Data Integration and Knowledge Discovery with the International Shark Attack Database

Department of Computer Science
College of Computing and Informatics

July 26, 2016
ITCS 6265 Summer 2016

Research Team

Sailesh Bhamidipati

Sonal Kaulker

Pown Arthi Thimiri Dayasagar

Jai Kiran Duvvu

Dr. Pamela Thompson

Statement of Purpose

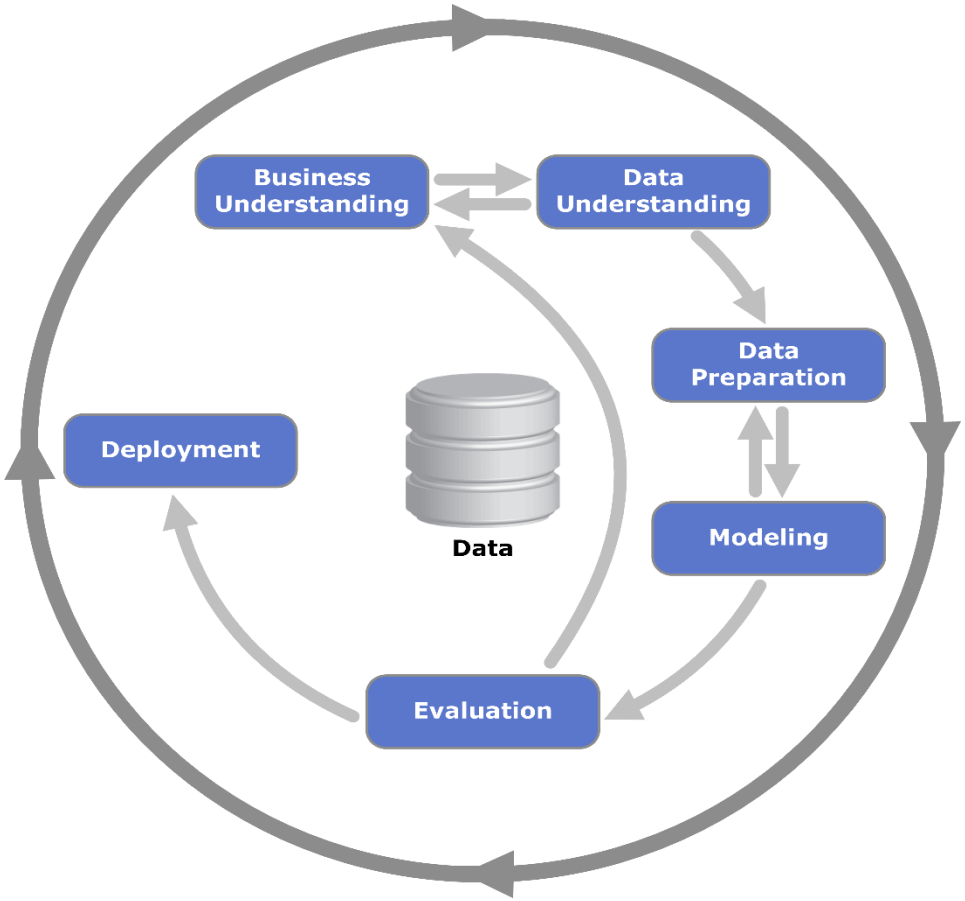
The objective of this research is to improve our understanding of the presence of sharks during tourist seasons in middle Atlantic and south eastern coastal waters, specifically North Carolina and South Carolina. Our study will focus on the analysis of existing data from the International Shark Attack database, weather and water data from NOAA, calculated moon phase dates, fish, crab and turtle populations. The quantitative analysis on this data will lead to new and interesting knowledge that may provide the basis for an app providing advanced information on the likelihood of sharks in coastal waters where tourists swim, surf and wade.

An additional focus of this research is to analyze social media activity relative to shark presence. Tweets will be mined with specific data ranges and locations in order to assess social media activity with respect to shark sightings, attacks, and even activities that are related to shark presence such as schools of fish. Once the nature of this activity is understood, a recommendation for a standardized way to tweet will be considered with interested and strategic partners in order to ultimately provide an additional feed to the app described above.

