Data Agnosticism: Feature Engineering Without Domain Expertise

SciPy2013
June 27th

Nick Kridler

nkridler@accretivehealth.com

About Me

Generalist

Applied Mathematician

Data Scientist

Kaggle Master

Level 50 Dragoon

Former Defense Scientist

New to Healthcare

http://www.nmfs.noaa.gov/pr/species/mammals/cetaceans/rightwhale_photos.htm

Today's Talk

Elaborate on
'Data Agnosticism'

Save Whales from
Ship Collisions

Illustrate my
data analysis process

Discuss Python's
impact on my work

Responsible Data Analysis and Quick Iteration Produce High-Performing Predictive Models

No Domain Expertise Required!

"In data science, domain expertise is more important than machine learning skill."

– Motion from the "Data Science Debate" at Strata CA 2012[1,2]

Subject Matter Experts Know Which Data and Features are Important

1. http://strata.oreilly.com/2012/03/machine-learning-expertise-google-analytics.html
2. http://medriscoll.com/post/18784448854/the-data-science-debate-domain-expertise-or-machine

Algorithms Don't Care

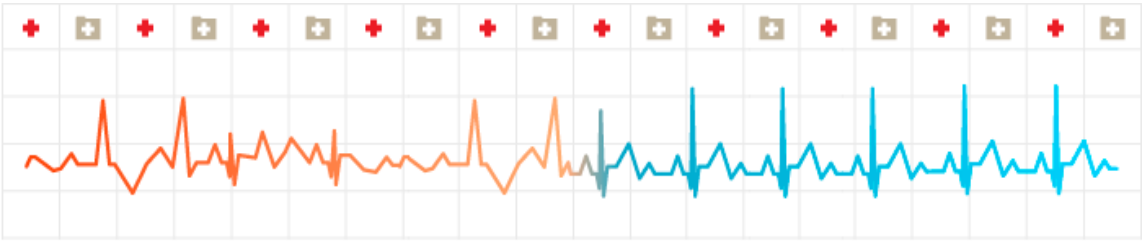
They don't know who generated the data

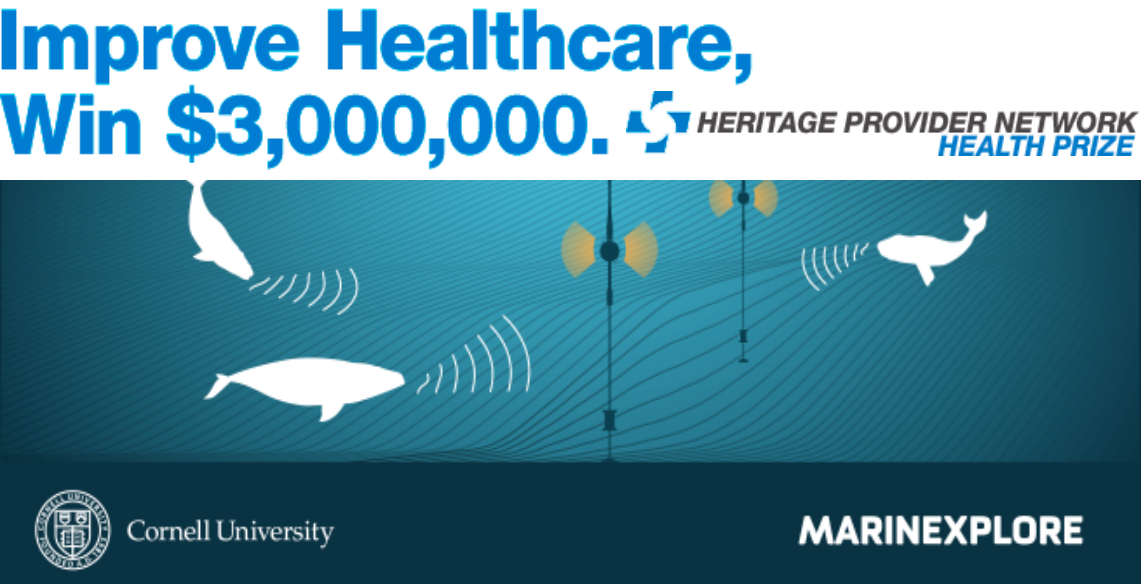
Just Like Garbage In → Garbage Out,
Bad Assumptions → Bad Model

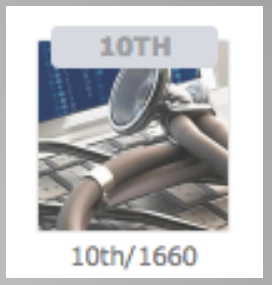
Models Will Help Us Find Features
(Without Domain Expertise!)

Kaggling With an Emphasis on Feature Engineering

No Prior Healthcare or Bio-Acoustics Experience







10th/1660

1st/249

1st/136

The Secret to My Success:

Responsible Data Analysis and Quick Iteration

…and a Lot of Terrible Ideas

Responsible Data Analysis?

- Look at samples, not just the aggregates
- Pay attention to sources of over-fitting
- Be skeptical of everything

Quick Iteration?

