CAPSTONE 1: THE FIGHT AGAINST MALARIA …AN EFFORT AT PREDICTING THE IMPACT OF MALARIA

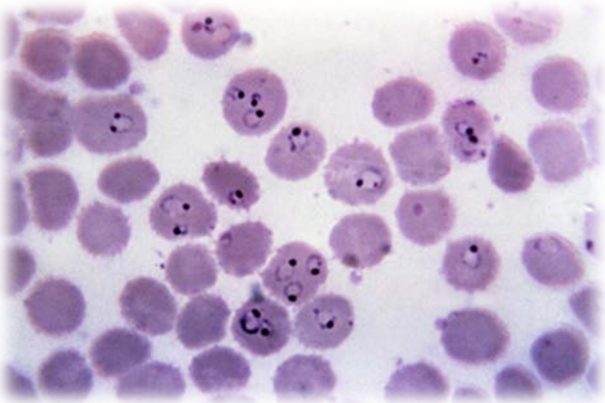
AUTHOR: ZACK ALLEN, MD, MBA

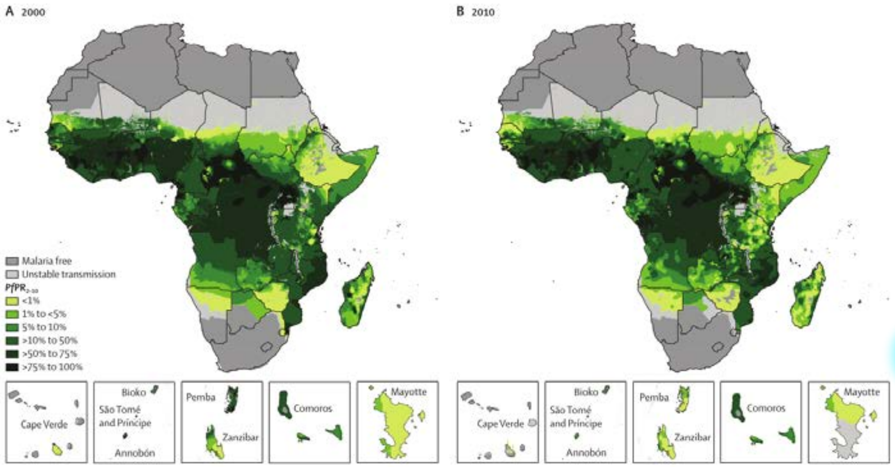
DATE: 1/18/2018

kaggle Springboard

OUTLINE

- Intro & Background
- Dataset and Data Acquisition
- EDA and Findings
- Statistical Inference
- Machine Learning
- Conclusions
- Next Steps

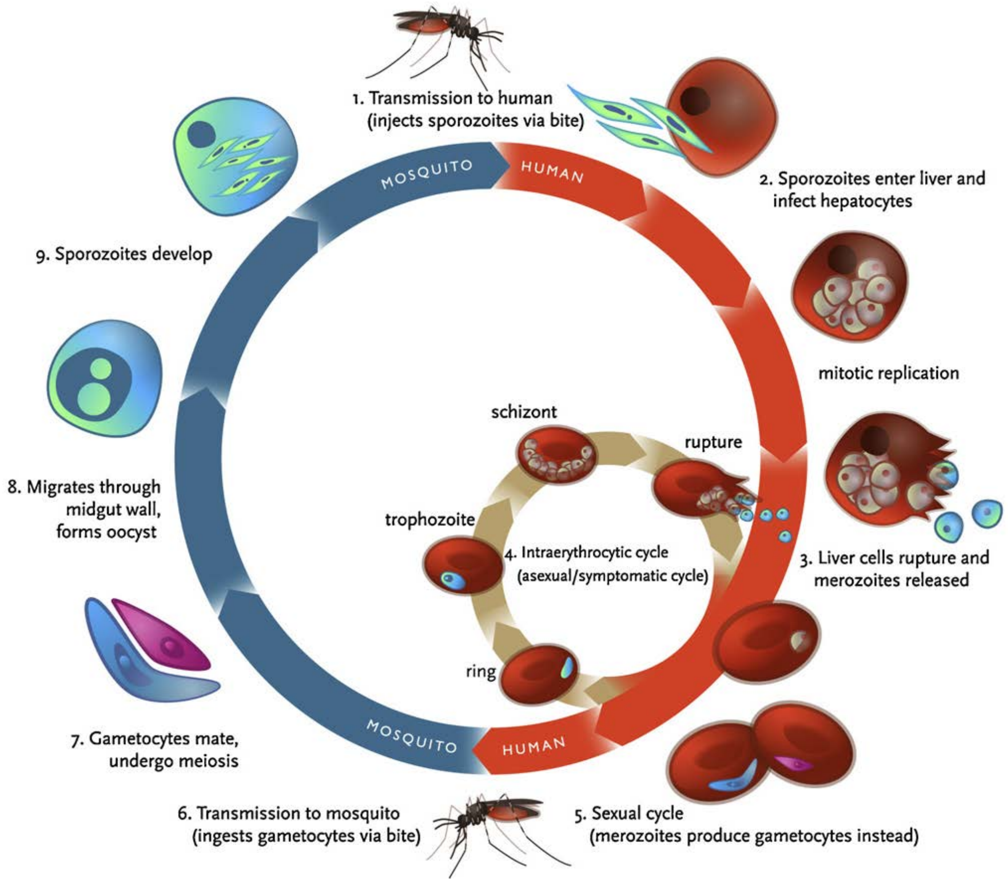




A LITTLE BACKGROUND ON THE TOPIC…

- Mosquito born disease
- Caused by parasite Plasmodium species
- Replicates in the blood stream
- Most cases can be treated/prevented with modern medication
  - Though resistance is a growing concern

