OPEN DATA SCIENCE CONFERENCE

#ODSC

@ODSC

Boston | April 30 - May 4, 2019

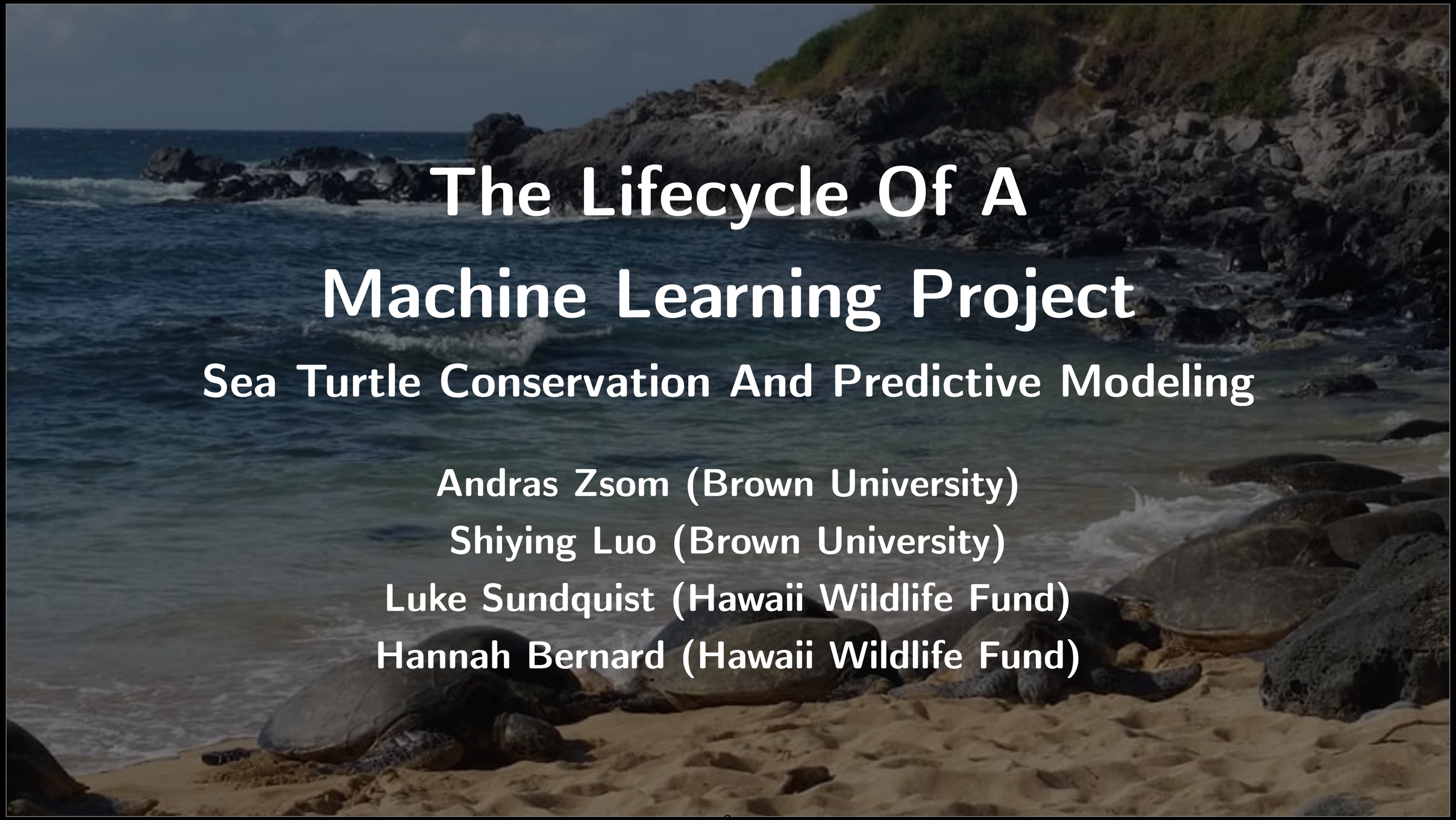
#ODSC

BOSTON
APR 30 – MAY 3

The Lifecycle Of A Machine Learning Project: Sea Turtle Conservation And Predictive Modeling

Andras Zsom, PhD

Lead Data Scientist,
Advanced Research Computing at CCV, Brown University



The Lifecycle Of A Machine Learning Project

Sea Turtle Conservation And Predictive Modeling

Andras Zsom (Brown University)
Shiying Luo (Brown University)
Luke Sundquist (Hawaii Wildlife Fund)
Hannah Bernard (Hawaii Wildlife Fund)

About me

- Born and raised in Hungary
- Astrophysics PhD at MPIA, Heidelberg, Germany
- Postdoctoral researcher at MIT (still in astrophysics at the time)
- Started at Brown in December 2015 as a data scientist

Data Science at Brown

- Center for Computation and Visualization
- Institutional Data group
    - Data-driven decision support and predictive modeling for Brown's administrative units
    - Academic research on data-intensive projects

Learning Objectives

- By the end of this tutorial, you will be able to
    - combine various data sources,
    - perform feature engineering on time series data,
    - select a cross validation method most appropriate to your problem,
    - develop an interpretable model and maximize your actionable/scientific insights.

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- Data collection and preprocessing
- Data exploration and feature generation
- Cross validation
- Results
- Lessons learnt

Overview

- **Project background and the Hawaii Wildlife Fund (HWF)**
- Problem description
- Data collection and preprocessing
- Data exploration and feature generation
- Cross validation
- Results
- Lessons learnt

How this project came about?

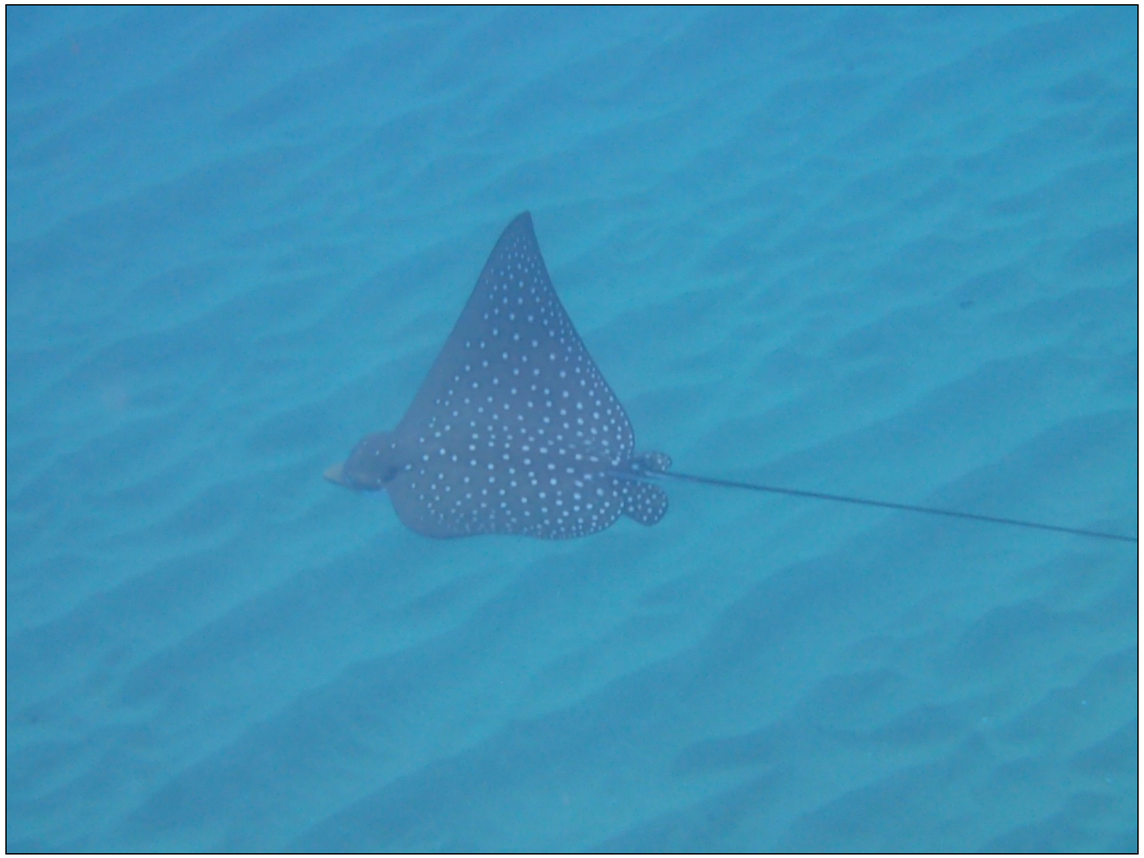
Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui



How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

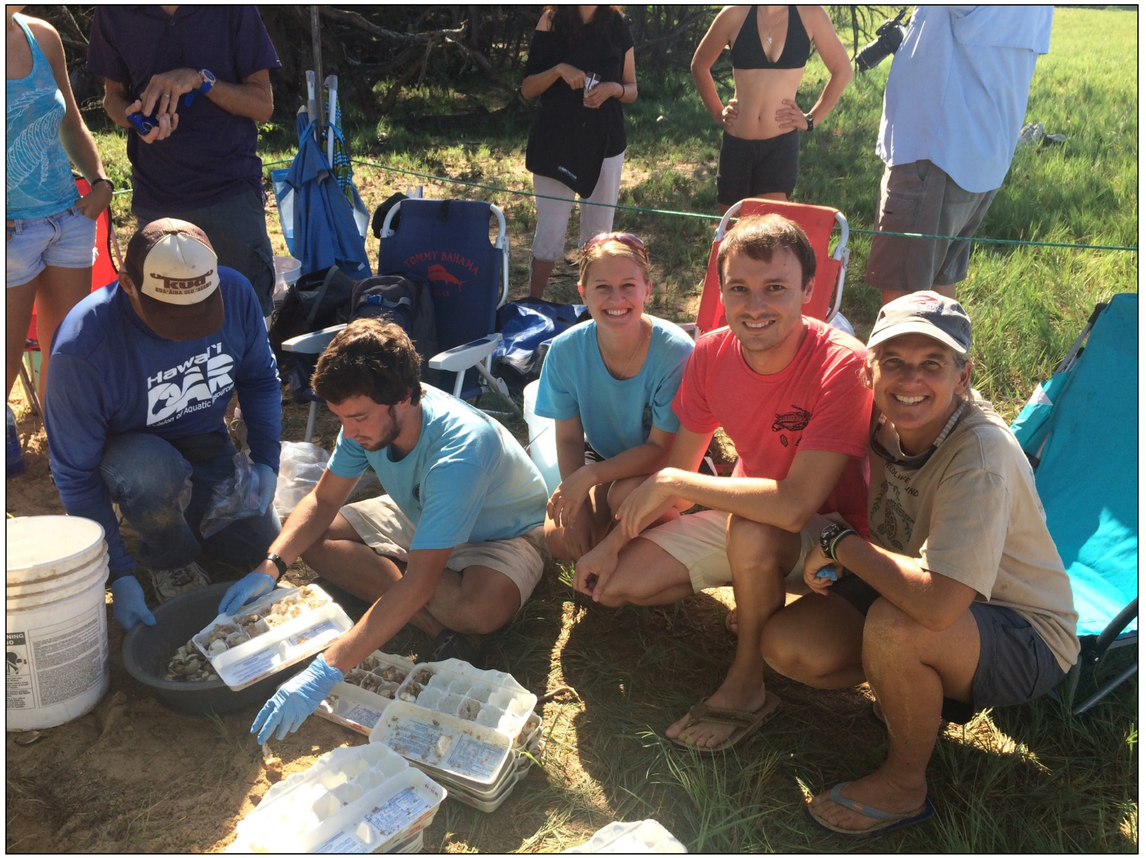
Vacation on Kauai and Maui

How this project came about?

Vacation on Kauai and Maui

How this project came about?

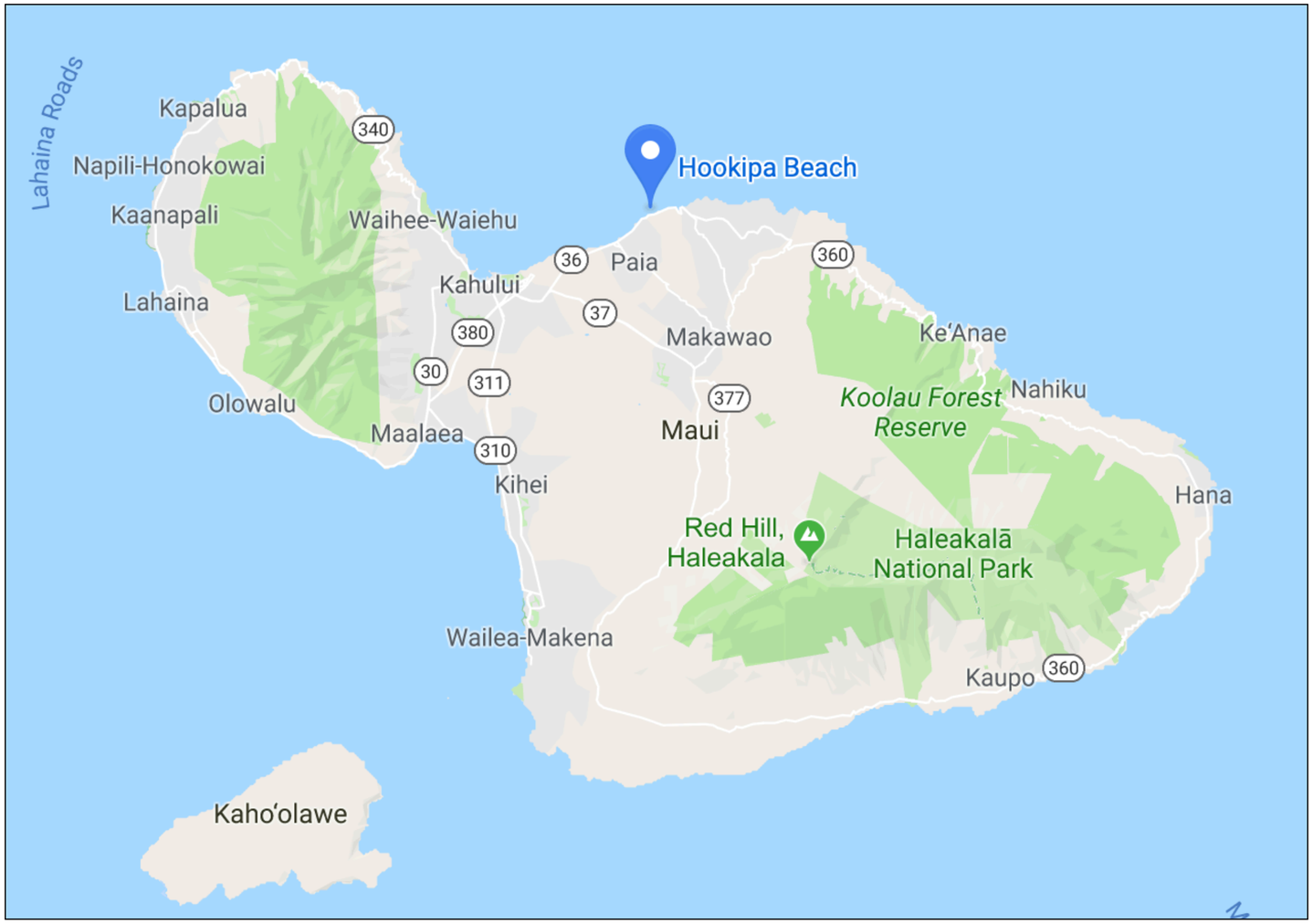
Vacation on Kauai and Maui



Hawaii Wildlife Fund (HWF)

- Honu watch program (Honu = green sea turtle)
- Increased basking behavior on select beaches
- Popular wildlife experience on Maui, distraction to lifeguards
- Educate visitors, manage disturbance
- Count the basking turtles (#**T**) and the visitors (#**H**) every 30 min from 2pm to sunset - for years!

