Understanding Machine Learning Model Expertise

Kiri L. Wagstaff

Jet Propulsion Laboratory,
California Institute of Technology

kiri.wagstaff@jpl.nasa.gov

HORSE Workshop

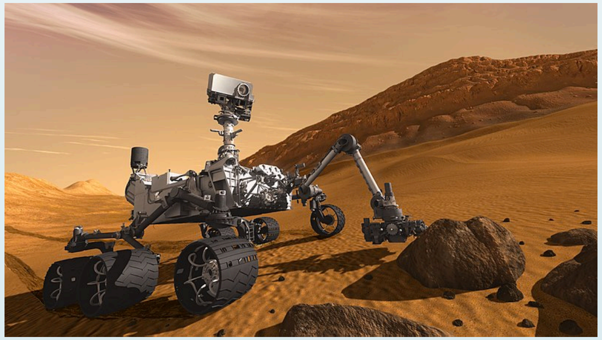
September 20, 2017

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

Machine learning is hot

- People tracking
- Music/clothing/product recommendation
- Fraud detection – my trip to Ireland

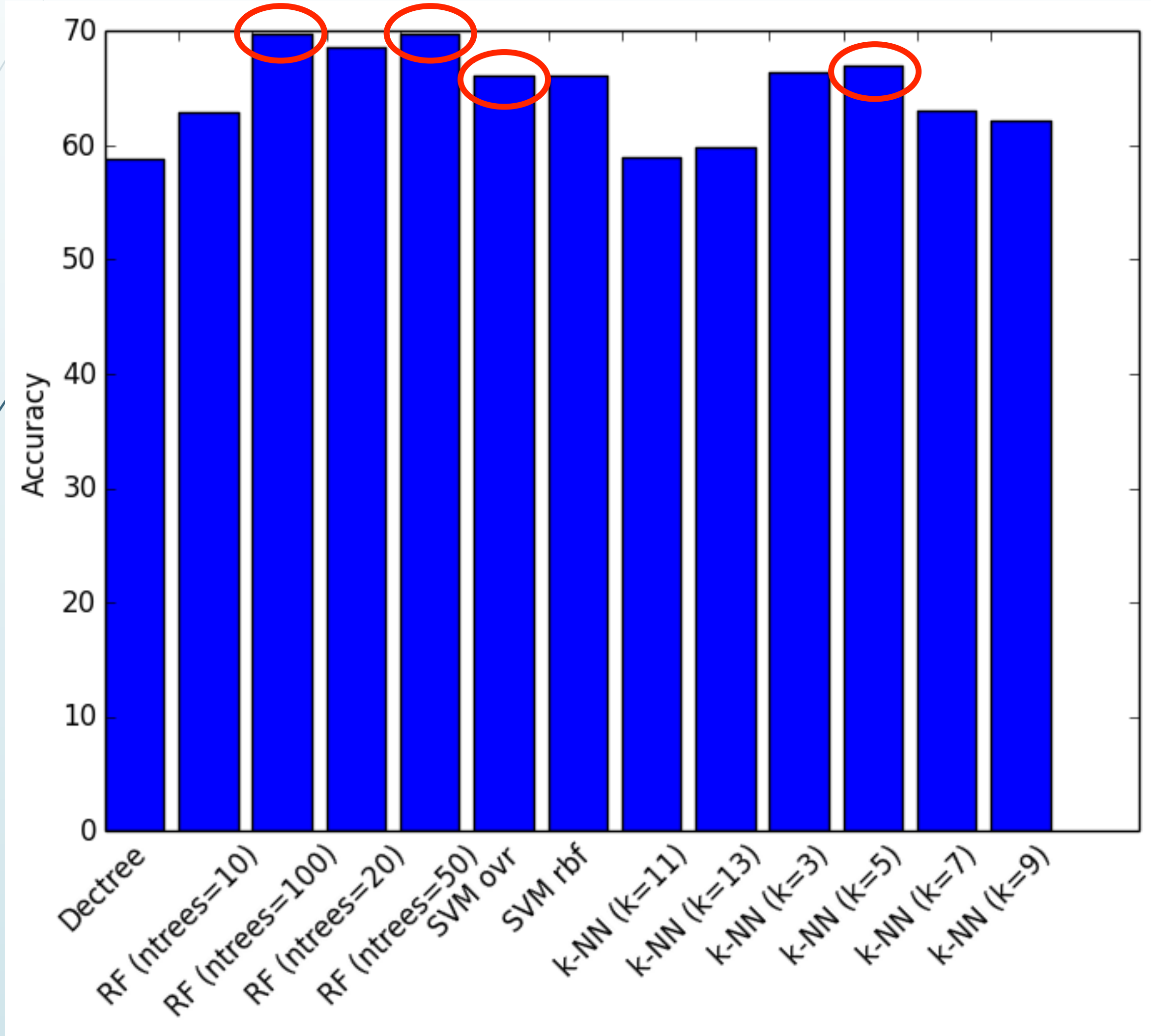
- Mars rovers



- Real world: the task is a moving target
- When can you trust a learned model?

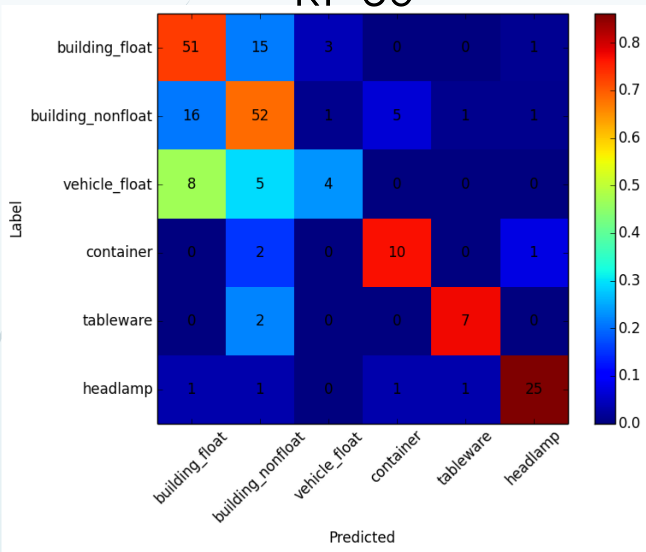
Classification accuracy

5-fold CV on UCI "glass" data set [Lichman, 2013]