Life Cycle of the Malaria Parasite

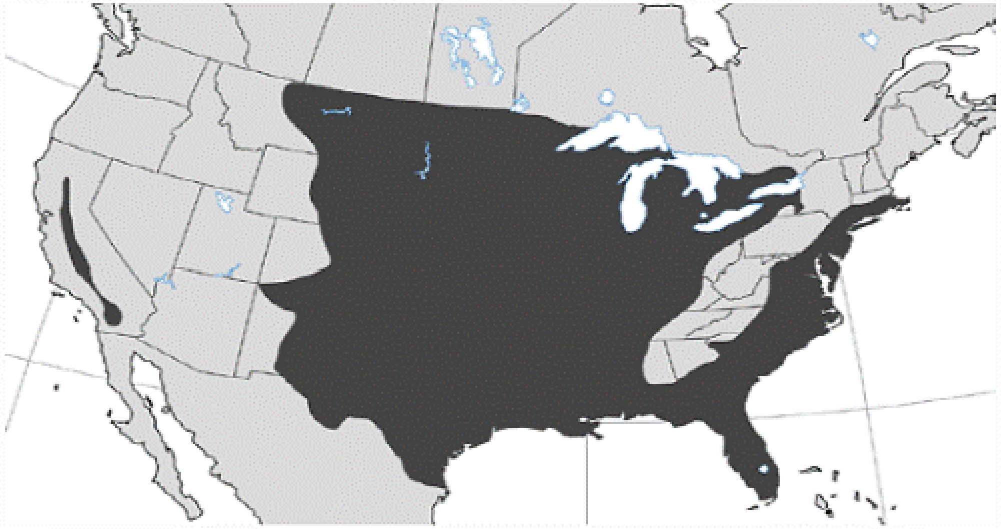


1. Transmission to human (injects sporozoites via bite)
2. Sporozoites enter liver and infect hepatocytes
mitotic replication
3. Liver cells rupture and merozoites released
4. Intraerythrocytic cycle (asexual/symptomatic cycle)
schizont
rupture
trophozoite
ring
5. Sexual cycle (merozoites produce gametocytes instead)
6. Transmission to mosquito (ingests gametocytes via bite)
7. Gametocytes mate, undergo meiosis
8. Migrates through midgut wall, forms oocyst
9. Sporozoites develop
MOSQUITO
HUMAN
MOSQUITO
HUMAN

