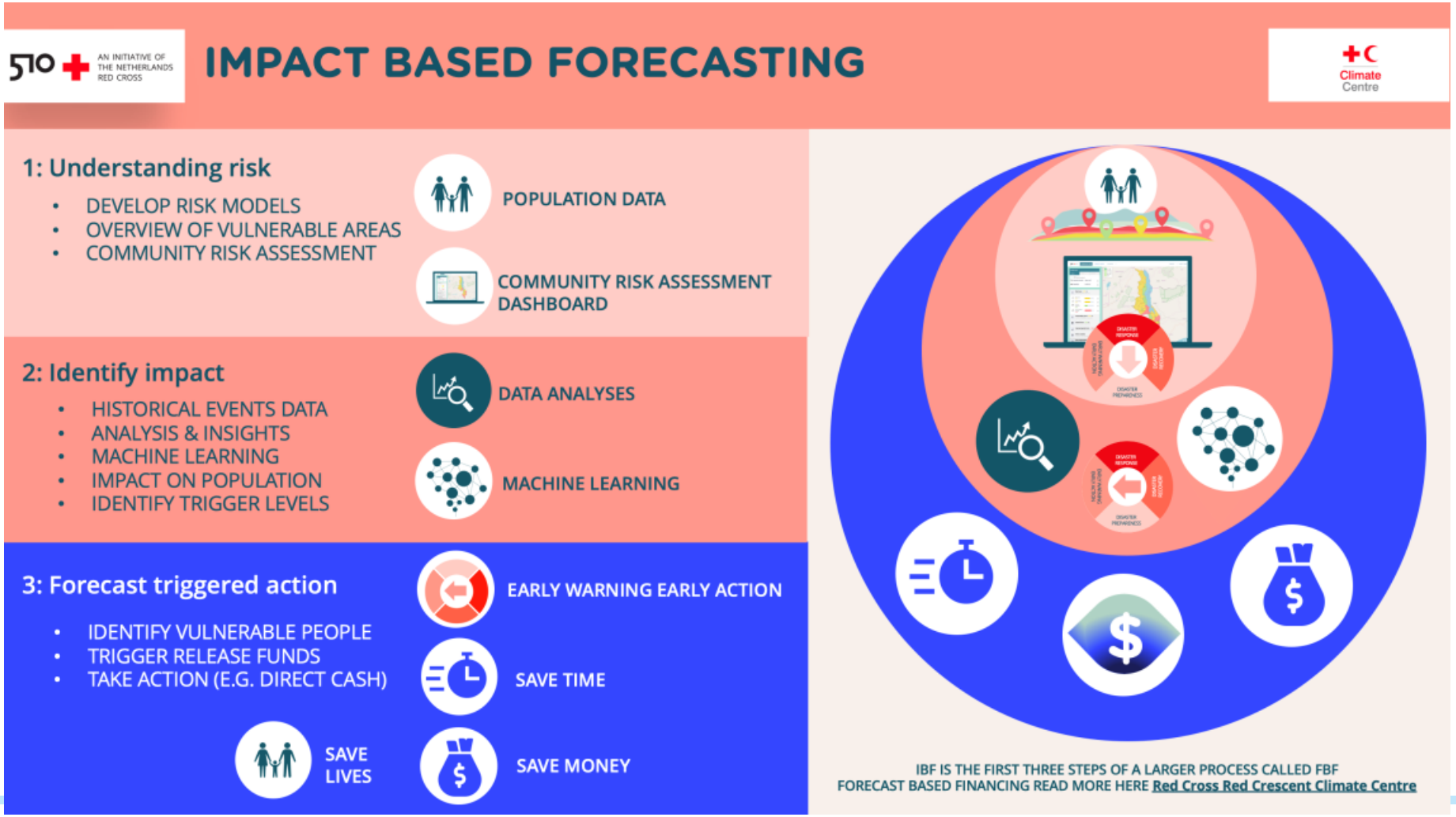
Impact data analysis Uganda

Veronique Verhees, 04-07-2019