Hawaii Wildlife Fund (HWF)



Hawaii Wildlife Fund (HWF)

Why do the turtles bask?

- Unclear
- Several theories: avoid predators, thermoregulation, helps metabolism and digestion (Whittow and Balazs, 1982)
- Previous studies on Galapagos Islands and Hawaii: thermoregulation is likely (Maxwell *et al.* 2014, Van Houtan, Halley, and Marks 2015)

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- **Problem description**
- Data collection and preprocessing
- Data exploration and feature generation
- Cross validation
- Results
- Lessons learnt

Science Goals

- **Data:** date time, #T, #H
- Generally improve our understanding of the basking behavior
- What environmental variables influence it?
- Are there differences between the basking behavior of Hawaiian and Galapagos turtles?

Management Goals

- **Data:** date time, #T, #H
- Can we help HWF better manage their human resources?
- Are there peaks in turtle and human counts requiring increased HWF presence on the beach?
- Larger presence on the beach at those times

Approach

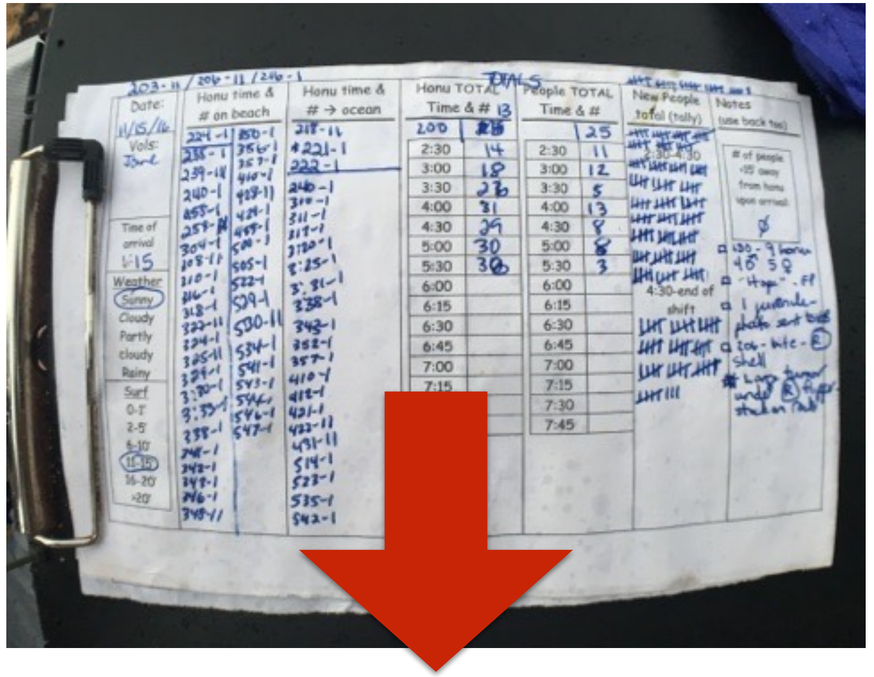
Develop regression models to predict #T and #H one day ahead

- If model has predictive power, we look into the ML black box
- Interpretability is key for science and to build trust in the model
- Using python (numpy, scipy, pandas, sklearn, matplotlib)

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- **Data collection and preprocessing**
- Data exploration and feature generation
- Cross validation
- Results
- Lessons learnt

HWF data



DATE	#of vols	Arrival time	Weather	Surf	# of honu at arrival	Time of first Honu	2:30	3:00	3:30	4:00	4:30	5:00
6/1/15	5	2:30	Cloudy, Sunny	2-5'	6	2:45	6	9	9	10	10	9
6/2/15	2											
6/3/15	1	2:30	Cloudy	0-1'	15	2:50	15	14	14	15	15	15
6/4/15	2	3:00	Sunny	0-1'	3	4:28	n/a	3	3	3	3	3
6/5/15	2	2:00	sunny	1-10'	10	32 2:31	10	6	6	6	7	7

HWF data

- Reformat data to be better suited for ML
  - date time, nr of volunteers, #T, #H
  - Drop other fields
- Check for outliers to fix obvious typos

DateTime	Nr_Volunteers	TurtleNumber	PeopleNumber
12/3/13 14:30	3	7	2
12/3/13 15:00	3	7	6
12/3/13 15:30	3	14	13
12/3/13 16:00	3	15	13
12/3/13 16:30	3	16	15

External data

- **Atmospheric properties:** temperature, relative humidity, winds
- **Sea properties:** tides, waves
- **Tourism data:** daily number of incoming tourists
- No data on biological environment!

External data

Ho'okipa beach

(turtles)

Kahului Airport

(weather station)

Kahului Harbor

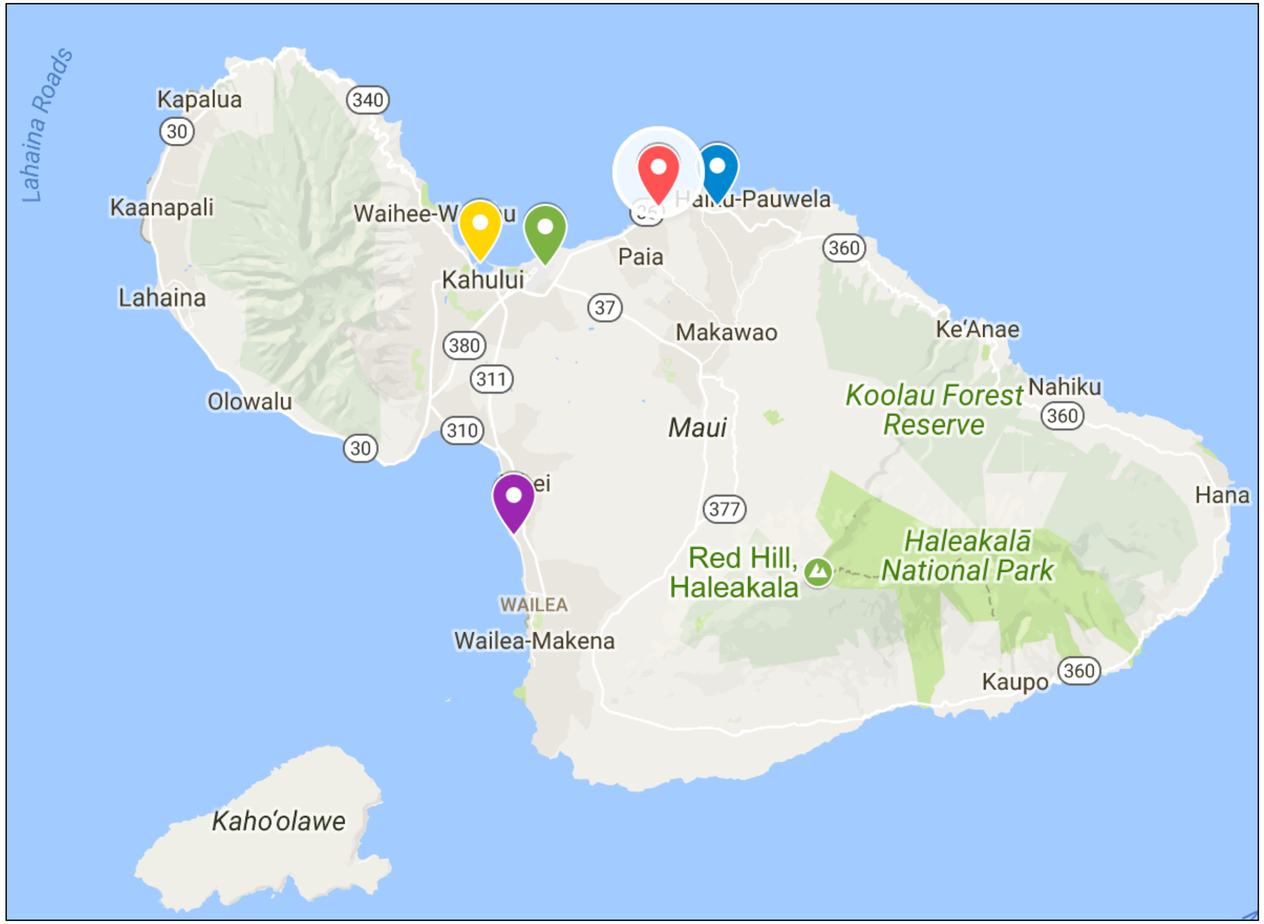
(tide data)

Pauwela

(wave data)

Kalama Park

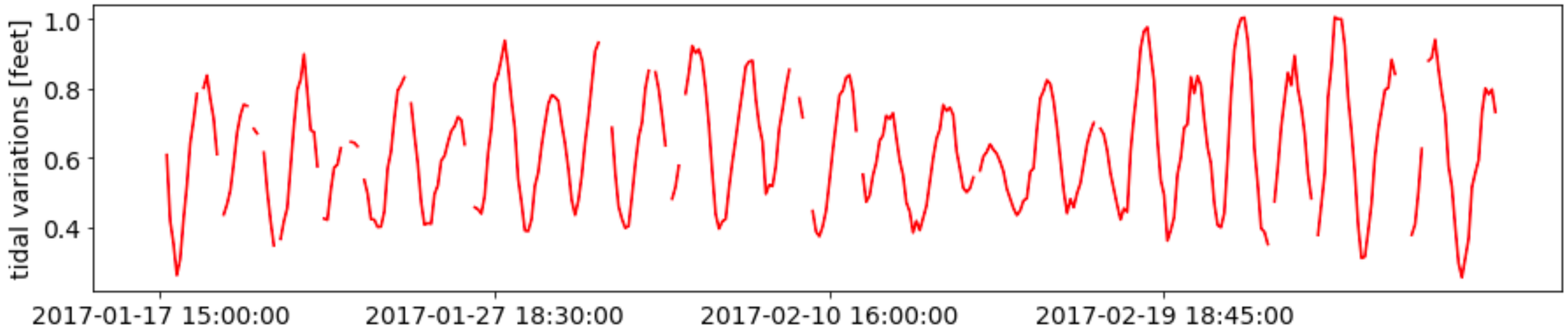
(sea data)



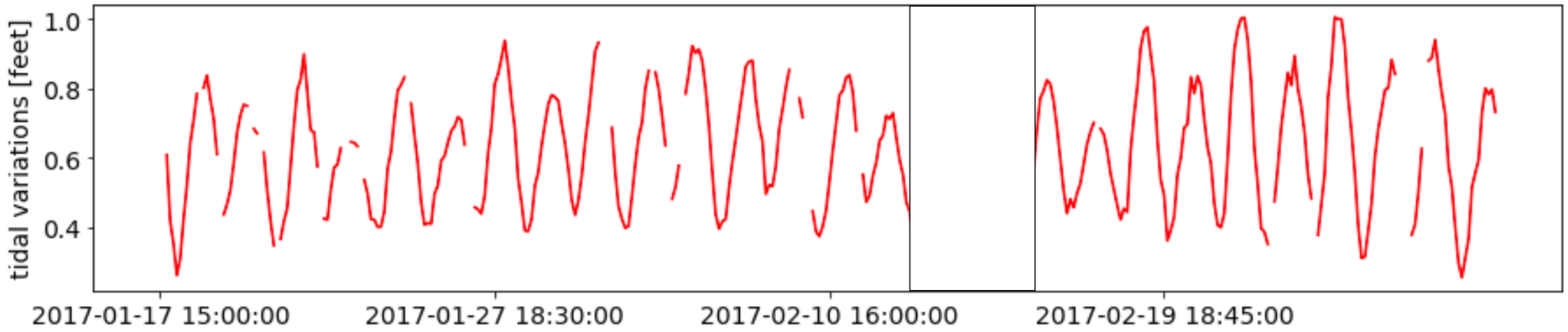
External data

- External data measured at different times than #**T** and #**H**
  - Interpolate to #**T** and #**H** date times
- External data can be missing for some period of time
  - Interpolate if ***dt*** of the missing period is short
  - What to do with missing values?

