SAND2016-5848C

RON prediction models for the New Fuels and Vehicles Systems

Leanne Whitmore

Corey M. Hudson

- Sandia National Labs

RON Prediction

Goal: To engineer and distribute a high-quality **fully open** software, using **publicly available** resources to predict **fuel properties.**

Stretch: Allow internal (closed source) datasets and tools to be added to the prediction framework.

Available Training Datasets

152 RON Compounds

- Collected from Al-Fahemi et. al (2014), ASTM (1958), Balaban et al. (1992) and Bluock et al. (1995)

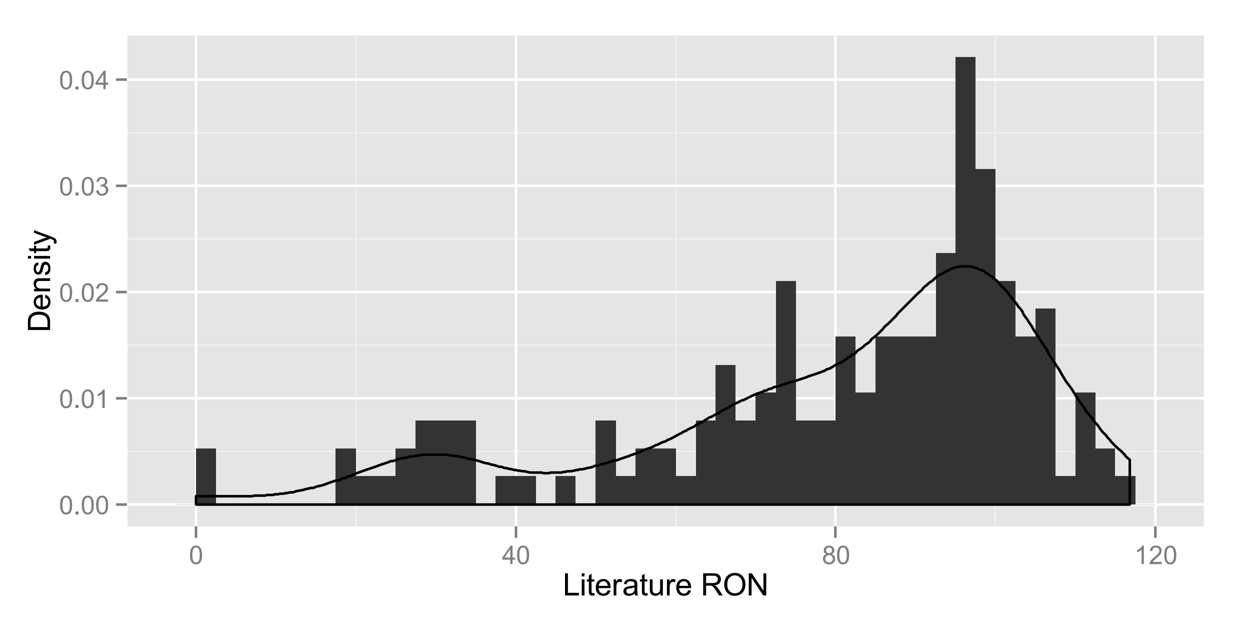
Available Features (Public)

- NCBI PubChem (881 structural features)
- NCBI Experimental (~20 common features)
- PaDEL-Descriptor (1875 QSAR Descriptors)