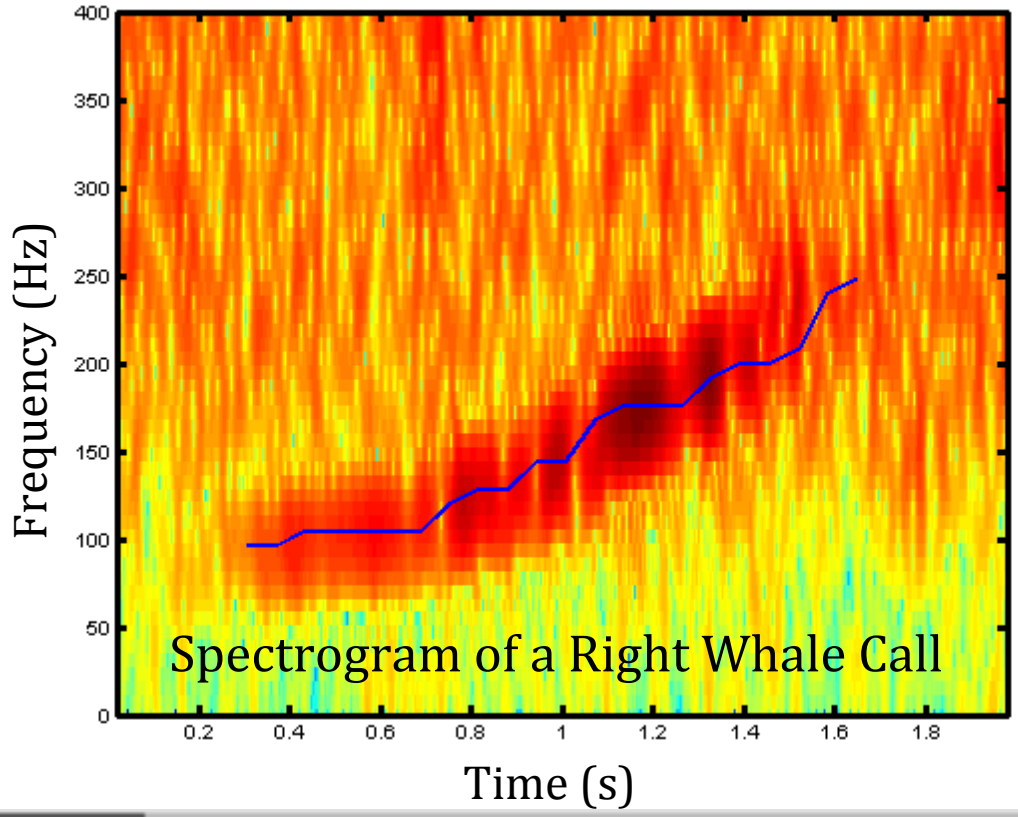
- Reduce the advantage of having domain expertise
- How else are you going to get through all those bad ideas?

A Process For Finding Whales

http://www.birds.cornell.edu/brp/research

North Atlantic Right Whale Up-Call Detection

Determine the Probability a 2 Second Audio Clip Contains a Whale Call
Maximize Area Under Curve (AUC) Metric



Marine Mammal Acoustics

Signal Processing

Audio Spectrograms

Mel-Frequency Cepstral Coefficients

Cornell University Benchmark

0.72141

Where Do We Start?

Google ‘Whale Detection’?

Found Competition Website

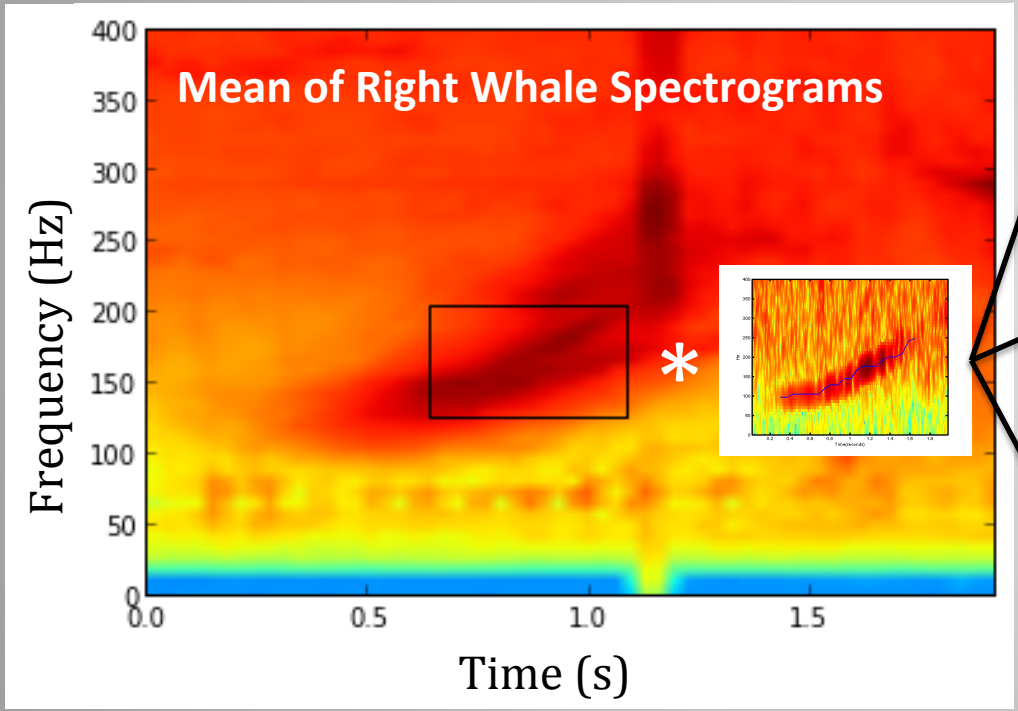
Throw The Whole Thing In A Random Forest?

What Would We Do After That?

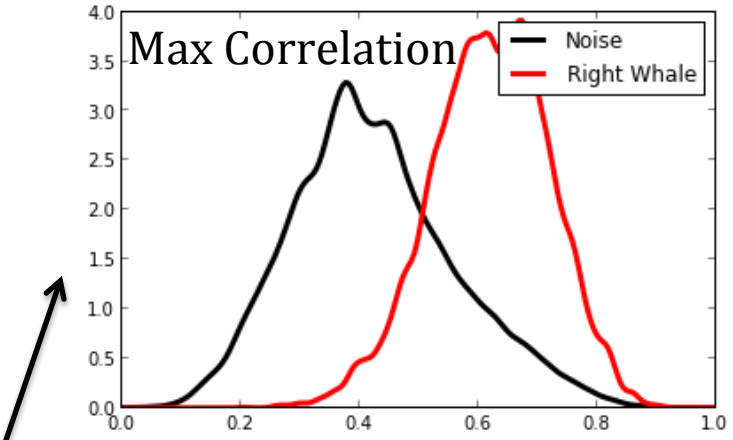
What can we do in a few hours?

