UC San Diego
SCHOOL OF MEDICINE

Department of BioMedical Informatics

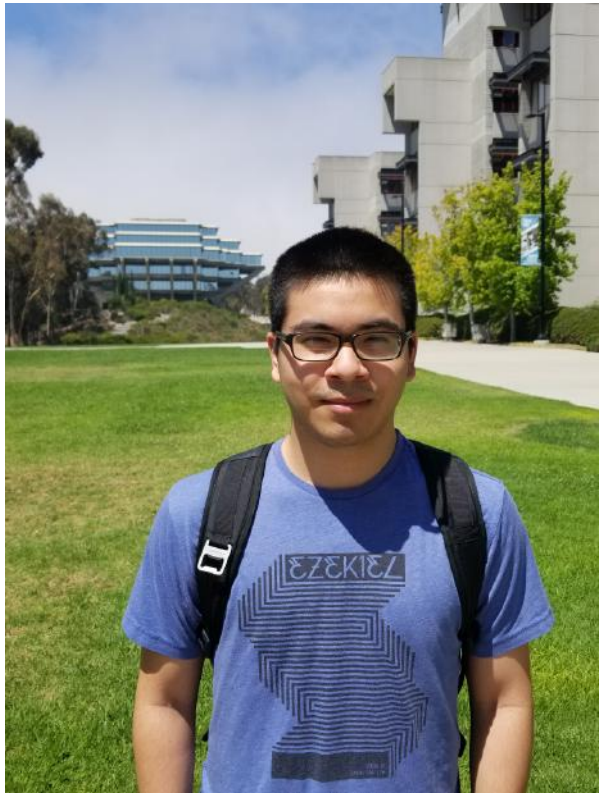
Justin Castro
Dr. Cinnamon Bloss & Dr. Jejo Koola

Applying Supervised Learning to Aid the Coding of Qualitative Data

This research was supported by grant T15LM011271

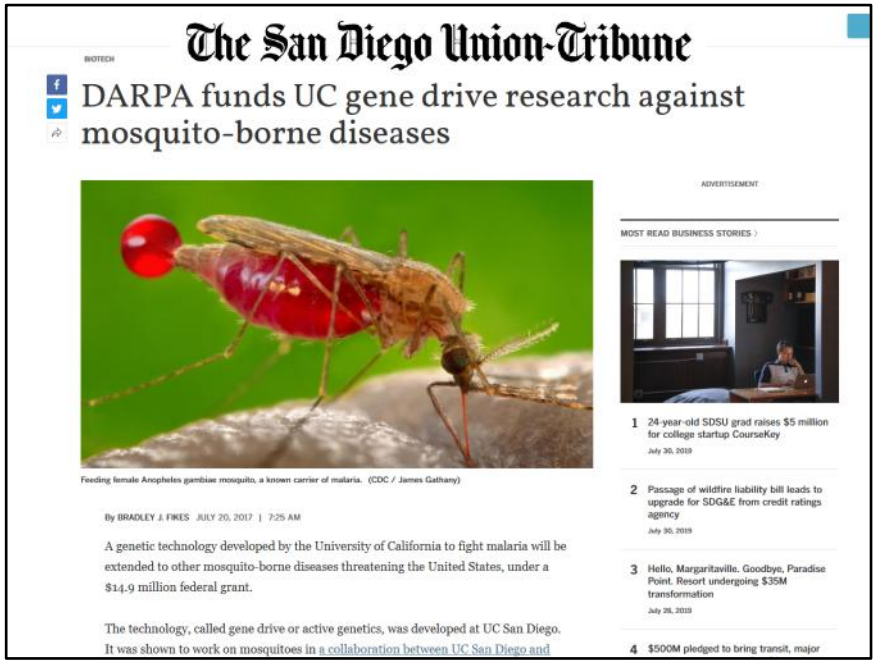
About Me

- 3rd Year Data Science Major
- Worked alongside Dr. Bloss's lab.
- Applied to the DBMI program to learn more about Data Science in a clinical setting and machine learning.