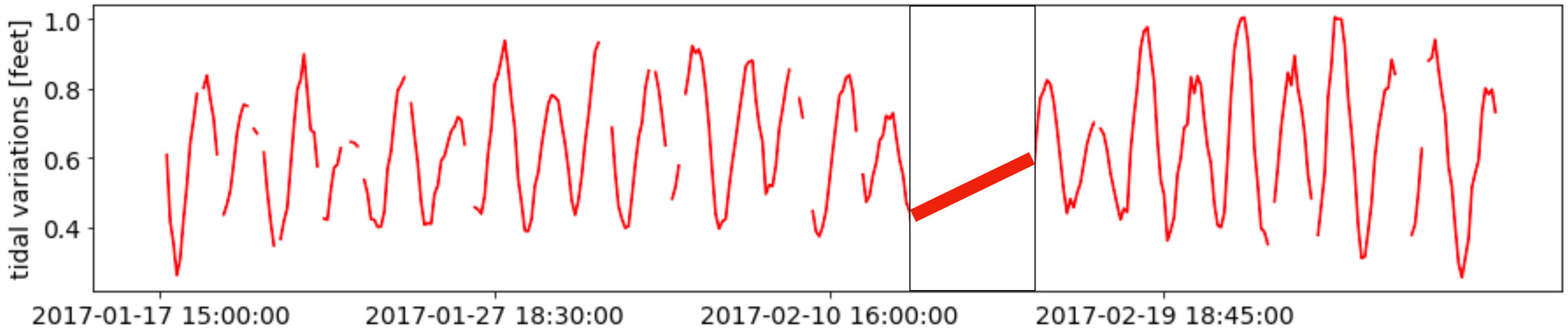
Interpolation



Interpolation



Interpolation



External data

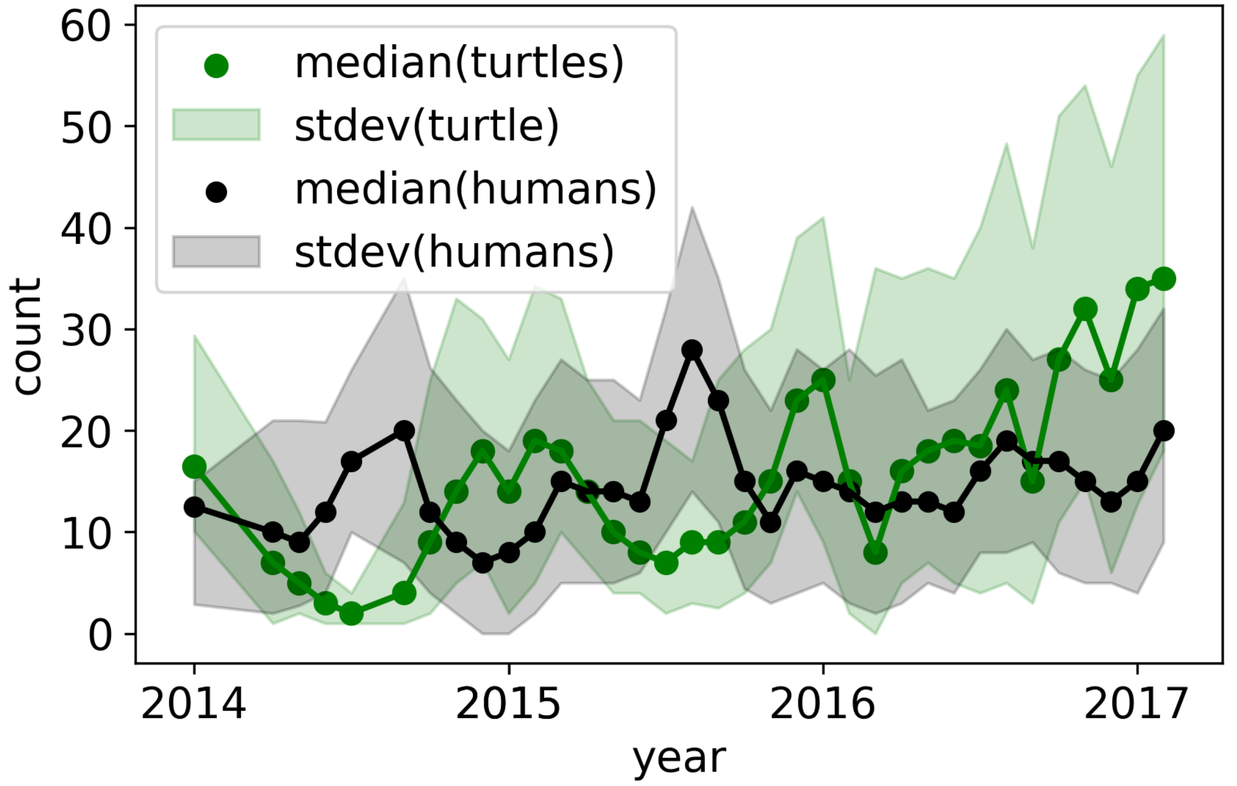
```
from scipy.interpolate import interp1d
from bisect import bisect

def interpolate(t_HWF,t_ext,y_ext,max_dt=2): # max_dt in  hours

    f = interp1d(t_ext,y_ext,kind='linear') # linear interpolation
    # f(t) - the value of the external variable at time t

    y_HWF = [] # collect result

    for t in t_HWF:
        indx = bisect(t_ext,t) # index where t should be inserted
        after = t_ext[indx]
        before = t_ext[indx-1]
        dt = np.abs(after - before)/3.6e12 # convert nsec to hour
        if dt <= max_dt:
            y_HWF.append(f(t)) # add interpolated value
        else:
            y_HWF.append(np.nan) # add nan

    return np.array(y_HWF) # return a numpy array
```

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- Data collection and preprocessing
- **Data exploration and feature generation**
- Cross validation
- Results
- Lessons learnt

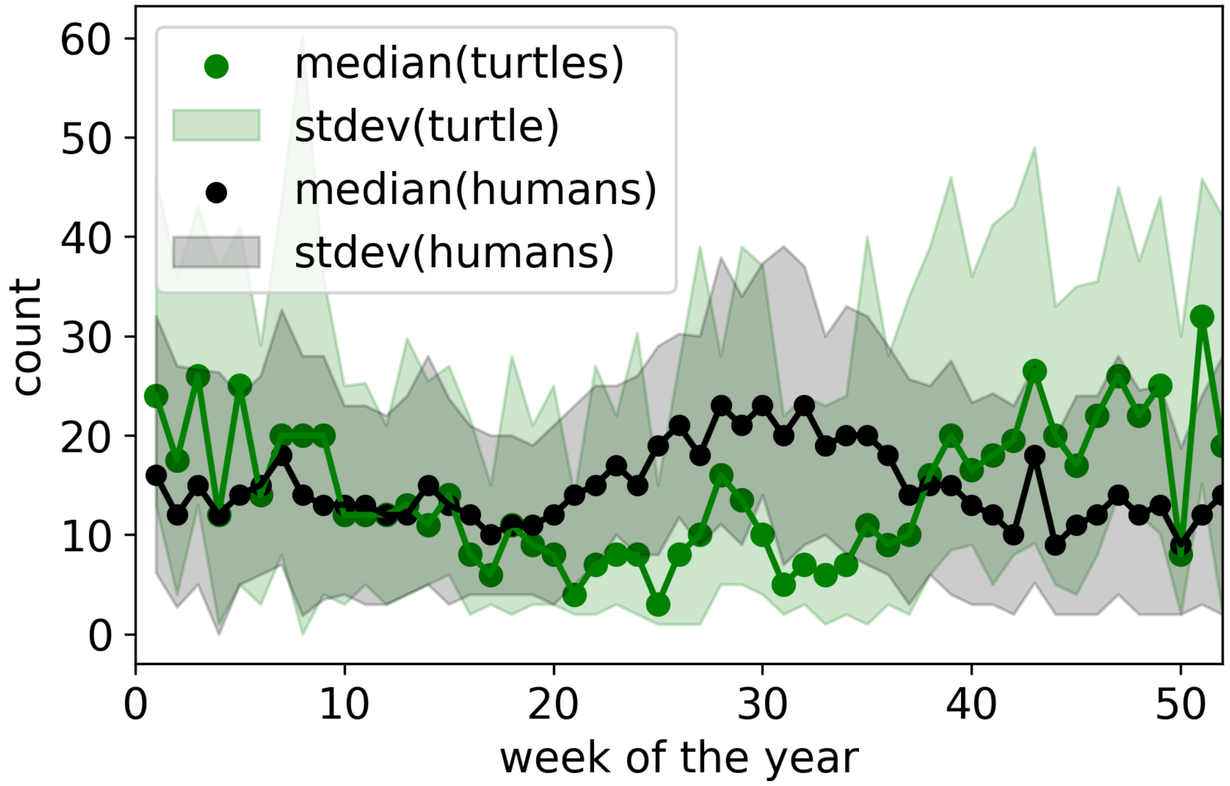
Target variables

More and more turtles over the years!



Target variables

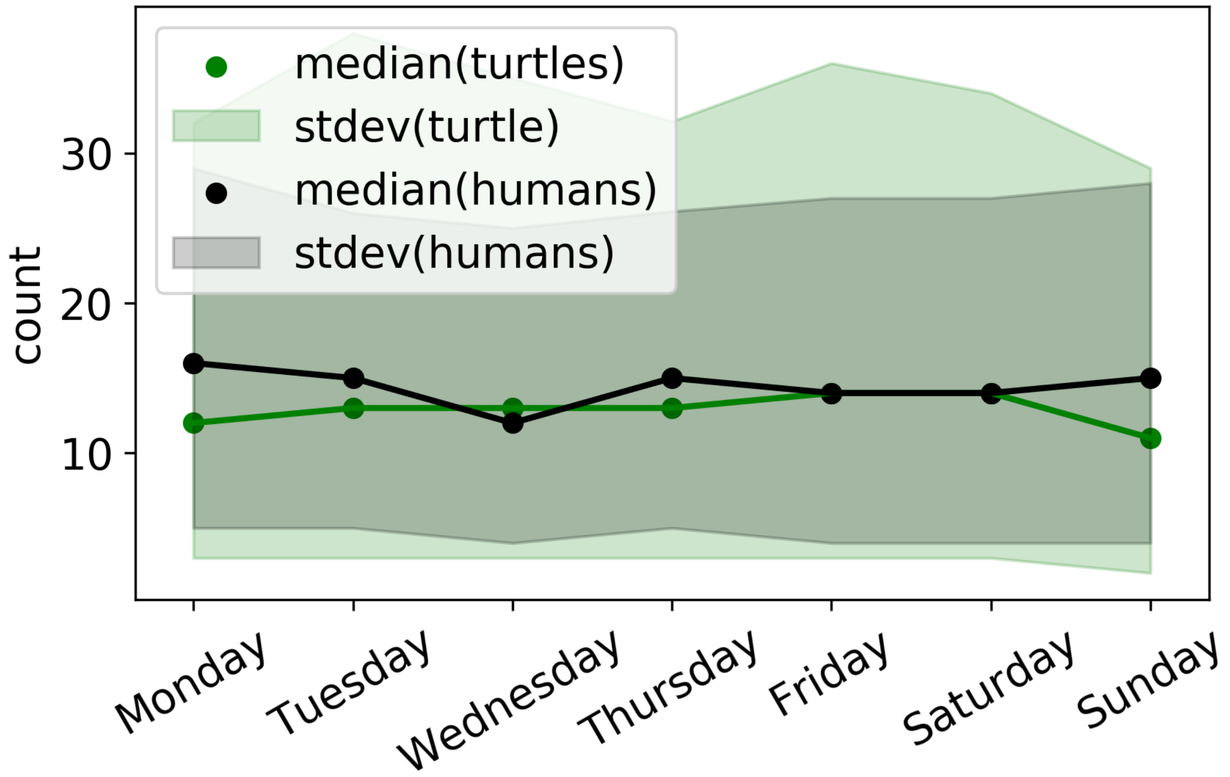
More humans during the summer, more turtles during the winter.



Target variables

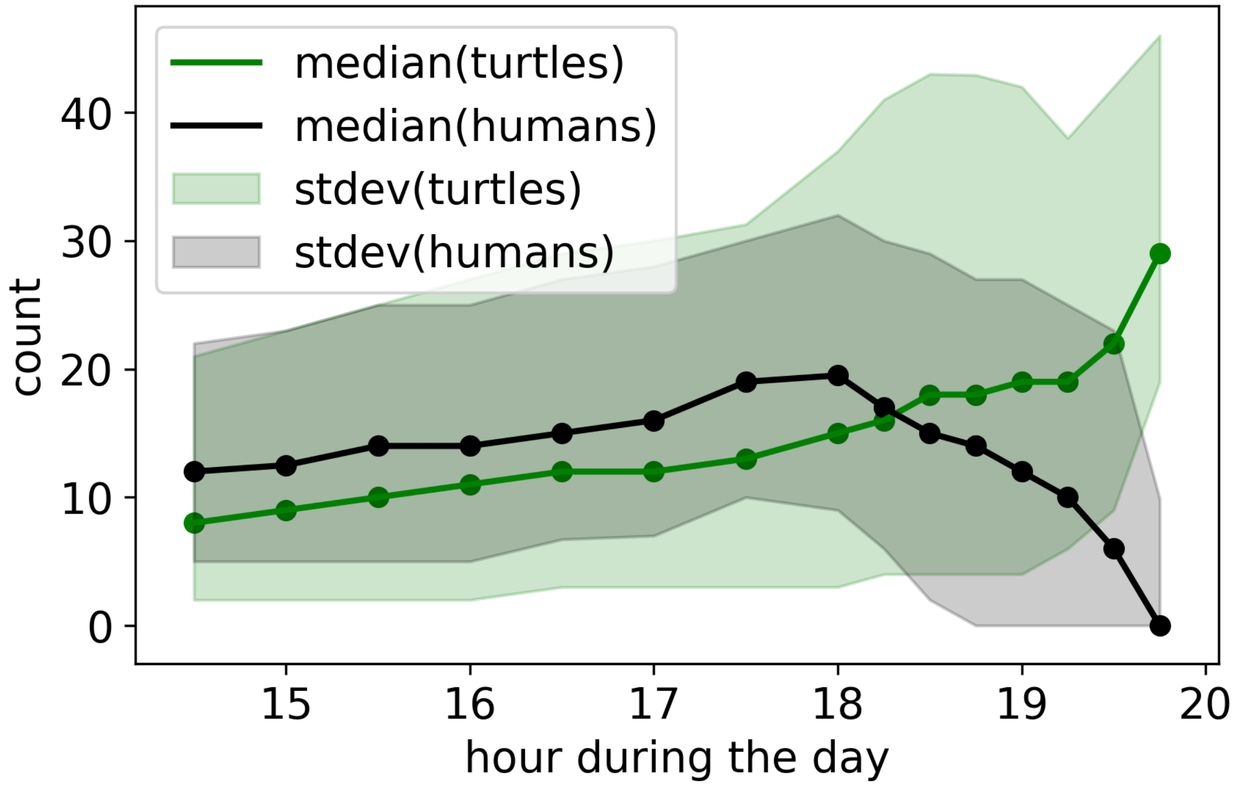
Human presence is not stronger during the weekend.