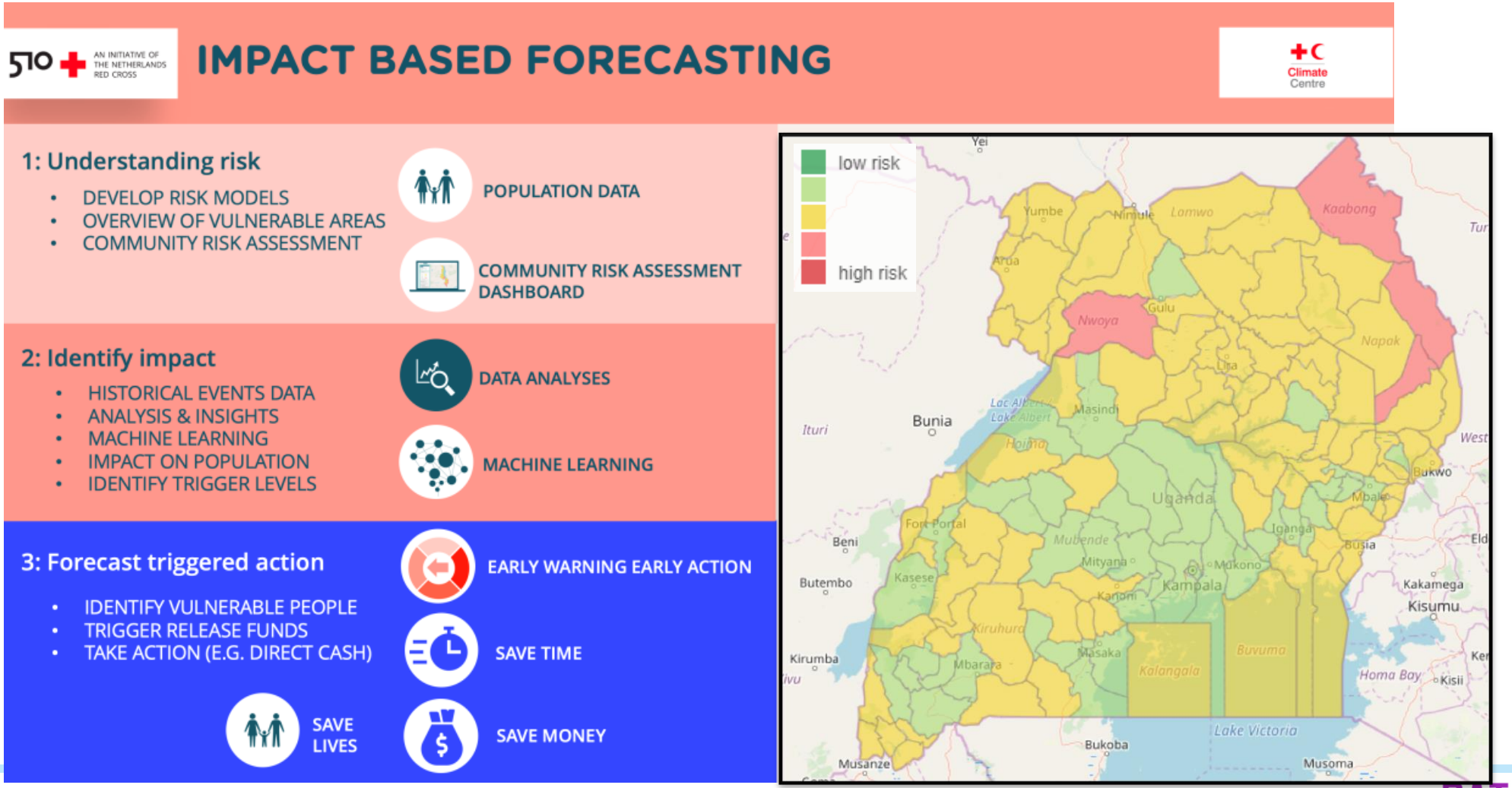
AN INITIATIVE OF
THE NETHERLANDS
RED CROSS

Impact based forecasting:



DATA SCIENCE WITH PURPOSE

Impact based forecasting: Uganda



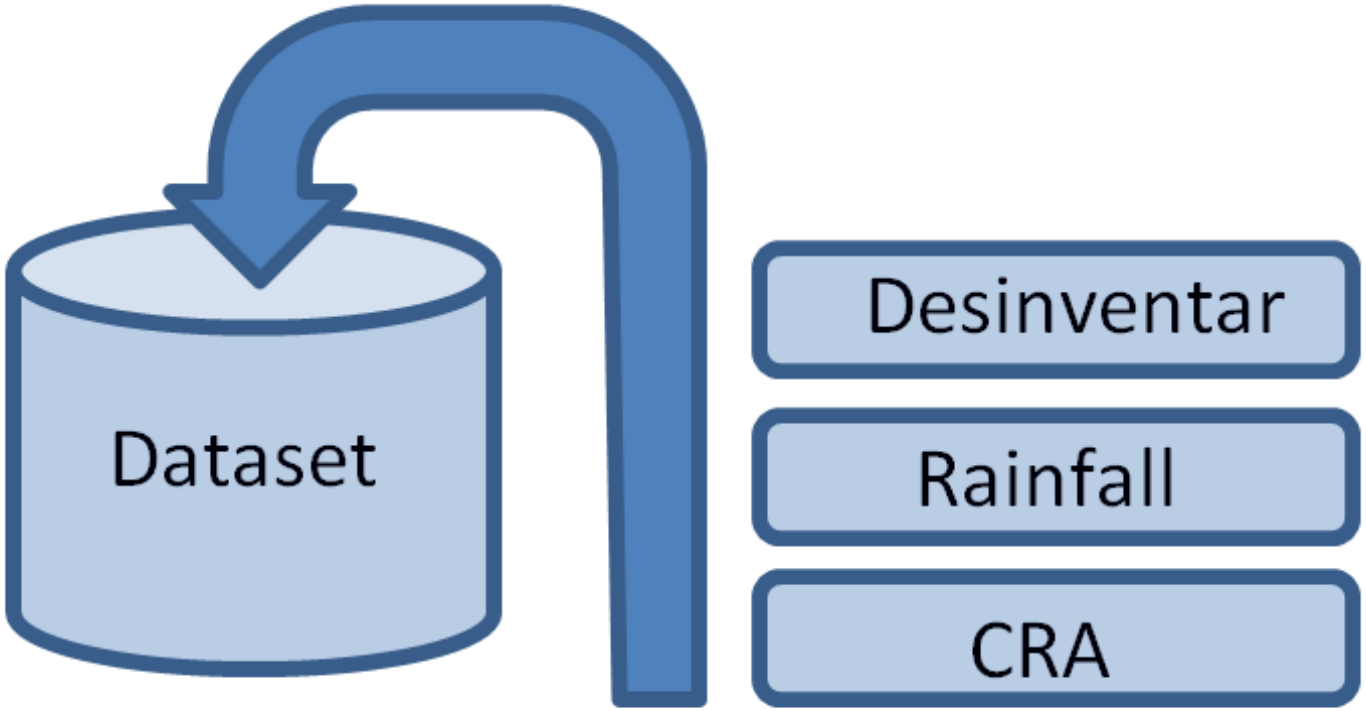
DATA SCIENCE WITH PURPOSE

Research question:

“How accurate can we predict the impact of future floods in Uganda at district-level based on historical data (i.e. historical impact and historical rainfall) and Community Risk Assessment data?”

DATA SCIENCE
WITH PURPOSE

Datasets:



Data preparation:

- **Merge three datasets**

district	date	DI_deaths	DI_injured	DI_.....	RAIN_at_day	RAIN_1day_before	RAIN_...	CRA_employed	CRA_literacy	CRA_...
ABIM	2011-04-11	0	0	....	16.53387	38.02542	....	0.9187	0.5891	....
ABIM	2012-07-23	0	0	....	8.214587	10.58498	....	0.9187	0.5891	....
....	....	....	....	....	....	....	....	....	....	....
ZOMBO	2012-06-09	600	3	....	0.000000	6.254879	....	0.9009	0.5305	....
ZOMBO	2017-04-13	0	0	....	2.145846	32.65487	....	0.9009	0.5305	....

Impact variables
(dependent variables)

Rain variables
(independent variables)

CRA variables
(independent variables)

Data preparation:

- **Aggregate floods in the same district on the same day or within several days**

district	date	DI_deaths	DI_injured	DI_houses_destroyed	....
ABIM	2007-07-29	0	0	0	....
ABIM	2007-07-30	0	0	0	....
ABIM	2007-07-30	300	0	1000	....
ABIM	2007-08-02	600	3	350	....
ABIM	2007-08-02	0	0	0	....
.....	.....	.....	.....	.....	.....