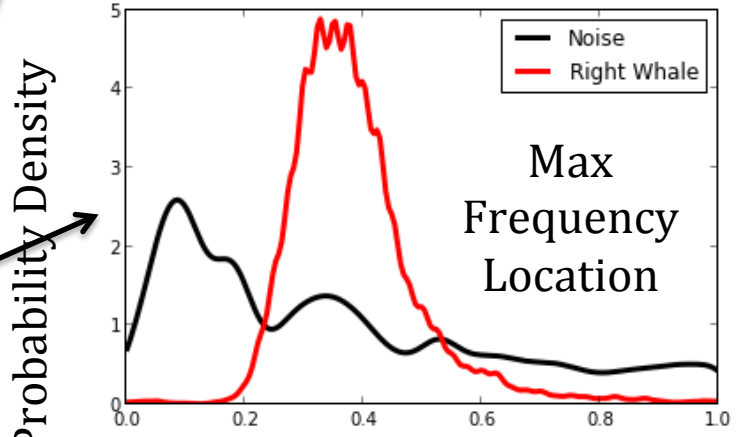
Start Simple

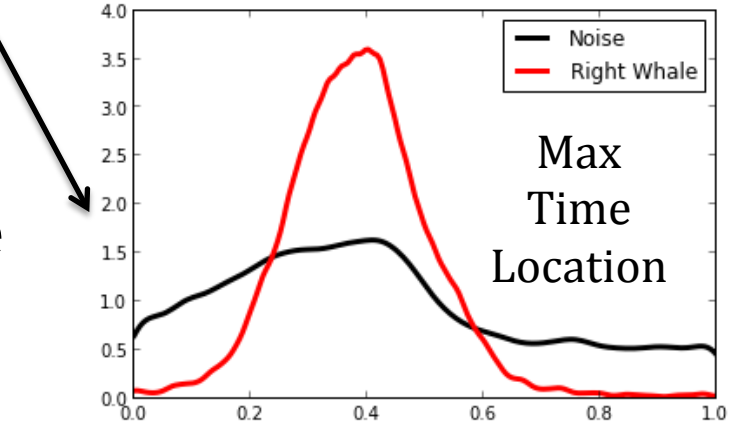
A Correlation-Based Model



How well does this chip correlate with the spectrograms from the audio clips?