UNC CHARLOTTE

The process: CRISP-DM

Cross Industry Standard Process – Data Mining



UNC CHARLOTTE

Domain Understanding

Summer of 2015 brought a record number of attacks.

Summer of 2015 – many attacks

Researchers described “perfect storm” of conditions relative to many theories:

- Population – lots of tourists, increased human activity
- Moon phases – this is disputed
- Weather, global warming
- Water conditions
- Shark preservation efforts
- Fish migration (food source)

Domain Understanding

"Every time we go to the beach, we are invading a natural world that is already occupied by animals and plants in that area. We need to remember that when we enter the sea it's not the equivalent of going to the YMCA pools . . . It's a wild world out there.“

George H. Burgess, director of the International Shark Attack File at the University of Florida's Florida Museum of Natural History

Domain Understanding

"The population has been going up and the number of people going in the water is always increasing. The risk of any shark bite is already incredibly low—far less likely than drowning or many other rare risks. But, the more people you have going into the water, the better the odds are that something bad is going to happen, whether it's a shark bite or getting pulled under on a riptide."

Dr. Chuck Bangley, ECU Shark Researcher

(from http://www.scientificamerican.com/article/shark-bites-are-up-but-attack-risk-is-down/)

Domain Understanding

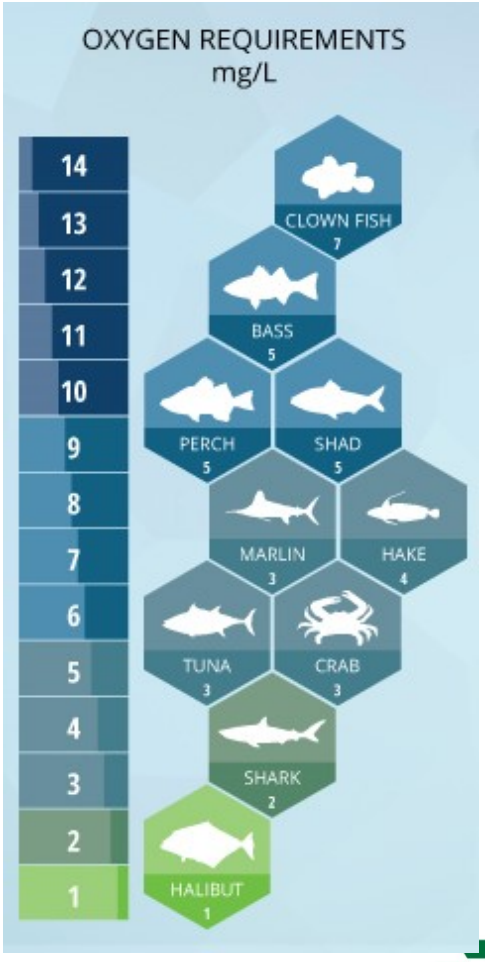
- Visit http://www.profthompson.net for interesting articles and resources related to shark attacks

Domain Understanding

- George Burgess also mentions parts of North Carolina have been abnormally dry or have experienced moderate drought conditions for several weeks.
- The salinity, or salt content, of ocean water close to shore is higher than usual, which is being cited as a possible reason for the surge in the shark attacks.
- Baby and other sea turtles may draw sharks to the North Carolina shores.
- Is it the warming ocean causing the sharks to follow prey that are migrating due to warmer temperatures perhaps caused by climate change?
- The annual migration of menhaden fish, a favorite shark food, appears linked to water temperature, which jumped 10 degrees in a week during the heat wave of 2015 (Burgess, Scientific American)
- What is true? What are the other factors? Is there hidden knowledge that can be discovered?

