Chicago Beaches Water Quality Prediction

Open Source Civic Tech Collaborative Data Science Project

Rebecca Jones

Overview

- Context of project : ChiHackNight collaboration
- Description of problem : Why E. Coli prediction is both important and difficult
- Modeling approach : Ensemble of GBM and Random Forest models
- Results : Performance comparisons
- Reflections : Pros and cons of collaborative data science work

CHI HACKNIGHT

- Chicago's weekly event to build, share & learn about civic tech.
- Every Tuesday 6-9pm at Merchandise Mart 8th floor, Braintree offices.
- Brings together people with technical skills and people with knowledge of aspects of civic life that can be improved by better use of data and technology.
- https://chihacknight.org/

City of Chicago

- Tom Schenk, Chief Data Officer.
- https://data.cityofchicago.org/

Escherichia coli : monitored as indicator of microbial contamination

Pathways for bacterial contamination

Local

Geese and other birds. Animal waste.

Runoff from streets and waterways. Washout of bacterial load buildup on beachfront.

Non functioning storm drains. Illicit discharges. Sewer overflows.

Escherichia coli :
monitored as indicator of microbial contamination

Pathways for bacterial contamination

Local

Geese and other birds. Animal waste.

Runoff from streets and waterways. Washout of bacterial load buildup on beachfront.

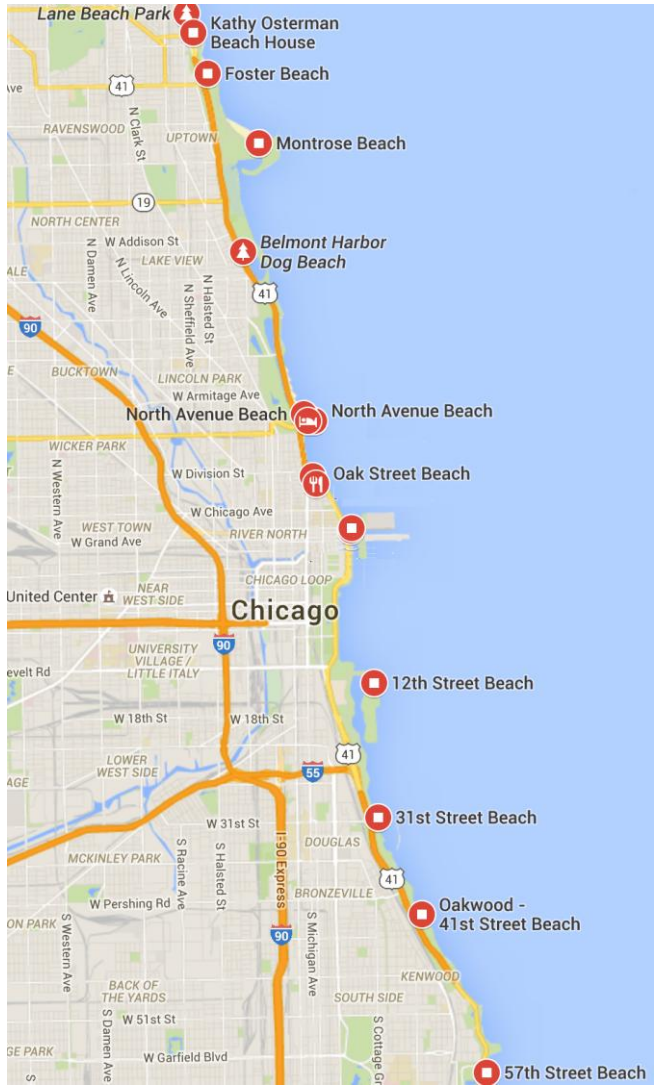
Non functioning storm drains. Illicit discharges. Sewer overflows.

Regional

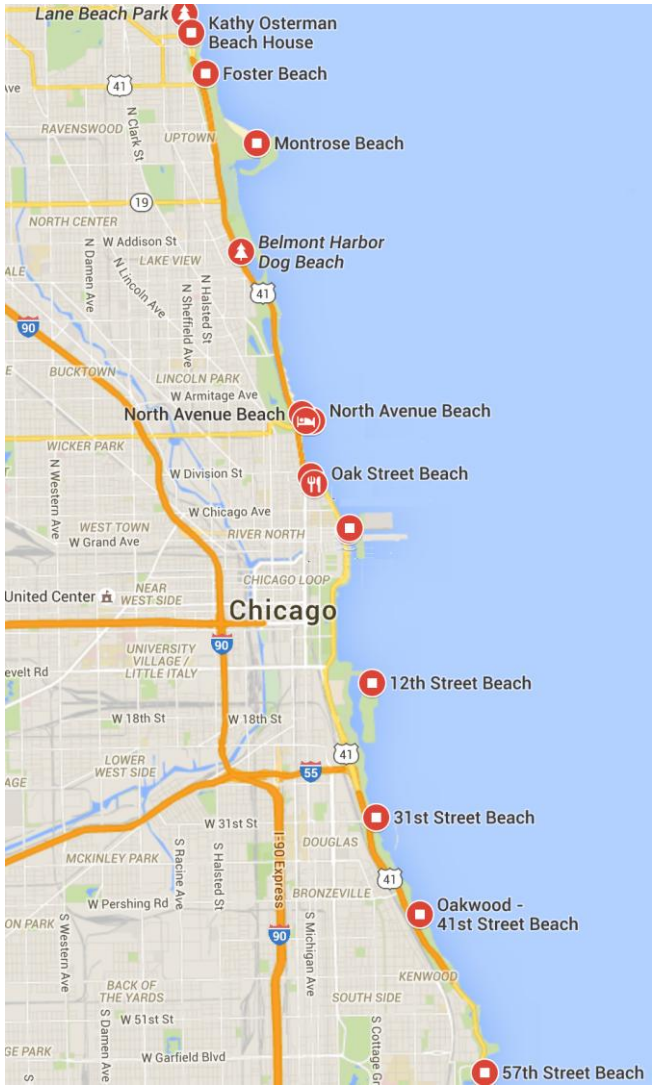
Storm surges. Sediment resuspension.

Turbidity currents. Flows from rivers.

Chicago Beach Monitoring:



Chicago Beach Monitoring:



www.cpdbeaches.com/beaches/Montrose-Beach/

Search

Home | Beach List | Keep Our Beaches Clean | Stay Connected

Montrose Beach

SWIM STATUS

NO RESTRICTIONS

WATER QUALITY INFORMATION

Water Quality Model Prediction:	112.8
Culture Based Test Results:	54
Rapid Test Method Results:	331

What do these numbers mean?