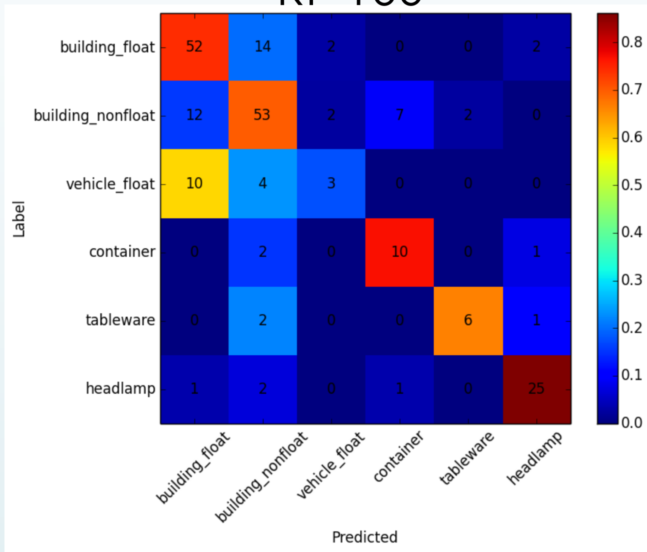


Confusion matrices

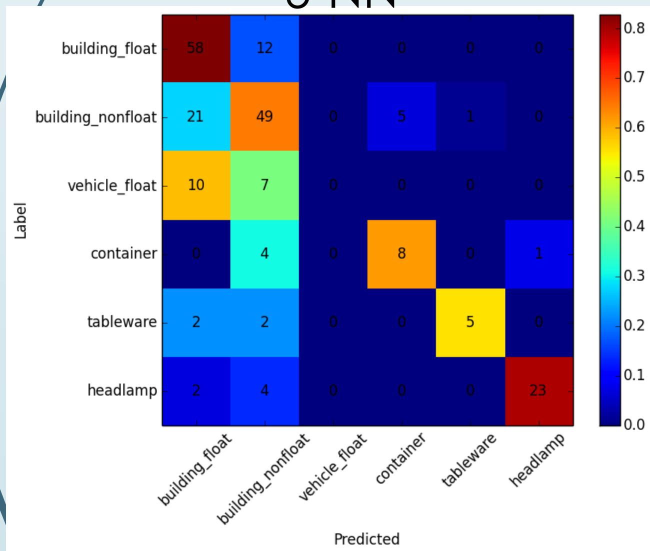
RF-50



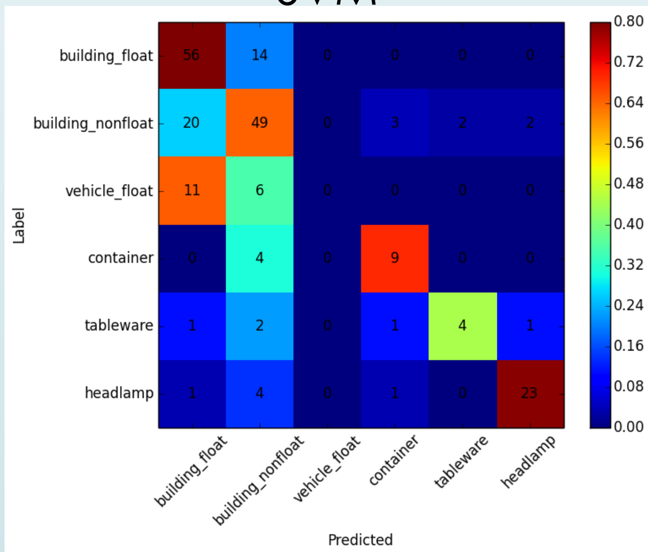
RF-100



5-NN



SVM



Limitations on generalization

- What if data is not i.i.d.?
  - Spatial correlations
  - Temporal correlations
- What if validation set not representative of the future?