Domain Understanding: Water

- Salinity, Turbidity, Oxygen, Sea Water Temperature from East Cribbing Station in NC (used for all data, average readings during frequent attack times)
- Oxygen: Gulf coast experiences “dead zones” caused by low oxygen levels due to fertilizer run off
- Suspected as cause of higher shark attacks due to low fish populations
- Daily measure included in our research



UNC CHARLOTTE

Domain Understanding: Weather

- Temperature, Precipitation, Moving Average Precipitation, Wind Speed, Wind Direction
- Daily readings from NOAA
- Daily measures included in our research
- Sharks like warmer weather, will migrate as seasons change
- People go to the beach when it is hot

Domain Understanding: Crabs

Sharks eat crabs.

Crabs have more frequent movement during full and new moon phases.

Crab landings data for NC (daily) is available courtesy of Alan Bianchi, North Carolina Division of Marine Fisheries

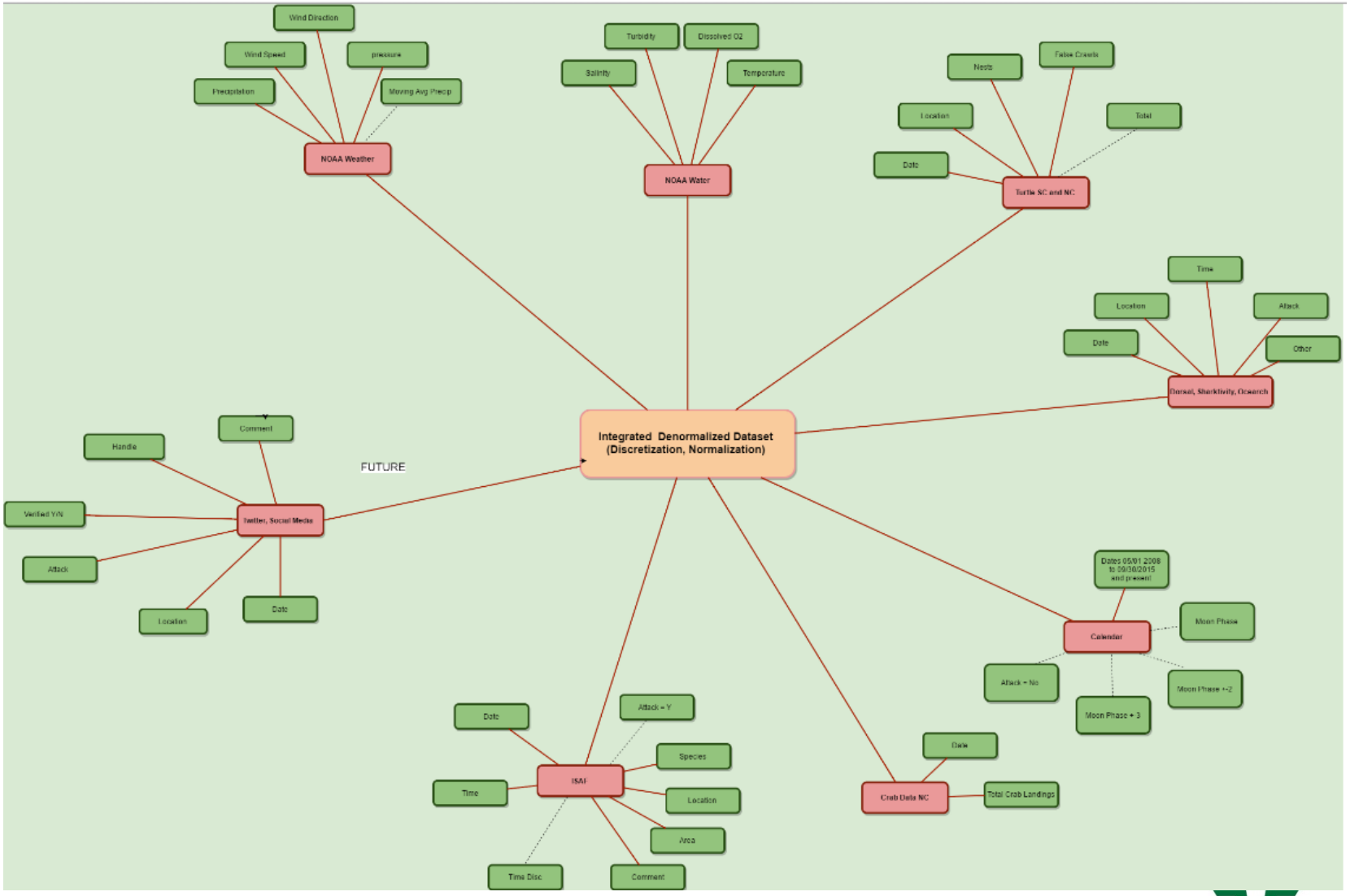
Domain Understanding: Turtles

- Sharks eat turtles
- NC and SC have many nests
- Turtles move from ocean to beaches for nesting and false crawl (they go to beach but don't lay eggs)
- Turtle data for NC and SC courtesy of
- Michelle Pate and Dr. Matthew Godfrey, State Coordinators, SC and NC Wildlife Resources Commissions
- Privacy and security concerns are always important!