This research was supported by grant T15LM011271

Controlling Vector-Borne Disease with Genetic Engineering



The San Diego Union-Tribune

BIOTECH

DARPA funds UC gene drive research against mosquito-borne diseases

Feeding female Anopheles gambiae mosquito, a known carrier of malaria. (CDC / James Gathany)

By BRADLEY J. FIKES JULY 20, 2017 | 7:25 AM

A genetic technology developed by the University of California to fight malaria will be extended to other mosquito-borne diseases threatening the United States, under a $14.9 million federal grant.

The technology, called gene drive or active genetics, was developed at UC San Diego. It was shown to work on mosquitoes in a collaboration between UC San Diego and

ADVERTISEMENT

MOST READ BUSINESS STORIES

1 24-year-old SDSU grad raises $5 million for college startup CourseKey
July 30, 2019

2 Passage of wildfire liability bill leads to upgrade for SDG&E from credit ratings agency
July 30, 2019

3 Hello, Margaritaville. Goodbye, Paradise Point. Resort undergoing $35M transformation
July 26, 2019

4 $500M pledged to bring transit, major

https://www.sandiegouniontribune.com/business/biotech/sd-me-darpa-ucsd-20170719-story.html

- Part of a larger collaboration to research solutions to vector borne disease.
  - GE Sterile Males
  - Gene Drive
- Focusing on the social and ethical perspective.

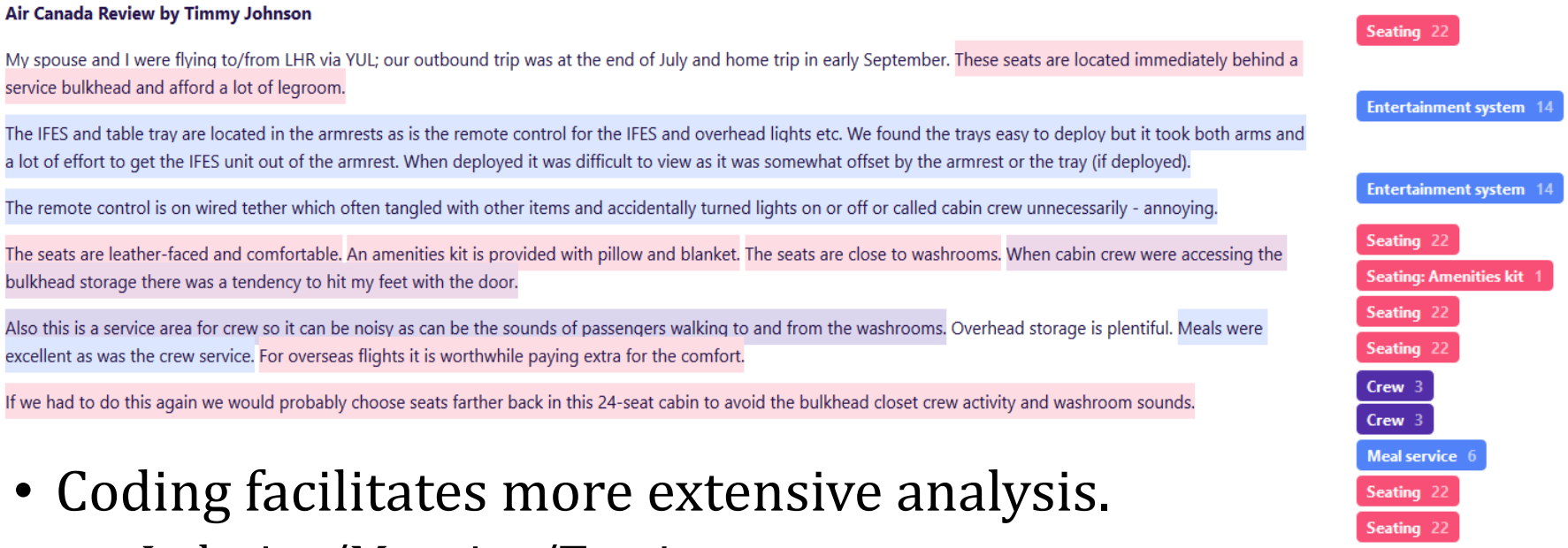
How can researchers incorporate concerns into the design of gene editing technologies?