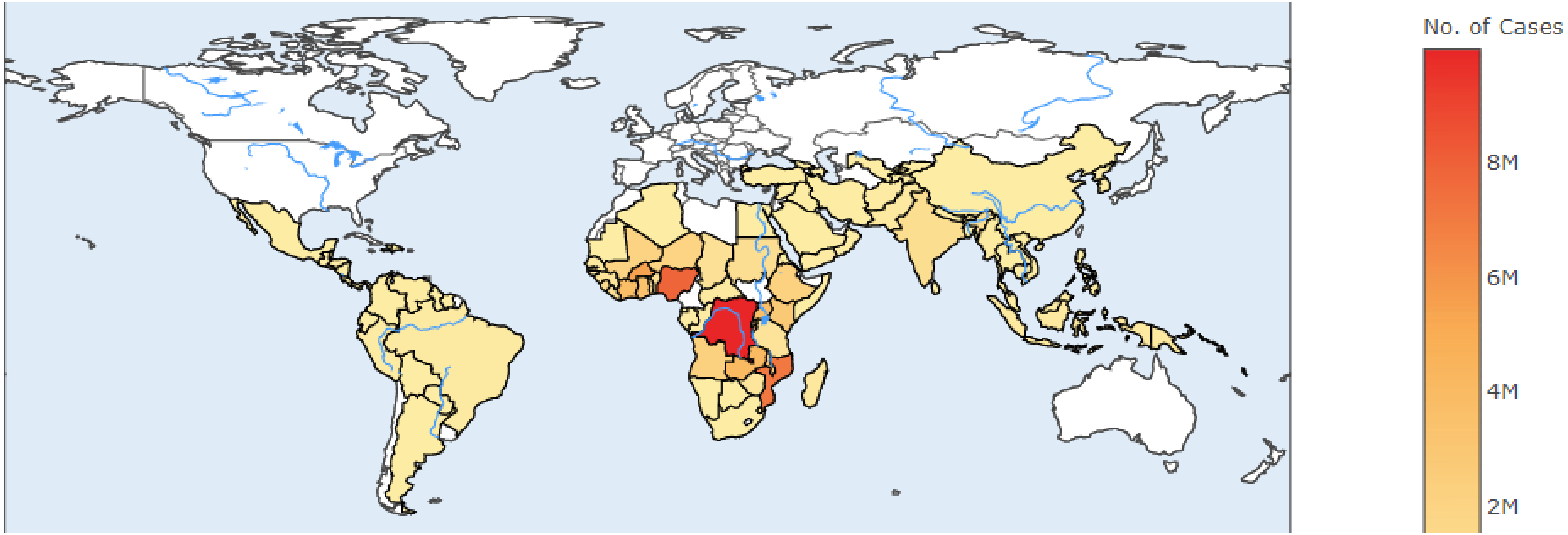
THE PROBLEM

- Geographically similar regions have large divergence in impact of malaria.
- It is difficult to assess true impact of aid efforts especially at scale and in an objective manner.
- The ability to accurately anticipate outbreaks has been elusive.

Malaria occurrence in the United States, 1880s



In the late 19th century malaria was endemic in shaded regions,
***Source:** Reiter, 2001.*



REPORTED MALARIA CASES IN 2014

The geographic spread of malaria is residing, slowly.

THE GOAL

- Provide a model by which the impact of malaria can be estimated and localized. Thus, allowing more efficient allocation of finite resources.
- Assess the effectiveness of aid efforts by tracking and evaluating features of each region and the type of efforts provided e.g. malarial medications and/or insect nets.

DATA ACQUISITION & WRANGLING

World Health Organization

IHME GHDx

THE WORLD BANK
IBRD • IDA

Data Acquisition:

- Kaggle Competition
  - Information on the distribution of aid efforts
- World Health Organization
  - Statistics on number of cases and mortality
- The World Bank
  - Environmental data
- Institute for Health Metrics and Evaluation
  - Data on the incidence and number of cases

Data Wrangling:

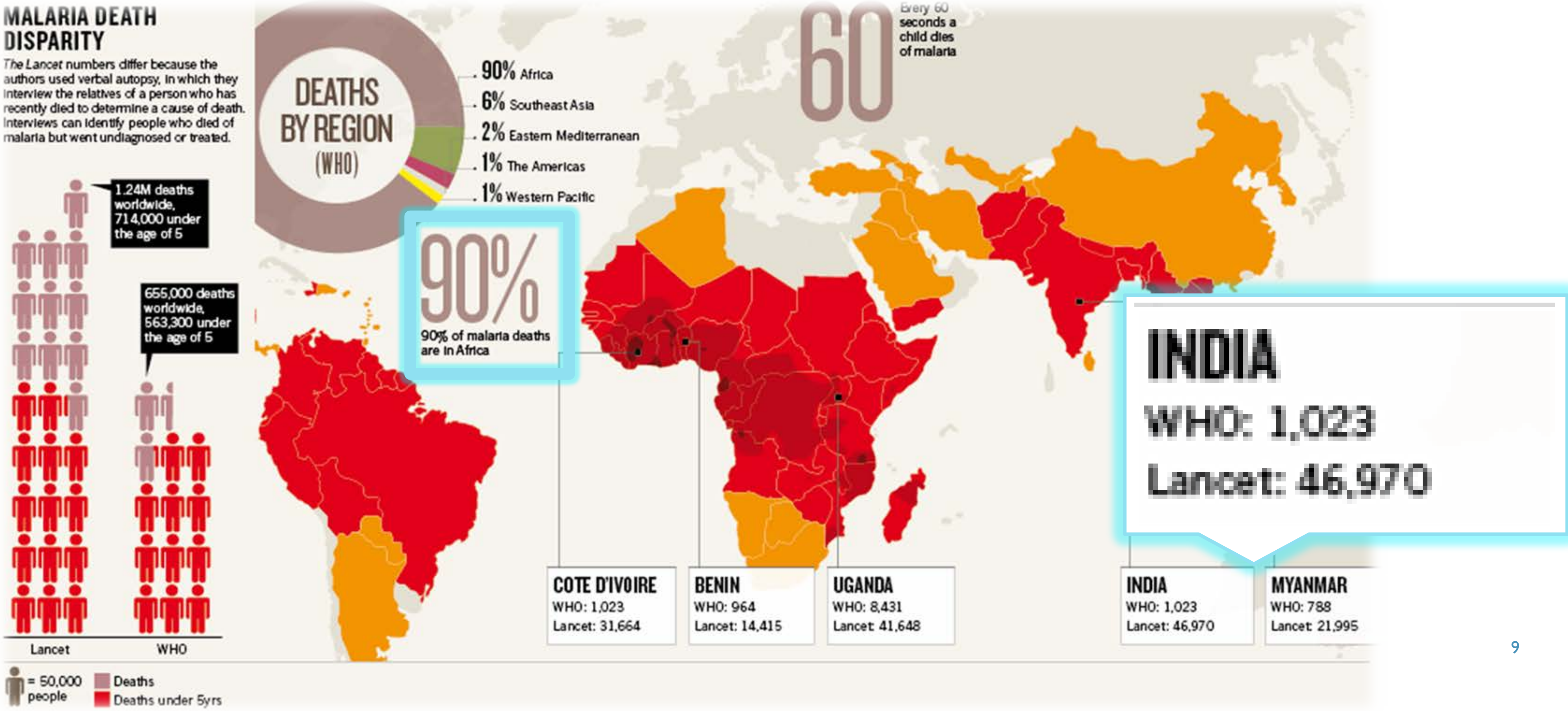
- Missing values – removed
- Duplicate values – removed
- The data is unbalanced i.e. not all countries have entries for all years
- Country code was treated as a categorical variable

```
1023 entries
Data columns (total 15 columns):
Unnamed: 0          1023 non-null int64
year                1023 non-null int64
country_code        1023 non-null object
reported_cases      1023 non-null float64
region_x            1023 non-null object
reported_deaths     1023 non-null float64
rainfall            1023 non-null float64
temperature         1023 non-null float64
population          1023 non-null float64
percent_agg         1023 non-null float64
percent_urb         1023 non-null float64
gdp_per_cap         1023 non-null float64
country_name        1023 non-null object
pop_density         1023 non-null float64
incidence           1023 non-null float64
```