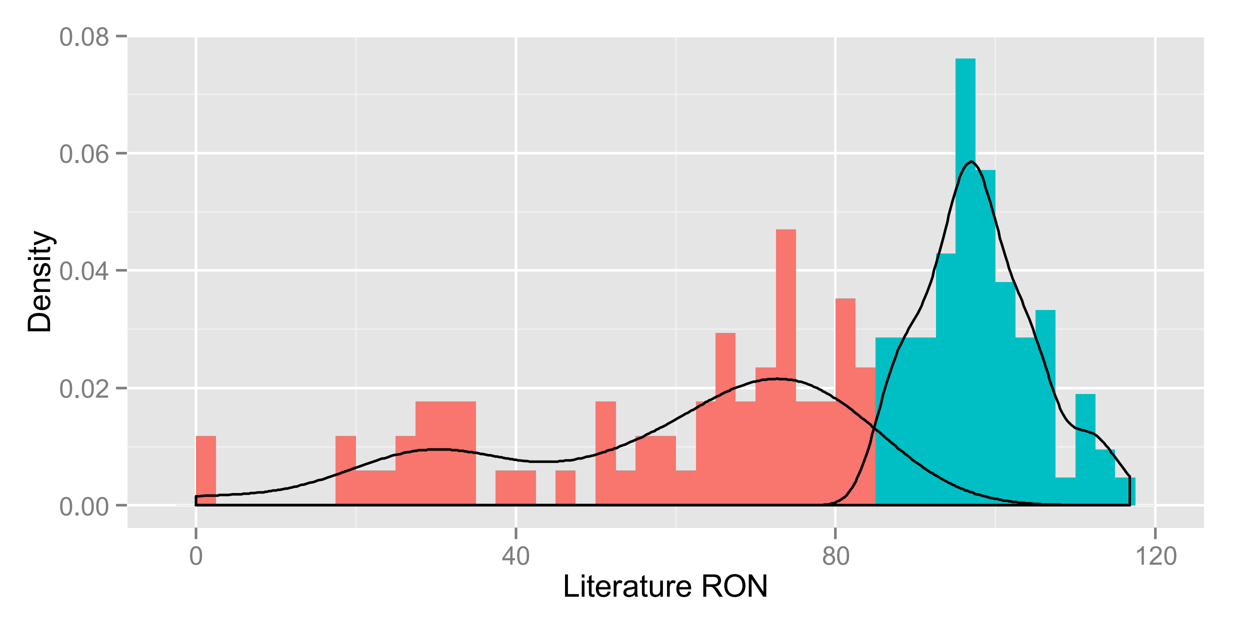
Private

- Collected from licensed software or data stores (ACD, EPI, etc.)

Literature RON Distribution



RON for Compounds of Interest

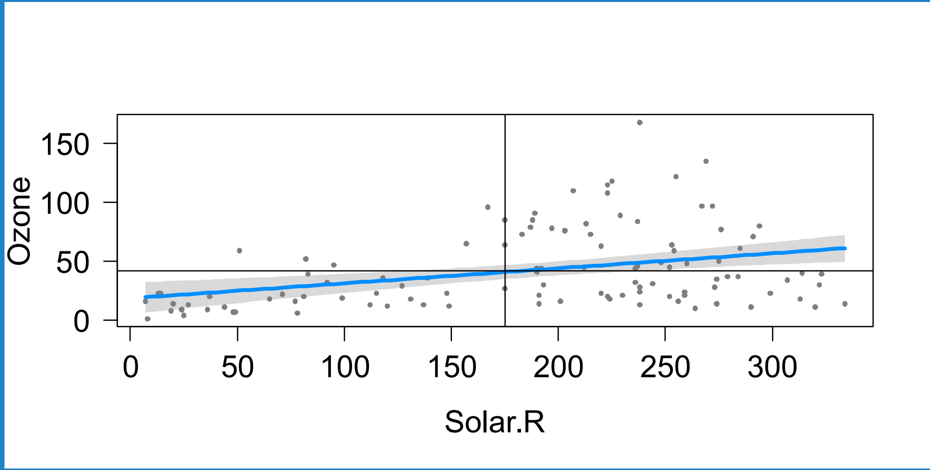


Classification vs. Regression

- **Machine Learning Classification** dramatically decreases the problem of overprediction.
- **Reason** information content.