Before aggregation

district	date	DI_deaths	DI_injured	DI_houses_destroyed	....
ABIM	2007-08-02	450	3	675	....
....	....	....	....	....	....

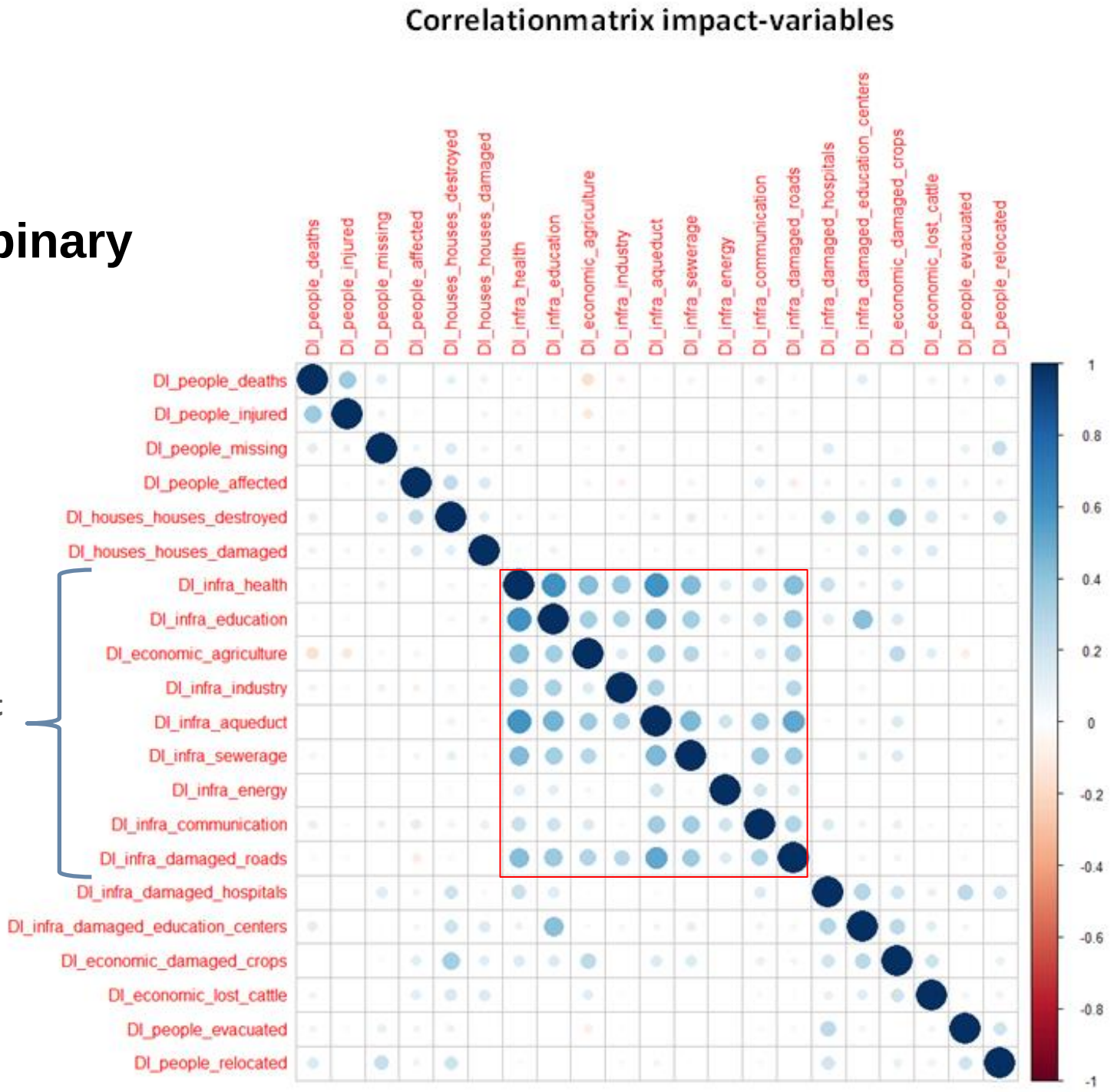
After aggregation

Data preparation:

- **Create one total binary impact variable (impact yes/no)**
- **Based on only the 9 <u>binary</u> impact variables**

Data preparation:

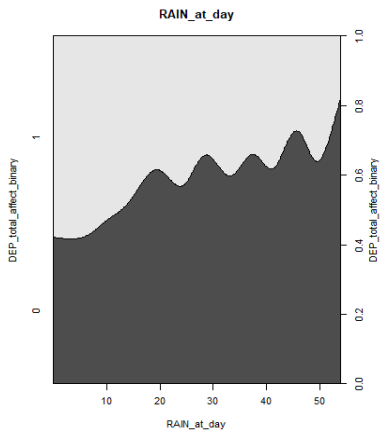
1. **Higher correlation between binary impact variables**



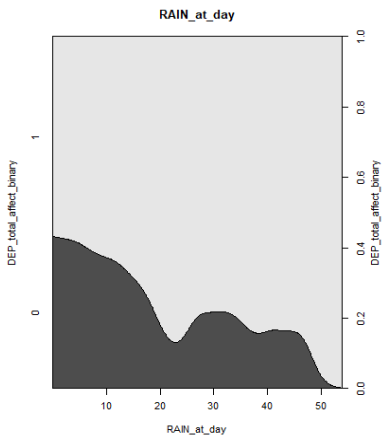
9 binary impact variables

Data preparation:

1. **Higher correlation between binary impact variables**
2. **Positive relationship binary impact variables vs. rainfall**



Total binary impact created based on 12 ***continuous*** impact variables



Total binary impact created based on 9 ***binary*** impact variables

DATA SCIENCE WITH PURPOSE

Data preparation:

1. **Higher correlation between binary impact variables**
2. **Positive relationship binary impact variables vs. rainfall**
3. **More info available for binary impact variables**

Data preparation:

- **Remove independent variables with more than 85% NA's**

Rain variables:
- Rain _at_day
- Rain_1_day_before_cumulative
- Rain_2_days_before_cumulative
- Rain_3_days_before_cumulative
- Rain_4_days_before_cumulative
- Rain_1_day_before
- Rain_2_days_before
- Rain_3_days_before
- Rain_4_days_before
- Rain_5_day_before

CRA variables:
- CRA_violent_incidents
- CRA_drought_exposure
- CRA_earthquake_exposure
- CRA_flood_exposure
- CRA_disability
- CRA_employed
- CRA_literacy
- CRA_mosquito_nets
- CRA_orphans
- CRA_poverty
- CRA_roof_type
- CRA_wall_type
- CRA_subsistence_farming
- CRA_drinking_water
- CRA_educational_facilities
- CRA_time_to_city
- CRA_electricity
- CRA_health_facilities
- CRA_sanitation
- CRA_internet_access
- CRA_mobile_access
- CRA_land_area
- ~~CRA_displaced_persons~~
- ~~CRA_displaced_local_population~~
- CRA_elevation
- CRA_population_density
- CRA_population
- CRA_general_coping
- CRA_general_risk
- CRA_general_hazard
- CRA_general_vulnerability

Data preparation:

- **Remove incorrect variables**

Rain variables:
- Rain _at_day
- Rain_1_day_before_cumulative
- Rain_2_days_before_cumulative
- Rain_3_days_before_cumulative
- Rain_4_days_before_cumulative
- Rain_1_day_before
- Rain_2_days_before
- Rain_3_days_before
- Rain_4_days_before
- Rain_5_day_before

CRA variables:
- CRA_violent_incidents
- CRA_drought_exposure
- CRA_earthquake_exposure
- CRA_flood_exposure
- CRA_disability
- CRA_employed
- CRA_literacy
- CRA_mosquito_nets
- CRA_orphans
- CRA_poverty
- CRA_roof_type
- CRA_wall_type
- CRA_subsistence_farming
- CRA_drinking_water
- CRA_educational_facilities
- CRA_time_to_city
- CRA_electricity
- CRA_health_facilities
- CRA_sanitation
- CRA_internet_access
- CRA_mobile_access
- CRA_land_area
- ~~CRA_displaced_persons~~
- ~~CRA_displaced_local_population~~
- CRA_elevation
- ~~CRA_population_density~~
- ~~CRA_population~~
- CRA_general_coping
- CRA_general_risk
- CRA_general_hazard
- CRA_general_vulnerability

Data preparation:

- **Remove unimportant variables:**
  - **Lasso logistic regression: variables of which coefficients is shrunken to zero**
  - **Stepwise logistic regression: variables not selected by model**
  - **Random forest: variables with lowest mean decrease in accuracy and/or Gini**