What concerns do citizens have?

Does being more knowledgeable about GE technology change how people view it?

This research was supported by grant T15LM011271

Analyzing Qualitative Data: Coding



- Coding facilitates more extensive analysis.
  - Indexing/Mapping/Tagging
  - Labeling your text allows you to revisit it in the analysis process
  - Makes the actual analysis much easier
- 30+ codes and 2300+ chat messages over 13 FG's.
- **Can we use ML to aid this tedious process?**

This research was supported by grant T15LM011271

Preprocessing

- Qualitative data is complicated and messy

130: people on nex door ocmplaining aobut being bit during day - i think so cal peopel just arent' used to mosquitoes

711: i am 711 am i in

714: 711 has me craving a slurpee. ;)

711: thats funny

- Are all words equally important?

"724: I am not sure of the risks associated with **Gene Driving.** It might be a solution ..... or it **might also create another problem.**"

"702: I'm for the most **cost effective and like that they are** not locally confined so I guess Sustained **Gene Drive"**

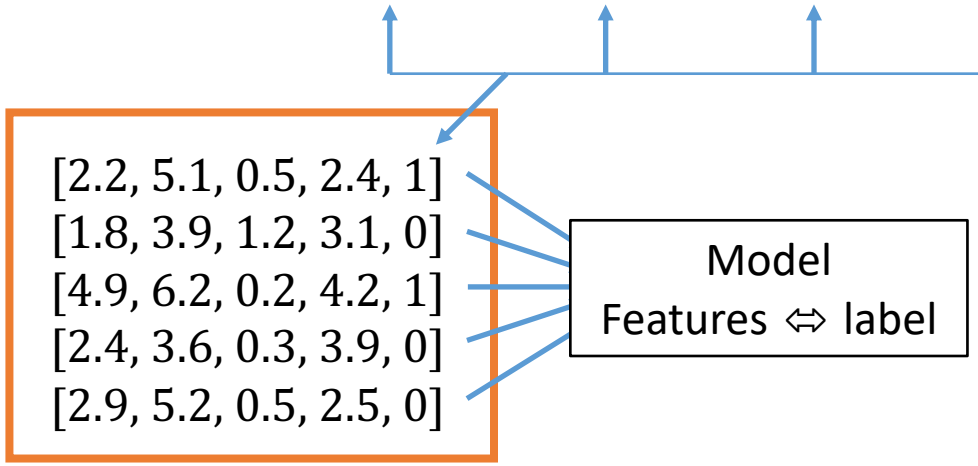
"711: would the **gene drive still** let them be a **part of the chain"**

- Computers do not understand language

This research was supported by grant T15LM011271

Feature Selection and Vectorization: Quantitative Data

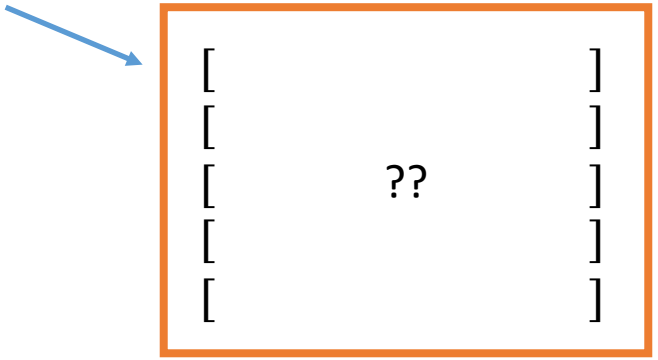
Flower type	Collector ID	Petal Width	Petal Length	Sepal Width	Sepal Length
sunflower	2	2.2	5.1	0.5	2.4
dandelion	2	1.8	3.9	1.2	3.1
sunflower	4	4.9	6.2	0.2	4.2
dandelion	4	2.4	3.6	0.3	3.9
dandelion	4	2.9	5.2	0.5	2.5