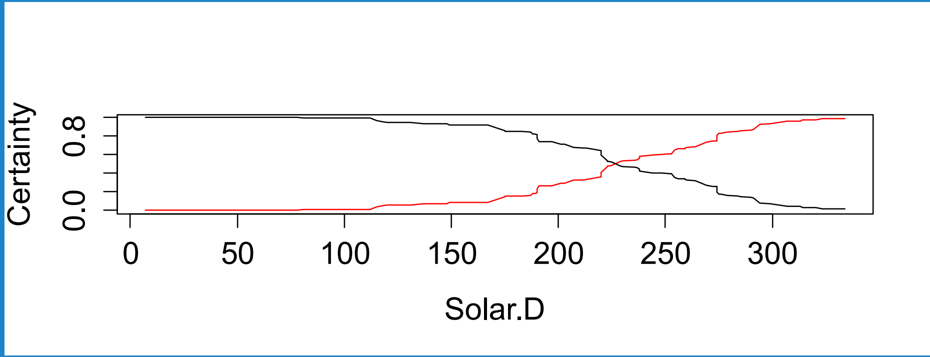
Regression Uncertainty

Regression residuals are biased toward the highest/lowest values



Classification Uncertainty

Classification residuals are biased toward the median values



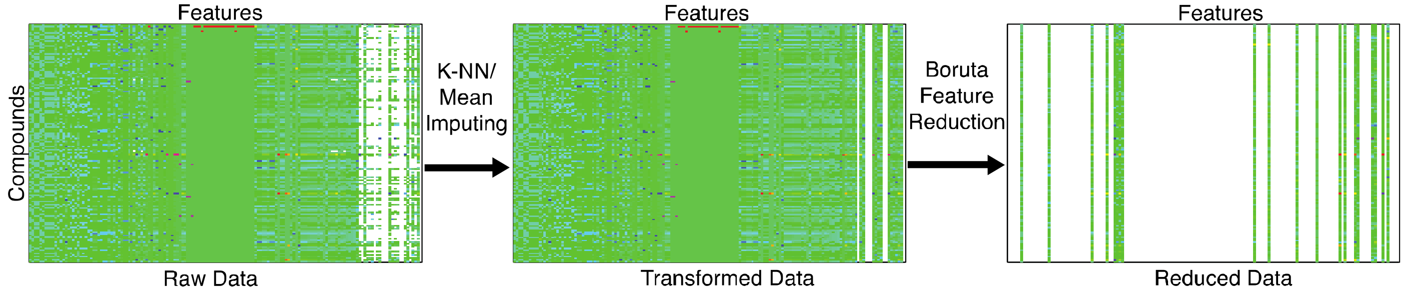
RON Classification

- For the purposes of screening, RON is a classification problem.
- High RON Chemicals are useful for drop-in blendstocks in ignition engines.

Machine Learning Methodology

- **Random Forests**
  - Classification
  - Fast
  - Scalable
  - Robust
- **Tanimoto**
  - Clustering
  - General
  - Agnostic to Imputed Features

Procedures for Feature Selection/Reduction



Boruta Feature Selection

Boruta algorithm

Original data

F1	F2	F3	F4
1	1	2	3
3	0	2	1
2	1	3	2

\+

Shadow features

S1	S2	S3	S4
3	1	2	1
1	1	3	2
2	0	2	3

Random Forest

Sample	Class
1	1
2	0
3	0

	F1	F2	F3	F4	S1	S2	S3	S4
MDI	.2	.05	.01	.15	.001	.02	.009	**.09**
Hit	+1	0	0	+1	-	-	-	**max**

Scale of Feature Reduce

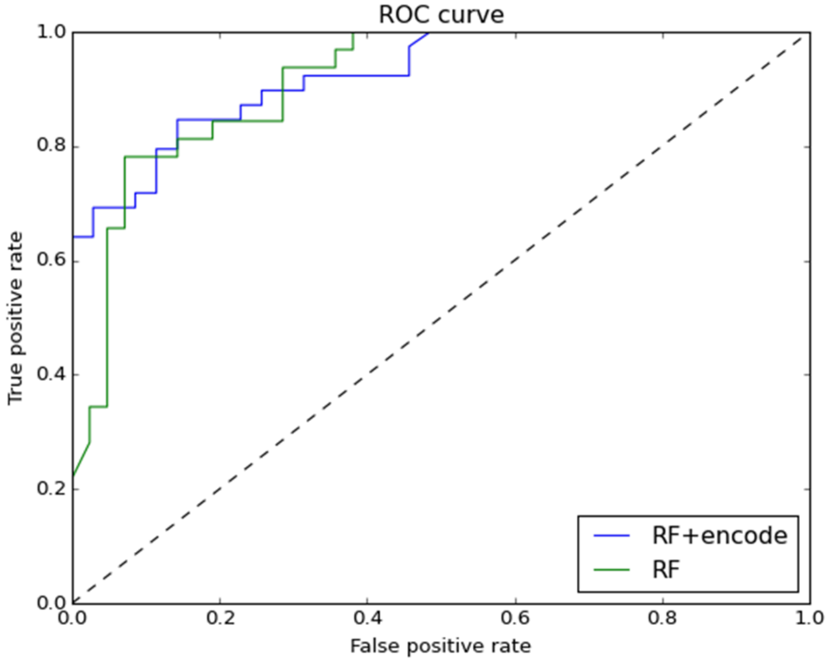
- 926 original features (experimental/NCBI/ACD-EPI)
- 147 variable features
- 19 after KNN-Imputation, followed by Boruta Features Selection
- Empirical estimates of accuracy improved 4%, precision improved 5%