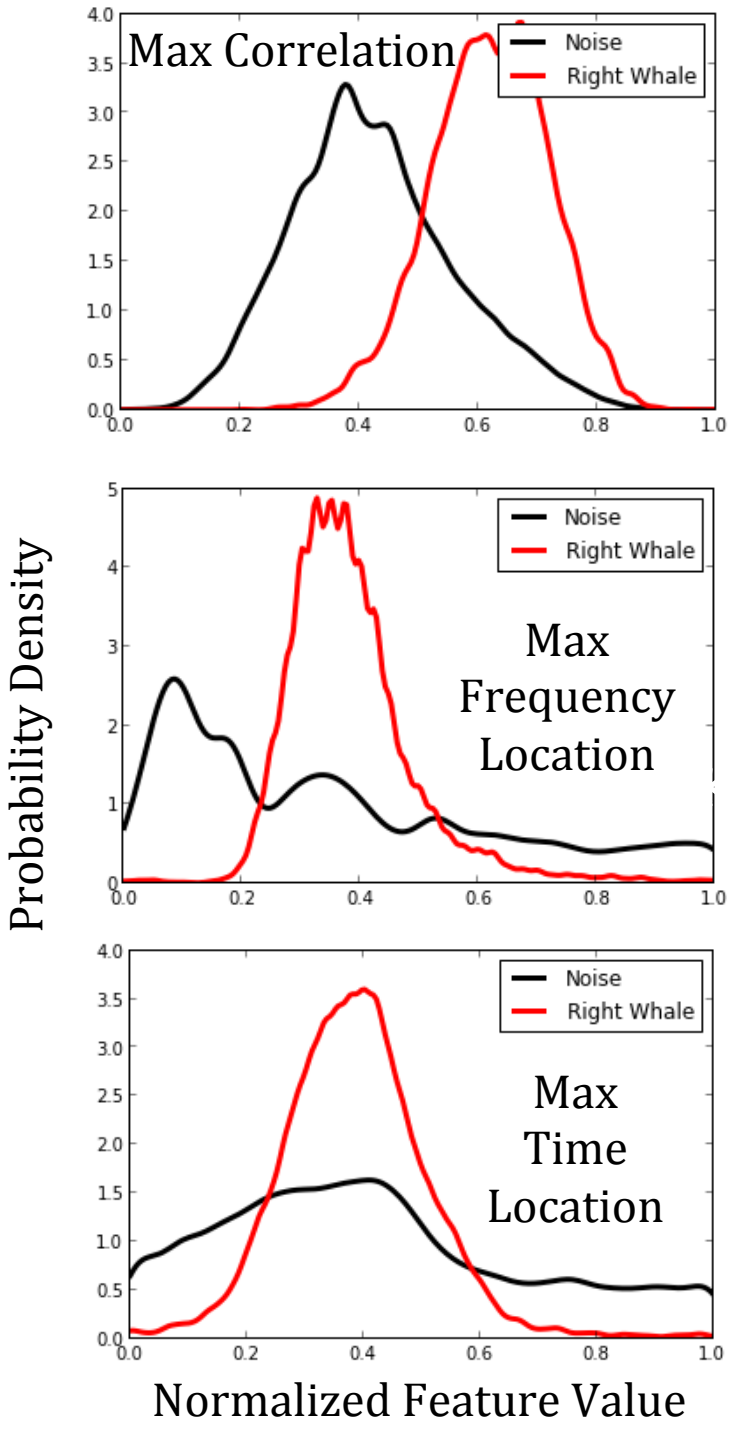




Probability Density

Normalized Feature Value

Leverage Great Work



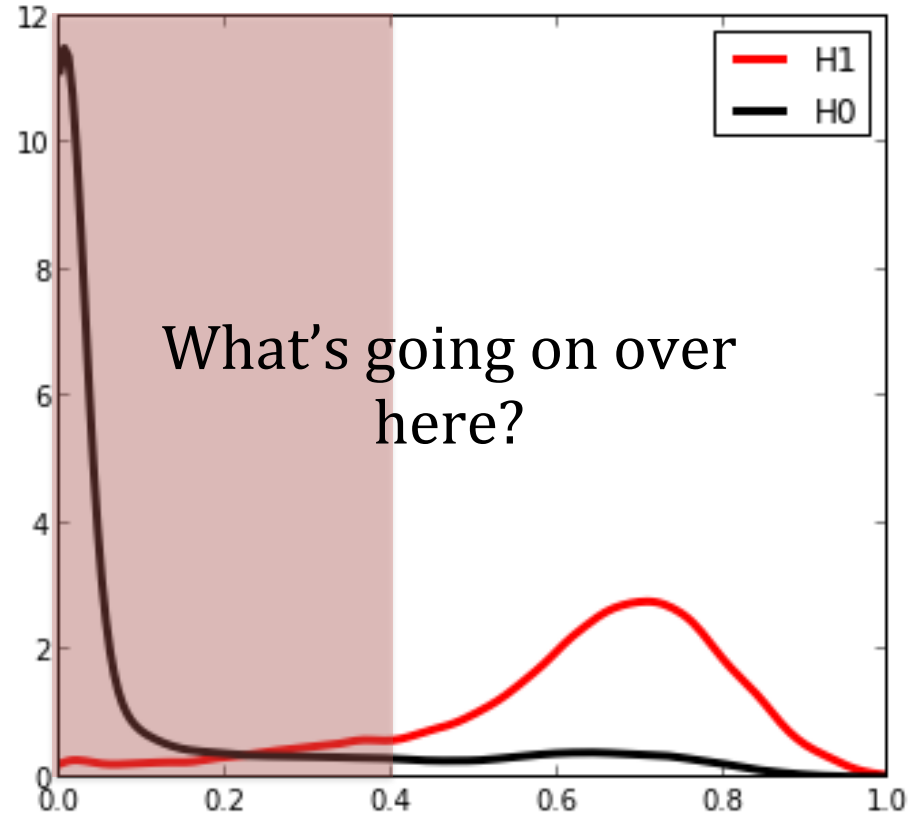
scikit learn

Random Forest

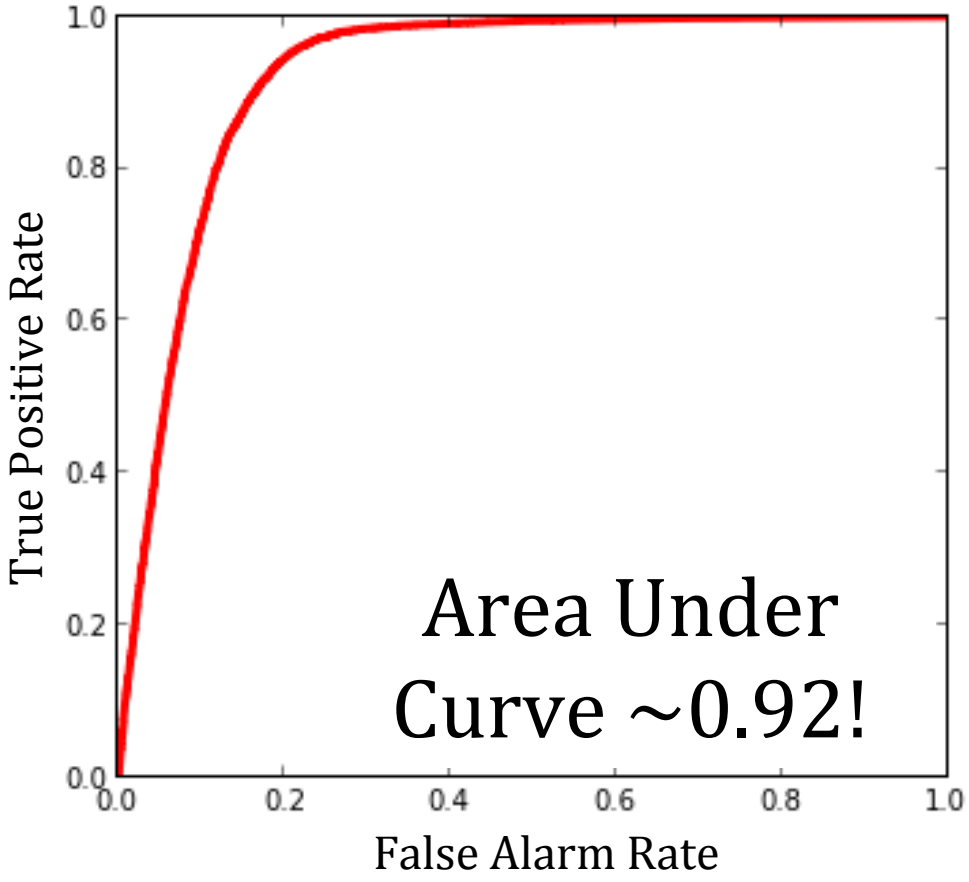
Random Forests are Quick, High Performing, and Easy to Interpret

How'd We Do?