DATA CHALLENGES & CONCERNS

- Inconsistency in reported numbers between institutions.
- Difficulties in obtaining accurate and timely data from regions most affected.
- Fluidity of borders e.g. people infected in one region and developing symptoms in another.

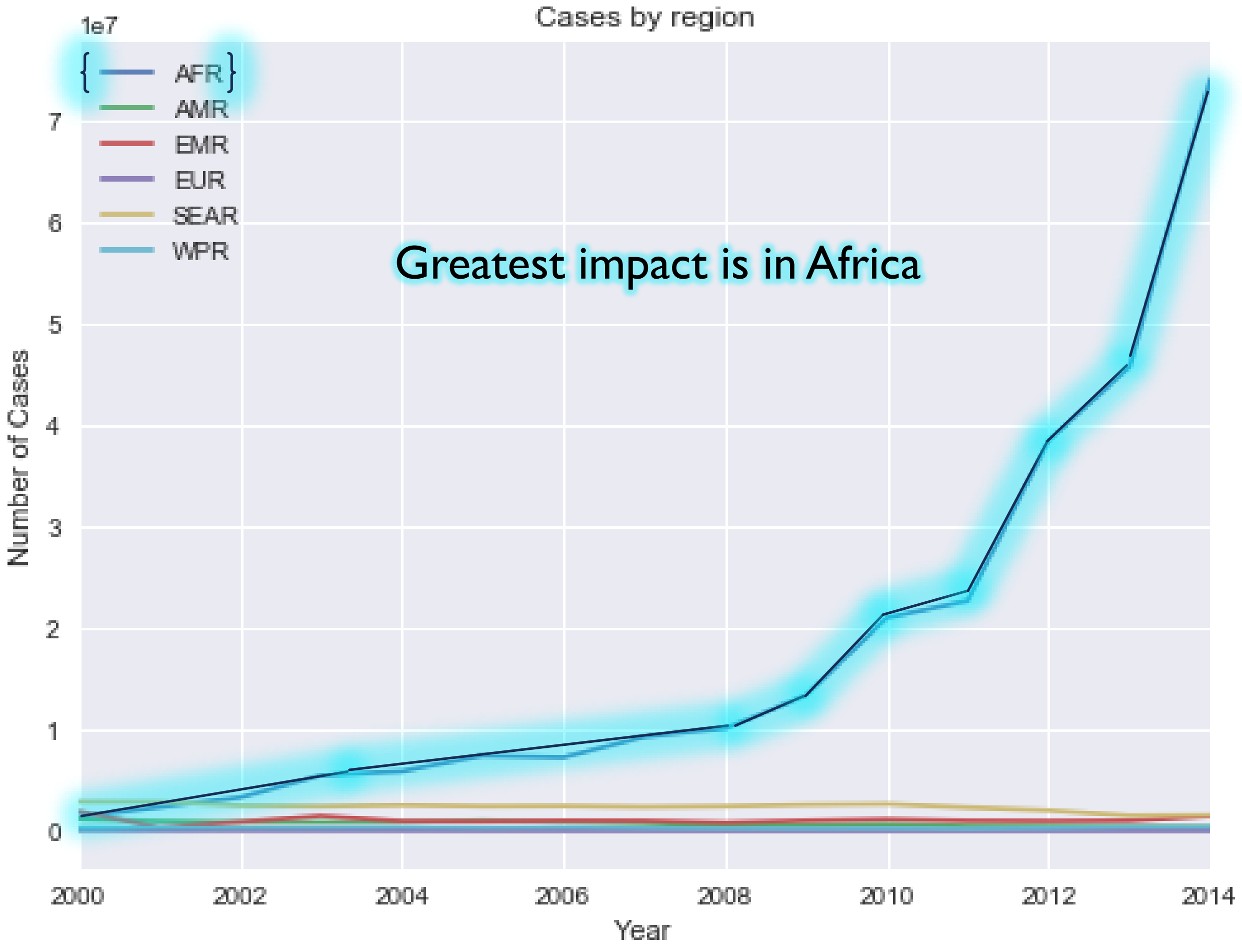
INCONSISTENT NUMBERS …BUT OVERALL TRENDS REMAIN



MALARIA DEATH DISPARITY

The Lancet numbers differ because the authors used verbal autopsy, in which they interview the relatives of a person who has recently died to determine a cause of death. Interviews can identify people who died of malaria but went undiagnosed or treated.