Data preparation:

- **Remove unimportant variables**

Rain variables:

- Rain _at_day
- Rain_1_day_before_cumulative
- Rain_2_days_before_cumulative
- Rain_3_days_before_cumulative
- Rain_4_days_before_cumulative
- ~~Rain_1_day_before~~
- ~~Rain_2_days_before~~
- ~~Rain_3_days_before~~
- ~~Rain_4_days_before~~
- ~~Rain_5_day_before~~

CRA variables:

- ~~CRA_violent_incidents~~
- ~~CRA_drought_exposure~~
- ~~CRA_earthquake_exposure~~
- ~~CRA_flood_exposure~~
- CRA_disability
- ~~CRA_employed~~
- CRA_literacy
- CRA_mosquito_nets
- CRA_orphans
- ~~CRA_poverty~~
- CRA_roof_type
- CRA_wall_type
- CRA_subsistence_farming
- ~~CRA_drinking_water~~
- ~~CRA_educational_facilities~~
- ~~CRA_time_to_city~~
- CRA_electricity
- CRA_health_facilities
- ~~CRA_sanitation~~
- ~~CRA_internet_access~~
- CRA_mobile_access
- ~~CRA_land_area~~
- ~~CRA_displaced_persons~~
- ~~CRA_displaced_local_population~~
- ~~CRA_elevation~~
- ~~CRA_population_density~~
- ~~CRA_population~~
- ~~CRA_general_coping~~
- CRA_general_risk
- ~~CRA_general_hazard~~
- CRA_general_vulnerability

Data preparation:

- 1 dependent variable (total binary impact variable)
- 17 independent variables → standardized

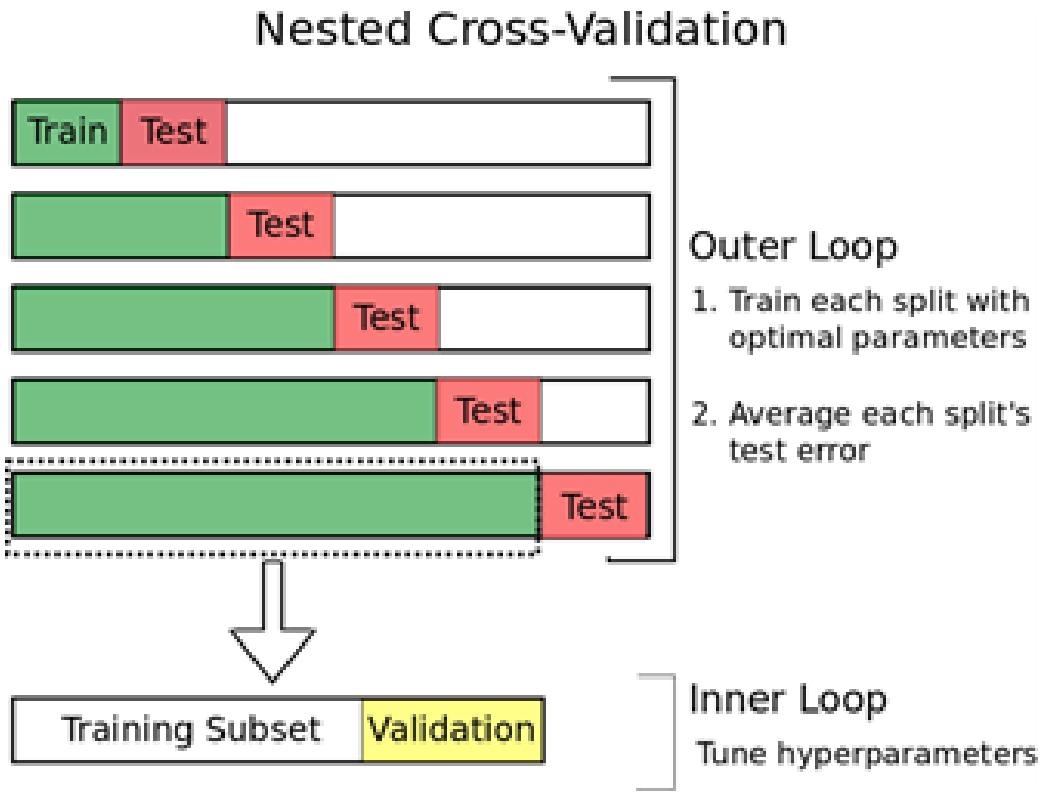
5 rain variables:
- Rain _at_day
- Rain_1_day_before_cumulative
- Rain_2_days_before_cumulative
- Rain_3_days_before_cumulative
- Rain_4_days_before_cumulative