Probability Density

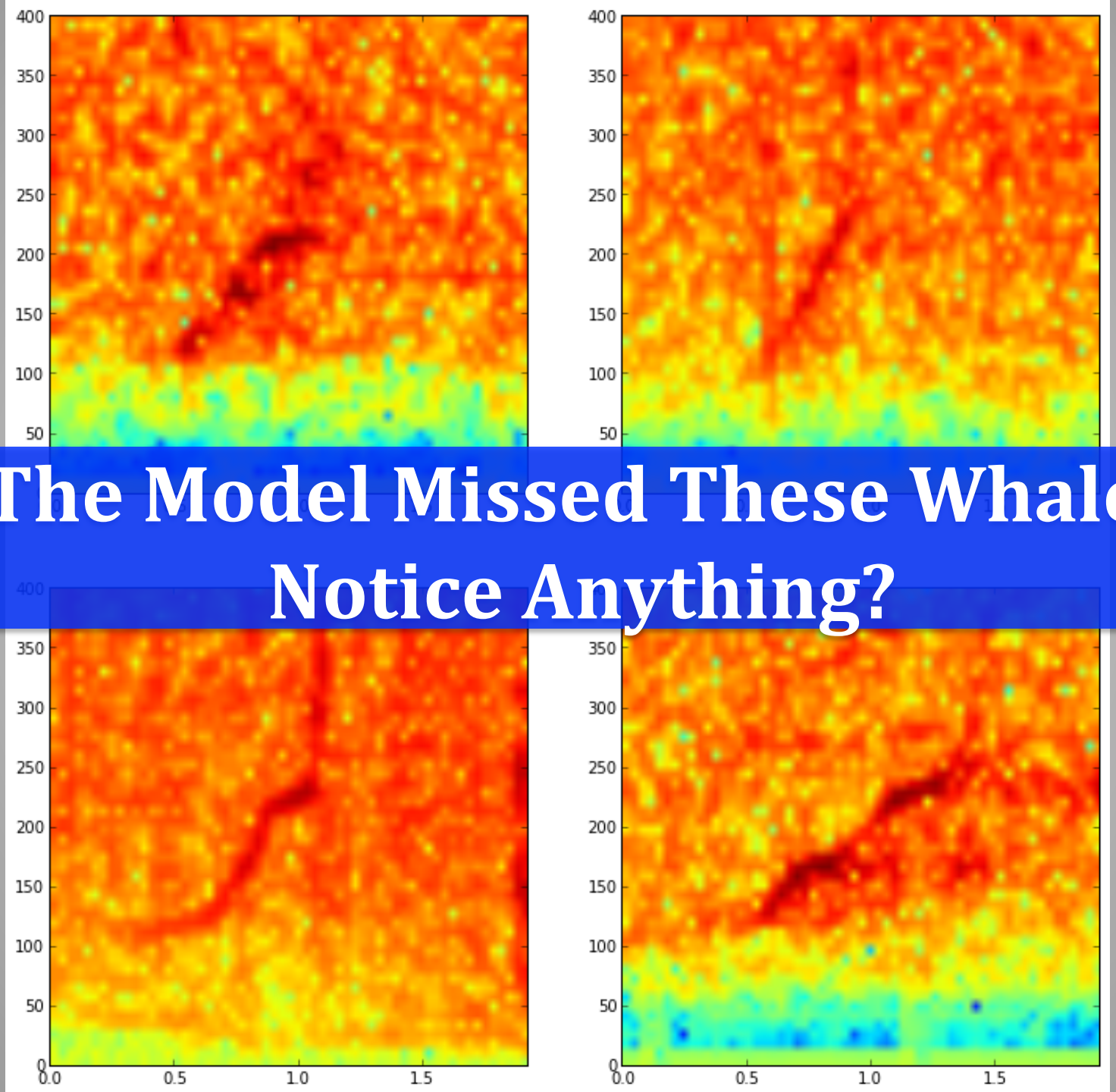


Receiver Operating Characteristic Curve



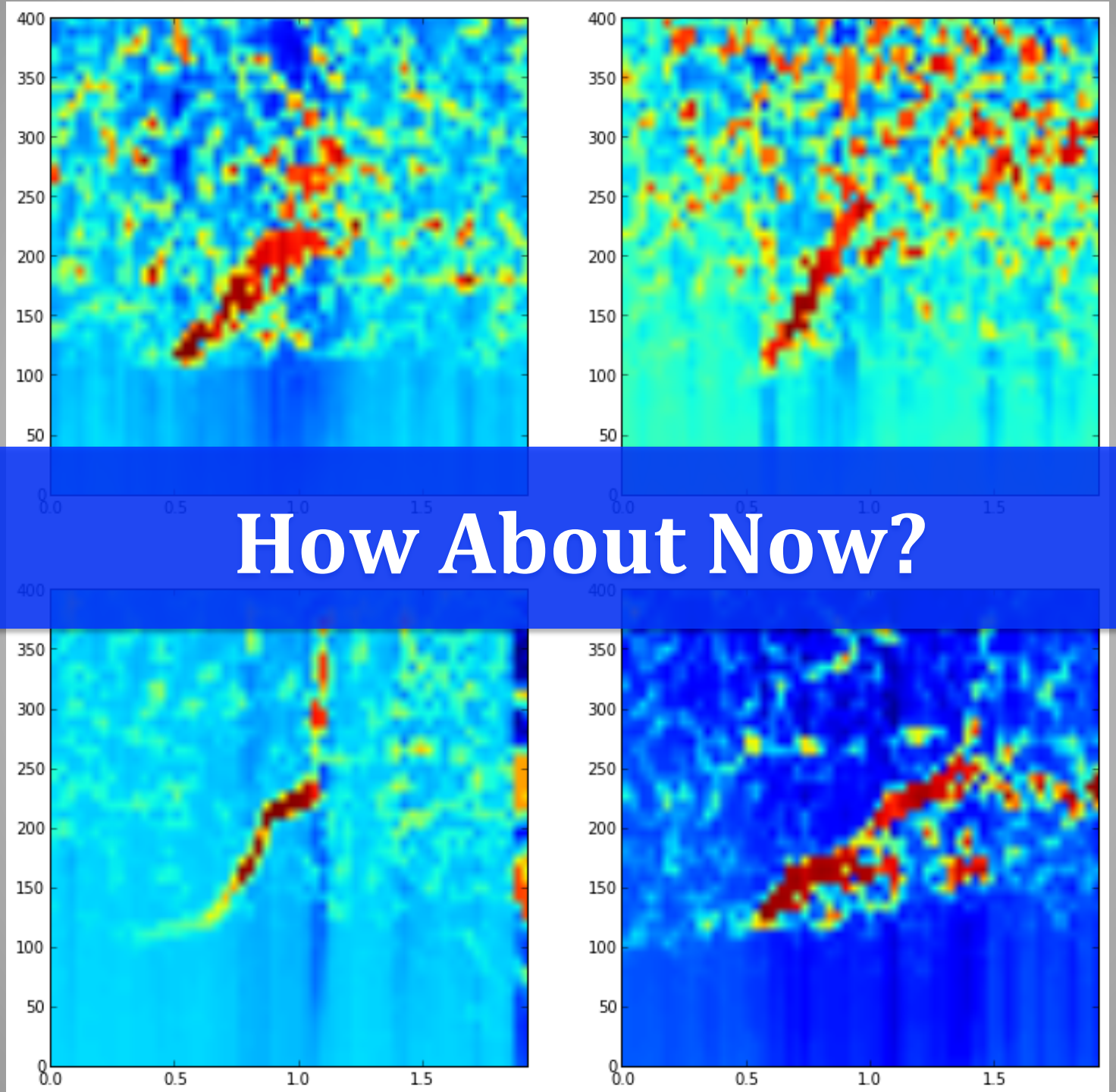