AFRICA

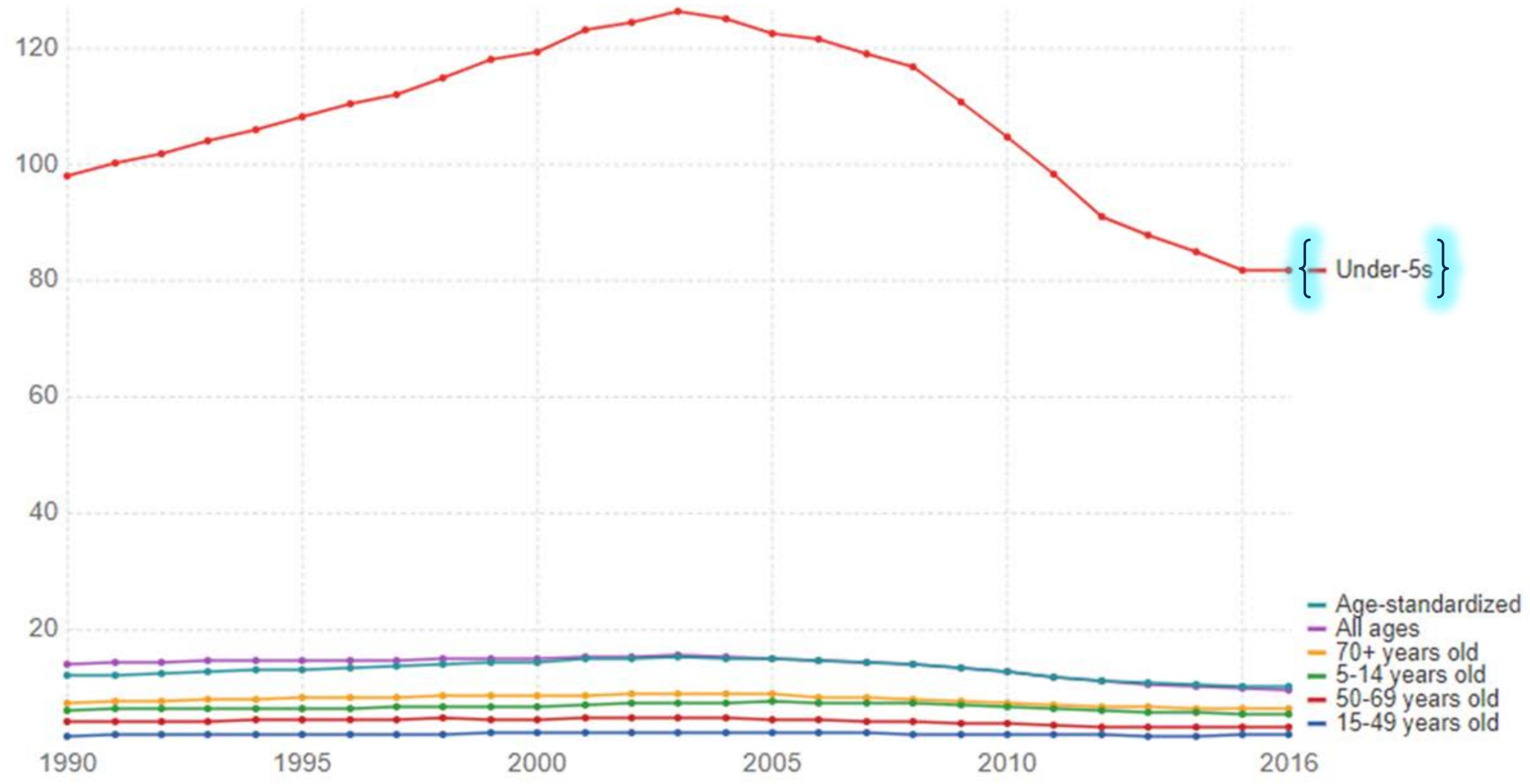


Greatest impact is in Africa

CHILDREN UNDER 5

Malaria death rates by age (per 100,000), World

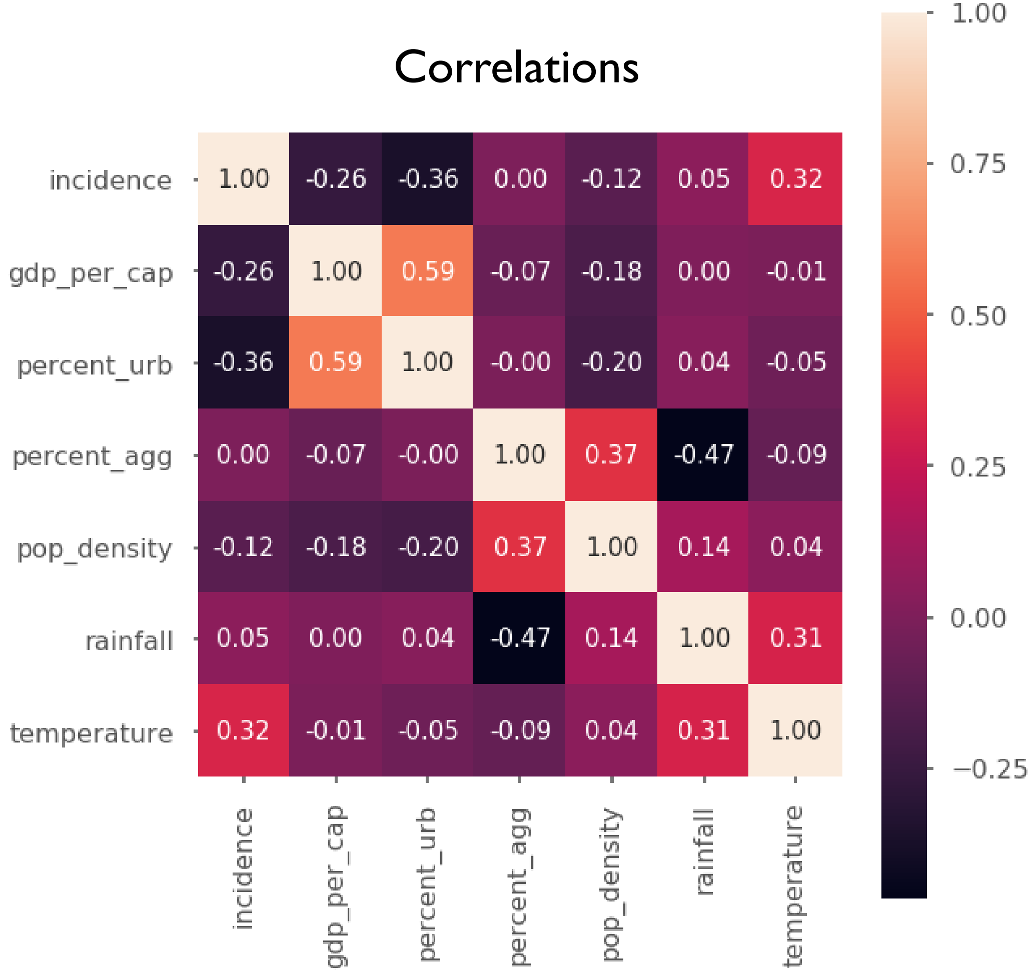
Death rates from malaria measured per 100,000 individuals across various age categories. Also shown is the total death rate across all ages (not age-standardized) and the age-standardized death rate. Age-standardization assumes a constant population age & structure to allow for comparisons between countries and with time without the effects of a changing age distribution within a population (e.g. aging).



Source: IHME, Global Burden of Disease (GBD)

LOOKING AT THE FEATURES

- Incidence of malaria
- GDP per capita
- Percent Urban
  - Percent of the total population living in an urban environment.
- Percent Agricultural land
  - Percent of the countries total land area that is used for agriculture. A marker of those with an outdoor occupation.