Performance of classifier

- 100 sub-sampled cross-validations (with 50% leave out)

Metric	Mean value	Std. dev
Accuracy	0.84	0.08
Precision	0.85	0.14
Sensitivity	0.83	0.17
Receiver Operator Characteristic (AUC)	0.93	0.06

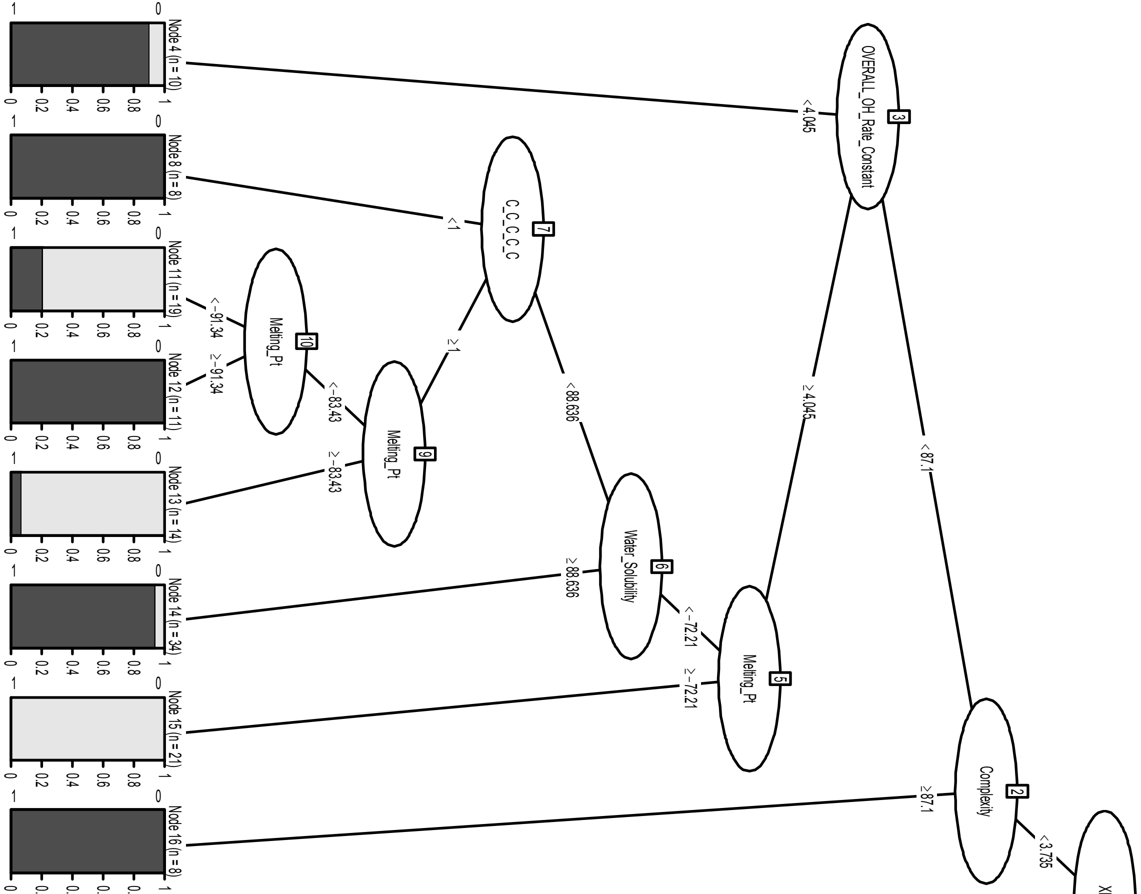
ROC Curve


ROC curve
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
RF+encode
RF

Most Heavily Weighted Features in Random Forest Classifier

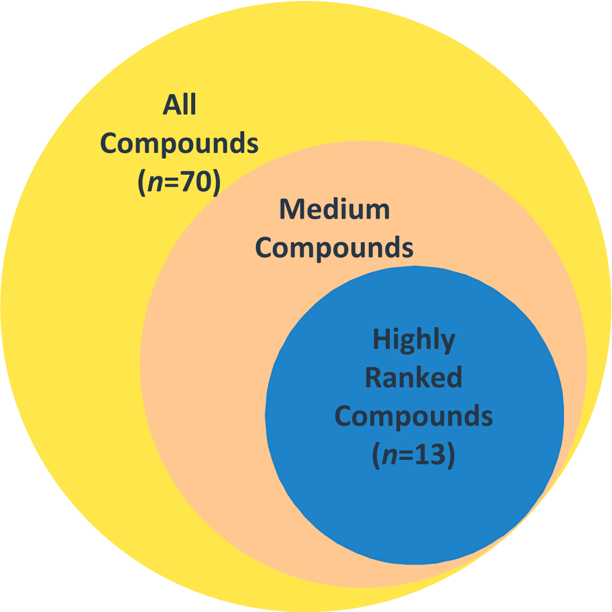
Features	Weight	Type
XLogP3 (Lipidocity)	0.1277	Physical
Log KOA (Air Partitioning)	0.0797	Physical
SMARTS Pattern: C-C-C-C-C-C	0.0781	Structural
Auto-Ignition	0.0772	Physical
Water Solubility	0.0767	Physical
Melting Point	0.0403	Structural
Boiling Point	0.0392	Physical
Surface Tension	0.0324	Physical
OH Rate Constant	0.0288	Physical
Complexity	0.0283	Physical