Cornell University Benchmark

0.72141



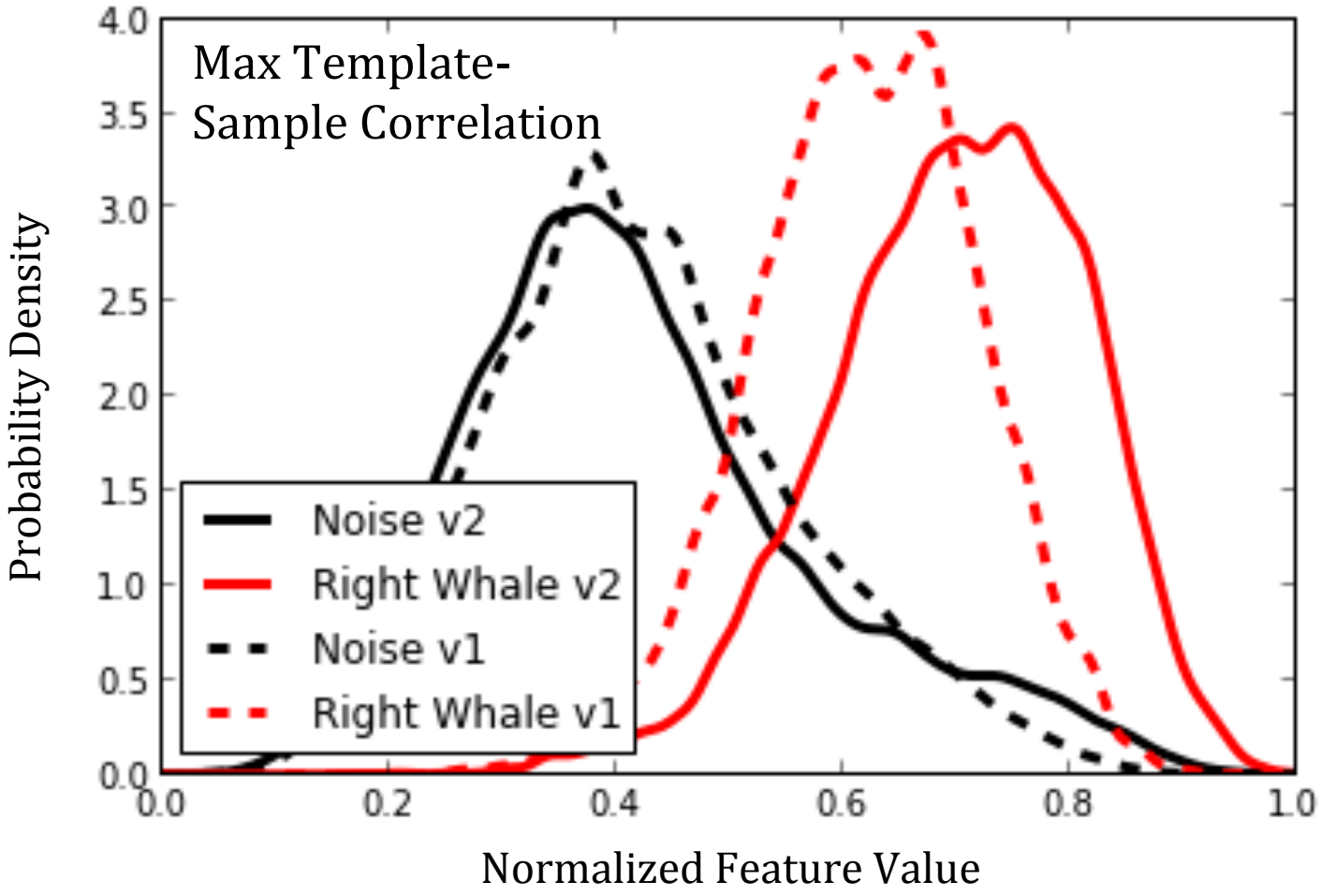
The Model Missed These Whales

Notice Anything?