- Population Density
- Rainfall
  - Scaled to a yearly average
- Temperature
  - Scaled to a yearly average

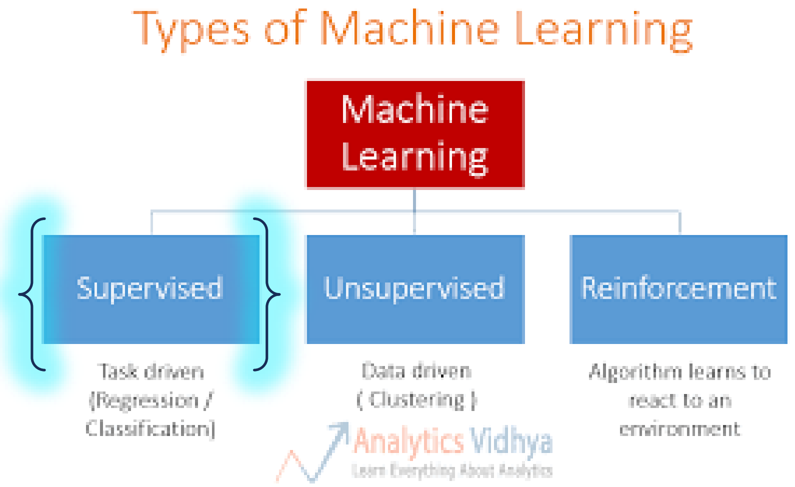


Correlations

	incidence	gdp_per_cap	percent_urb	percent_agg	pop_density	rainfall	temperature
incidence	1.00	-0.26	-0.36	0.00	-0.12	0.05	0.32
gdp_per_cap	-0.26	1.00	0.59	-0.07	-0.18	0.00	-0.01
percent_urb	-0.36	0.59	1.00	-0.00	-0.20	0.04	-0.05
percent_agg	0.00	-0.07	-0.00	1.00	0.37	-0.47	-0.09
pop_density	-0.12	-0.18	-0.20	0.37	1.00	0.14	0.04
rainfall	0.05	0.00	0.04	-0.47	0.14	1.00	0.31
temperature	0.32	-0.01	-0.05	-0.09	0.04	0.31	1.00

MACHINE LEARNING

- Approach:
  1) Test algorithms: Linear Regression, Bayesian Ridge, Random Forrest.
  2) Perform feature analysis using Random Forrest and tune parameters as necessary.
  3) Examine models and identify the best overall model.
  4) Use models to predict incidence and examine results.


Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task driven (Regression / Classification)
Data driven (Clustering)
Algorithm learns to react to an environment
Analytics Vidhya
Learn Everything About Analytics

MACHINE LEARNING: MODEL PERFORMANCE

	Train/Test Split	Rounds of Validation	Average Model Score (1.0 is best)
Random Forrest Regression	Yes	5	0.68
Bayesian Ridge Regression	Yes	5	0.66
Linear Regression	Yes	5	0.60

TRAINING STRUCTURE

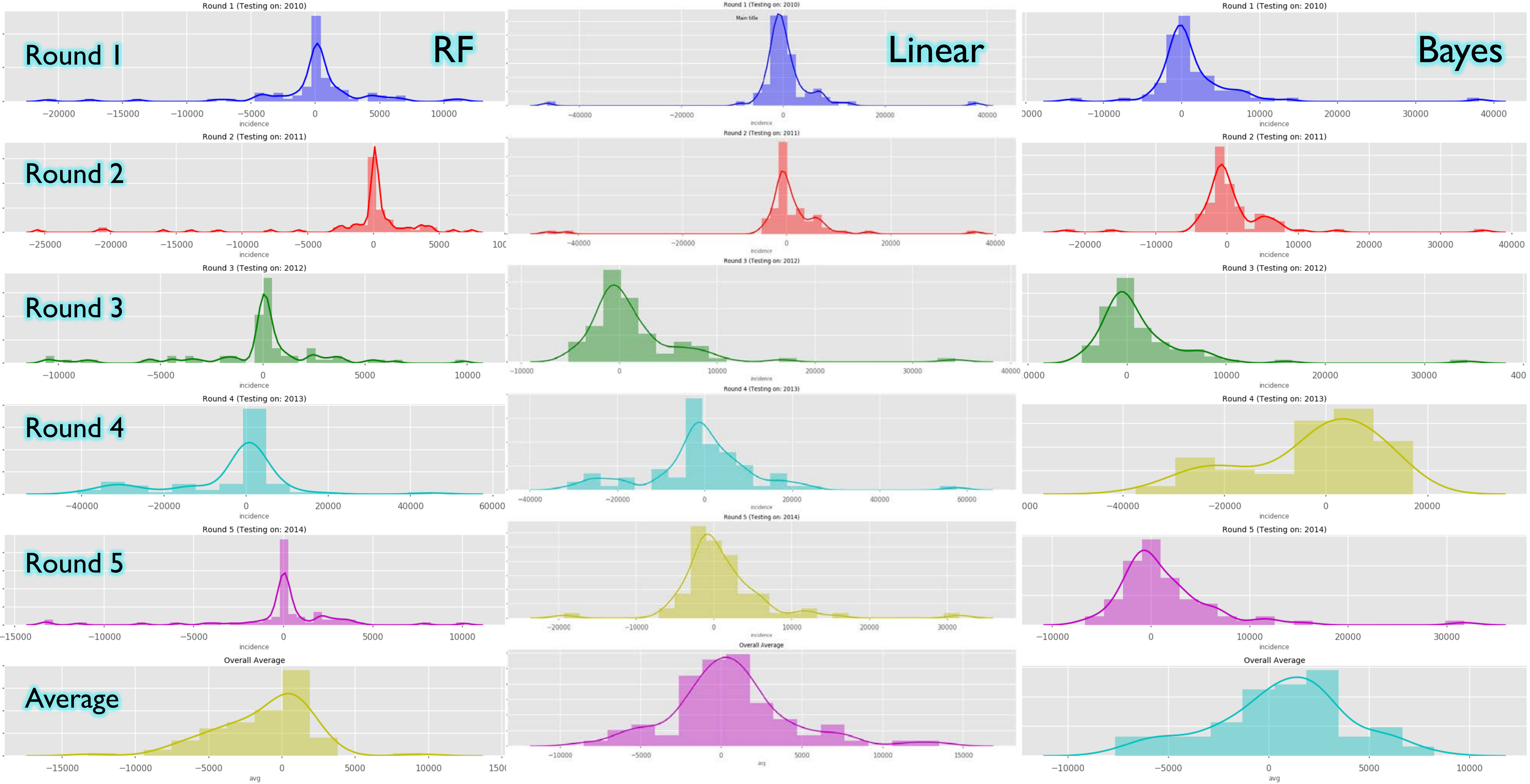
2000 - 2009	2010	2011	2012	2013	2014
Train	Test				
Train	▨	Test			
Train	▨	▨	Test		
Train	▨	▨	▨	Test	
Train	▨	▨	▨	▨	Test

SINGLE YEAR ANOMALY IN 2013 PREDICTIONS

We can see abnormally low accuracy for all models on the 2013 test data.