This research was supported by grant T15LM011271

Feature Selection and Vectorization: Qualitative Data

Response	Cost	Diseases	…	Ecosystem
“724: I am not sure of the risks associated with Gene Driving. It might be a solution ..... or it might also create another problem.”	0	0	…	0
“702: I'm for the most cost effective and like that they are not locally confined so I guess Sustained Gene Drive”	1	0	…	0
…	…	…	…	…
“711: would the gene drive still let them be a part of the chain”	0	0	…	1



This research was supported by grant T15LM011271

Feature Selection and Vectorization: Qualitative Data

a_1	a_2	a_3	…	a_n
1	4	1	…	2
0	3	1	…	1
0	2	0	…	0
1	2	0	…	1
1	2	0	…	0
0	3	0	…	1
0	2	0	…	0
…	…	…	…	…
0	1	0	…	0

Responses

Unique Words

- Bag of Words representation, Term Document Matrix, Document Vector Matrix.
- One way of representing documents mathematically.
  - Word Embeddings (Kevin)

Columns = All unique words across ALL responses (the corpus)

Rows = Individual responses

Values = Times that word appears in the document

This research was supported by grant T15LM011271

Feature Selection and Vectorization: Qualitative Data

a_1	a_2	a_3	…	a_n
.4251	.4947	.5692	…	.4928
0	.4256	.5692	…	.4297
0	.3985	0	…	0
.4251	.3985	0	…	.4297
.4251	.3985	0	…	0
0	.4256	0	…	.4297
0	.3985	0	…	0
…	…	…	…	…
0	.2718	0	…	0

Responses

Unique Words

- Term Frequency – Inverse Document Frequency (TF-IDF) Vectorization
- Gives more weight to words that are unique to that document.

$$tf_{i,d} \times \log \frac{Total\ \#\ of\ Documents}{df_{i,}}$$

tf = term frequency

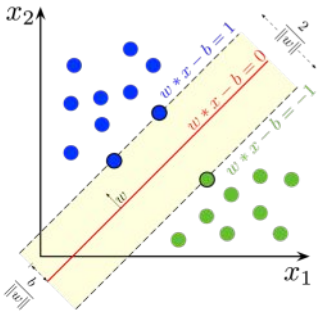
df = document frequency

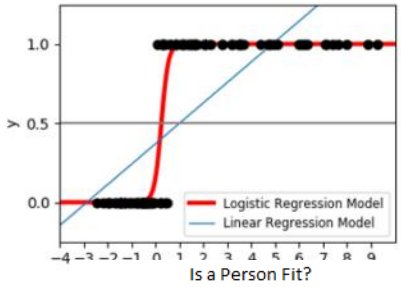
This research was supported by grant T15LM011271

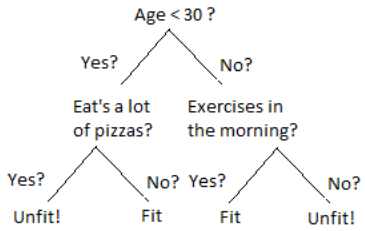
Machine Learning Models

- Disease vs. Not Disease
  - Mentions of diseases or negative health consequences.
- Models
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