Visualizing a Single Decision Tree



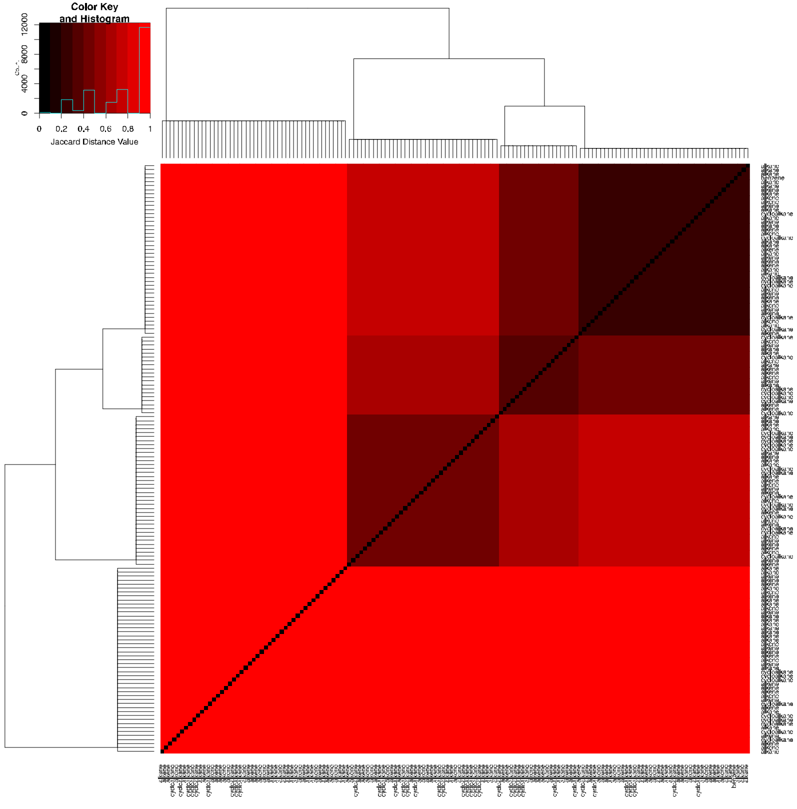
Ranking Compounds of Interest

- Initial Ranking for 70 Compounds of Interest

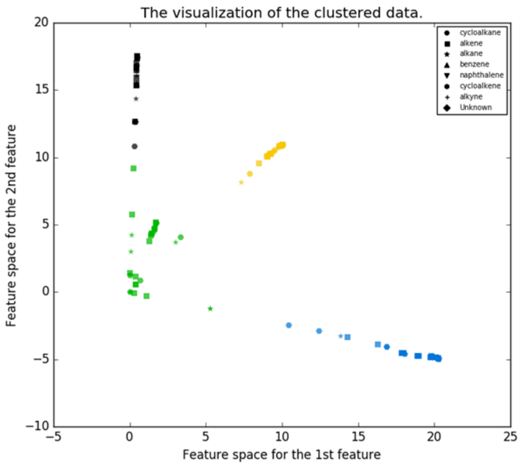


Compounds predicted to have RON > 85 in order of probability in class	
1. Isooctane	8. Isoprenol
2. Methylcyclopentane	9. Isobutanol
3. Ethanol	10. 3 methyl 1 butanol
4. Methyl butyrate	11. Butyl acetate
5. Ethyl isobutyrate	12. Toluene
6. Methyl 2-methylbutyrate	13. Isoamyl acetate
7. 2-methyl 2 butanol	