  - Has anyone ever applied a glass classifier to new glass samples?
- What if classes are imbalanced?
- What if costs are imbalanced?
- What if the reliability of each item is not equal?
  - Labels
  - Feature values

What we really want to know:

"What was learned?"

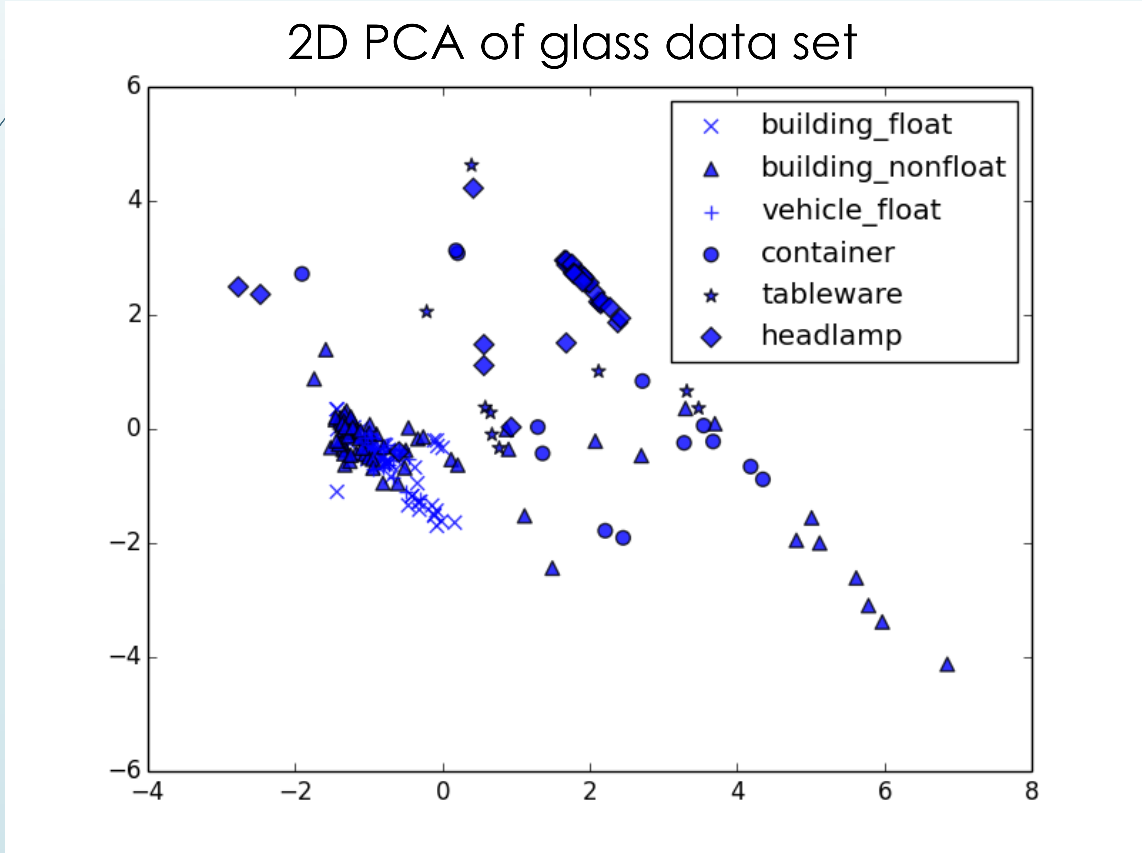
"What WASN'T learned?"