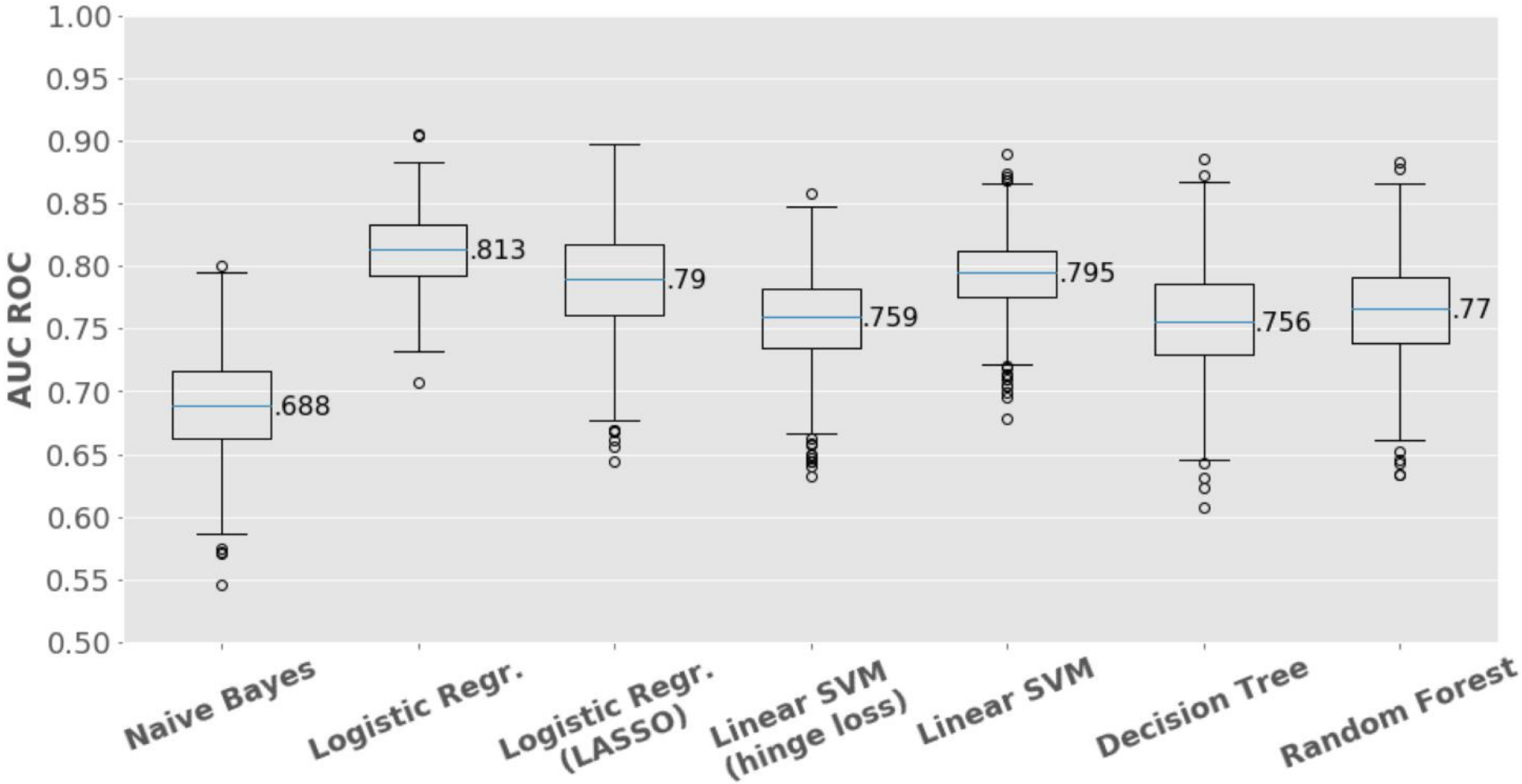




- Internally Validated with bootstrapping.
- Externally Validated with FDA comment dataset.
- Area Under ROC Curve (AUC ROC) primary metric of model discrimination.

This research was supported by grant T15LM011271

Comparing Model Performance



This research was supported by grant T15LM011271

Logistic Regr. and Naïve Bayes: a closer look

Logistic Regr.

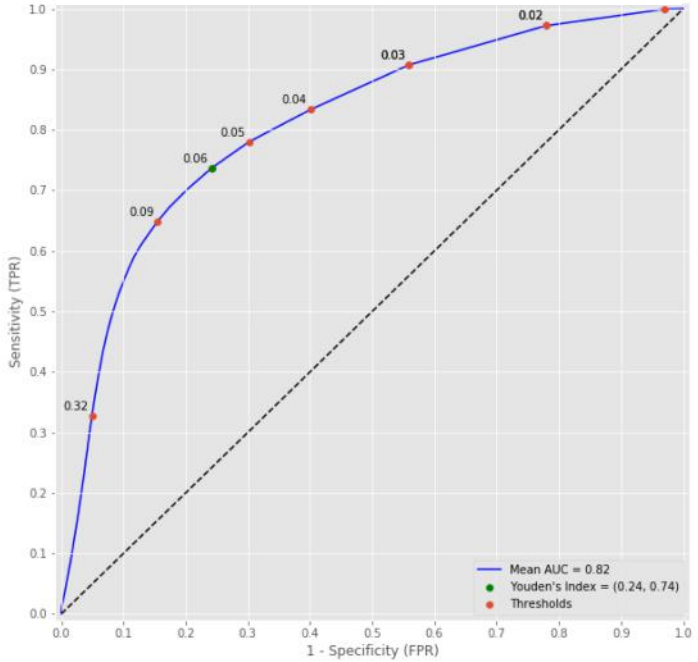
word	β
disease	4.69
health	2.18
carry	1.81
life	1.01
eliminate	1.01
zika	1.01
malaria	1.01
risk	.992
spread	.955