Most beachgoers are tourists?

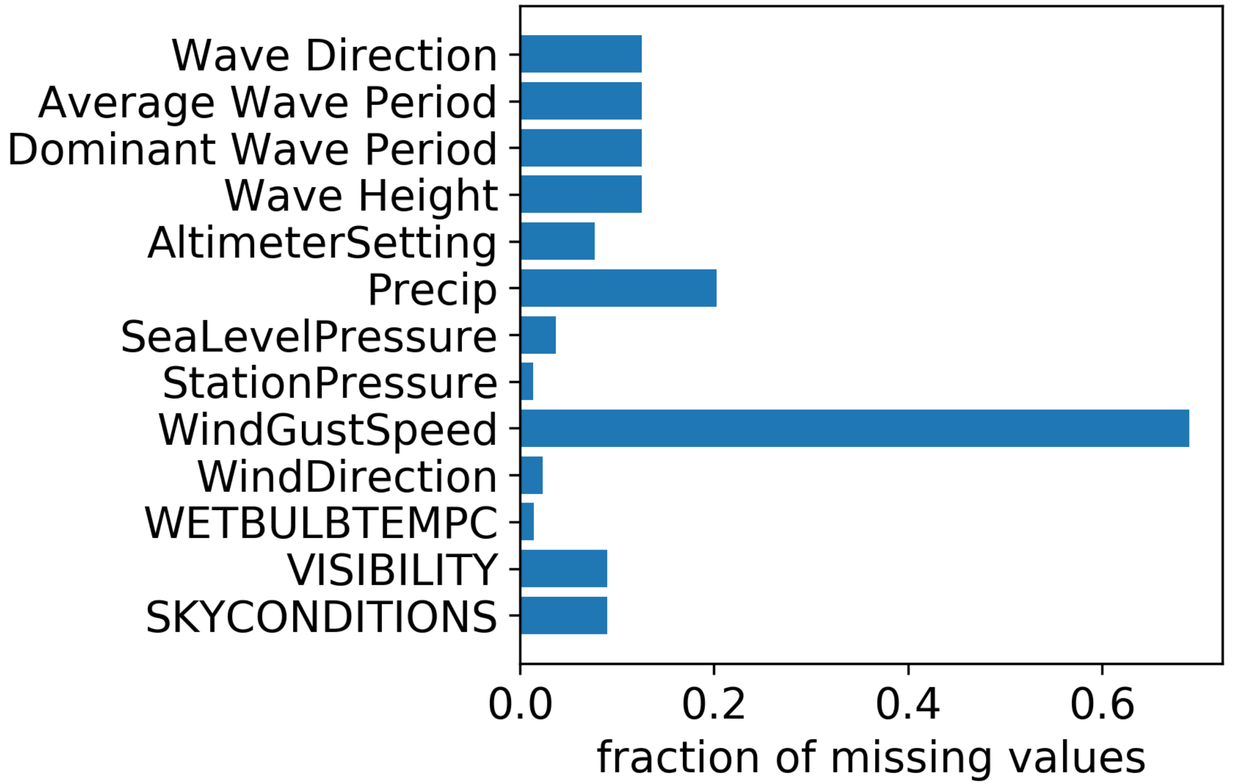


Target variables

Turtles come to shore towards the end of the day.

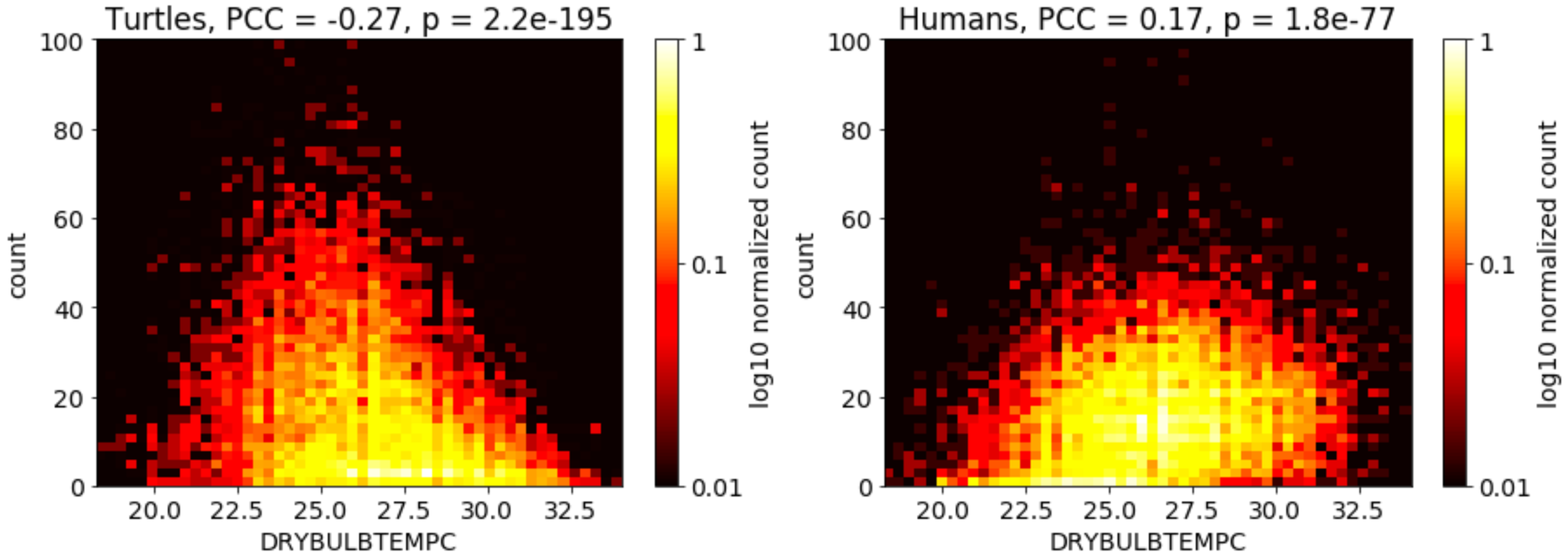


Missing values



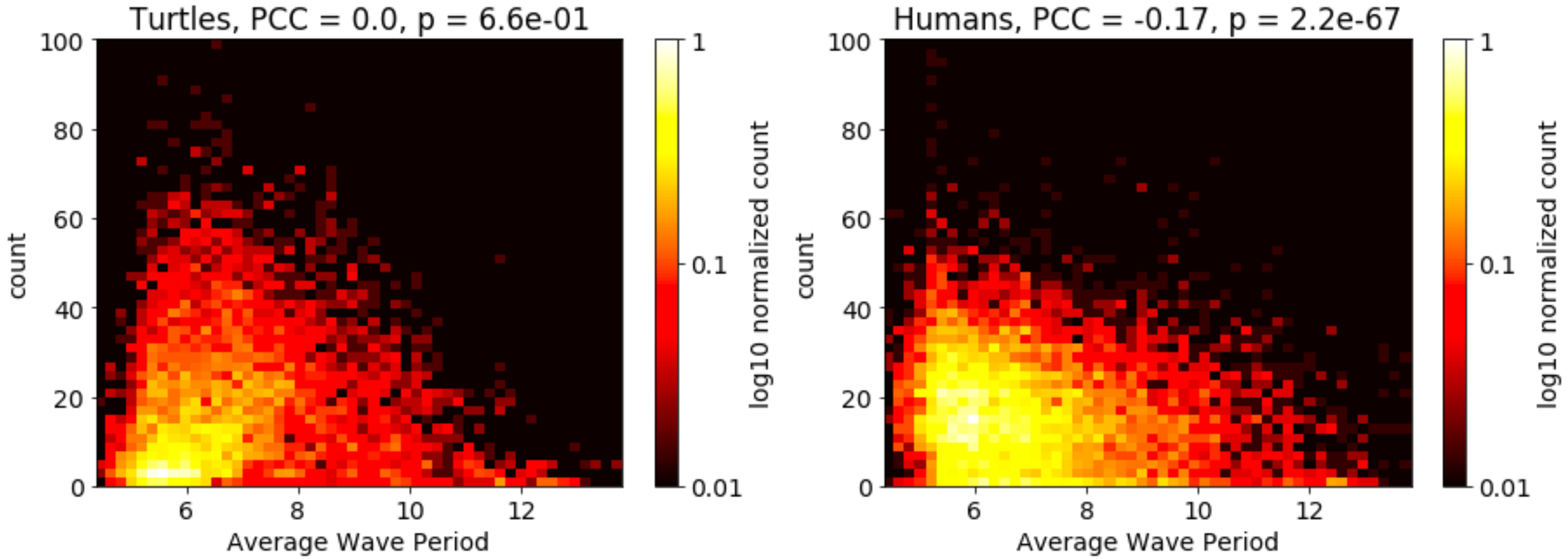
Correlations

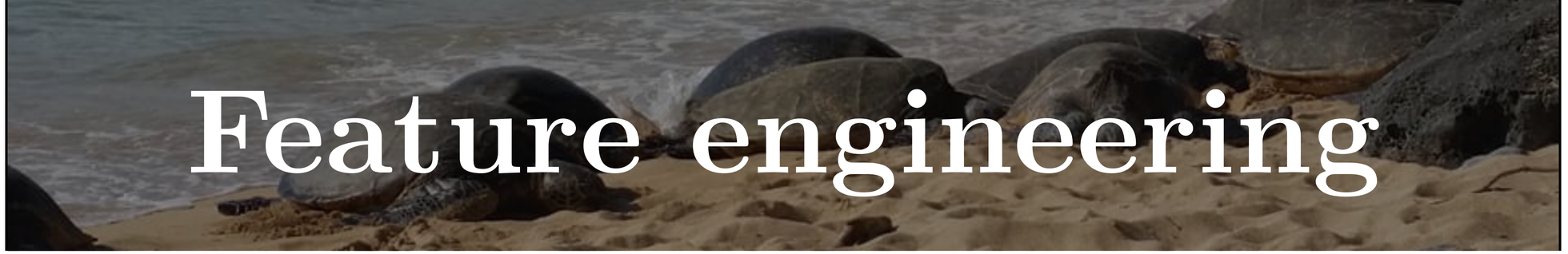
With 11000 points, any correlation with |PCC| $>$ 0.03 is significant with p $<$ 0.01.



Correlations

Turtles don't care much about the waves, but the humans do!