LOCATION | BEACH HOURS

LOCATION

4400 N Lake Shore Dr

Montrose Avenue and Lake Shore Drive

(773) 363-2225

Map and Directions

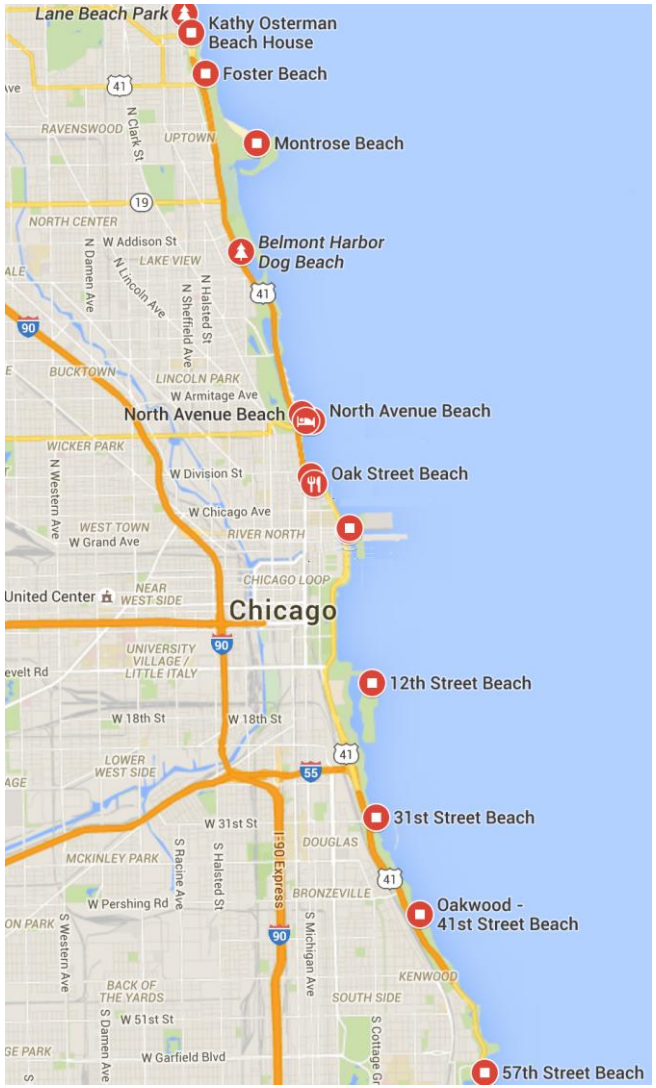
DISTANCE SWIMMING

tower 4 (north of boathouse), parallel to shore

CURRENT WEATHER

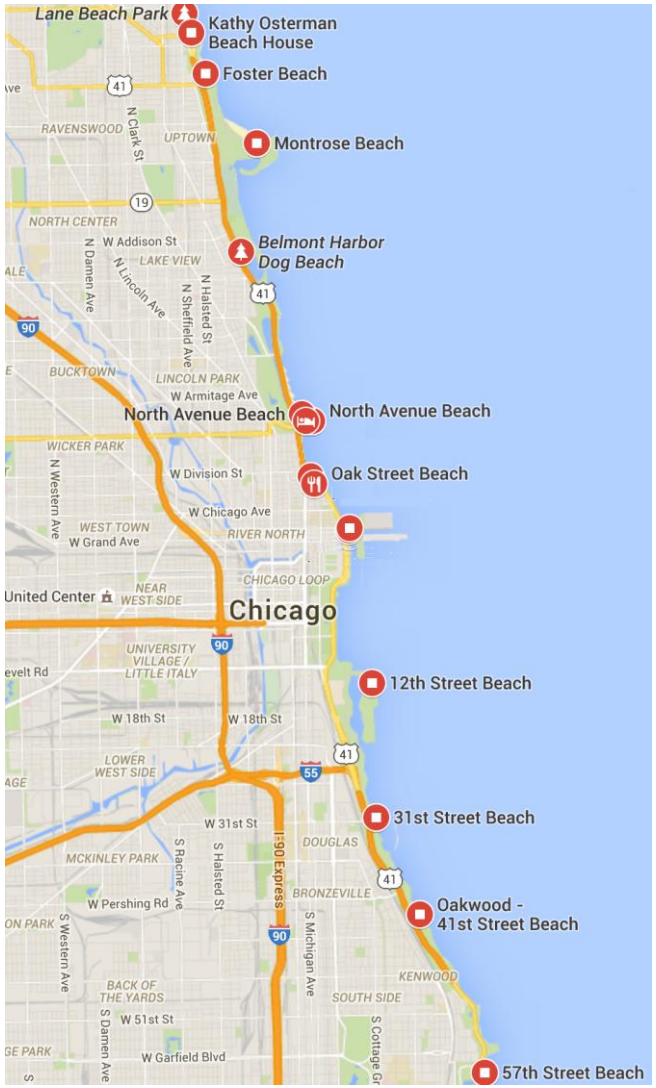
77° F/25° C

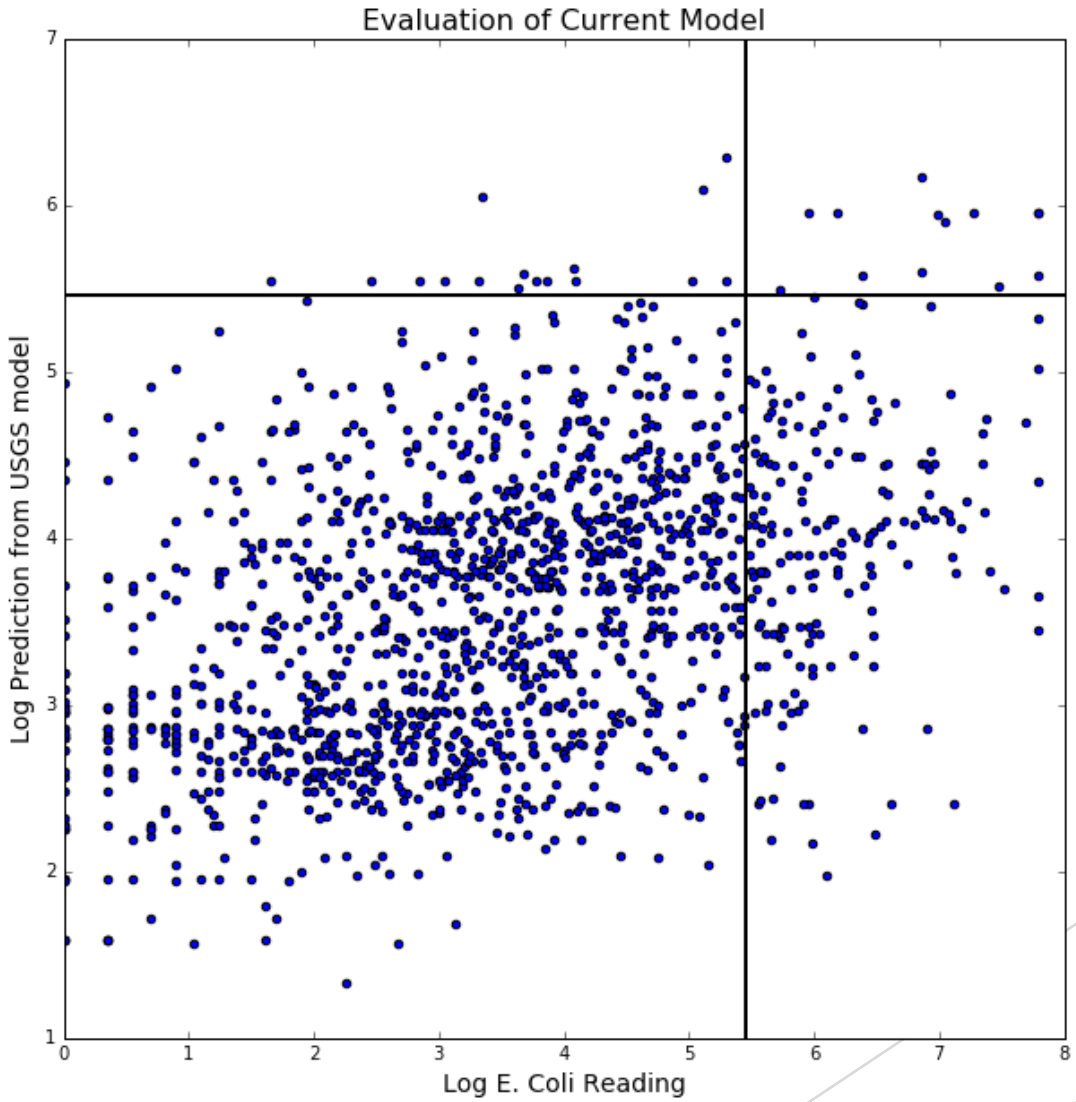
Chicago Beach Monitoring:



BEACH	incorrect_warning	correct_warning	missed_warning
Juneway	0	0	5
Rogers	0	0	7
Howard	0	0	7
Jarvis	0	0	2
Leone	0	0	4
Albion	0	0	6
Osterman	0	0	14
Foster	1	0	9
Montrose	1	5	26
North Avenue	0	0	5
Oak Street	0	0	1
Ohio	1	0	13
12th	0	0	11
31st	2	1	12
39th	2	0	6
57th	2	1	4
63rd	2	1	11
South Shore	2	1	13
Rainbow	3	4	17
Calumet	0	0	30

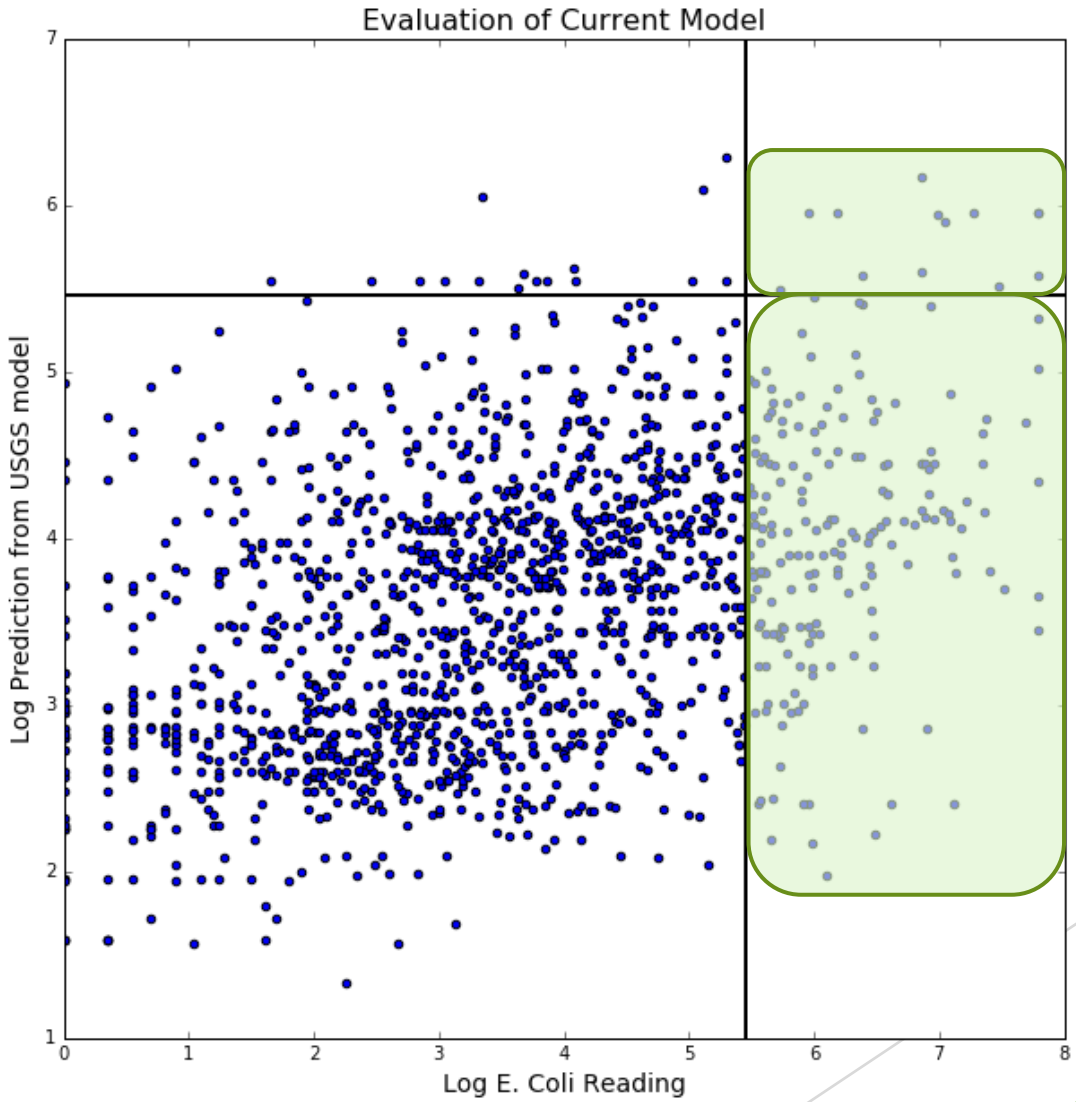
Chicago Beach Monitoring:





Chicago Beach Monitoring:

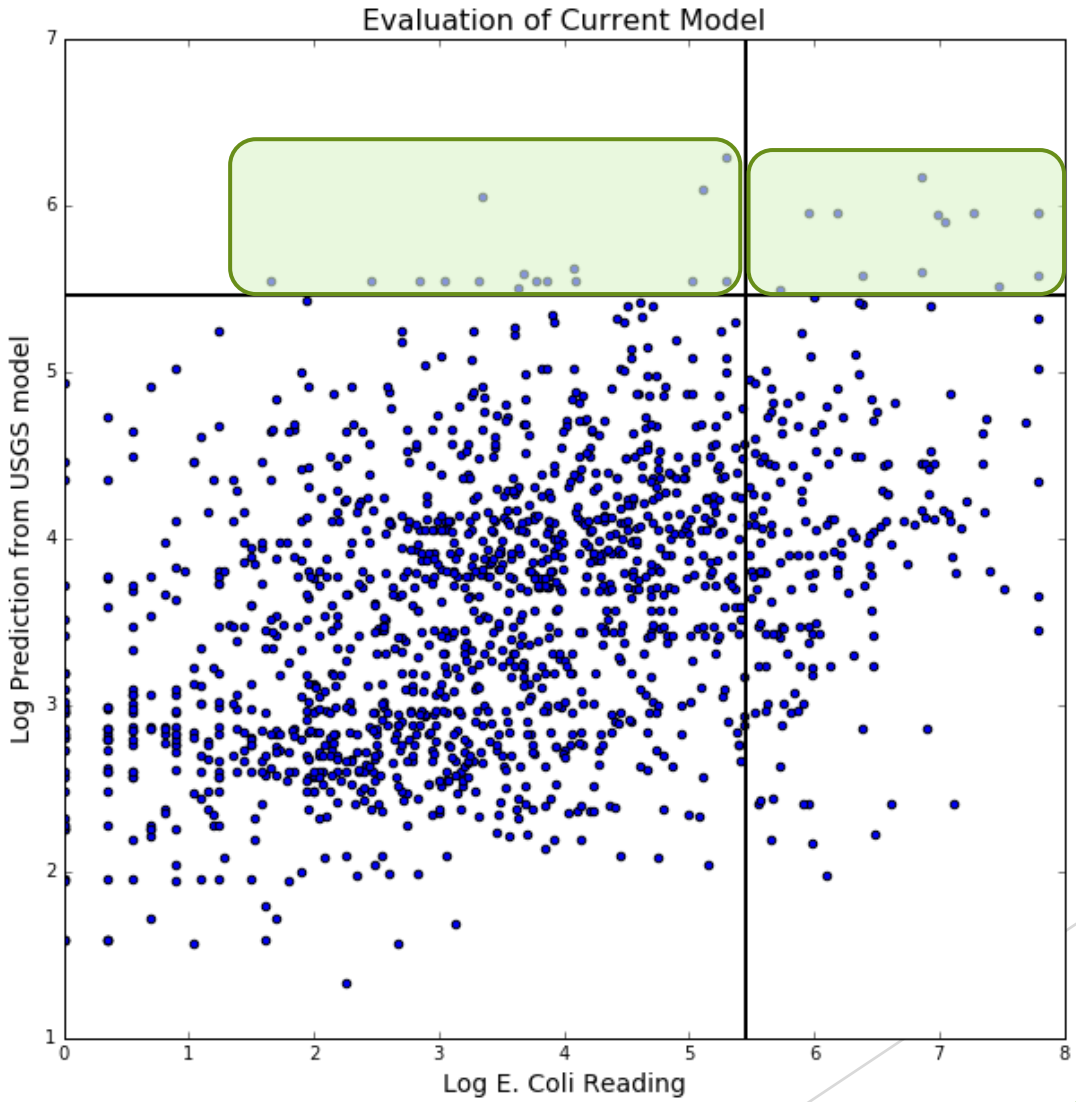
Specificity (Recall or TPR) **: ~6%**
proportion of correctly identified positives



Chicago Beach Monitoring:

Sensitivity (Recall or TPR) : ~6%
proportion of correctly identified positives

Precision : ~42%
proportion of identified that are positives

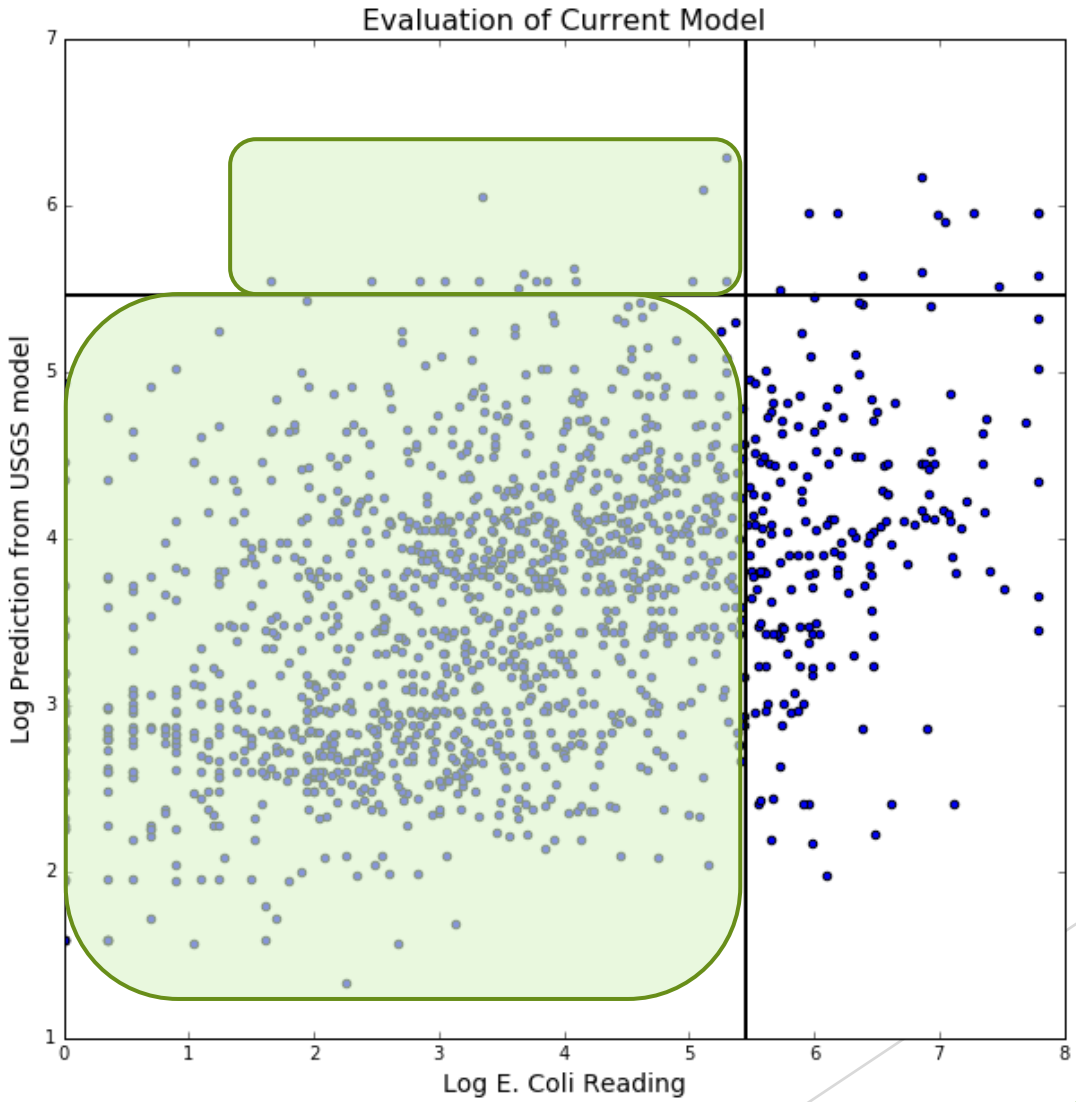


Chicago Beach Monitoring:

Sensitivity (Recall or TPR) : ~6%
proportion of correctly identified positives

Precision : ~42%
proportion of identified that are positives

Specificity (TNR): ~98%
Proportion of correctly identified negatives



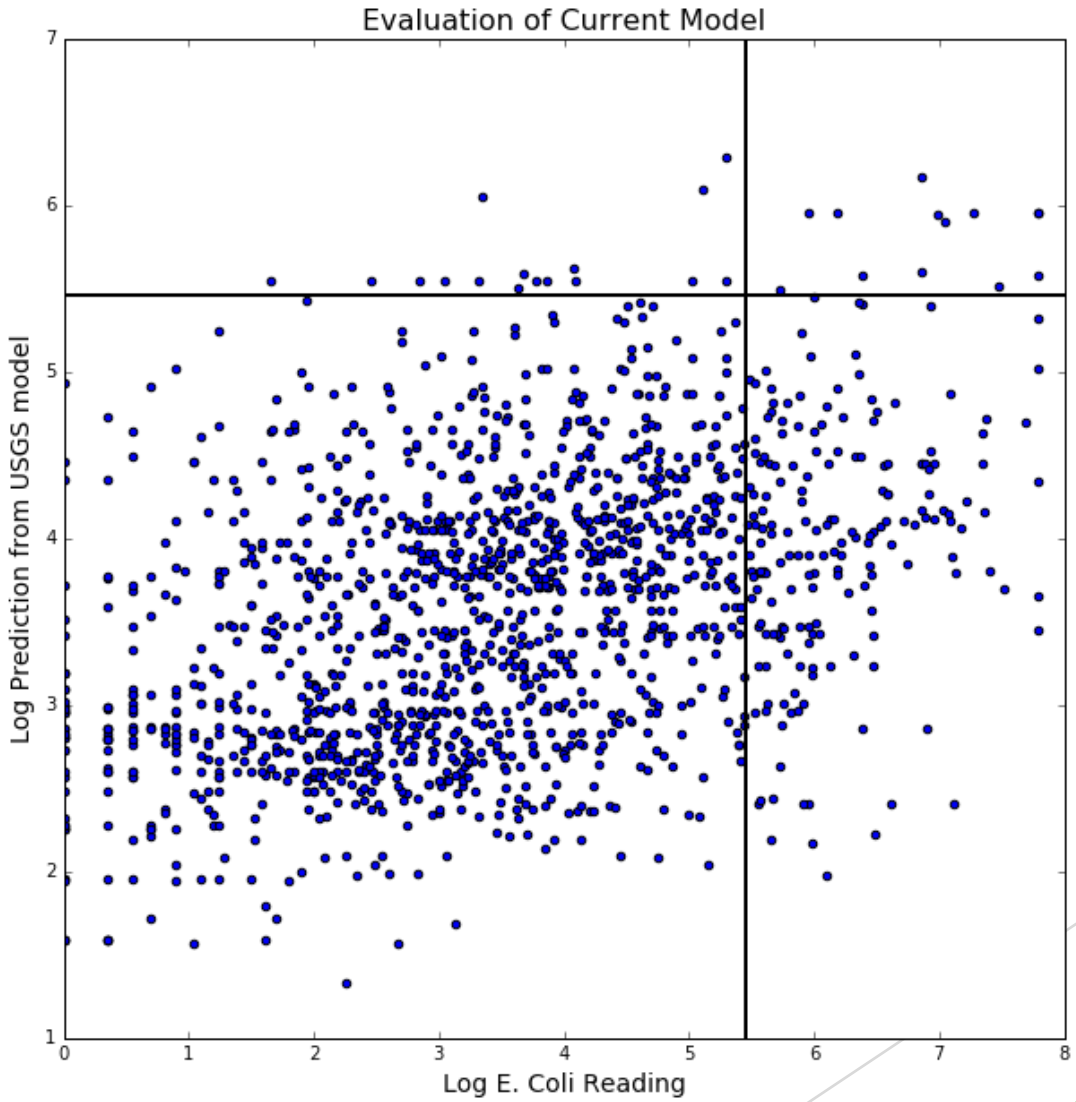
Chicago Beach Monitoring:

Sensitivity (Recall or TPR) : ~6%
proportion of correctly identified positives

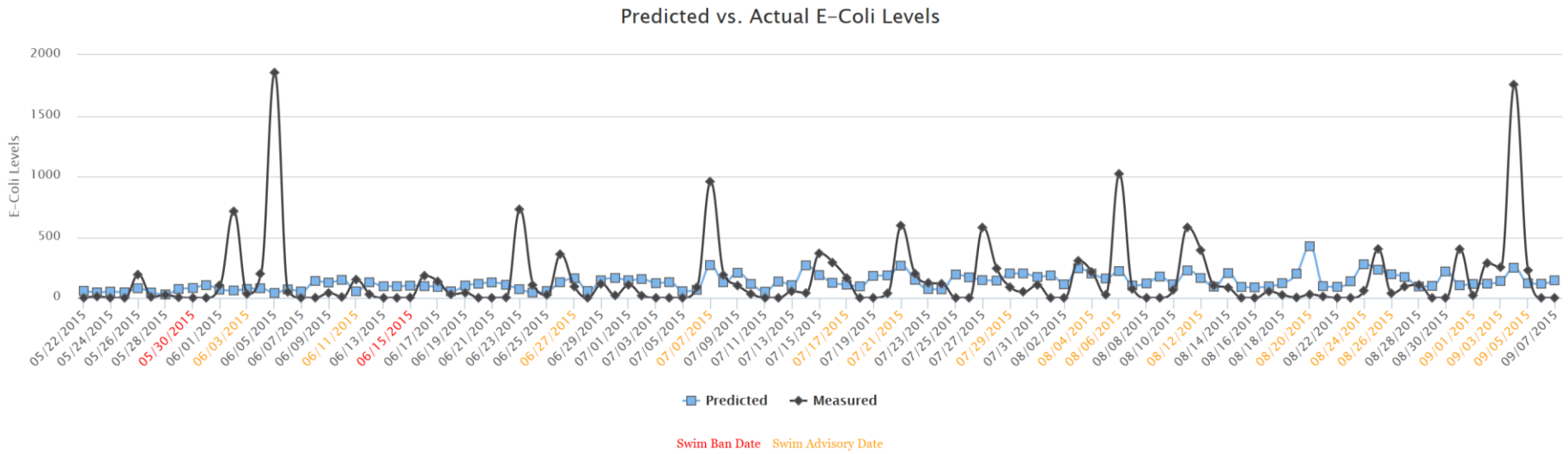
Precision : ~42%
proportion of identified that are positives

Specificity (TNR): ~98%
Proportion of correctly identified negatives

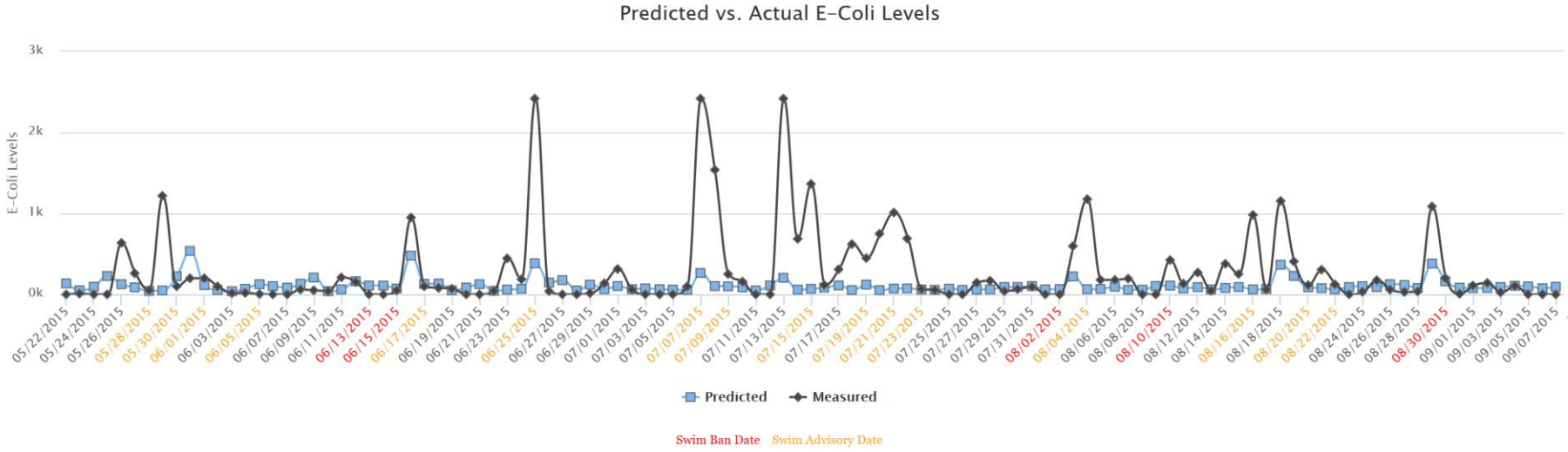
What do we want to maximize?
- Public safety
- Reputation of monitoring system
- Reputation of beaches



Rainbow beach

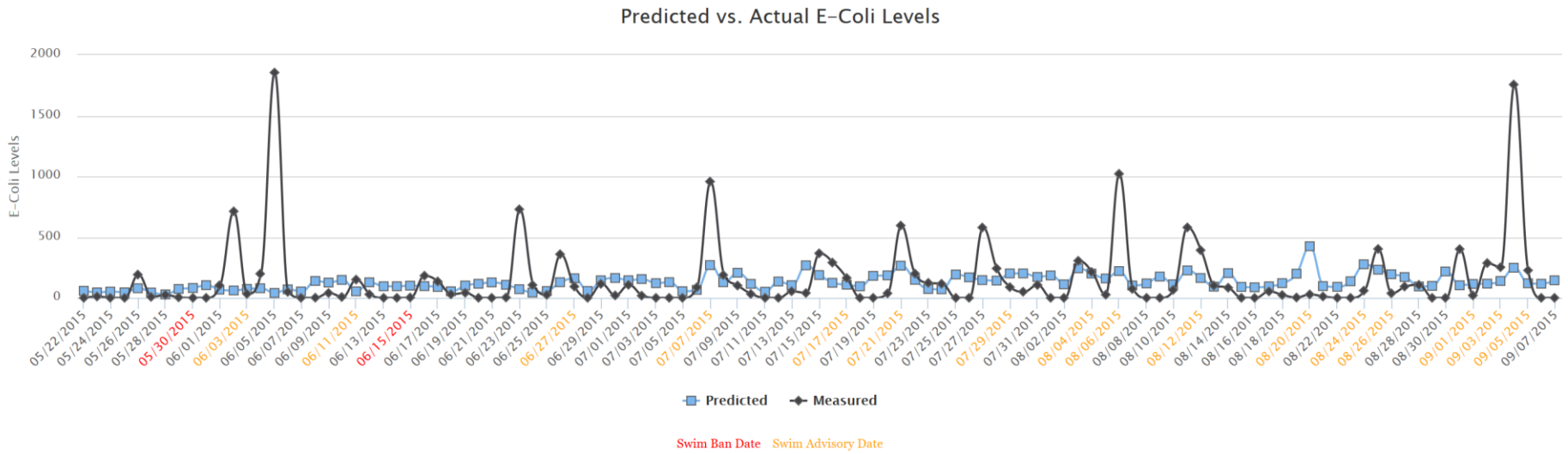


Montrose beach

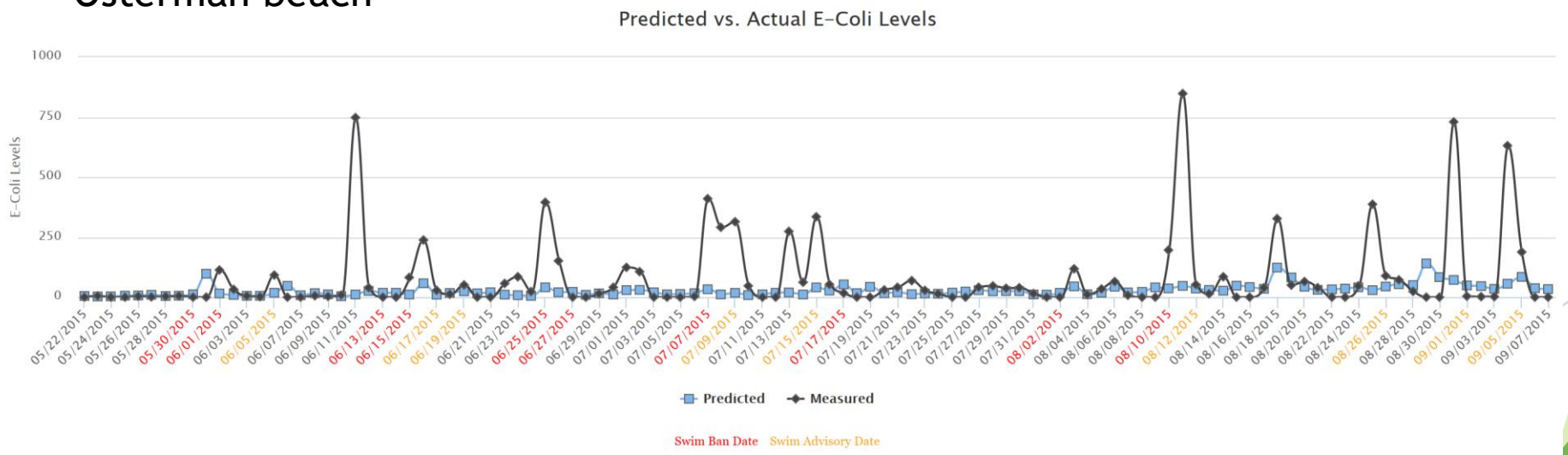


From website:
http://drekbeach.org/
by Scott Beslow

Rainbow beach



Osterman beach



From website:
http://drekbeach.org/
by Scott Beslow

Data

Response Variable:

E Coli readings : geometric mean of two readings at each beach.

20 beaches. Measured Mon-Fri. ~75days per year. 10 years of data.

Predictors:

Weather : temp, wind speed & direction, humidity, rain, pressure, dew point

Water Sensor Readings : turbidity, wave height, chlorophyll, ph levels

Data

Response Variable:

E Coli readings : geometric mean of two readings at each beach.

20 beaches. Measured Mon-Fri. ~75days per year. 10 years of data.

Predictors:

Weather : temp, wind speed & direction, humidity, rain, pressure, dew point

~~**Water Sensor Readings :** turbidity, wave height, chlorophyll, ph levels~~

Data

Response Variable:

E Coli readings : geometric mean of two readings at each beach.