12 CRA variables:
- CRA_disability
- CRA_literacy
- CRA_mosquito_nets
- CRA_orphans
- CRA_roof_type
- CRA_wall_type
- CRA_subsistence_farming
- CRA_electricity
- CRA_health_facilities
- CRA_mobile_access
- CRA_general_risk
- CRA_general_vulnerability

Statistical models:

(Nested) 5-fold cross-validation to get estimates of several performance metrics for 4 different models:

- **Stepwise logistic regression**
- **Lasso logistic regression**
- **Support vector machine (with radial basis kernel)**
- **Random forest**



DATA SCIENCE WITH PURPOSE

Results:

	Stepwise logistic regression	Lasso logistic regression	Support vector machine	Random forest
AUC	0.666	0.675	0.641	0.644
Accuracy	0.675	0.671	0.678	0.652
F1 score	0.774	0.780	0.794	0.748

	Actual: no impact (0)	Actual: impact (1)
Predicted: no impact (0)	13	10
Predicted: impact (1)	28	65

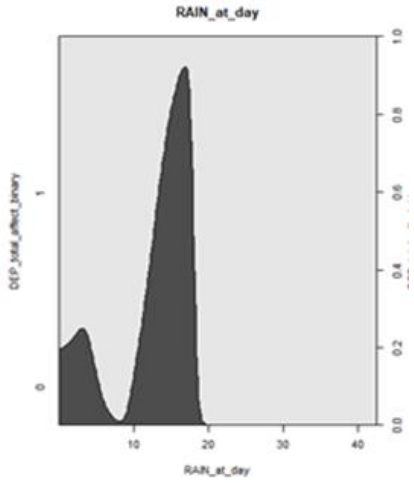
	Actual: no impact (0)	Actual: impact (1)
Predicted: no impact (0)	10	7
Predicted: impact (1)	31	67

	Actual: no impact (0)	Actual: impact (1)
Predicted: no impact (0)	6	3
Predicted: impact (1)	35	72

	Actual: no impact (0)	Actual: impact (1)
Predicted: no impact (0)	15	15
Predicted: impact (1)	25	60

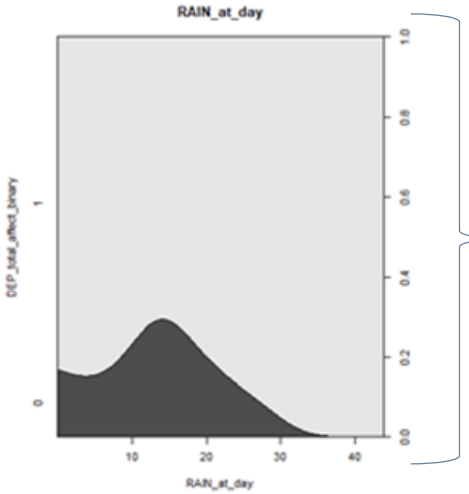
Future improvements:

- **Mean rainfall per catchment area of a district**



Rainfall predictor(s) created based on the mean rainfall per catchment area of a district

R = 0.13



Rainfall predictor(s) created based on the mean rainfall per district

R = 0.09

DATA SCIENCE WITH PURPOSE

Future improvements:

- Mean rainfall per catchment area of a district
- Add GloFAS dataset
- Add more accurate impact data
- Predict different impact variables (i.e. related to people, houses etc.)
- Create total impact variable based on expertise knowledge
- Select most important variables based on expertise knowledge
- Resample the minority class (= no impact)
- Tune the parameters of the models (even further)
- Make R-script more reproducible

Future improvements:

- Mean rainfall per catchment area of a district ✓
- Add GloFAS dataset ✓
- Add more accurate impact data
- Predict different impact variables (i.e. related to people, houses etc.)
- Create total impact variable based on expertise knowledge
- Select most important variables based on expertise knowledge
- Resample the minority class (= no impact)
- Tune the parameters of the models (even further)
- Make R-script more reproducible ✓

Questions…?