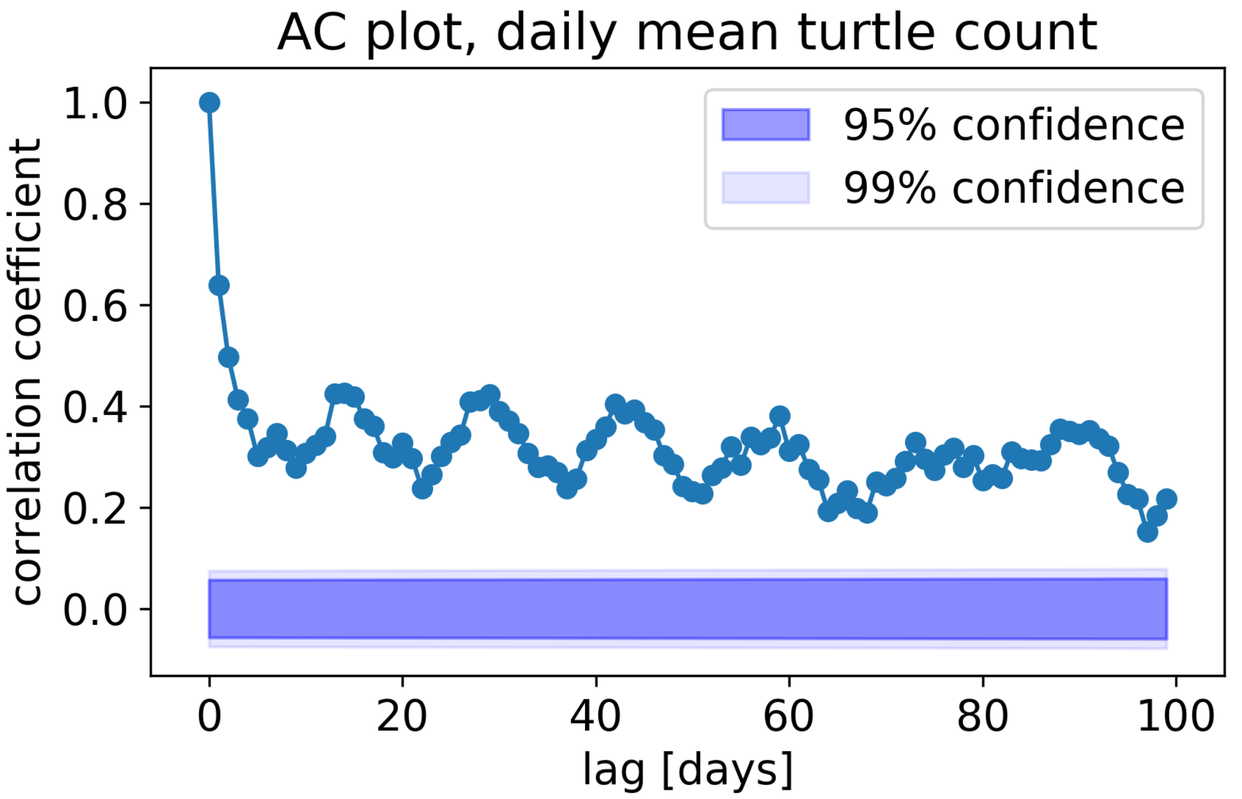


Feature engineering

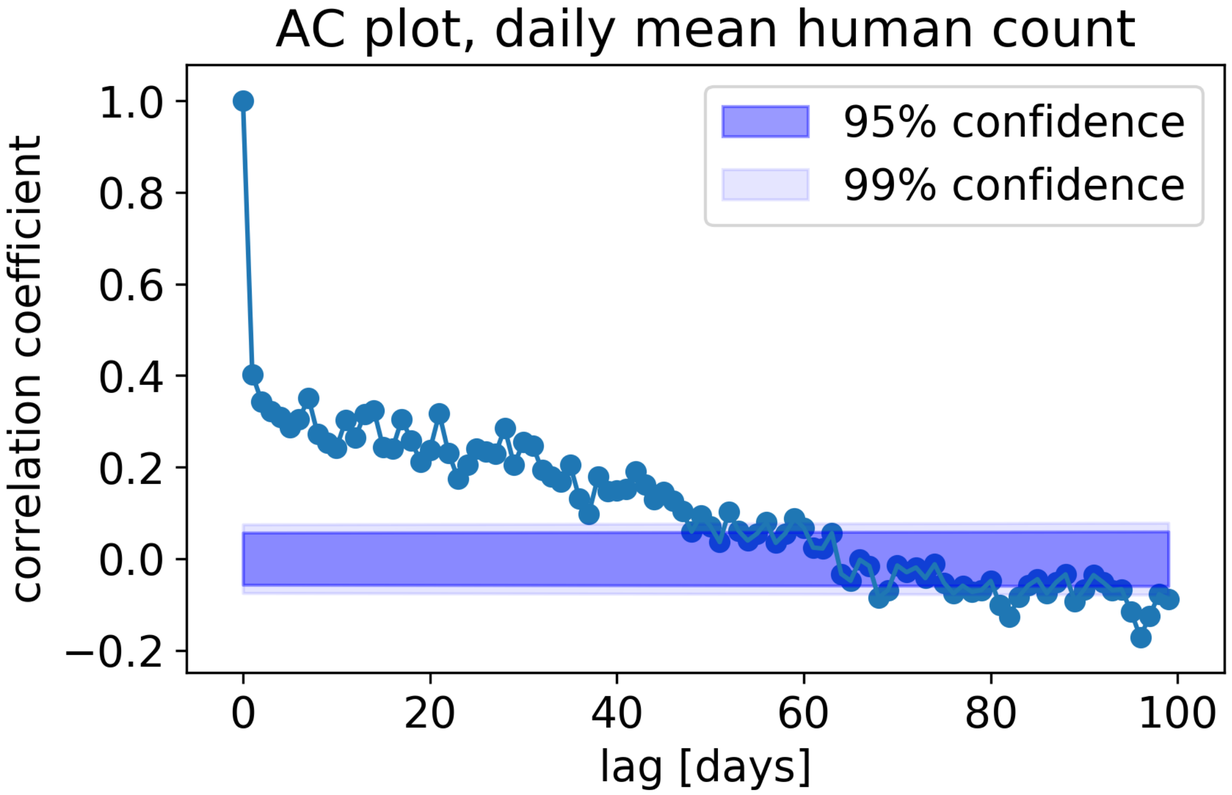
Develop regression models to predict #T and #H one day ahead

- Potentially predictive features are #T and #H a couple of days before.
- Let's look at autocorrelation!

Autocorrelation



Autocorrelation



Final feature set

- Atmospheric properties: 13
- Sea properties: 9 (tides) + 4 (waves)
- Tourism data: 35
- Autocorrelation features: 10 (turtles) + 4 (humans)
- Date and time features: 7
- Other: 1 (Nr. Of Volunteers)

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- Data collection and preprocessing
- Data exploration and feature generation
- **Cross validation**
- Results
- Lessons learnt

Cross Validation

Always mimic the intended use of the model!

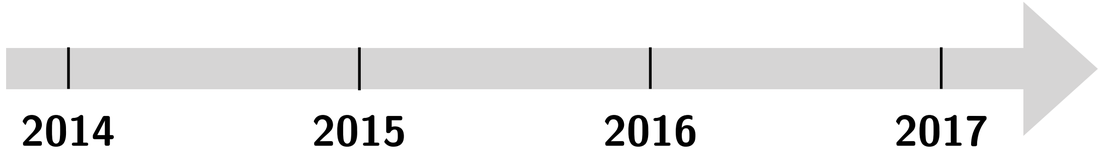
Develop regression models to predict #**T** and #**H** one day ahead

- Random splitting gives high R^2 (or low RMSE), but it does not generalize to new data
- Split data based on days!

Cross Validation

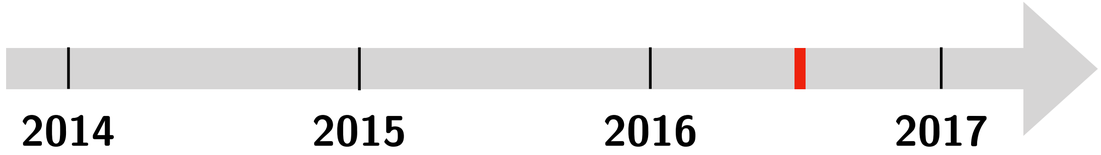
- **Should I use all the data I have to train models?**
- No
- The number of turtles increase over time.
- The model might be better, if only recent observations were used.
- How recent? - A parameter to tune!

Cross Validation



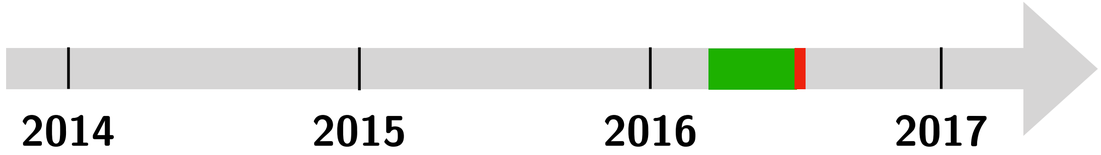
- Decide how many consecutive days (n) to use

Cross Validation



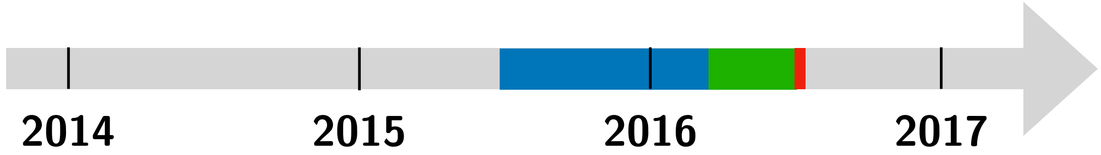
- Decide how many consecutive days (n) to use
- Randomly select one day - **holdout**

Cross Validation



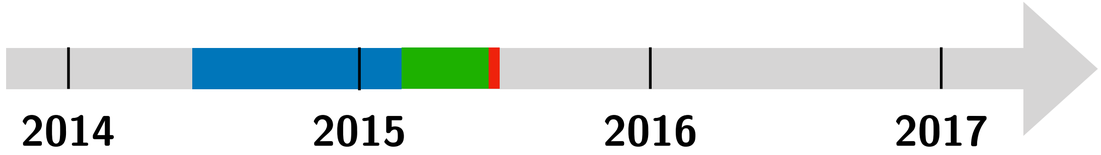
- Decide how many consecutive days (n) to use
- Randomly select one day - **holdout**
- 30% of days before holdout is **test**

Cross Validation



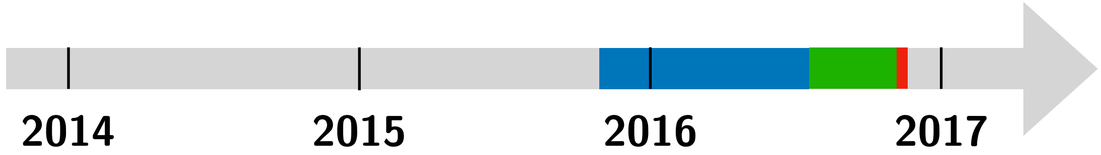
- Decide how many consecutive days (n) to use
- Randomly select one day - **holdout**
- 30% of days before holdout is **test**
- Rest (70% of days) is **train**

Cross Validation



- Decide how many consecutive days (n) to use
- Randomly select one day - **holdout**
- 30% of days before holdout is **test**
- Rest (70% of days) is **train**
- Repeat several times to do uncertainty estimates

Cross Validation



- Decide how many consecutive days (n) to use
- Randomly select one day - **holdout**
- 30% of days before holdout is **test**
- Rest (70% of days) is **train**
- Repeat several times to do uncertainty estimates

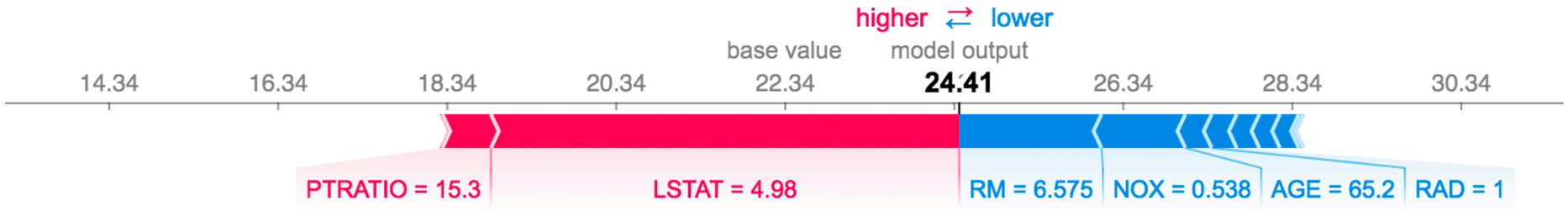
ML Pipeline

- Chen et al., 2016: **XGBoost**: A Scalable Tree Boosting System
  - No need to impute missing values
  - SHapley Additive exPlanation - SHAP values
- Evaluation Metric: RMSE
- Calculate and plot coefficient of determination R^2

SHAP values

- Lundberg et al 2019: Consistent Individualized Feature Attribution for Tree Ensembles

- Global and local feature importances



ML Pipeline

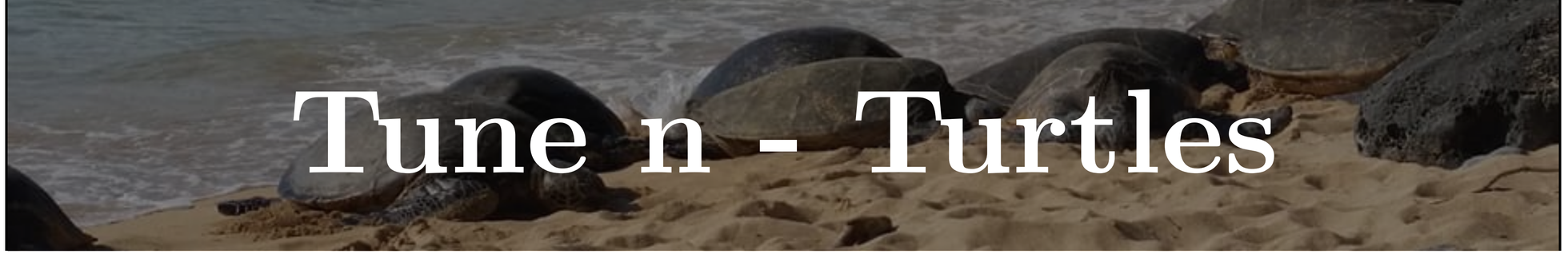
```
param_grid = {"learning_rate": [0.03],
              "objective":['reg:linear'],
              #"reg_alpha": [0e0,0.1,0.3,1e0,3e0,10e0],
              #"reg_lambda": [0e0,0.1,0.03,1e0,3e0,10e0],
              #"max_depth": [2,3,5,7],
              "n_jobs": [-1],
              "colsample_bytree": [0.7, 0.8, 0.9],
              "subsample": [0.6, 0.8],
              "missing": [np.nan],
              "n_estimators": [2000]}
```

ML Pipeline

```
def xgb_cv(X_tr,Y_tr,X_t,Y_t,X_h,Y_h):

    test_score = np.zeros(len(ParameterGrid(param_grid)))

    regressors = []
    for i in range(len(ParameterGrid(param_grid))):
        XGB = xgboost.XGBRegressor()
        g = ParameterGrid(param_grid)[i] # XGBoost parameter grid
        XGB.set_params(**g)
        XGB.fit(X_tr,Y_tr,early_stopping_rounds=50, eval_metric='rmse', \
                eval_set=[(X_t, Y_t)], verbose=False) # early stopping
        test_score[i] = XGB.best_score # collect test scores
        regressors.append(XGB)

    indx = np.where(test_score == np.min(test_score))[0][0]
    best_params = np.array(ParameterGrid(param_grid))[indx]
    XGB = regressors[indx]

    Y_pred = XGB.predict(X_h,ntree_limit=XGB.best_ntree_limit)
    X_shap = XGB.get_booster().predict(xgboost.DMatrix(X_h), \
                pred_contribs=True,ntree_limit=XGB.best_ntree_limit)

  return best_params,test_score[indx],regressors[indx],Y_pred,X_shap
```