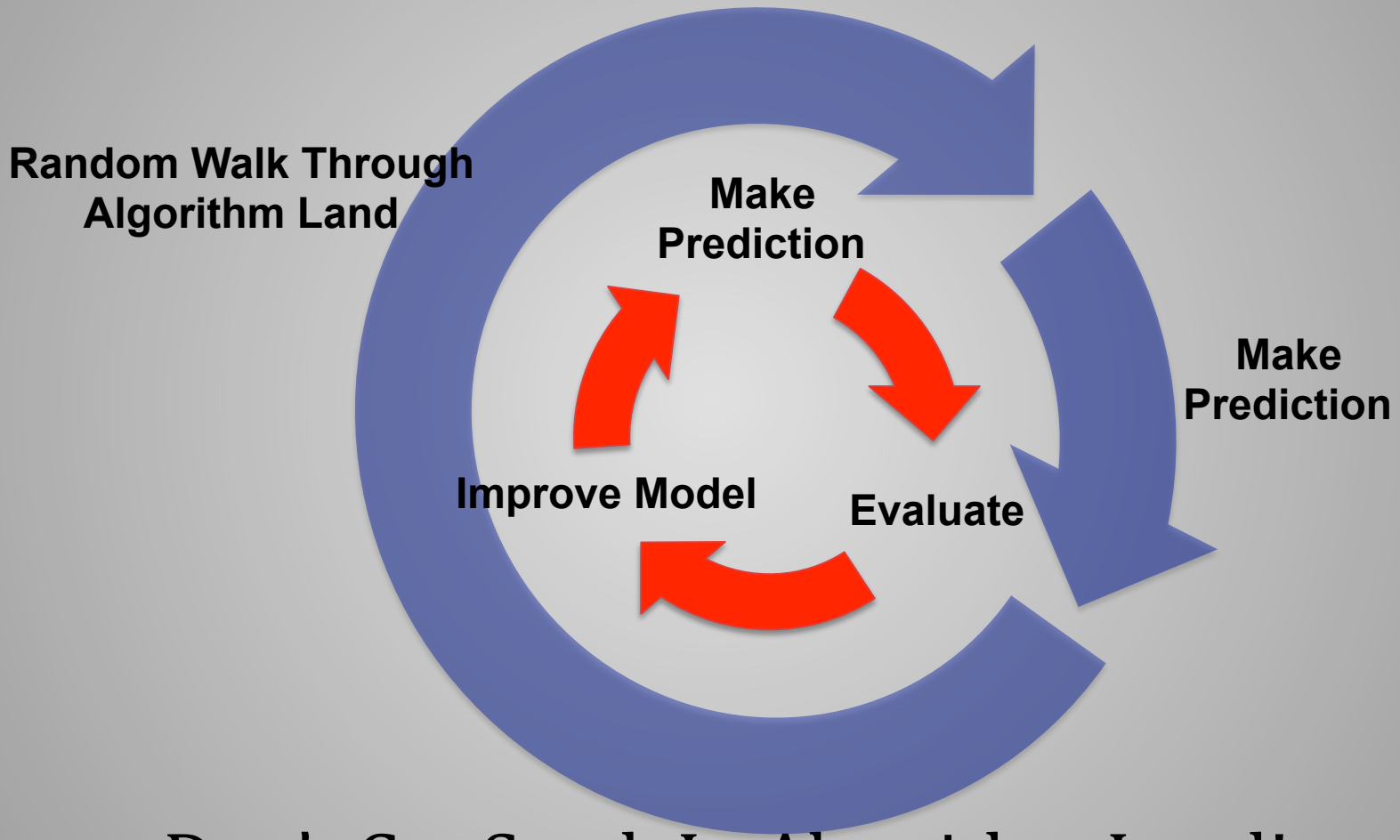
How About Now?

Do Our Changes Make A Difference?



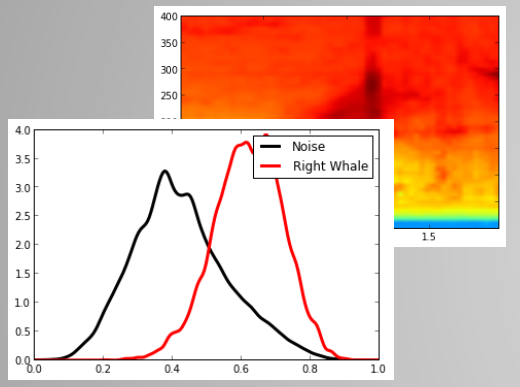
More Separation! AUC = 0.94

The Process (Good Vs. Bad)

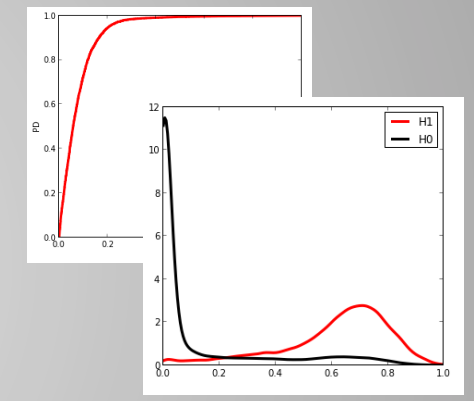


Don’t Get Stuck In Algorithm Land!
Focus on Putting Better Data in the Algorithm

How Do We Iterate Quickly?

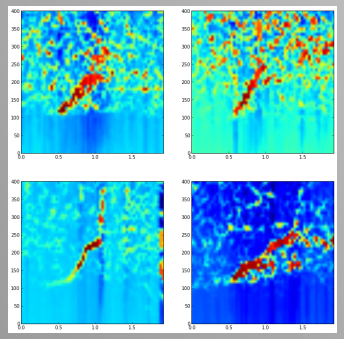


Choose an Algorithm, Generate a Model

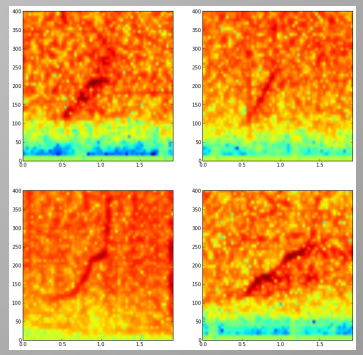


Make Prediction

Turn Ideas into Code!



