APPLIED MACHINE LEARNING: CLASSIFICATION OF CORONARY ARTERY DISEASE

The Kardiac Kids

"Data from the heart, data for good."

GOALS

- Analyze available data using ***supervised machine learning*** techniques to ***predict the presence of coronary artery disease*** (Yes | No).

- Explore ***feature selection*** techniques and compare results against ***existing research*** and ***domain expertise***.

DATA SOURCES – OVERVIEW

- Two primary heart disease studies have been used consistently throughout the research community: the *Cleveland Clinic Study* and the *Framingham Heart Study*. The Framingham data was not available for this effort due to access restrictions.
- The Cleveland Clinic database includes data sets collected from patients at four locations: The Cleveland Clinic, OH; V.A. Medical Center Long Beach, CA; University Hospital Zurich and Basel, Switzerland; and The Hungarian Institute of Cardiology, Budapest.
- This database contains seventy six attributes, but the majority of published experiments refer to using a subset of fourteen (thirteen estimators plus one predictor for the presence of heart disease).
- Acknowledgements:

1. *Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.*
2. *University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.*
3. *University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.*
4. *V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.*

DATA SOURCES – DETAILS

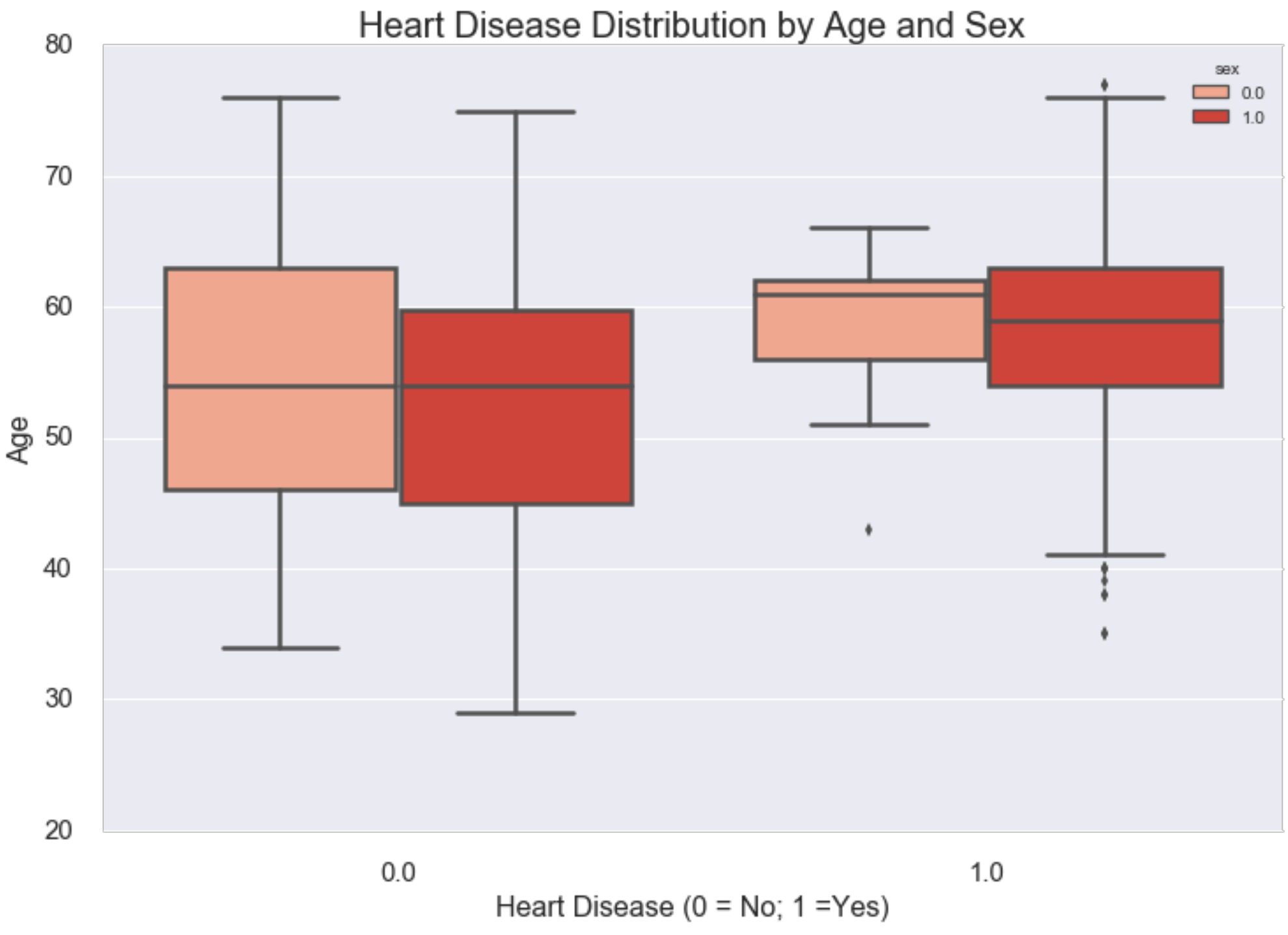
- Specifically, the Cleveland Clinic table is the most complete data set and has been previously "processed", but corrupted records and missing values still persist.
- Data was cleaned and pre-processed. K-Nearest Neighbors classification was used to predict missing values where feasible.
- The results presented herein use the combined data sets from the Cleveland Clinic and V.A. Medical Center in Long Beach.