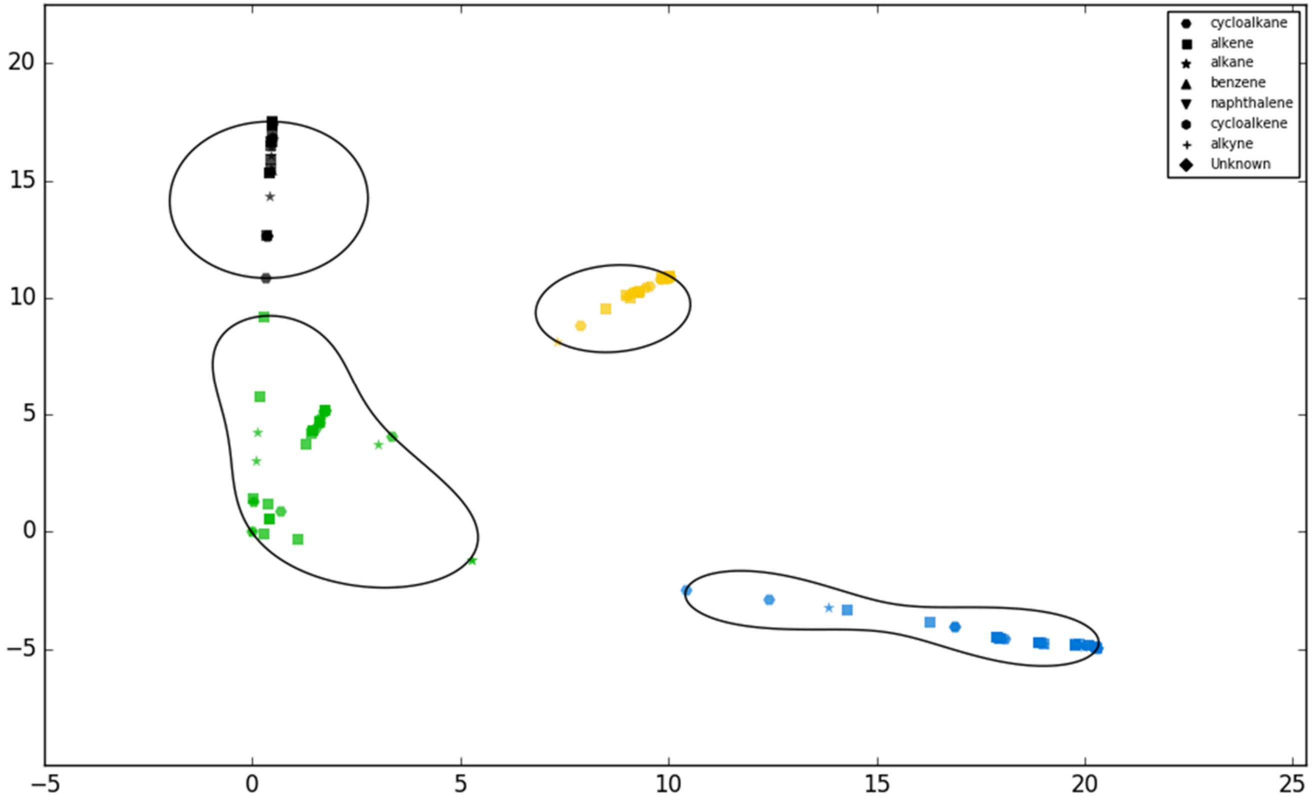
Challenge in Generalizing Approach



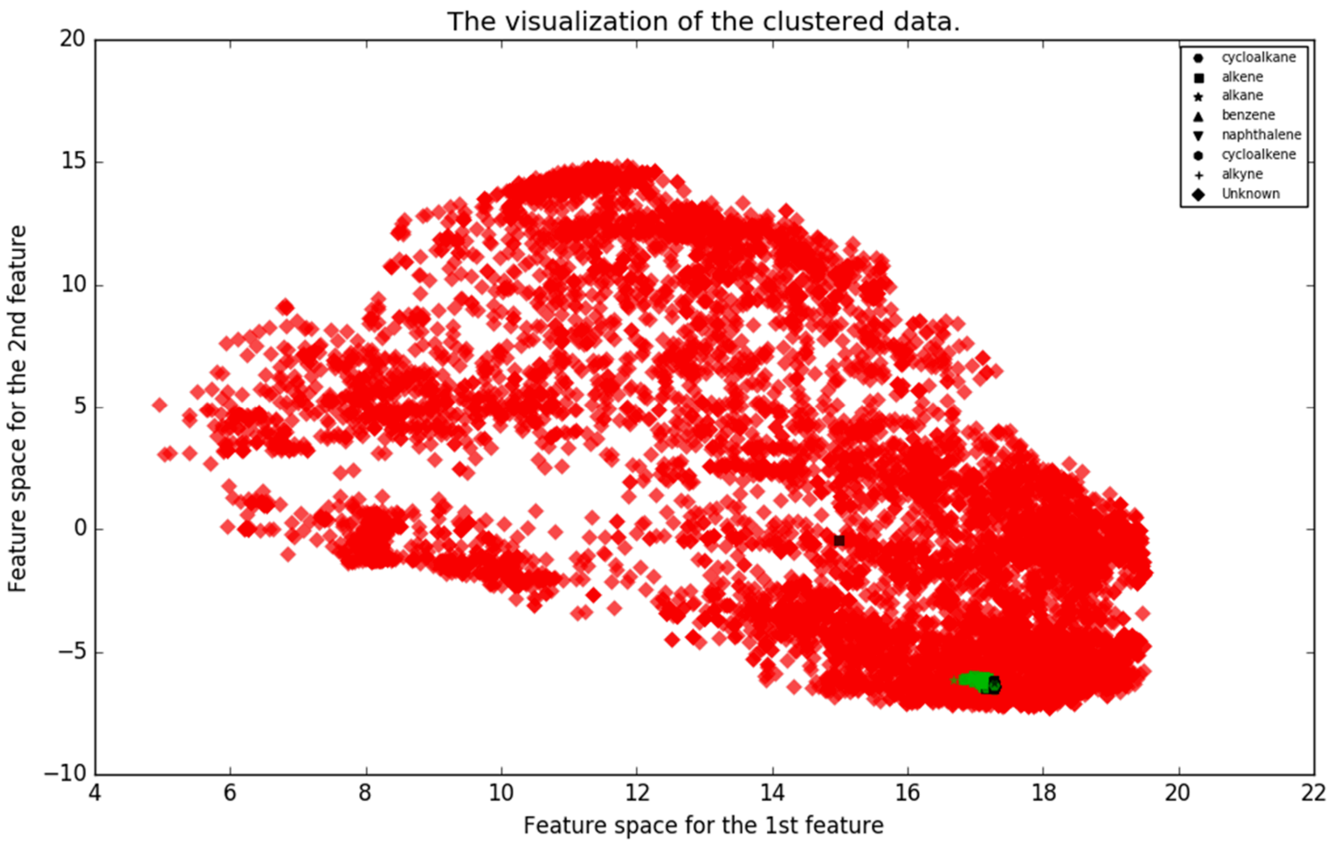
Clustering Training Data