Domain Understanding: Moon Phases

- Moon phases affect water levels
- Full moon and new moon cause levels to rise higher than normal during tides
- Effect starts 2-4 days before a phase and extends 2-4 days after
- Spring **tides** happen when the sun and **moon** are on the same side of the earth (New **Moon**) or when the sun and **moon** are on opposite sides of the earth (Full **Moon**).
- One study shows no effect: http://benthamopen.com/contents/pdf/TOFISHSJ/TOFISHSJ-6-71.pdf

Data Preparation



UNC CHARLOTTE

Data Preparation

Numeric values: Discretized, Normalized for new attributes (3 bin)

Precipitation: 5 day moving average added

Moon phases: extended for Full, New Moon

Crab Data: For attack = Y, added by date

Crab Data: For attack = N, Necessary imputations were made

NOTE: Class imbalance problem! Handled with *stratified* sampling of Attack=No Subset so that 1/3 of records remain with adequate representation for each date.

Modeling:
Exploratory Data Analysis

FIRST ROUND: with the discretized data base using R, WEKA

Classification: Naïve Bayes

Clustering: Simple EM (Expected Maximization)

Association Rule: Apriori

NOTE: See github http://www.github.com/AKDDResearch/Shark-Attack for complete Source of data and files for this research

Modeling: Classification (discretized dataset, reduced dimensionality)

Naive Bayes:

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     june_aug_weka_csv_discretized_reduced_to_1_3_random_no-
weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.unsupervised.attribute.Remove-R2-3-
weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R2-
weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R2-
weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R5-
weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R10-
weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Remove-R3
Instances:    285
Attributes:   11
              Turtle_Discretize
              Attack
              MoonPhaseExtended
              DissolvedO2_discretize
              salinity_discretize
              turbidity_discretize
              temperature_discretize
              pressure_discretize
              windspeed_discretize
              precipitationmva_discretize
              Crab_Landings_Discretize
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                          Class
Attribute                   No    Yes
```

Modeling: Clustering

Simple EM (Expected Maximization)

Clustering

Simple EM Maximization (see NOTE at end)

Results:

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Relation: june_aug_weka_csv_discretized_reduced_to_1_3_random_no-weka.filters.unsupervised.attribute.Remove-R1-3,5-6-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R2-4-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Remove-R3-4-weka.filters.unsupervised.attribute.Remove-R4-weka.filters.unsupervised.attribute.Remove-R8

Instances: 285

Attributes: 11

- Turtle_Discretize
- Attack
- MoonPhaseExtended
- DissolvedO2_discretize
- salinity_discretize
- turbidity_discretize
- temperature_discretize
- pressure_discretize
- windspeed_discretize
- precipitationmva_discretize
- Crab_Landings_Discretize

Test mode: evaluate on training data

=== Clustering model (full training set) ===

Attribute	Cluster 0 (0.11)	1 (0.06)	2 (0.17)	3 (0.65)
Turtle_Discretize				
Medium	6.1868	10.8648	20.8048	66.1436
Low	23.4523	3.2824	11.3272	106.9381
High	3.1305	6.7933	19.6859	11.3903
Very High	2.9729	1.1292	1.9987	4.8992
[total]	35.7424	22.0697	53.8166	189.3712
Attack				
No	31.965	15.8598	5.5238	170.6514
Yes	1.7775	4.21	46.2927	16.7198
[total]	33.7424	20.0697	51.8166	187.3712
MoonPhaseExtended				
Full	5.4434	5.4229	16.705	45.4288
Third quarter	2.7519	3.185	2.0362	25.0269
New	12.7777	1.9686	14.3133	54.9404
First quarter	4.9532	4.5346	9.5607	21.9515
Waning gibbous	1.8467	2.8068	4.4876	15.8588
Waxing gibbous	3.7436	3.1043	5.8625	8.2897
Waning crescent	1.8039	2.3745	1.9528	9.8688
Waxing crescent	6.422	2.6731	2.8985	12.0064
[total]	39.7424	26.0697	57.8166	193.3712
DissolvedO2_discretize				
Low	7.2264	3.2664	8.219	6.2882
Medium	17.1091	13.4138	41.5236	180.9535
High	10.4069	4.3895	3.074	1.1296
[total]	34.7424	21.0697	52.8166	188.3712
salinity_discretize				
High	1.5592	2.8721	19.4847	158.0839
Medium	31.9402	5.5168	31.3896	29.1534
Low	1.243	12.6808	1.9423	1.1339
[total]	34.7424	21.0697	52.8166	188.3712
turbidity_discretize				
Low	26.6803	15.8878	47.8739	184.558
High	2.0227	1.0415	1.0268	1.909
Medium	6.0394	4.1404	3.9159	1.9043
[total]	34.7424	21.0697	52.8166	188.3712
temperature_discretize				
High	24.126	9.5087	49.2901	186.0752
Medium	8.7728	10.504	2.5082	1.215
Low	1.8436	1.057	1.0183	1.0811
[total]	34.7424	21.0697	52.8166	188.3712

Modeling: Clustering

Simple EM (Expected Maximization)

```
pressure_discretize
  Medium                    23.7769   9.3064  31.2966 132.6201
  High                       5.9865   8.4276  19.3288  47.2571
  Low                        4.9791   3.3357   2.1912   8.494
  [total]                   34.7424  21.0697  52.8166 188.3712
windspeed_discretize
  Low                       22.4374   8.862    6.6265 129.0741
  Medium                     9.5352  11.1864  28.6779  57.6005
  High                       2.7698   1.0213  17.5122   1.6966
  [total]                   34.7424  21.0697  52.8166 188.3712
precipitationmva_discretize
  Low                       30.6367  17.9498   49.55  161.8635
  Medium                     1.2302   2.0038   2.2526  24.5134
  High                       2.8755   1.1162   1.0139   1.9944
  [total]                   34.7424  21.0697  52.8166 188.3712
Crab_Landings_Discretize
  High                       6.0063   9.2566  33.8044  41.9328
  Medium                    14.3684   2.2614   8.8114  73.5588
  Low                       14.3678   9.5517  10.2008  72.8797
  [total]                   34.7424  21.0697  52.8166 188.3712