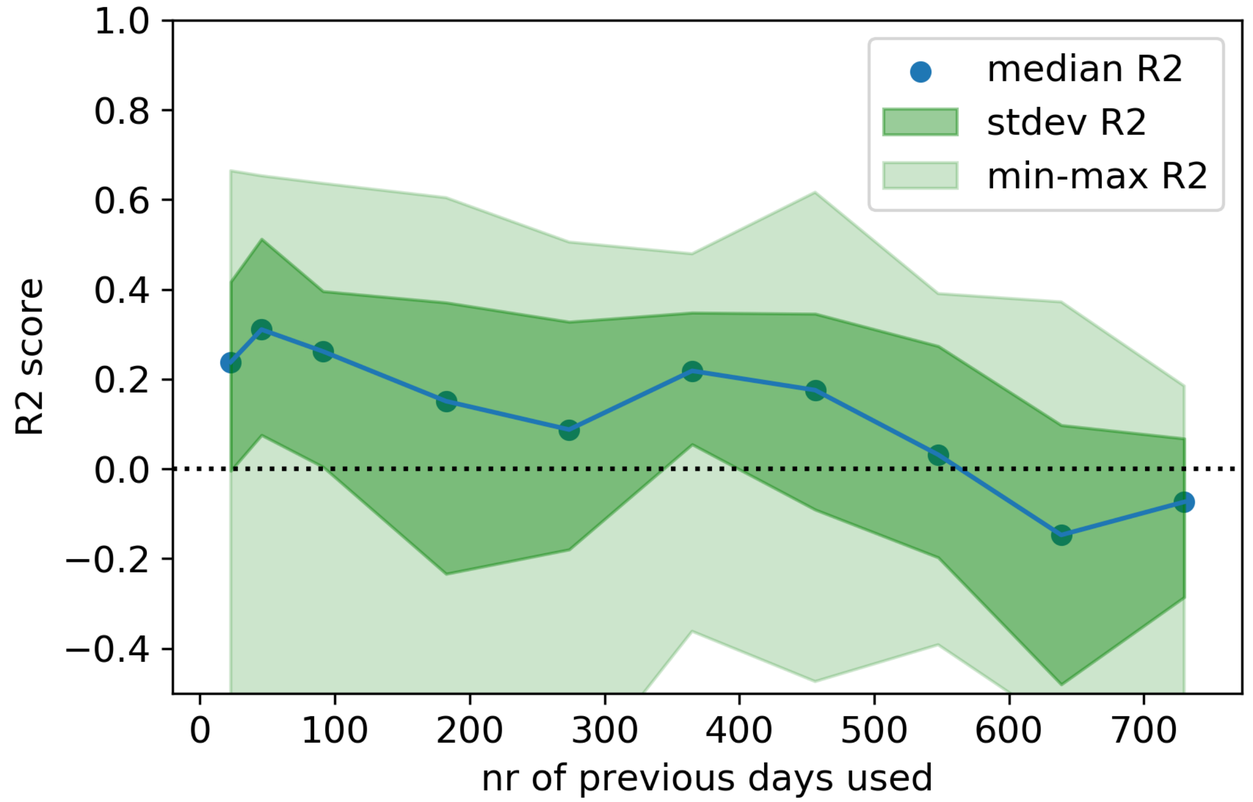
Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- Data collection and preprocessing
- Data exploration and feature generation
- Cross validation
- **Results**
- Lessons learnt