Creating Clusters using Training Hydrocarbons

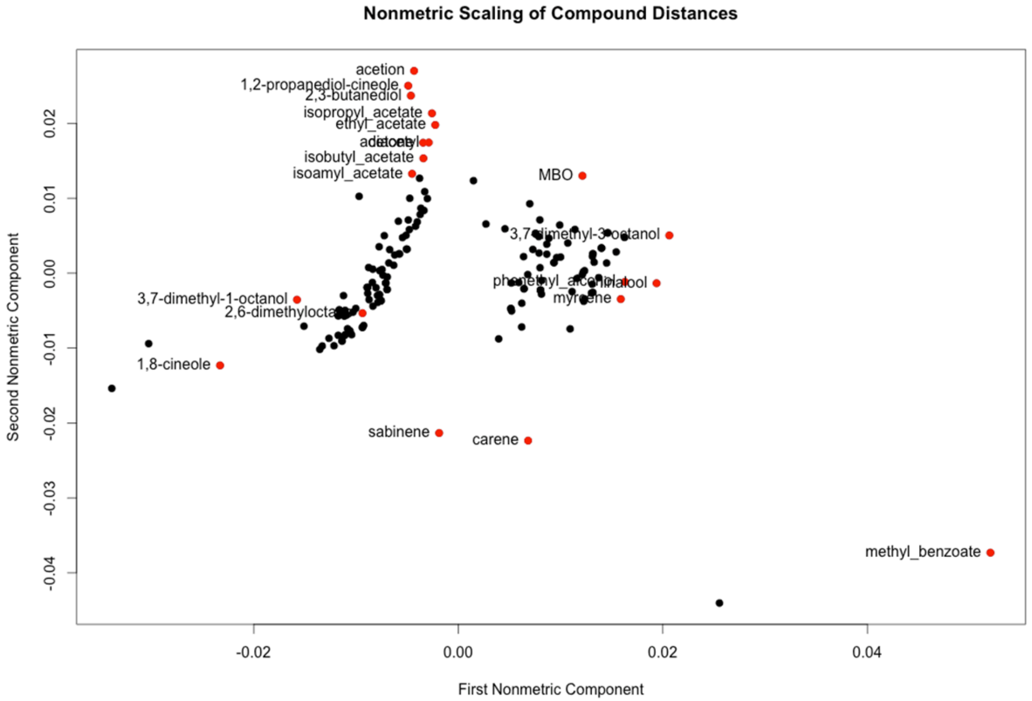


Comparison with Total Dataset

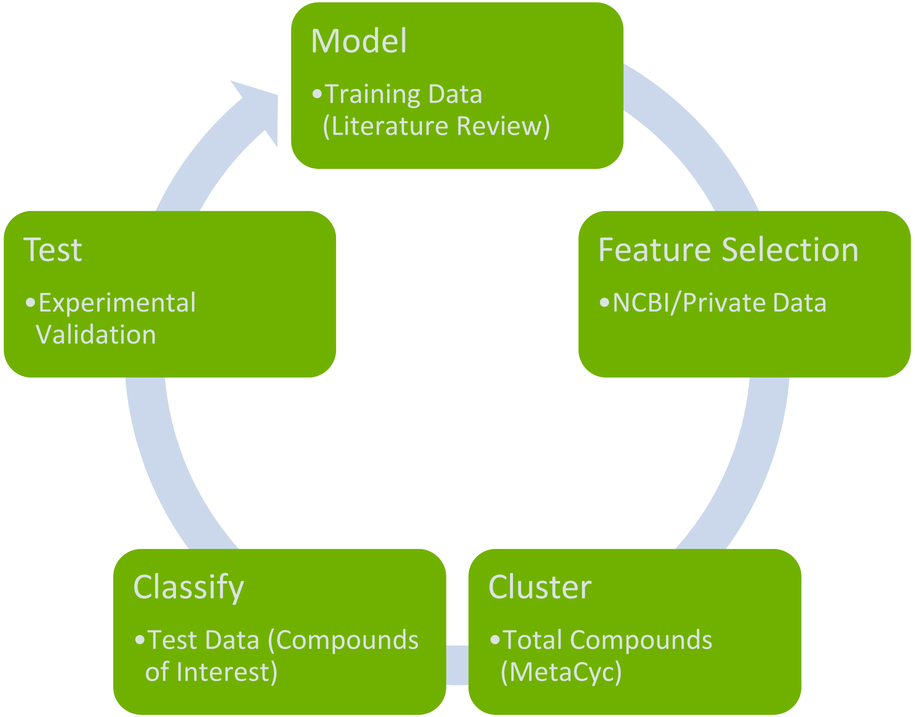


Structure of Software

Testing Model with Measured RON Values from SWRI



Machine Learning Process



Distribution of Software