Time taken to build model (full training data) : 5.74 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       25 (  9%)
1       15 (  5%)
2       48 ( 17%)
3      197 ( 69%)
```

Modeling: Association Rule Mining

As water temp and salinity increase, Dissolved oxygen levels decrease. Temp and Salinity removed for Association Rule Mining. Best rules:

5. MoonPhaseExtended=New DissolvedO2_discretize=Low 7 ==> Attack=Yes 7 conf:(1)

11. Turtle_Discretize=High MoonPhaseExtended=New 6 ==> Attack=Yes 6 conf:(1)

18. MoonPhaseExtended=New DissolvedO2_discretize=Low turbidity_discretize=Low 5 ==> Attack=Yes 5 conf:(1)

Dissolved Oxygen levels and "jubilee" effect:
http://oceanservice.noaa.gov/education/kits/estuaries/media/supp_estuar10d_disolvedox.html

Association Rule Mining

=== Run information ===

Scheme: weka.associations.Apriori -N 15000 -T 0 -C 0.5 -D 0.05 -U 0.05 -M 0.01 -S 0.05 -A -c 2

Relation: june_aug_weka_csv_discretized_reduced_to_1_3_random_no-
weka.filters.unsupervised.attribute.Remove-R1-3,5-6-weka.filters.unsupervised.attribute.Remove-R2-
weka.filters.unsupervised.attribute.Remove-R2-4-weka.filters.unsupervised.attribute.Remove-R2-
weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Remove-R3-4-
weka.filters.unsupervised.attribute.Remove-R4-weka.filters.unsupervised.attribute.Remove-R8-
weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.Remove-R10-
weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.Remove-R7-
weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R6

Instances: 285

Attributes: 5

Turtle_Discretize

Attack

MoonPhaseExtended

DissolvedO2_discretize

turbidity_discretize

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.01 (3 instances)

Minimum metric <confidence>: 0.5

Significance level: 0.05

Number of cycles performed: 20

Generated sets of large itemsets:

Preliminary Results

Attributes that are interesting:

Turbidity

Turtle Discretized

Moon Phase

Temperature

Wind Speed

Wind Direction

Twitter Hash Tag Analysis

Twitter Mining on Shark Attacks

North Carolina and South Carolina

Recommendations

- DATA – continue to improve the data collection and analysis with location awareness where possible (we used East Cribbing data for both NC and SC)
  - Attributes of interest:
    - Weather: Wind speed, Direction, Temperature, Turbidity
    - Environment: Crab Landings, Turtles
  - Sightings: A feed from social media, as close to real time as possible (need better classifier)
    - Other sources: coordination with apps, rescue personnel, beach patrols
    - STANDARDIZED HASH TAG
      - Collaboration (https://twitter.com/DorsalAus, Sharktivity, other researchers)
      - Publicity
      - Use:
        - Shark sightings (see @sharkreports, @dorsalau on Twitter)
        - Turtle activity
        - Schools of fish (draw sharks)
- CONTINUED STUDY
  - Research – work on binnings, class imbalance, Chlorophyll A, fish, social media feed
  - COLLABORATION – shark researchers
  - App for warnings, sightings – collaboration with DORSAL, others
  - Grant

Recommendations

- Some Considerations
  - Will better awareness of conditions and sightings affect tourism? (Dr. Craig Depken, Economics at UNCC, investigating media reports of attacks and tourism
  - Will individuals try to “game” the system?
  - Are we ever able to say it is “safe” to swim?
  - How about liability?

Questions?

"Our results indicate that investing in increasing and communicating our understanding of the behavior, distribution and ecological role of sharks as well as the factors influencing the risk of shark bites, may ultimately be the most effective way to increase safety of people. If people learn to avoid being near shark food during feeding times, we become far less likely to end up an as accidental appetizer."

Dr. Francesco Ferretti, Stanford, commenting on a recent study in http://www.scientificamerican.com/article/shark-bites-are-up-but-attack-risk-is-down/

Selected References

Scientific American: Shark Bites are Up, But Attack Risk is Down? http://www.scientificamerican.com/article/shark-bites-are-up-but-attack-risk-is-down/

Hashtag Standards for Emergencies: https://docs.unocha.org/sites/dms/Documents/TB%20012_Hashtag%20Standards.pdf

Lunar Cycle Effects: http://benthamopen.com/contents/pdf/TOFISHSJ/TOFISHSJ-6-71.pdf

What 3 Words: http://what3words.com/

Dr. Pamela Thompson blog: http://www.profthompson.net

Dorsal app: https://www.dorsalapp.com/

Sharktivity map and app: http://www.atlanticwhiteshark.org/sharktivity-map/