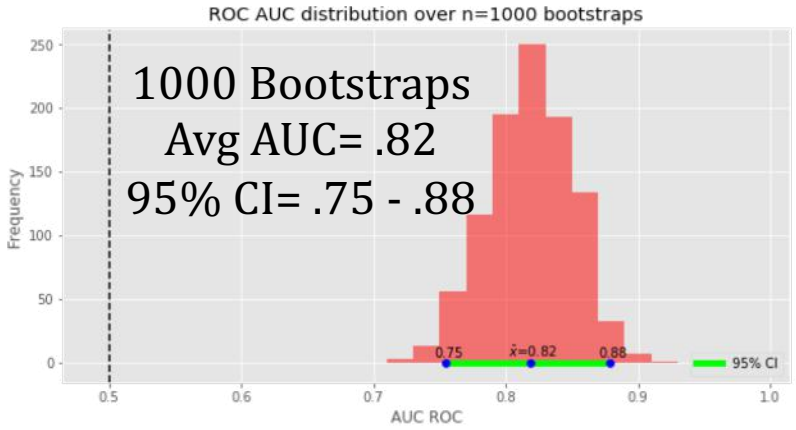
Naïve Bayes

word	weight
disease	-5.07
health	-5.93
carry	-5.98
mosquito	-6.19
get	-6.24
control	-6.26
spread	-6.30
eliminate	-6.31
people	-6.33

Higher = more indicative of "Disease"

Logistic Regr. ROC curve





This research was supported by grant T15LM011271

But is the Model Actually Generalizable? External Validation

- FDA Comments Dataset
  - 2600 individual comments sent to the FDA in 2014.
  - Potential impact of Oxitec OX513A Mosquito deployment in Monroe County, Florida.
    - Would release modified male mosquitoes whose offspring would carry a gene that produced a lethal protein.

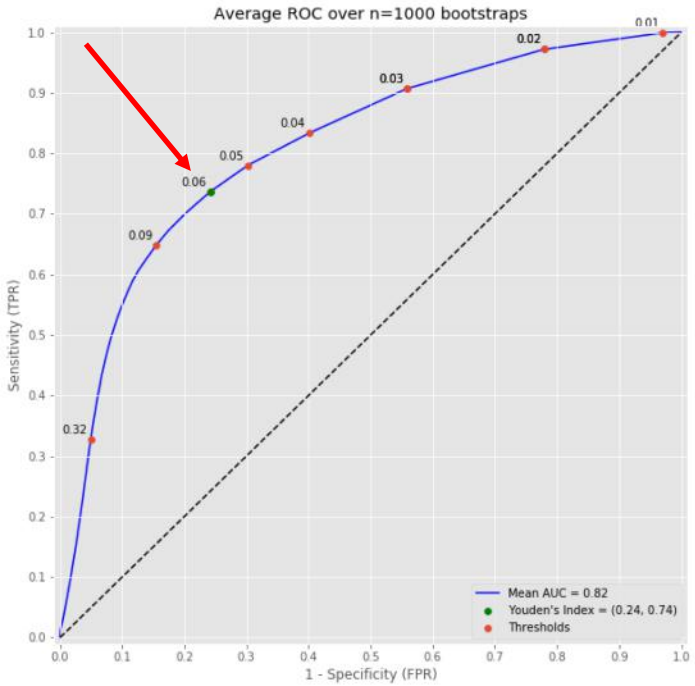
"Use the GMO mosquitos NOW, before there's thousands of cases of microcephaly here in Florida. Waiting is not an option."