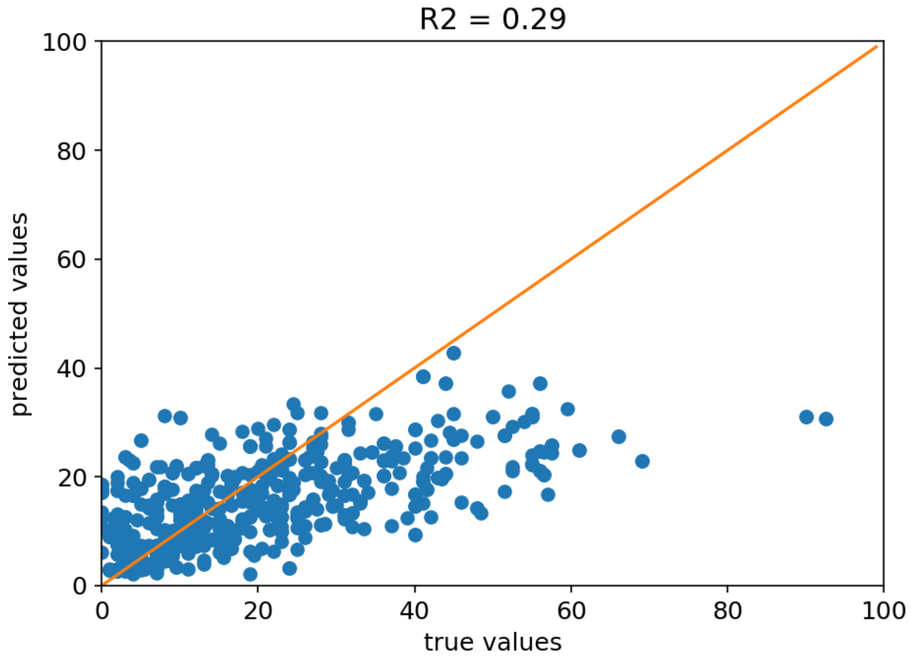


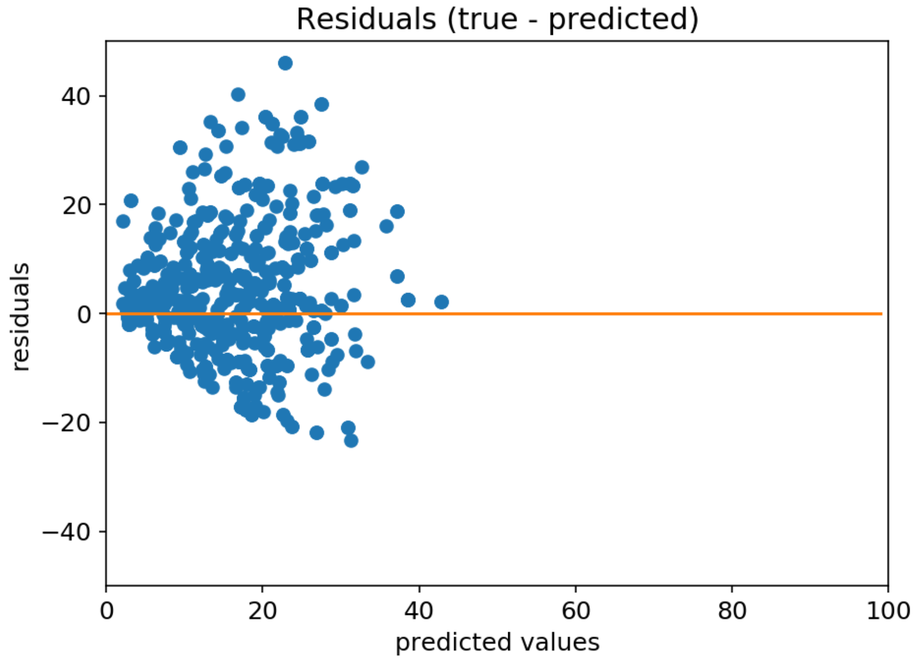
Tune n - Turtles

- n - number of consecutive days used to train model



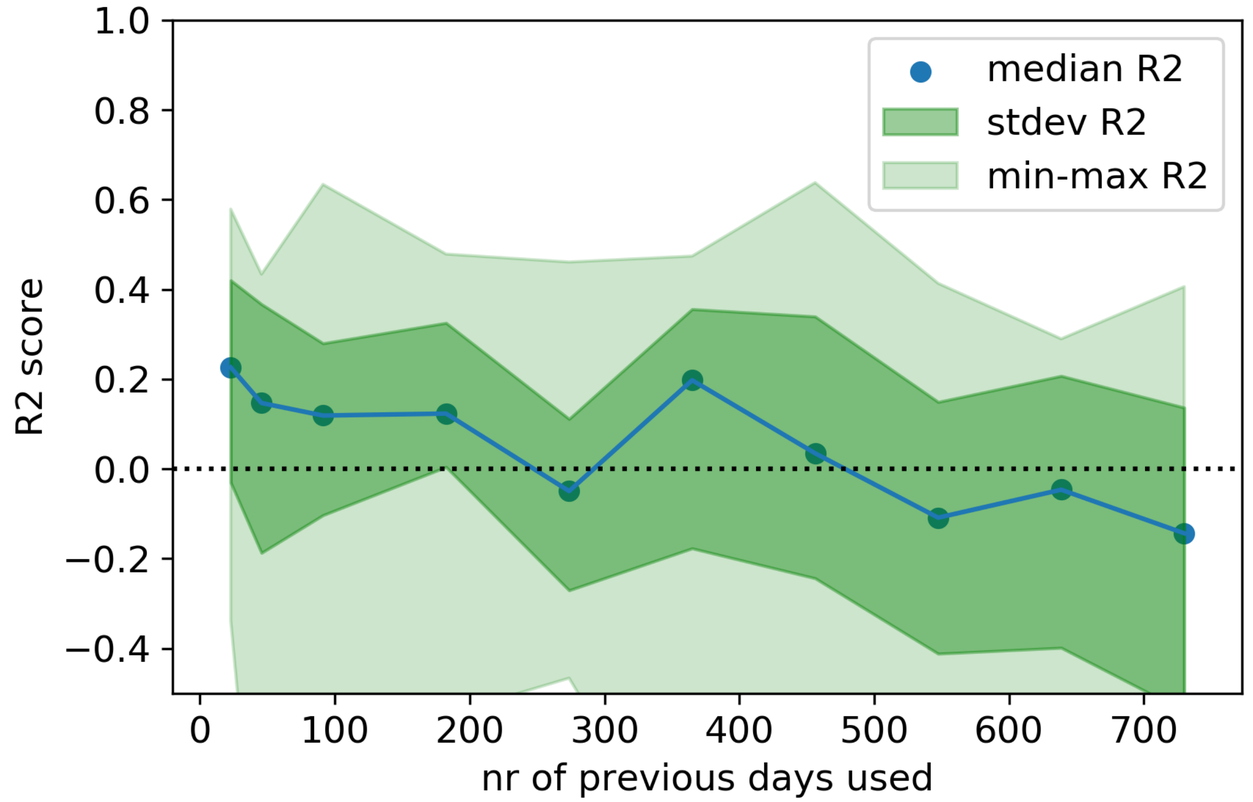
Residuals - Turtles



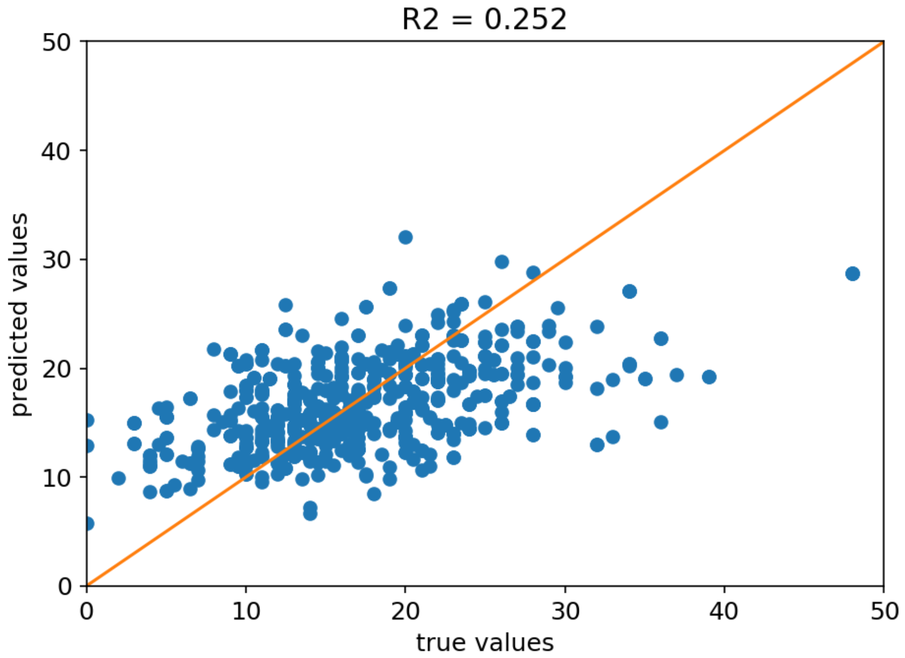


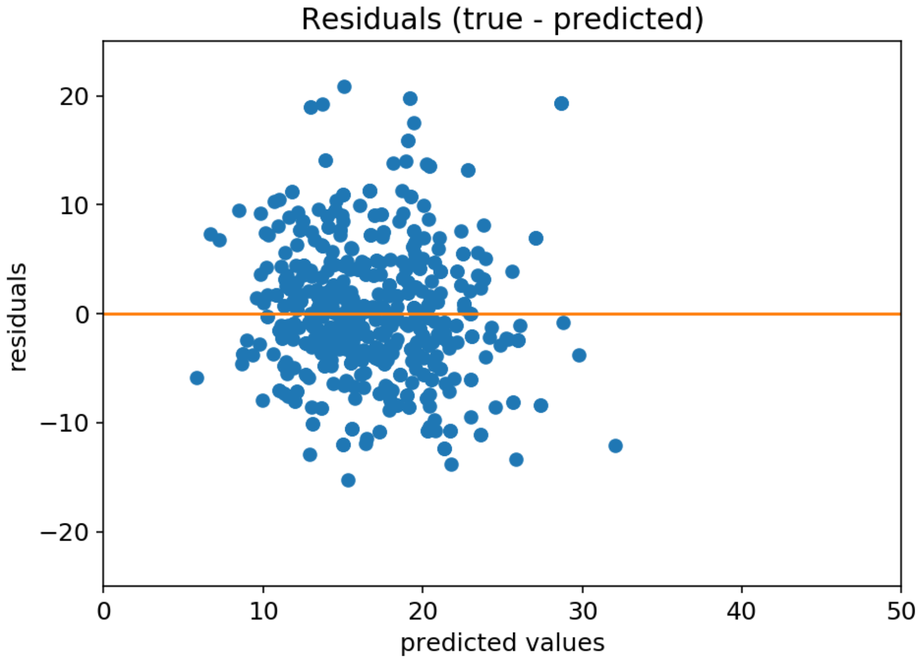
Tune n - Humans

- n - number of consecutive days used to train model

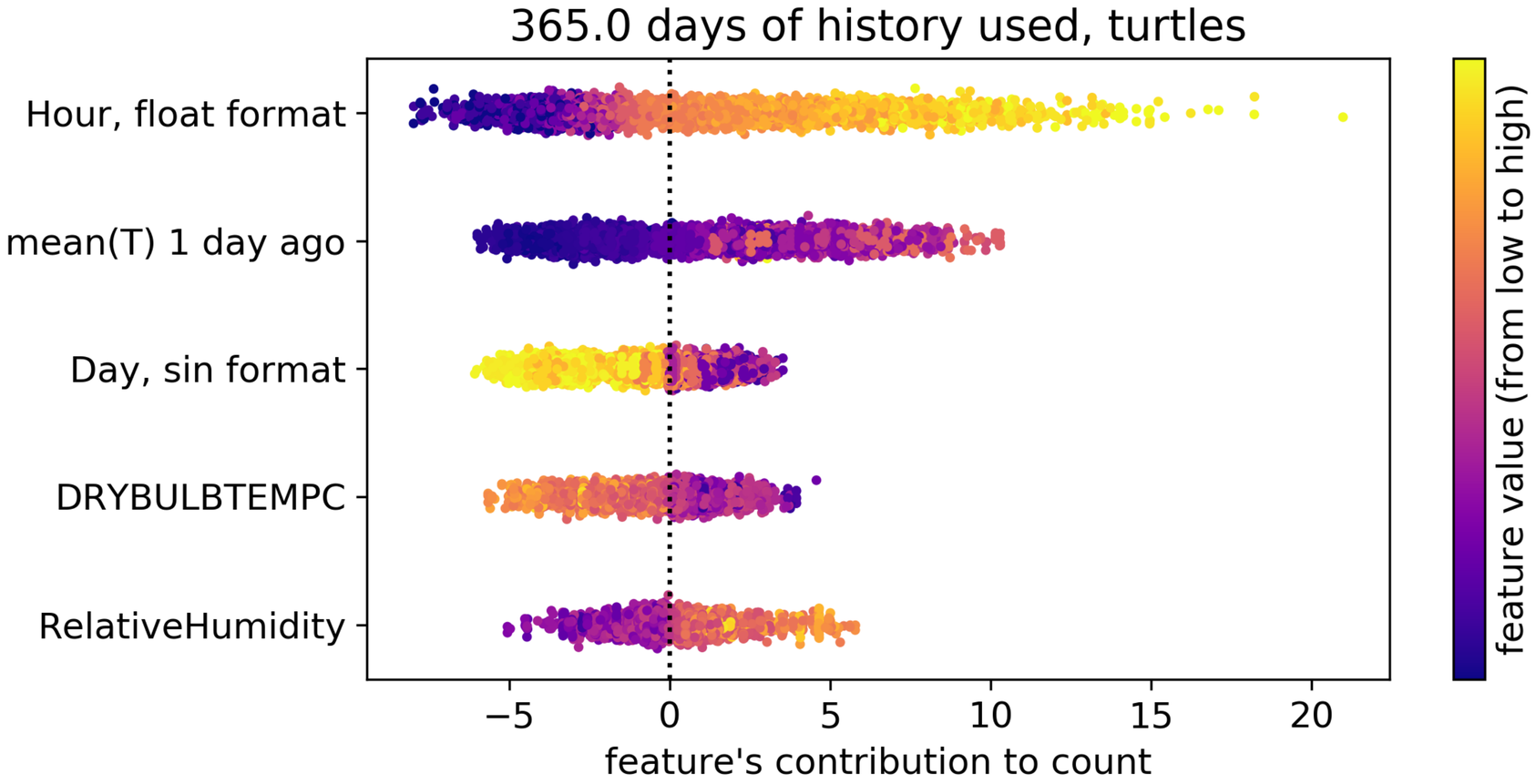


Residuals - Humans

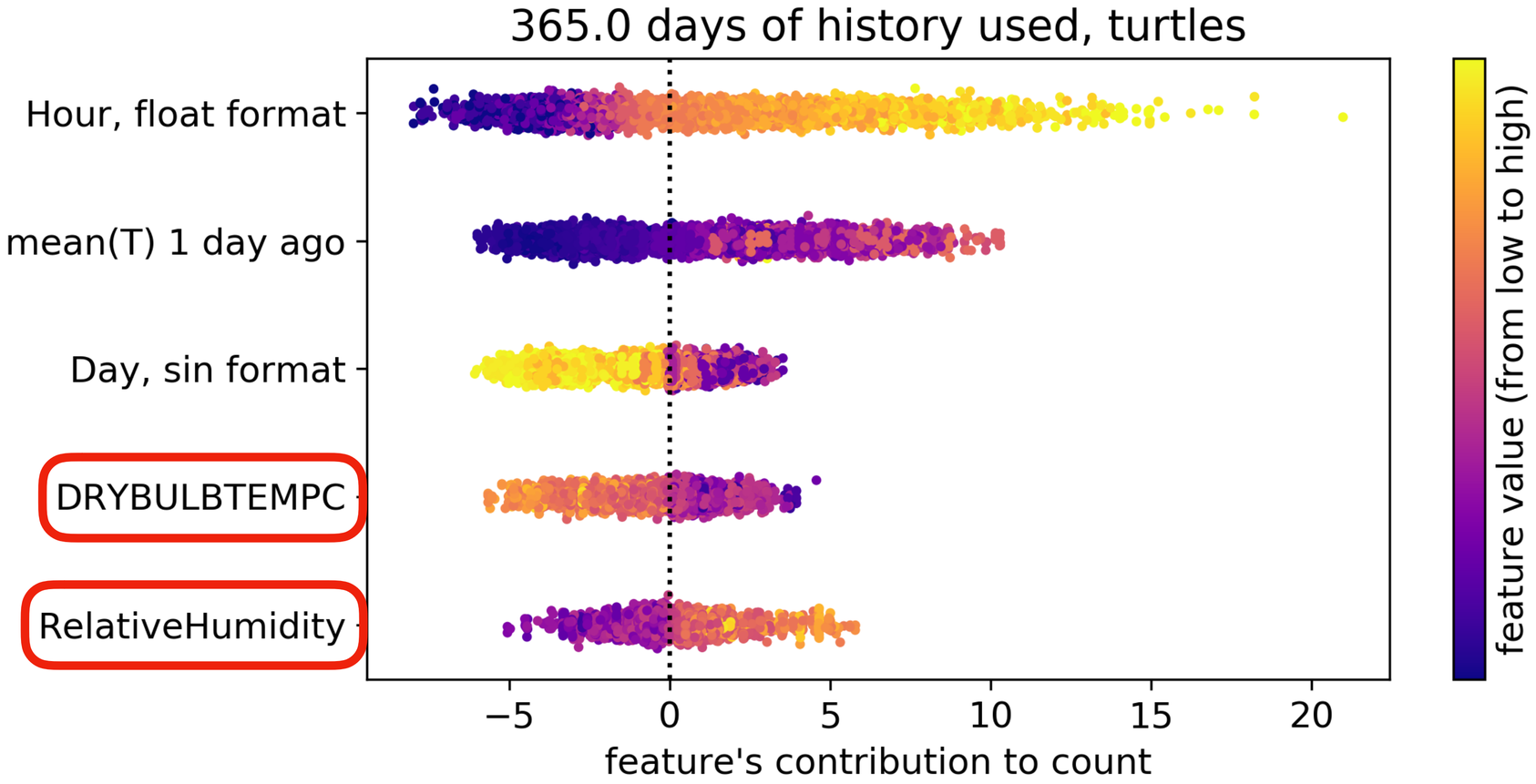




SHAP values - Turtles



SHAP values - Turtles

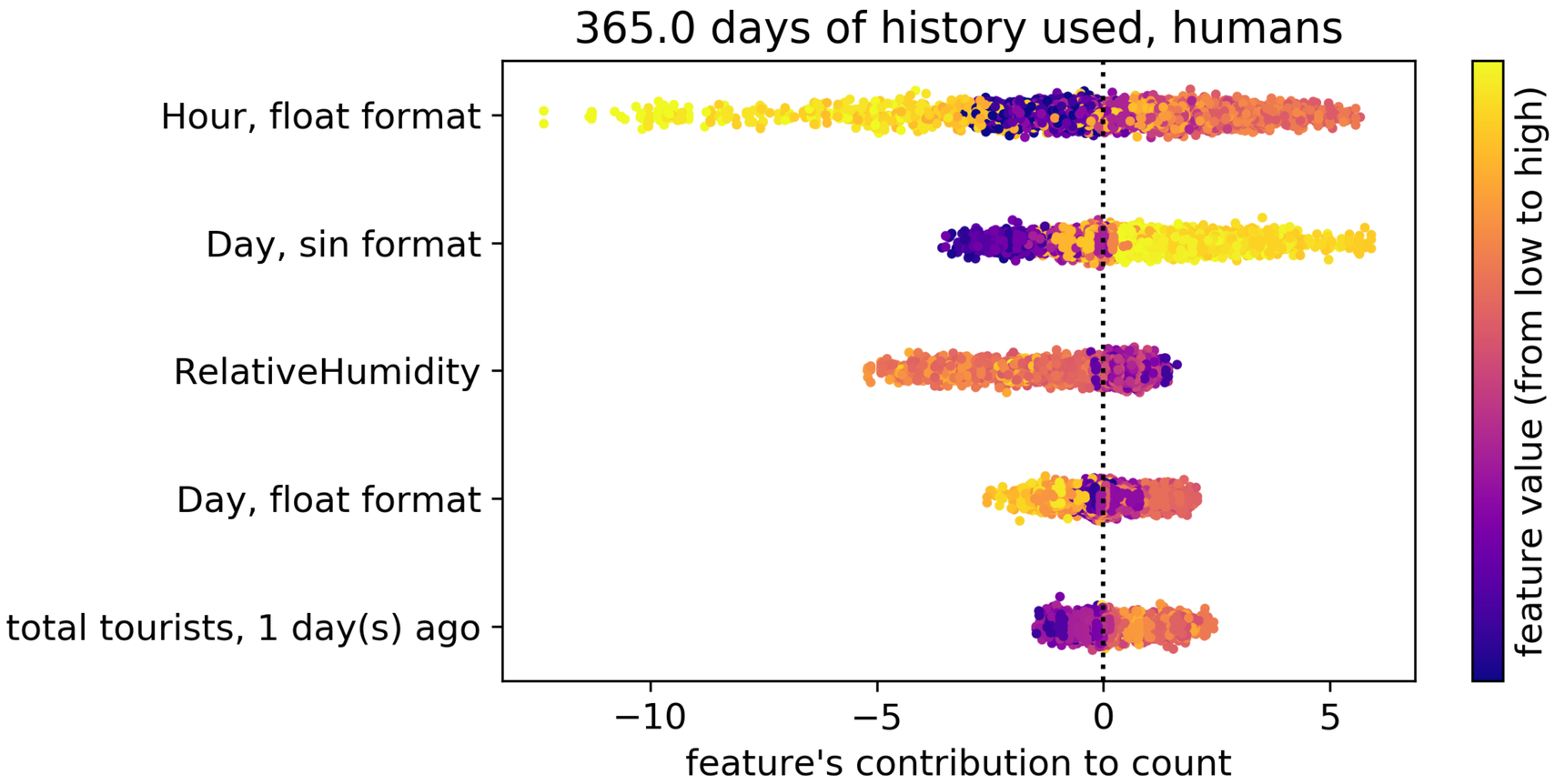


Comparison to other studies

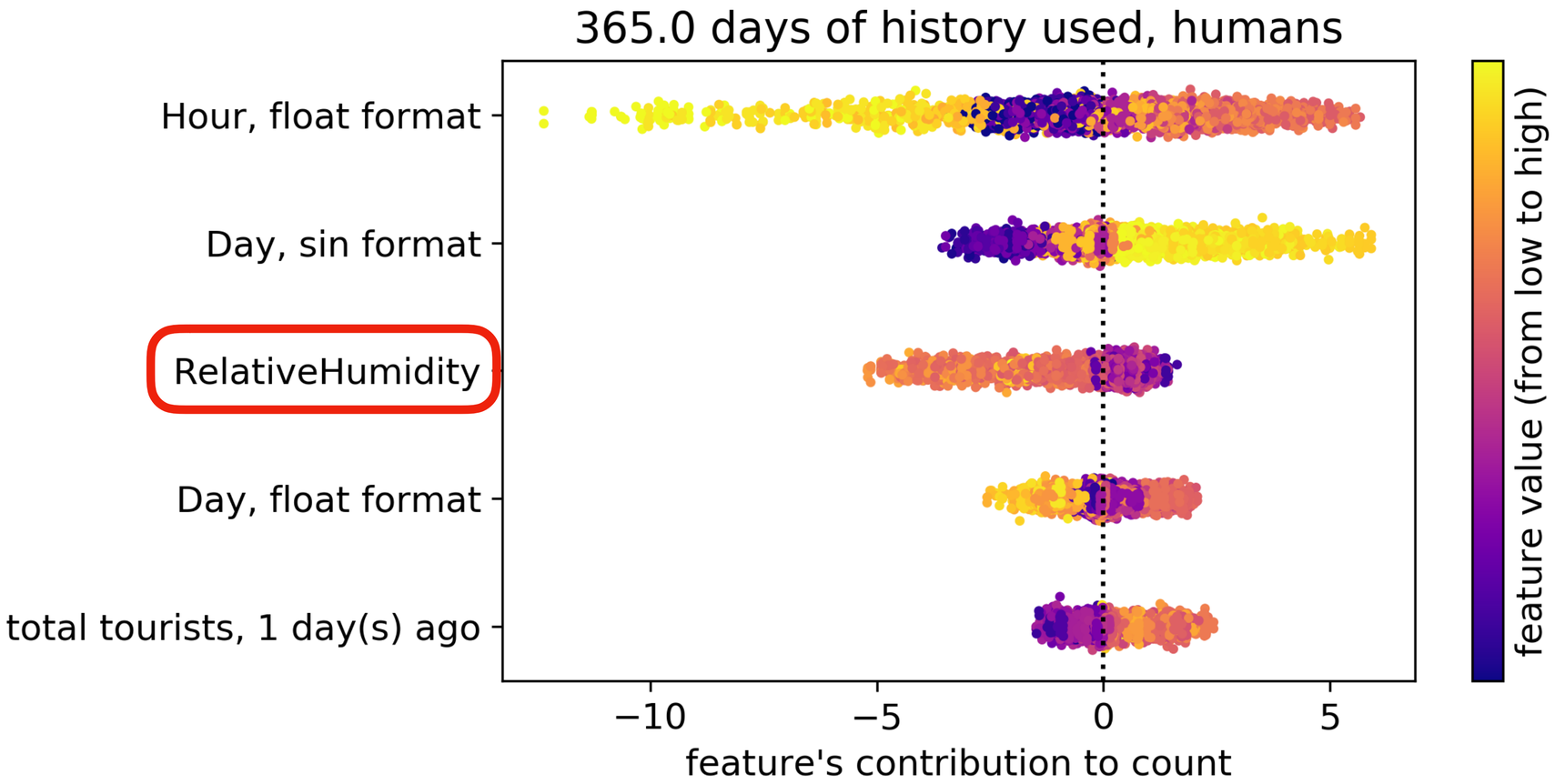
- **Green turtles of Galapagos** (Maxwell *et al.*, 2014):
  - More turtles at warm temperatures (> 30C) and sunny skies
- **Green turtles of Maui**
  - This study: More turtles at cold temperatures and high humidity
  - Van Houtan et al., 2015: More turtles at low sea surface temperatures

Explanation unknown ¯_(ツ)_/¯

SHAP values - Humans



SHAP values - Humans



Improve predictive power?