  482 records; 36 different attributes

  274 cases of heart disease; 208 cases of no heart disease

  Males: 385; Females: 97

A LOOK AT THE OVERALL DISTRIBUTION



TOOLKIT

- Data Acquisition and Pre-processing:
    - Manual Inspection and Familiarization w/ Data Description Document
    - PostgreSQL on Amazon EC2: convert data to relational database and enable distributed access for the team
- Machine Learning Algorithms - Classification:
    - from sklearn import <everything>
- Data Visualization:
    - Matplotlib and Seaborn
    - D3.js

METHODOLOGY

- Comparison of classification methods across three primary feature sets:

    1. Prior Research
    2. Best Features Selected via Statistical Methods
        a. Recursive Feature Elimination
        b. Model Based Ranking
    3. American Heart Association Risk Factors

Method	Features Selected
1	age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
2	age, sex, cp, chol, exang, fbs, restecg, thal, oldpeak, slope, proto, nitr, met, rldv5e, thalach, thalrest, thaldur, thaltime
3	age, sex, famhist, cigs, years, trestbps, chol

RESULTS – CAN YOU PICK A WINNER?

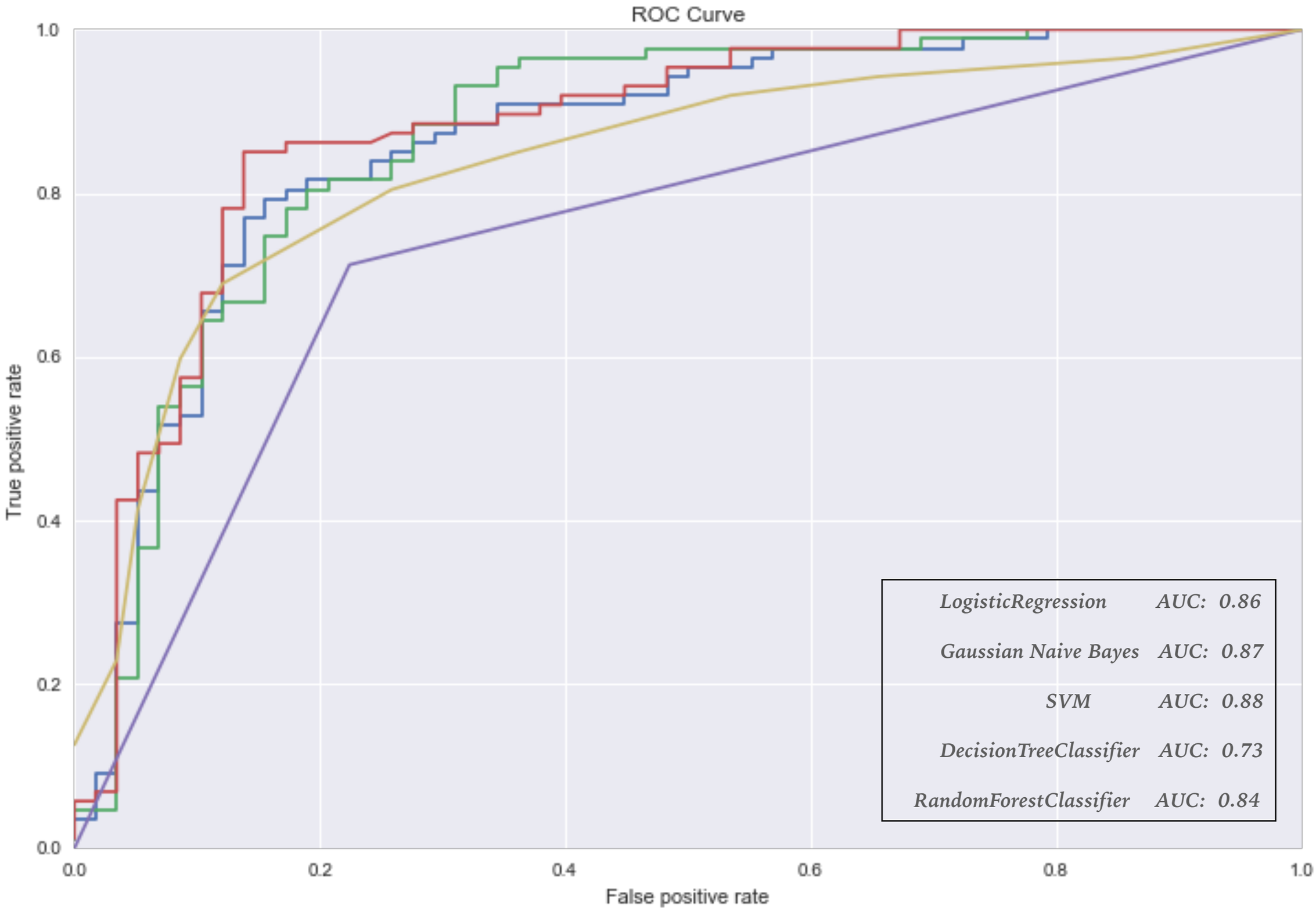
	Classifier	Accuracy	Precision	Recall
Prior Research	Logistic Regression	0.81	0.84	0.84
	Naive Bayes	0.81	0.86	0.80
	SVM	0.82	0.86	0.84
	Decision Tree	0.66	0.74	0.70
	Random Forest	0.68	0.77	0.66
Feature Selection	Logistic Regression	0.81	0.84	0.84