20 beaches. Measured Mon-Fri. ~75days per year. 10 years of data.

Predictors:

Weather : temp, wind speed & direction, humidity, rain, pressure, dew point

~~**Water Sensor Readings :** turbidity, wave height, chlorophyll, ph levels~~

Engineered features : lagged variables, trailing averages,
overnight pressure change, accumulated rain,
North/South wind speed, East/West wind speed

Modeling methods considered: Random Forests and Gradient Boosting

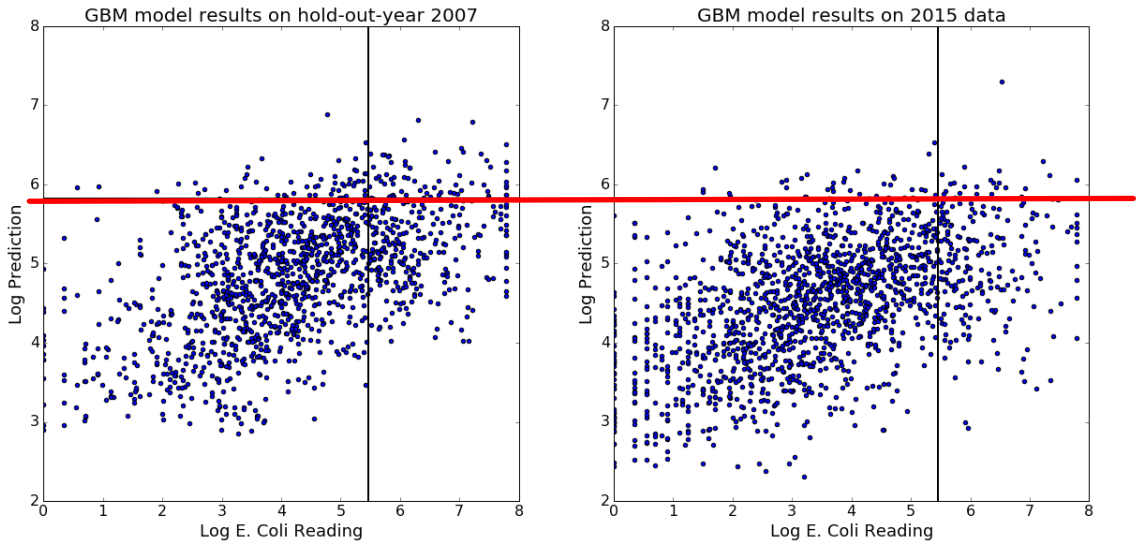
- Bagging (RF) - averaging across independent decision tree models
- Boosting (GBM) – iterative decision tree building to strengthen weak learners
- Use both in combination can lower variance, give more robust results.

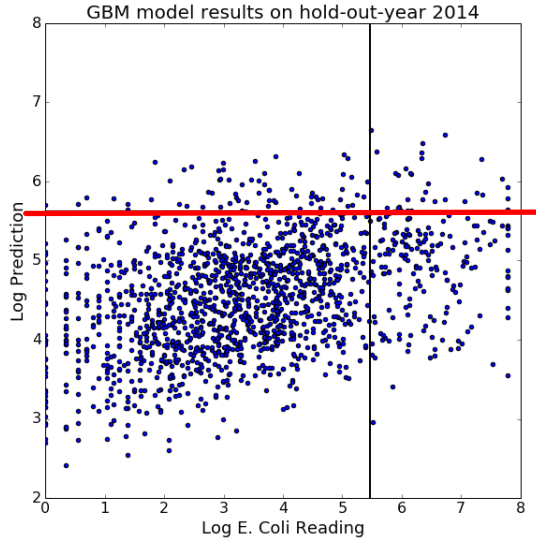
Modeling Approach:
Ensembles of Random Forest and Gradient Boosting Machine Regression Trees

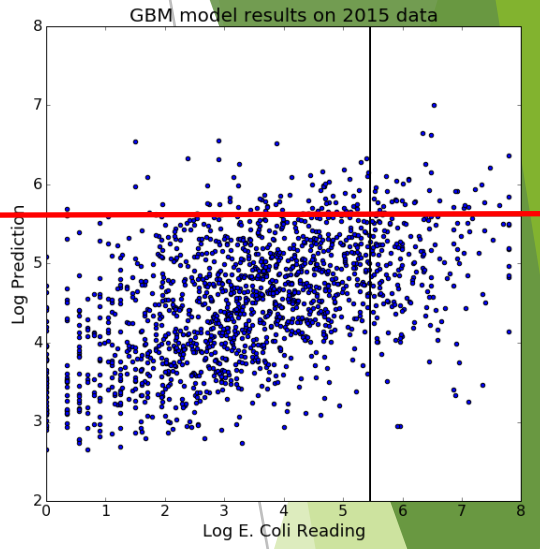
Process:

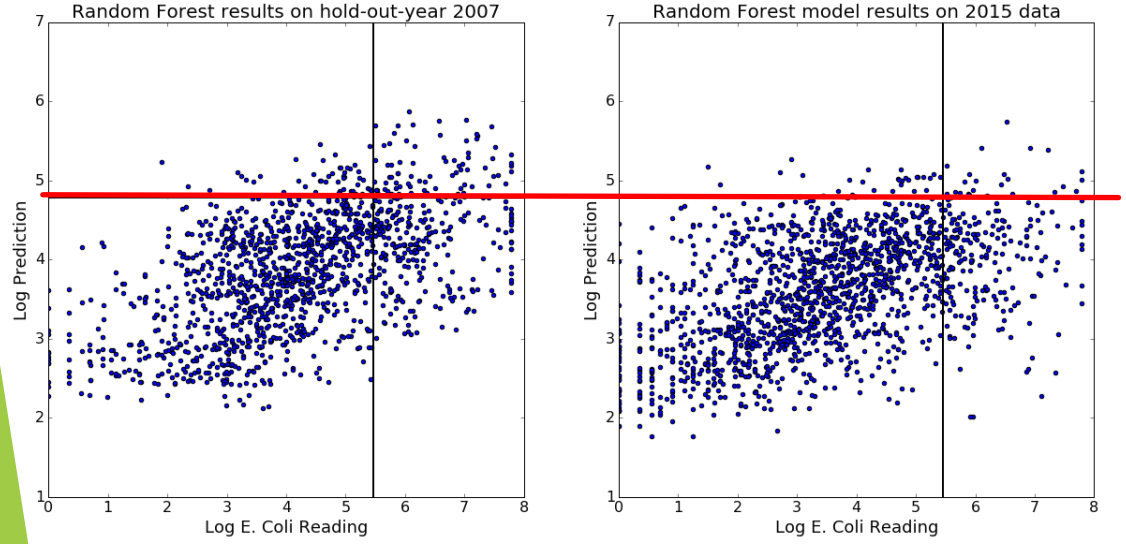
- Remove 2015 from training data. Keep as final test set.
- For each year 2006-2014, hold year of data out of training set and build model.
- Calibrate decision threshold using hold out year. (set FalseNegRate to 5%, max or 2% max)
- Test resulting set of models on 2015 data using calibrated decision thresholds.

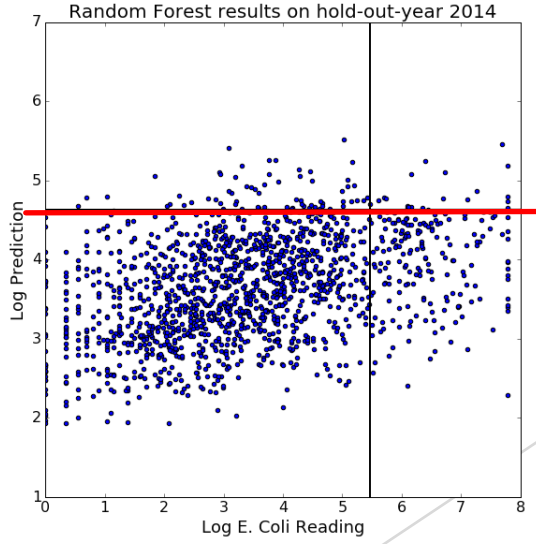
Two hold-out-year models. Example of setting and using decision threshold.