- More feature engineering
- Collect more data
    - Currently no data on biological environment
    - Sea surface temperature would be good to compare our study to Van Houtan et al., 2015
- Ensembling might improve R^2 slightly

Overview

- Project background and the Hawaii Wildlife Fund (HWF)
- Problem description
- Data collection and preprocessing
- Data exploration and feature generation
- Cross validation
- Results
- **Lessons learnt**

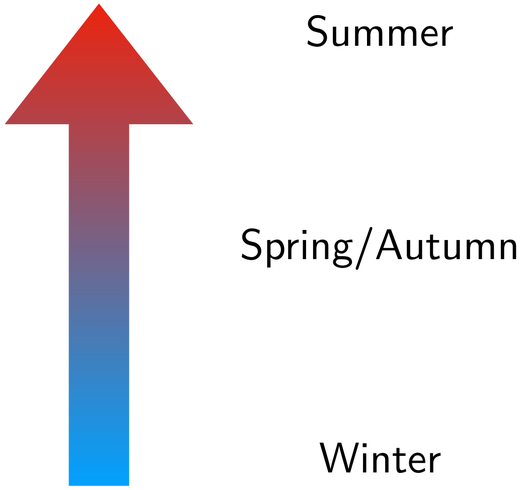
Lessons learnt

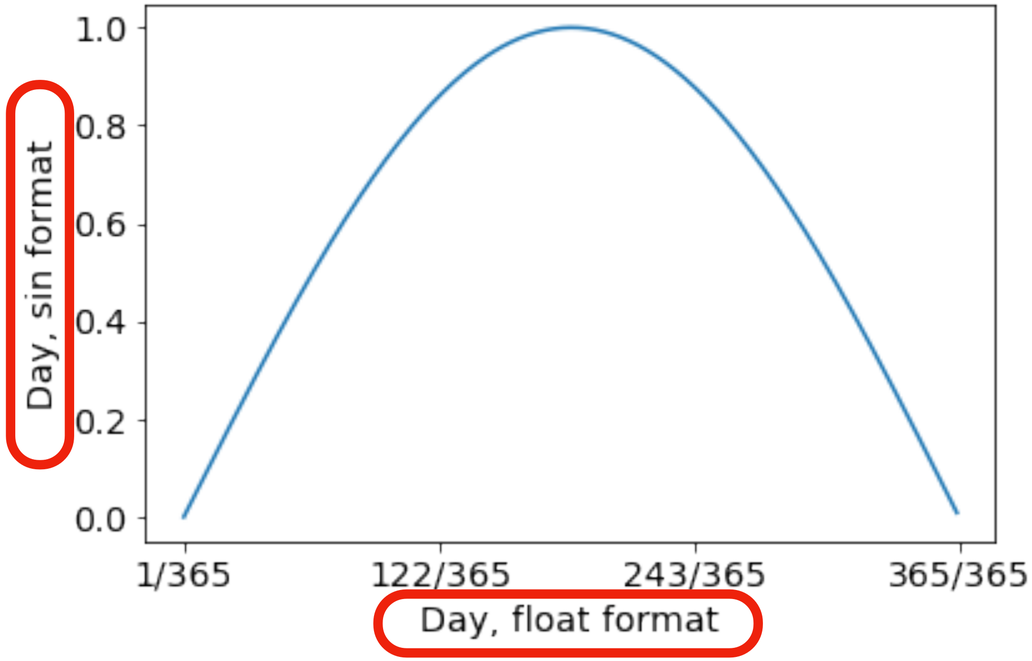
- By the end of this tutorial, you will be able to
    - combine various data sources,
    - perform feature engineering on time series data,
    - select a cross validation method most appropriate to your problem,
    - develop an interpretable model and maximize your actionable/scientific insights.

https://github.com/brown-ccv/turtles-HWF-Brown

Feature engineering

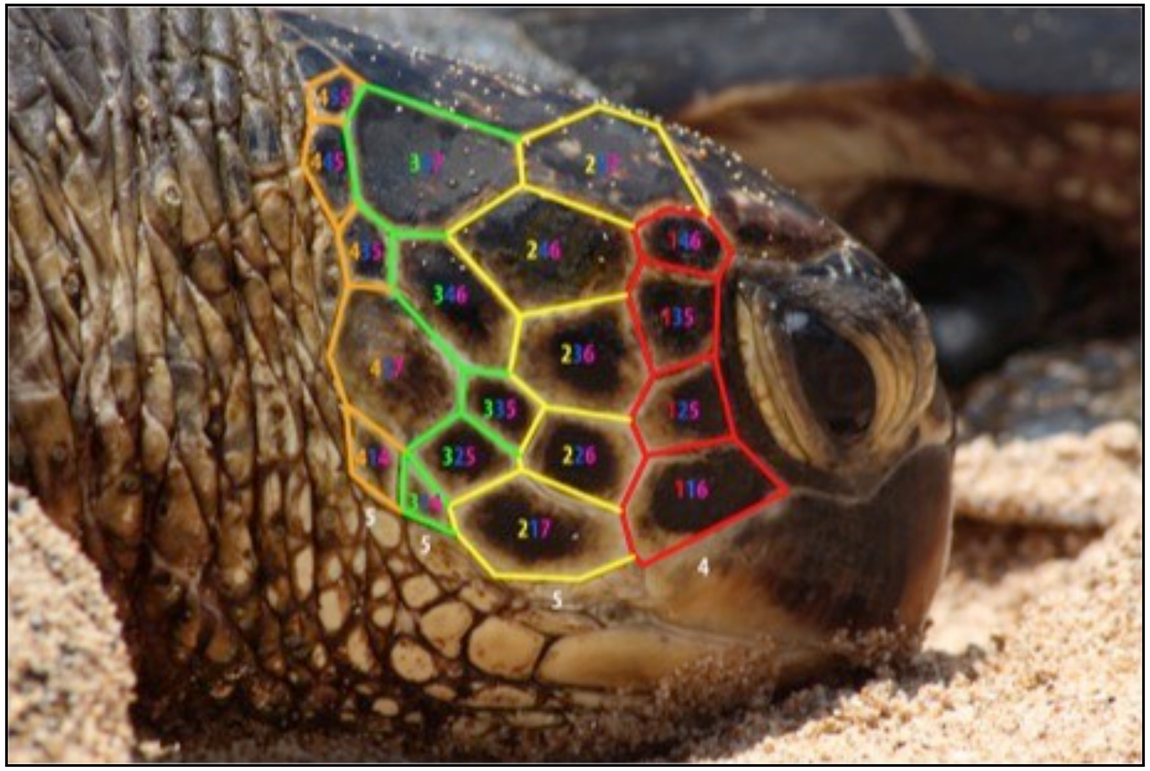
- Day of the year not good to represent winter-summer seasons





Other HWF project

- Face recognition for turtles



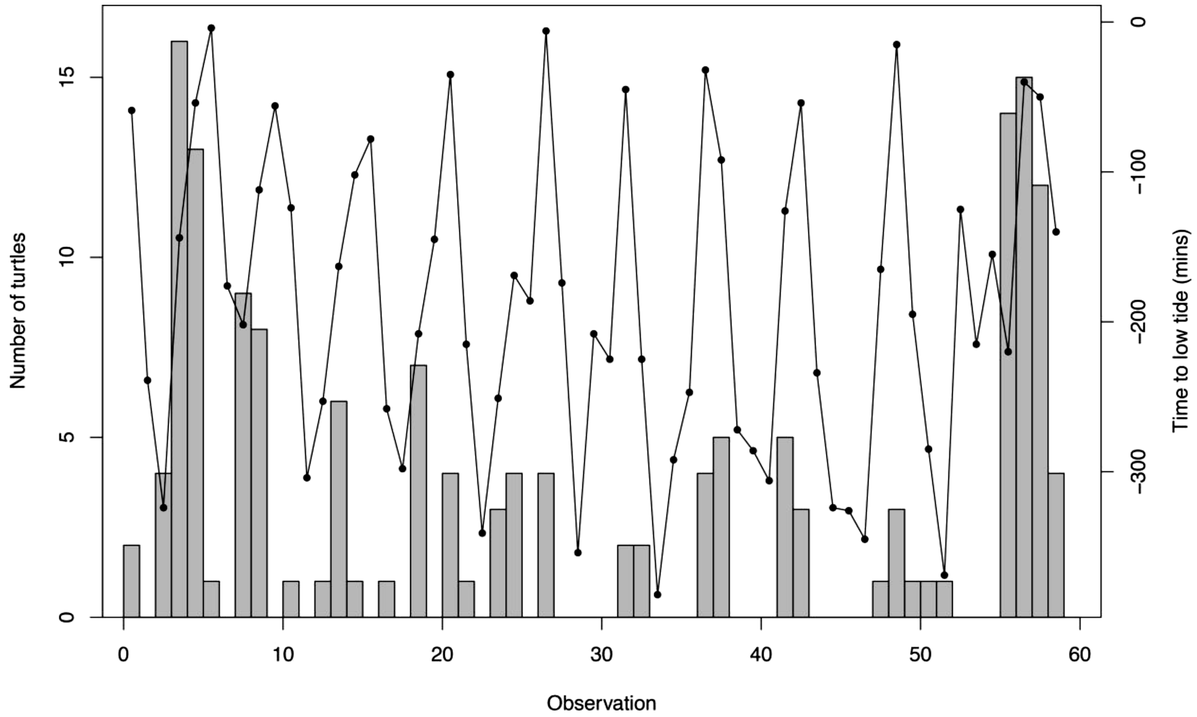
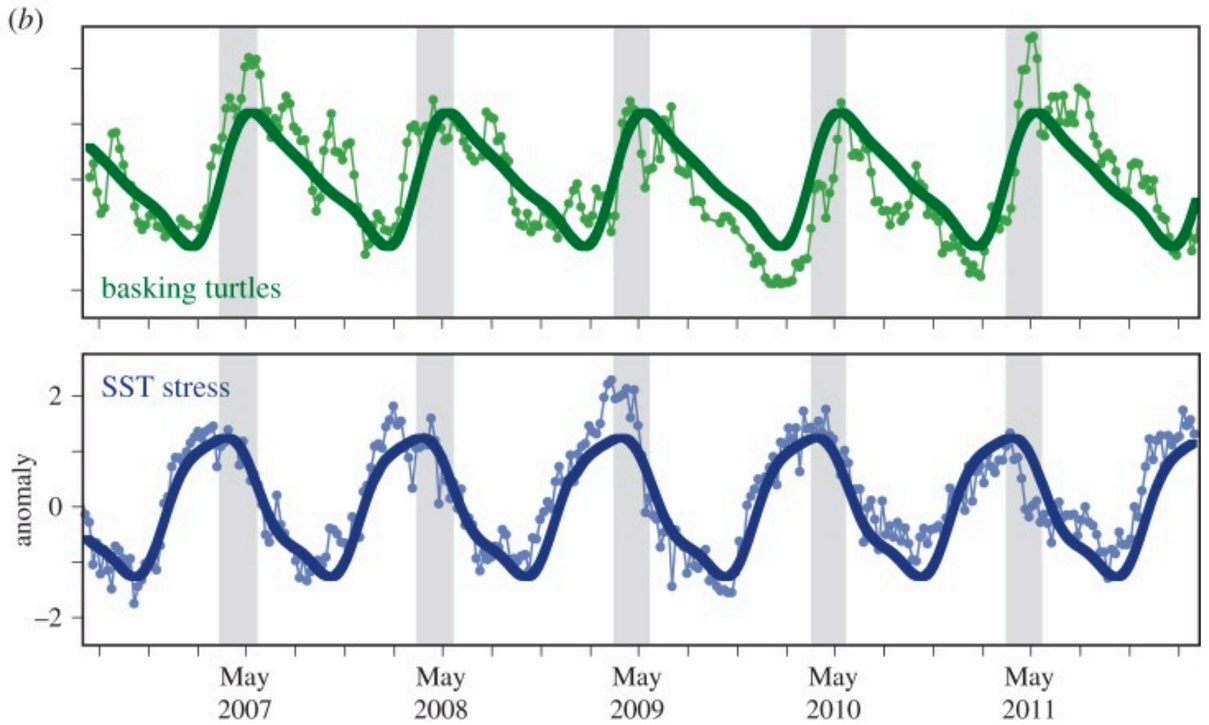


Figure 2. Number of turtles (grey bars) and tide level (black lines) across the 59 study observations. Tide level represents minutes since low tide (e.g., −200 min equates to 200 min since the low tide). Turtles are present in greater abundance during low tide. Note that observations are not evenly spaced in time (for details, see text).

"For daytime observations, air temperature ranged between 19.7C and 46.9C (mean = 30.7C, SD ± 7.72C)."

Van Houtan et al., 2015



"Data and models indicated basking peaks when SST is coolest."