"PLEASE don't release these GMO mosquitoes. Please STOP this madness. Nature knows best not science."

This research was supported by grant T15LM011271

Generalizing a Model: FDA comments

TP = True Positive FP = False Positive
FN = False Negative TN = True Negative



Expected:
FPR = .24
TPR = .74

Ground Truth

Predicted	Disease	Not Disease
D	**62** TP	**38** FP
ND	**27** FN	**73** TN

67.5% Accuracy
F1 Score = .66
FPR = .34
TPR = .697

This research was supported by grant T15LM011271

Takeaways and the Future

- Garbage in, Garbage out.
  - Preprocessing and Feature Selection is key.
- Is machine learning a reliable coding method?
  - Depends on aims of qualitative analysis.
  - Depends on how distinct codes are, distinguishing features are key.
- Limitations of BOW representations.
- New territory for ML aided coding.

. . .

This research was supported by grant T15LM011271

UC San Diego
SCHOOL OF MEDICINE

Department of BioMedical Informatics

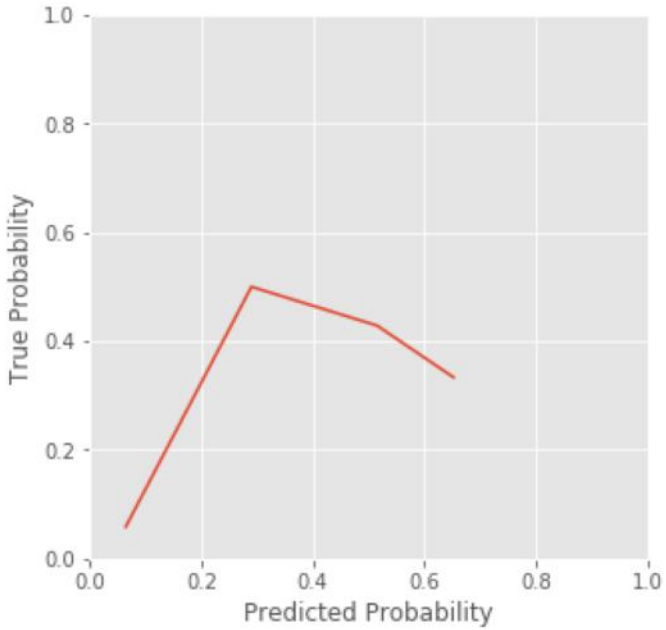
Acknowledgements

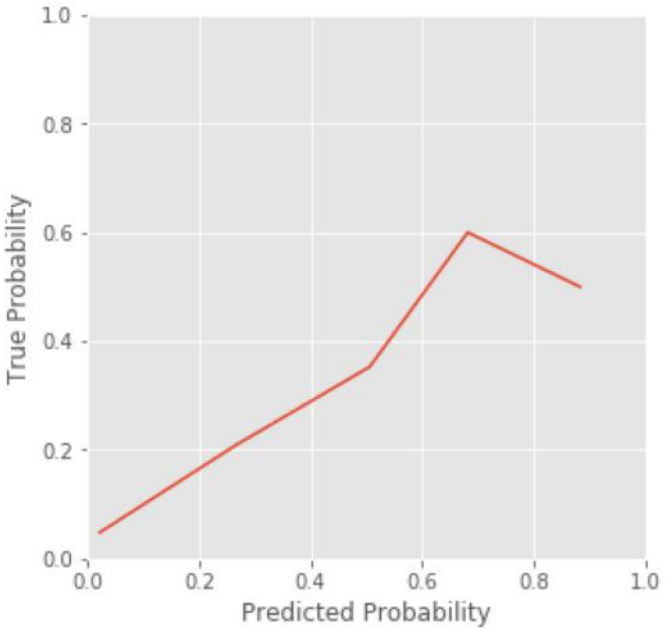
- My Mentors:
  - Dr. Cinnamon Bloss
  - Dr. Jejo Koola
- Kevin Ngo
- Cynthia Schairer Ph.D
- The Bloss Lab

Questions?

This research was supported by grant T15LM011271

Logistic Regr. Calibration with/without Isotonic Regression





This research was supported by grant T15LM011271