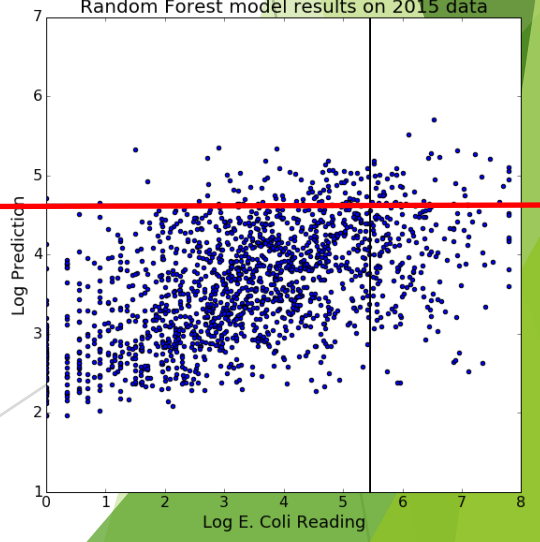




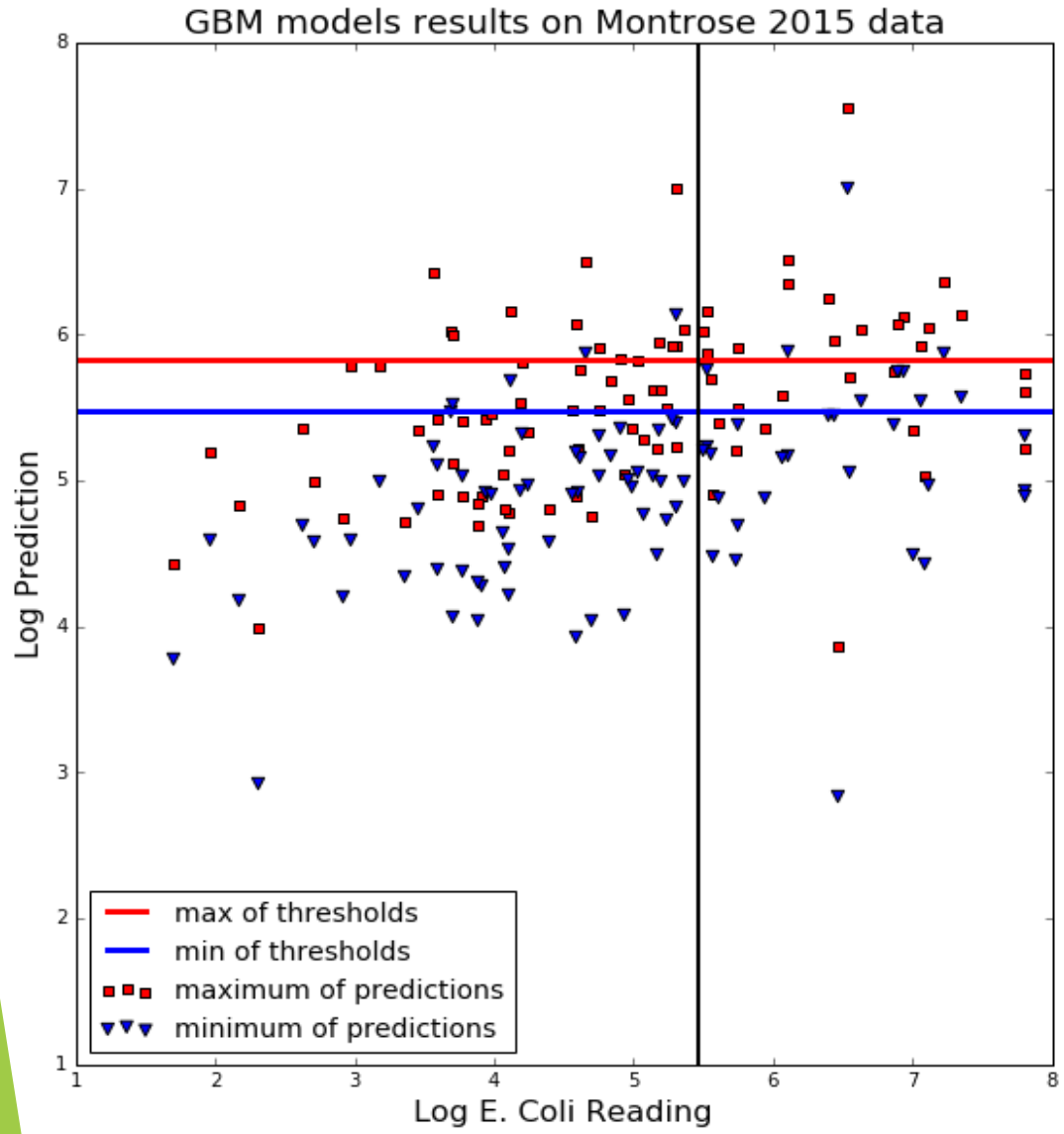


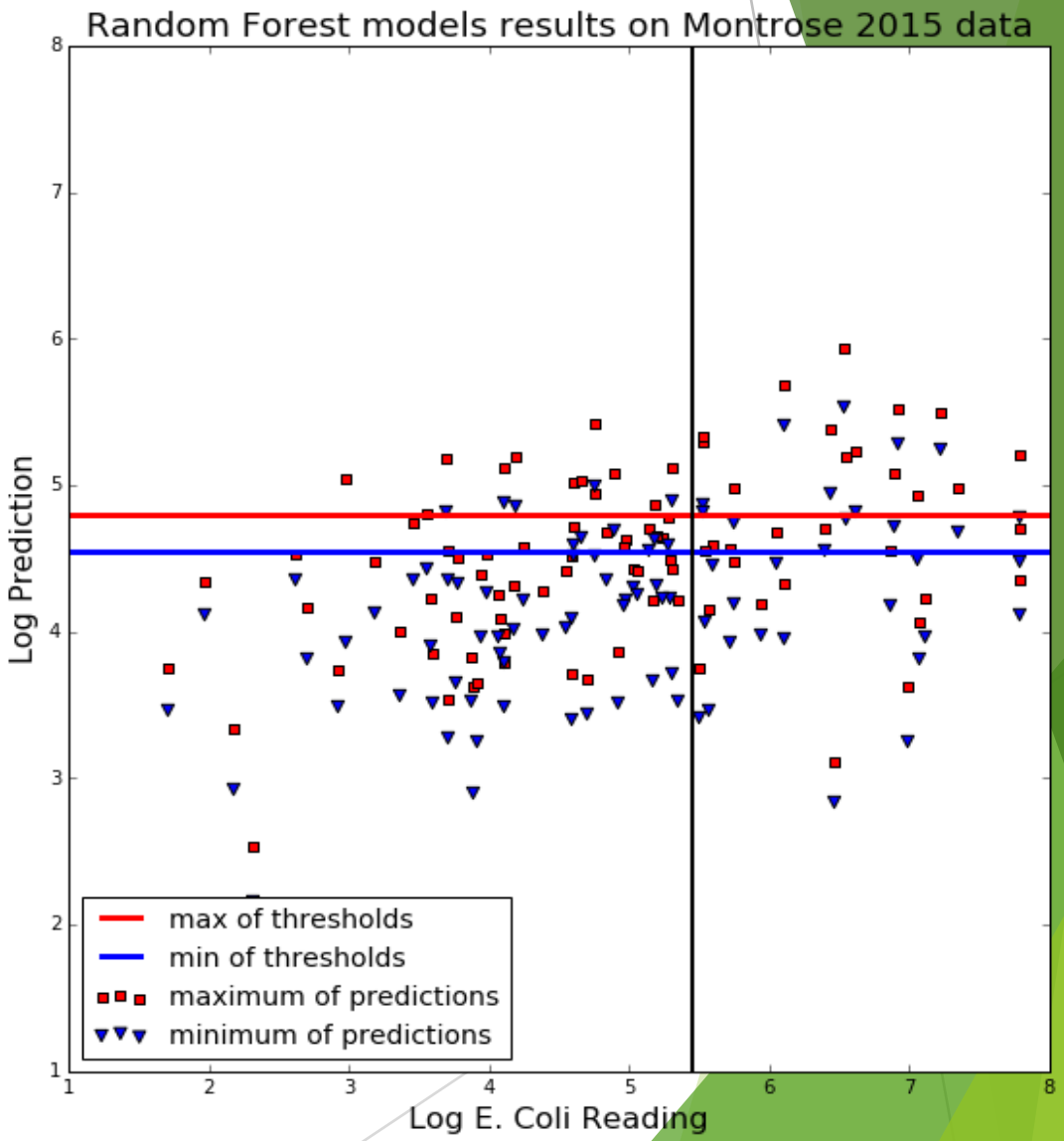




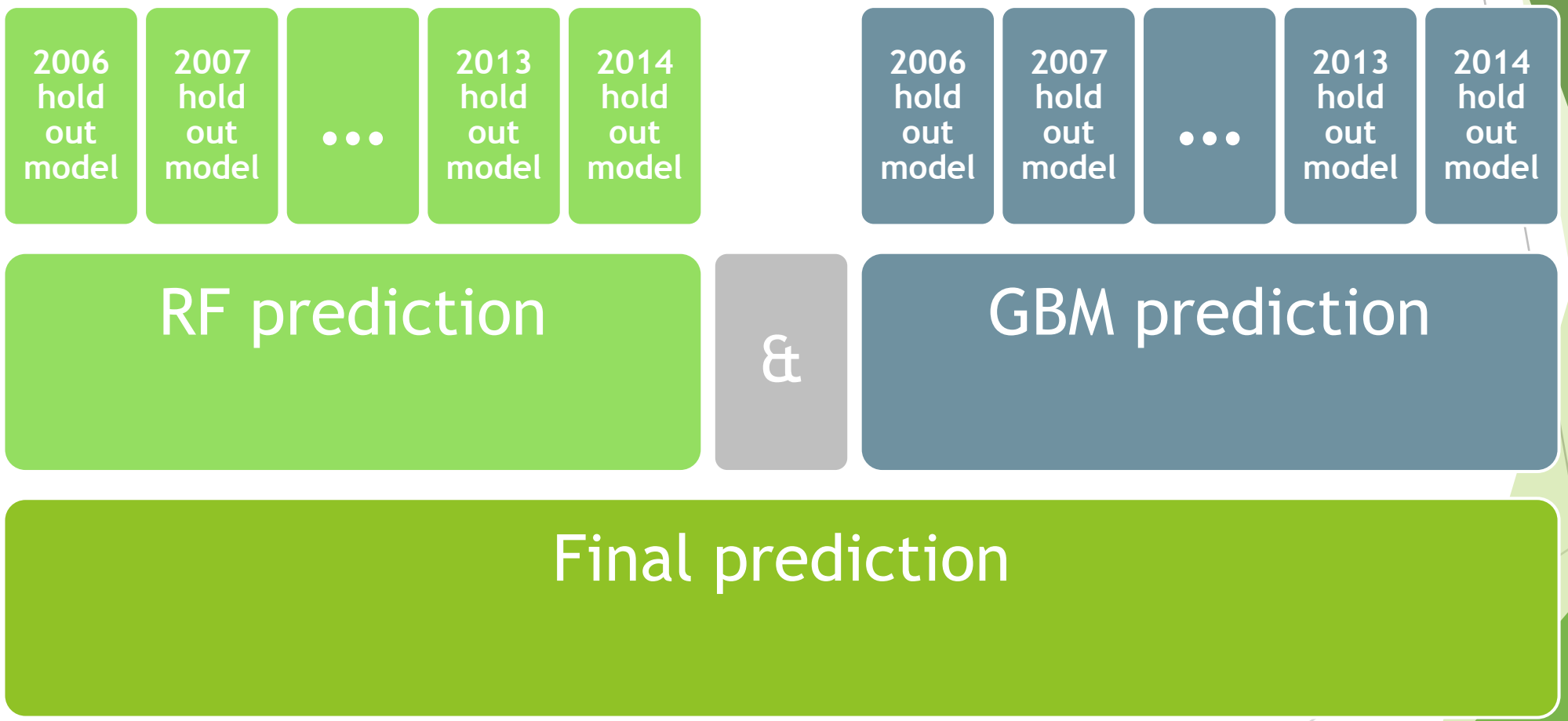


Composition and Comparison of Models

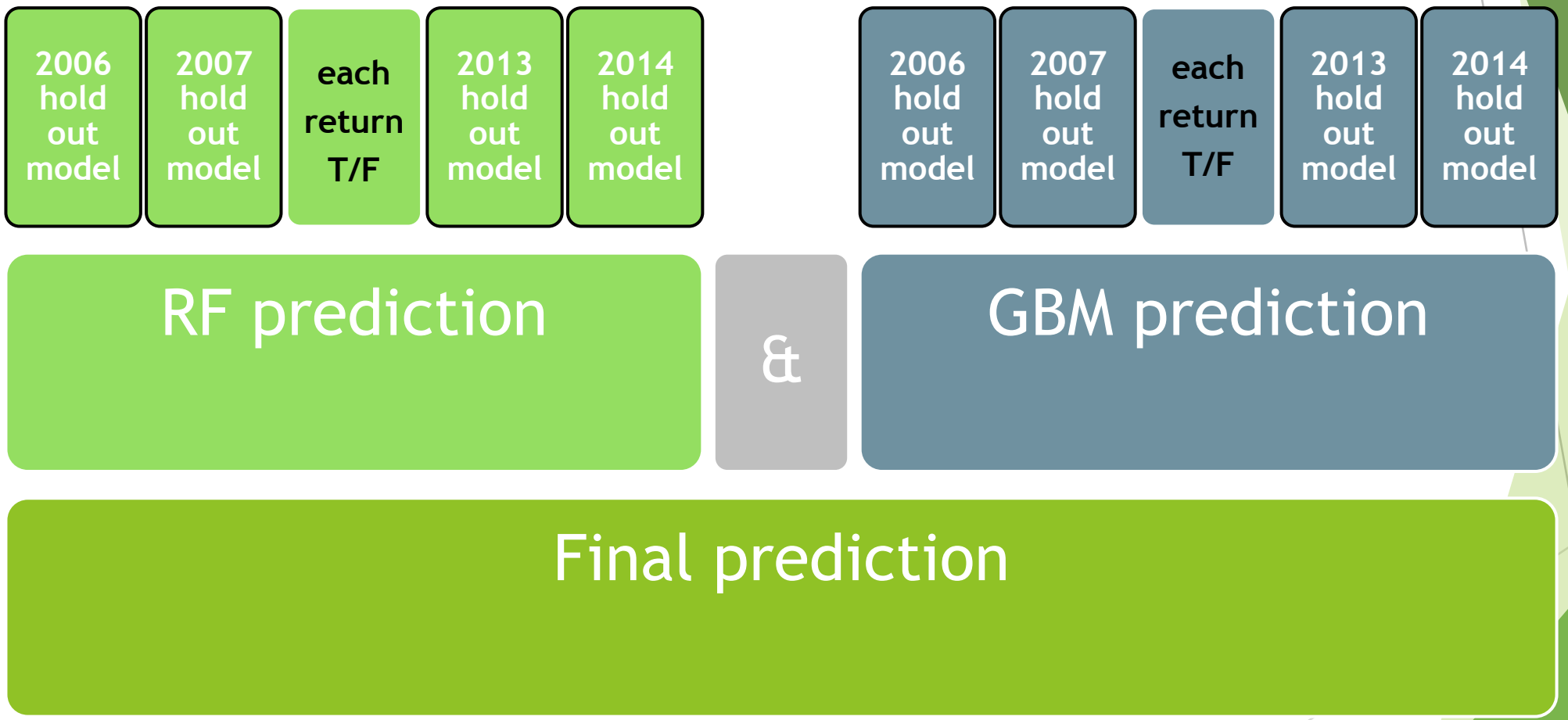




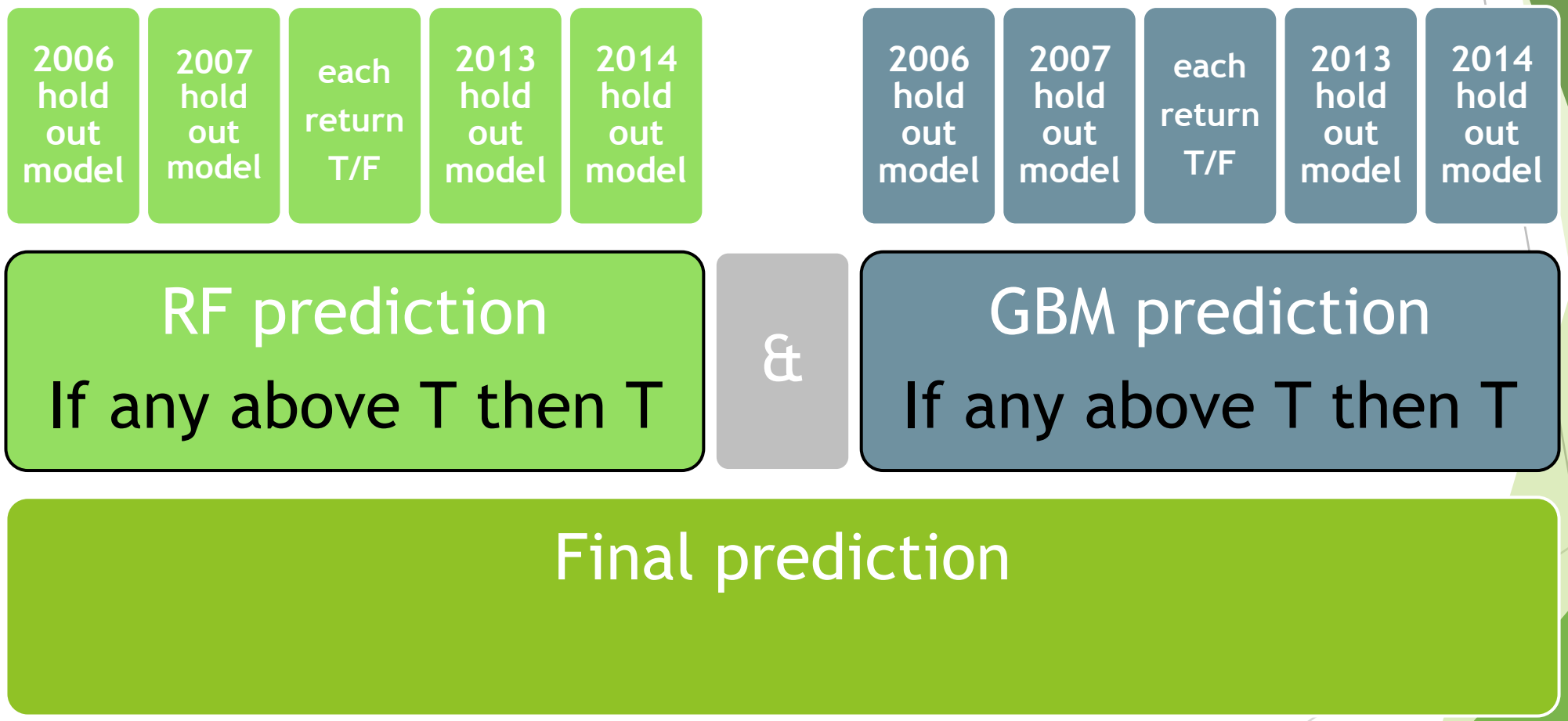
Ensemble of ensembles



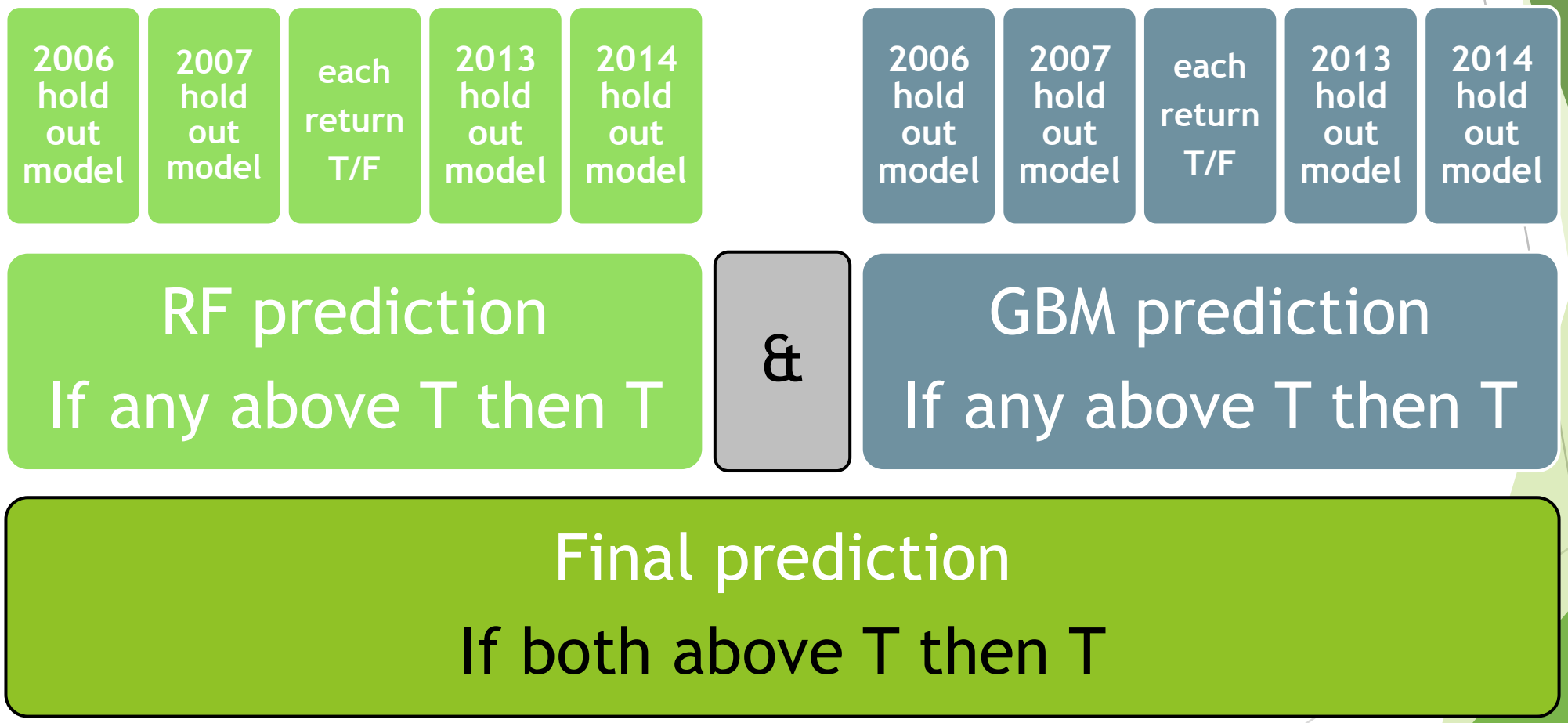
Ensemble of ensembles



Ensemble of ensembles



Ensemble of ensembles



Comparison of Current and Proposed models

Current Model predictions

BEACH	incorrect_warning	correct_warning	missed_warning
Juneway	0	0	5
Rogers	0	0	7
Howard	0	0	7
Jarvis	0	0	2
Leone	0	0	4
Albion	0	0	6
Osterman	0	0	14
Foster	1	0	9
Montrose	1	5	26
North Avenue	0	0	5
Oak Street	0	0	1
Ohio	1	0	13
12th	0	0	11
31st	2	1	12
39th	2	0	6
57th	2	1	4
63rd	2	1	11
South Shore	2	1	13
Rainbow	3	4	17
Calumet	0	0	30