	Naive Bayes	0.81	0.86	0.80
	SVM	0.82	0.86	0.84
	Decision Tree	0.73	0.82	0.70
	Random Forest	0.8	0.89	0.76
AHA Risk Factors	Logistic Regression	0.71	0.71	0.86
	Naive Bayes	0.77	0.77	0.87
	SVM	0.74	0.78	0.80
	Decision Tree	0.67	0.72	0.72
	Random Forest	0.64	0.73	0.63

Not surprisingly, Prior Research scored the highest.

Feature Selection outperformed AHA, but AHA did have higher Recall.

Overall Decision Tree and Random Forest were the lowest performing classifiers. KNN was also attempted, but the results were even lower and are not listed.

ROC CURVES FOR BEST FEATURES APPROACH



ANALYSIS

- The Numbers Don't Lie….Well

  Is The American Heart Association Wrong? NO

  Doctors are likely to sacrifice accuracy during the examination phase and defer to the results of further testing before making a diagnosis

  (Has Risk Factors? NO | YES = Refer for Further Testing)

- Feature Selection Needs Context

  The "best" features also happen to be some of the most invasive procedures - stress tests and catheterization while extremely accurate, are also cost prohibitive, resource intensive and time consuming to give to everyone

- Be Comfortable With Your Data

  There's good reason why a particular data set has been used so extensively. That is also why it is difficult to gleam any new information from it.

THANK YOU . . . FROM THE HEART

JASON SYPNIEWSKI
JASON.SYPNIEWSKI@GMAIL.COM