Model	2010	2011	2012	2013	2014
Random Forrest Regression	0.90	0.85	0.95	-0.19	0.93
Bayesian Ridge Regression	0.83	0.80	0.85	-0.002	0.82
Linear Regression	0.72	0.58	0.84	0.07	0.79

Distribution of Residuals

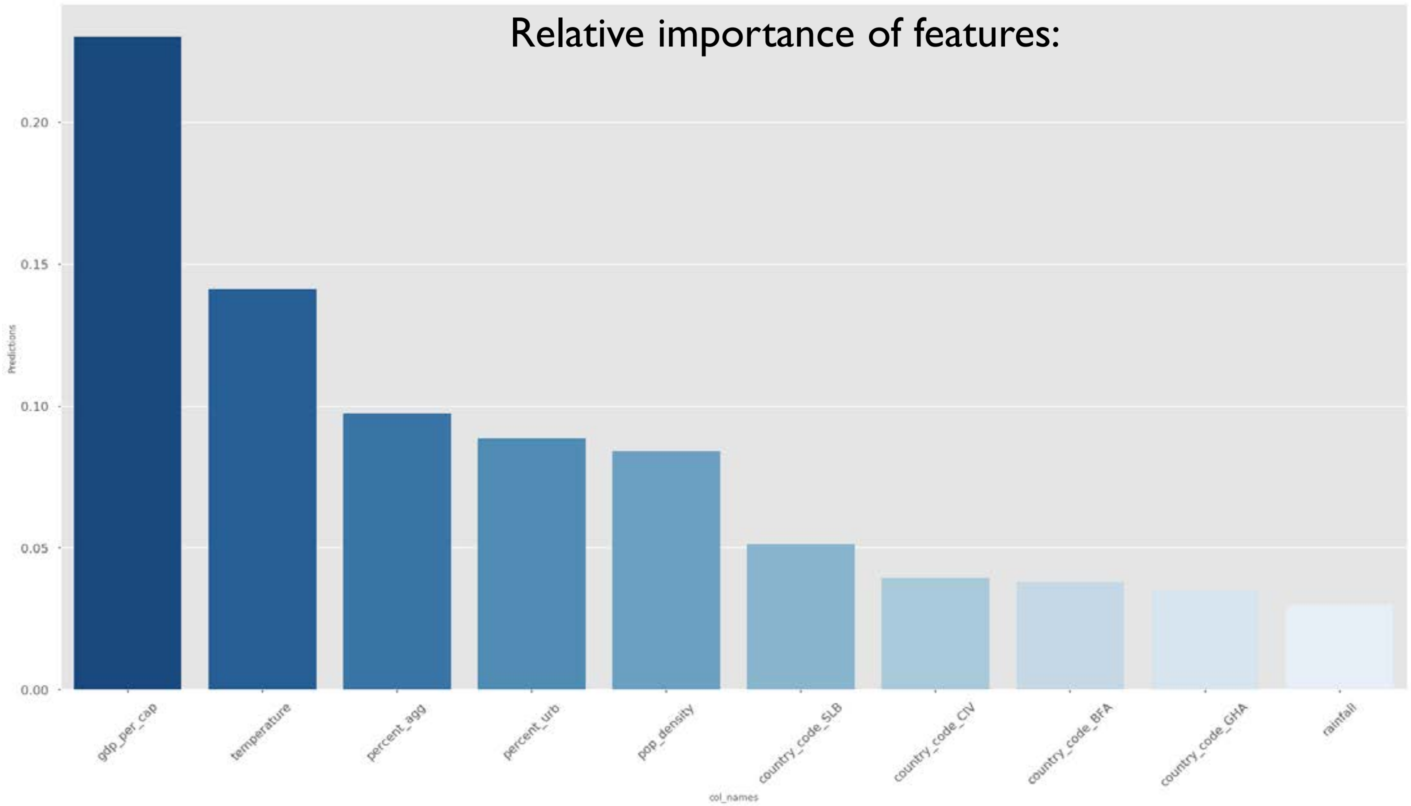


MACHINE LEARNING: FEATURE ANALYSIS

GDP per capita is the most influential feature of those examined.

<u>Features in order of importance:</u>

- GDP per capita
- Temperature
- Percent agricultural land
- Percent urban population
- Population density
- Country code


Relative importance of features:

KEY FINDINGS

- The Random Forrest model provided the most consistent accuracy.
- GDP per capita was the most influential of human factors and temperature was the most influential of environmental factors.
- In 4 out of 5 years the tested models were all able to give accurate (0.85 or greater for RF) and potentially usable prediction of the incidence from the given data.

LIMITATIONS

- Use of country and yearly level data limiting resolution.
  - Need access to monthly and county/province level data.
- Use of substitute markers e.g. percent agricultural land.
  - Probably more accurate to use actual numbers (estimates) of those with an outdoor occupation.
- Limited number of data points approximately 1000 entries used.
  - Increasing the number of years with complete data either by better data collection or imputing with the average could provide more robust results.

RECOMMENDATIONS & FUTURE IMPROVEMENTS

Recommendations:

- A ML based tool to predict the impact of malaria and where resources should be allocated appears feasible.
- Increasing resolution of data to a county/province level would allow use of additional factors such as elevation.
- Train the model on a greater number years worth of data, especially outlier years e.g. 2013.

Future improvements:

- Access to real/near-real time data to integrate into an active model would increase usefulness.
- Use of entirely new features such as satellite data and mosquito species.

RESOURCES

- https://www.kaggle.com/teajay/the-fight-against-malaria
- http://www.who.int/en/
- http://www.worldbank.org/