Evaluate the Model, Visualize The Failures

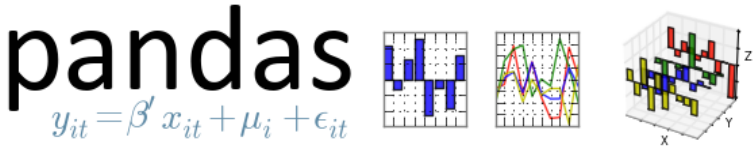


Improve Model

Evaluate

From Data to Model in No Time

Leveraging Great Work

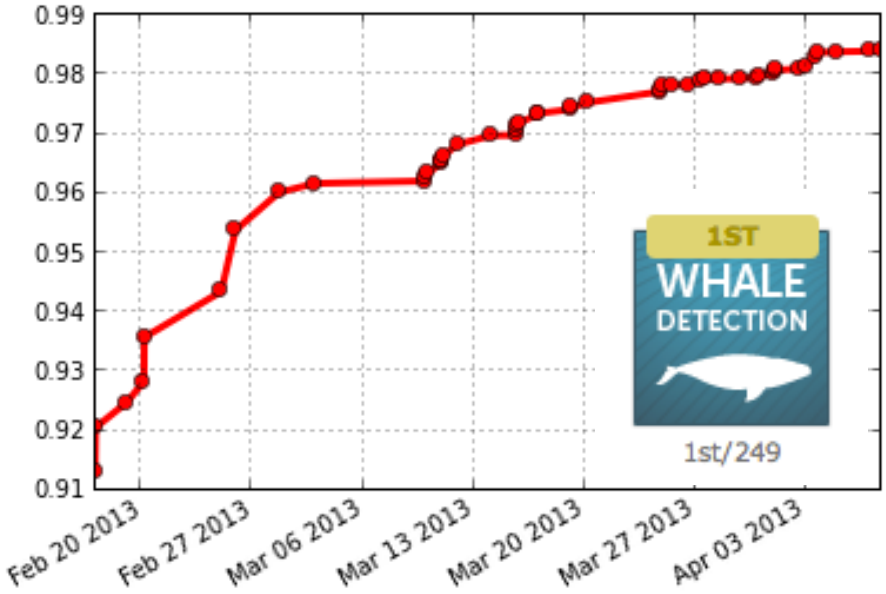


Shift the Focus From Algorithm Implementation to Data Analysis

Consistent Data-Driven Improvement

Public Leaderboard Scores

Heritage Health Prize

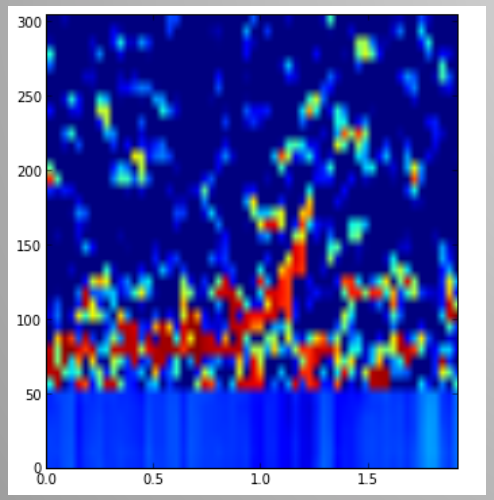


Marinexplore/Cornell
Whale Detection Challenge

ICML 2013 Whale Challenge
Right Whale Redux

1	-	**SluiceBox** * • Scott Dobson • Nick Kridler	0.99380	13

Marinexplore Challenge Code -> 0.98707

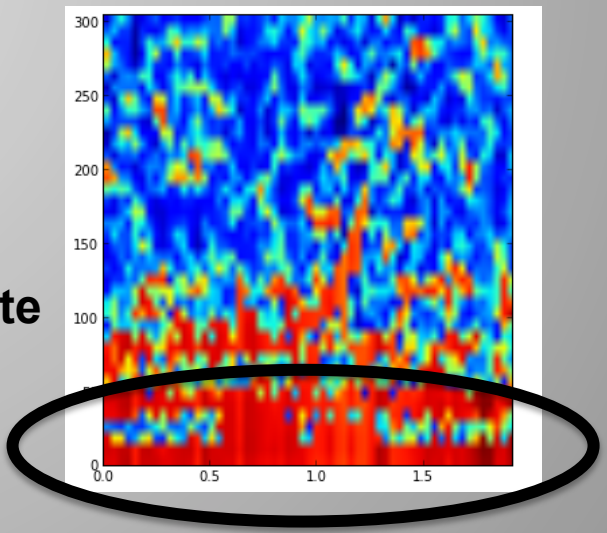


Just notch out low frequencies

Make Prediction

Evaluate

Improve Model



WTF is this?

Algorithms Only Care About Better Data
Data Agnosticism

Expert approaches involve…

- Audio Enhancement
- Mel-Frequency Cepstral Coefficients
- Zero Crossing Rate
- …and many other things I didn't use

Responsible Data Analysis and Quick Iteration Produce High-Performing Predictive Models

No Domain Expertise Required!

Kaggle Challenge Improves Cornell's Whale Detection Model to 98%

"This data challenge was game changing in the way that it helps us to focus on value-added tasks instead of technicalities."

— Dr. Christopher Clark

Contact Me

MASTER ?

Highest 27th

Current 28th /102,926

228,163.4 points
Joined 2 years ago

nkridler@accretivehealth.com

http://www.kaggle.com/users/7947/nick-kridler

http://www.linkedin.com/pub/nicholas-kridler/2b/a67/376

https://github.com/nmkridler

https://twitter.com/nmkridler

Accretive is Hiring!