Precision: 42% Recall: 6%

Ensemble of Ensemble predictions (FNR of 5%)

BEACH	incorrect_warning	correct_warning	missed_warning
Juneway	1	1	4
Rogers	0	1	6
Howard	1	1	6
Jarvis	1	1	1
Leone	2	0	4
Albion	0	1	5
Osterman	1	2	12
Foster	2	2	7
Montrose	11	17	14
North Avenue	1	0	5
Oak Street	0	0	1
Ohio	2	1	12
12th	3	0	11
31st	7	1	12
39th	4	0	6
57th	4	2	3
63rd	8	1	11
South Shore	7	5	9
Rainbow	16	8	13
Calumet	8	9	21

Precision: 40% Recall: 24%

Comparison of Current and Proposed models

Current Model predictions

BEACH	incorrect_warning	correct_warning	missed_warning
Juneway	0	0	5
Rogers	0	0	7
Howard	0	0	7
Jarvis	0	0	2
Leone	0	0	4
Albion	0	0	6
Osterman	0	0	14
Foster	1	0	9
Montrose	1	5	26
North Avenue	0	0	5
Oak Street	0	0	1
Ohio	1	0	13
12th	0	0	11
31st	2	1	12
39th	2	0	6
57th	2	1	4
63rd	2	1	11
South Shore	2	1	13
Rainbow	3	4	17
Calumet	0	0	30

Precision: 42% Recall: 6%

Ensemble of Ensemble predictions (FNR of 2%)

BEACH	incorrect_warning	correct_warning	missed_warning
Juneway	0	0	5
Rogers	0	0	7
Howard	0	0	7
Jarvis	0	0	2
Leone	0	0	4
Albion	0	0	6
Osterman	1	1	13
Foster	1	1	8
Montrose	3	13	18
North Avenue	0	0	5
Oak Street	0	0	1
Ohio	1	0	13
12th	1	0	11
31st	2	1	12
39th	2	0	6
57th	2	2	3
63rd	3	1	11
South Shore	5	2	12
Rainbow	7	7	14
Calumet	3	3	27

Precision: 50% Recall: 14%

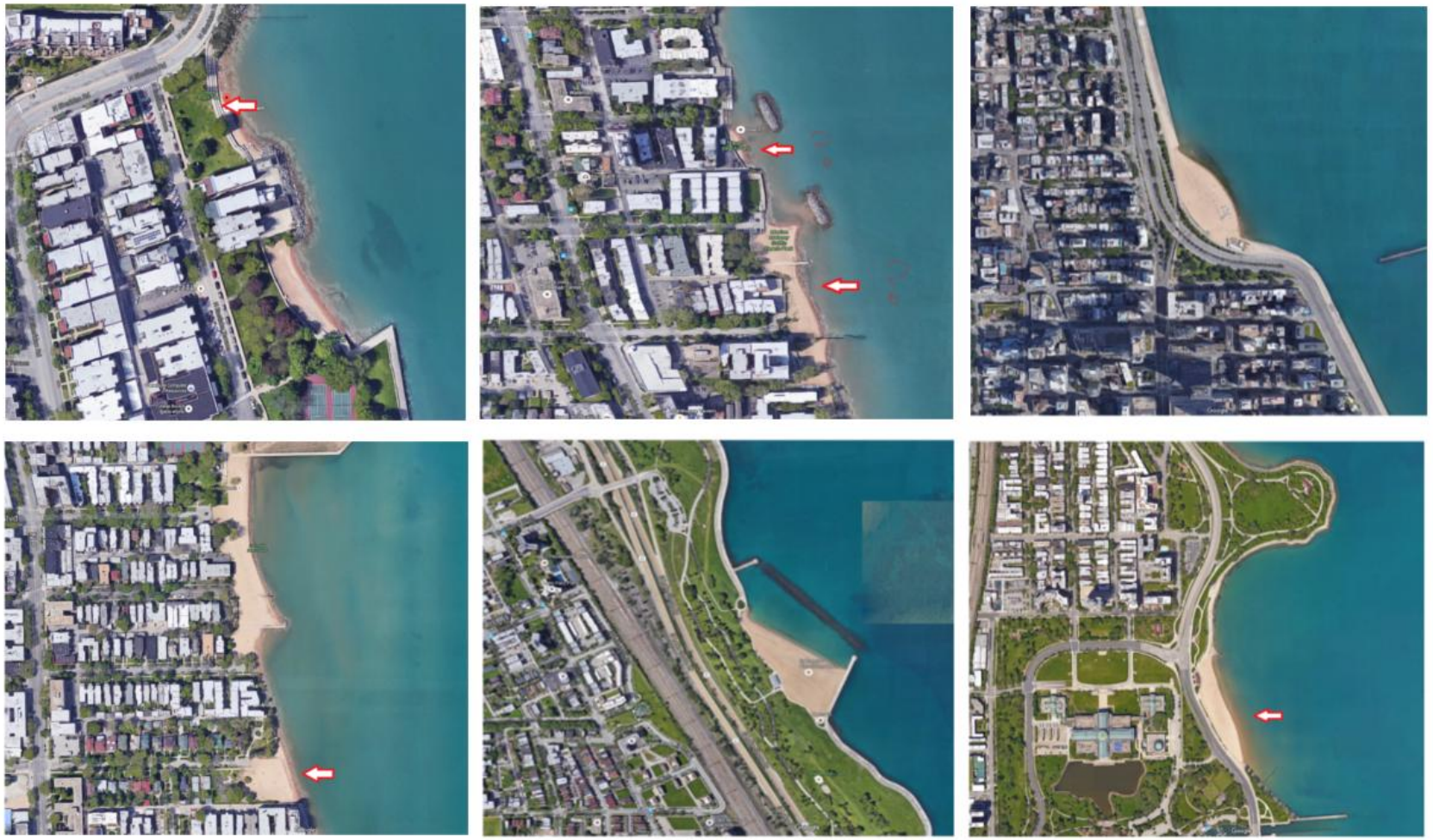
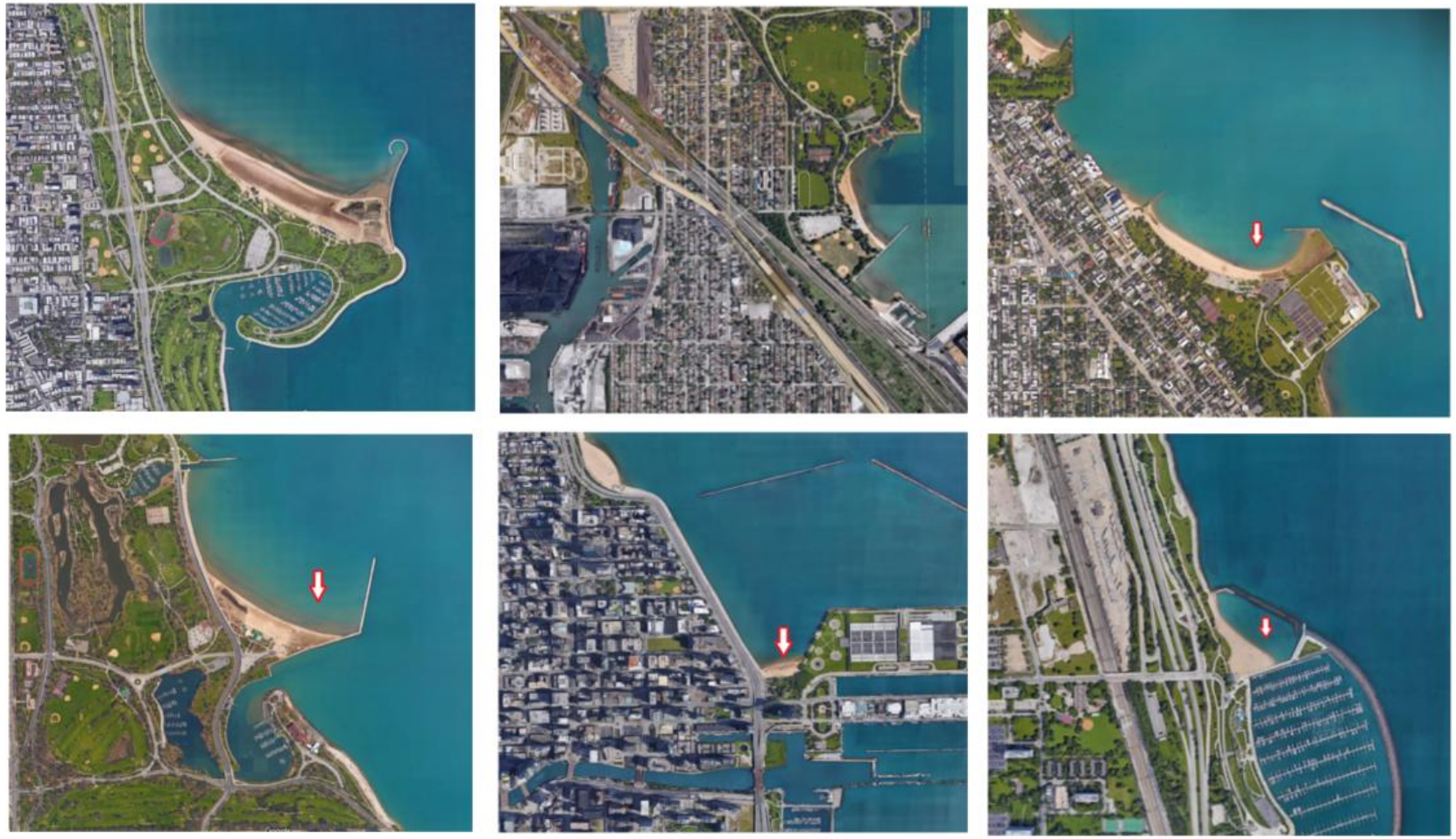
Collaborative Data Science : Pros

- Coding improved by collaboration
  - Reinforces better documentation habits
  - Learn from style of more advanced python users
- Opportunity to debate questions of data science
  - Learn from others
  - Teach others
- Familiarity with collaborative tools improved
  - Github, slack, waffle.io

Collaborative Data Science : Cons

- Redundancy of effort
  - Try something, doesn't work. No report that it was tried.
  - Code not uploaded and shared. Or not documented well enough to make it useable.
- Difficulty bringing people "on-board" mid-project
  - Documentation scattered.
  - Fossilized idiosyncrasies of project.

Beaches with few E. Coli exceedances



Beaches with many E. Coli exceedances