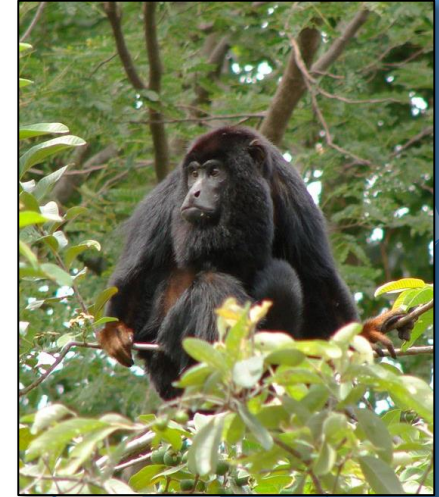
White-fronted  
Capuchin  
(*Cebus albifrons*)



Brown  
howler  
(*Alouatta guariba*)



Cotton-top  
tamarin  
(*Saguinus oedipus*)



Red-handed  
howler  
(*Alouatta belzebul*)

# Trait profiles of high-risk animals

Variable name	Importance	mean.low	mean.hi
logGR_Area_km2	23.75	8.21	13.38
logAgeatFirstBirth_d	23.39	6.57	7.09
logNeonateBodyMass_g	23.07	3.28	5.28
X26.6_GR_MinLong_dd	22.93	25.76	21.12
logHomeRange_km2	22.69	-4.07	-1.13
logSexualMaturityAge_d	22.17	6.17	7.15
X26.5_GR_MaxLong_dd	21.67	26.91	37.50
logInterbirthInterval_d	21.66	5.63	6.07
logHuPopDen_Min_n.km2	21.35	2.38	1.24
logBodyMass_g_Resid	20.95	0.02	0.08

- Larger animals that have high body mass, less offsprings and larger social groups seem to be more likely reservoirs

# What to do with the outputs?

- ▶ Work with disease ecology researchers to collect blood samples from these monkeys and test for presence of Zika virus;
- ▶ If a new reservoir is detected, focus on mosquito eradication efforts around the animal's habitat;
- ▶ Provide a much needed empirical baseline for future similar studies regarding a proactive approach towards infectious disease management.



# Future work

- ▶ Integrate with the Prospector tool ([Krause, Perer and Ng, 2016](#)) to understand how risk scores are affected by different levels of a covariate, i.e. partial dependence plots;
- ▶ Modify outcomes for unknown reservoirs based on their geographic range overlap with known reservoirs, as well as incorporate primate-mosquito interactions;
- ▶ Extend the underlying model to incorporate information on the different viruses;
- ▶ Build a unified framework for simultaneously imputing missing data and modelling the outcomes.

# References

- ▶ First detection of Zika virus in neotropical primates in Brazil: a possible new reservoir. Favoretto, S.; Araujo, D.; Oliviera, D.; Duarte, N.; Mesquita, F.; Zanotto, P. and Durigon, E. Available in bioRxiv: <http://dx.doi.org/10.1101/049395>, April **2016**.
- ▶ Han, B.; Schmidt, J. P.; Bowden, S. E. and Drake, J. M. Rodent reservoirs of future zoonotic diseases. *Proc. Natl. Acad. Sci.*, **2015**, 112(22): 7039-7044.
- ▶ Raghunathan, T.W.; Lepkowski, J.M.; Van Hoewyk, J. and Solenbeger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **2001**, 27: 85–95.
- ▶ Rai, P.; Hu, C.; Henao, R. and Carin, L. Large-Scale Bayesian Multi-Label Learning via Topic-Based Label Embeddings. In: *Advances in Neural Information Processing Systems 29 (NIPS 2015)*, **2015**, 3222-3230.
- ▶ Krause, J.; Perer, A. and Ng, K. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*, **2016**, 5686-5697.

# Acknowledgements

- ▶ Saška Mojsilović and Kush R. Varshney;
- ▶ Barbara Han, Cary Institute of Ecosystem Studies;
- ▶ Social Good mentors: Flavio D. Calmon, Abhishek Kumar, Raya Horesh, Adam Perer, Elisa B. Von Marschall, Dennis Wei;
- ▶ Prasanna Sattigeri and Karthi Ramamurthy.

The background features abstract, overlapping geometric shapes in various shades of blue, primarily on the right side and bottom, creating a modern, layered effect. The central area is a plain white space.

THANK YOU!