Not just "How well does it work?"

Existing methods

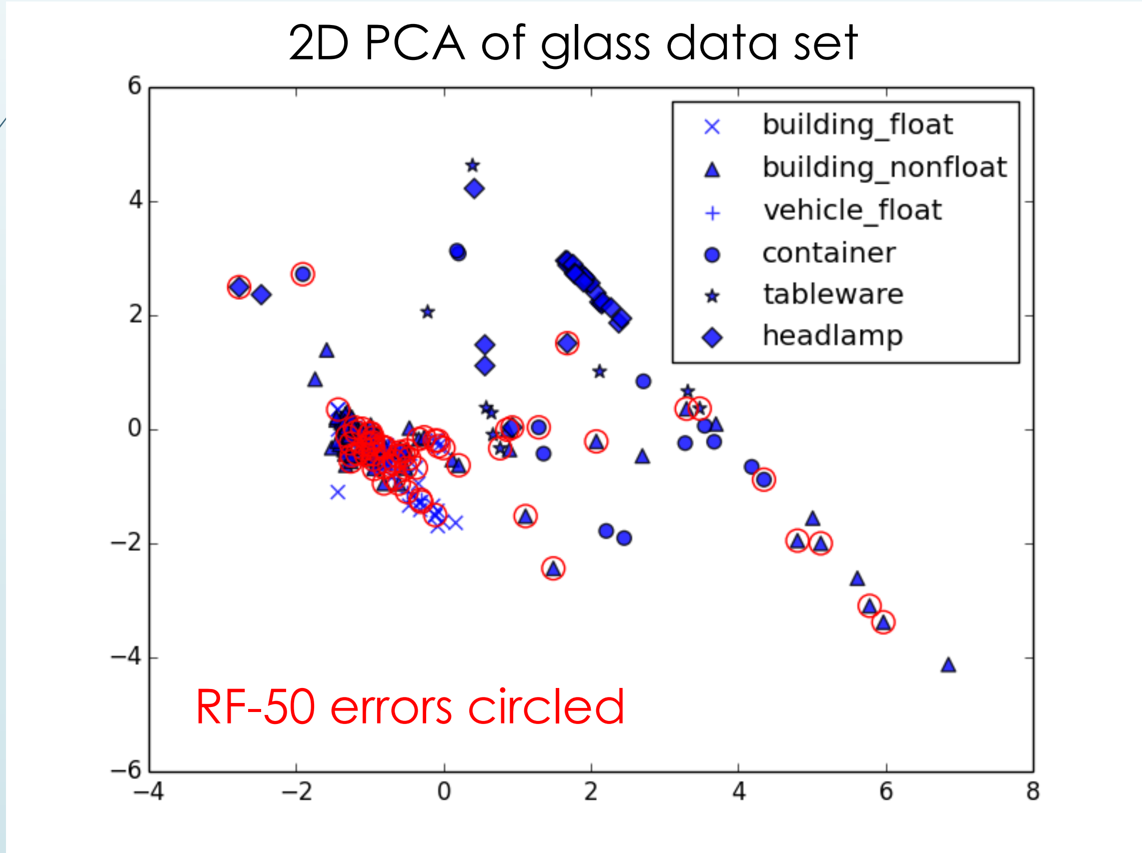
- Mostly in the context of ensembles (which classifier should classify x?)
  - Find subset of examples that classifier gets right; use referees to arbitrate decisions [Ortega et al., 2001]
  - Delegating classifiers (cascade) [Ferri et al., 2004]
- ROC isometrics (when to abstain) [Vanderlooy et al., 2009]
- We want to know what a classifier learned
  - Look deeper at "behavior" [Sturm, 2013]

Visualize errors in context



Visualize errors in context

- What makes these items difficult?
  - Inadequate features? Mislabeled items? Erroneous values?



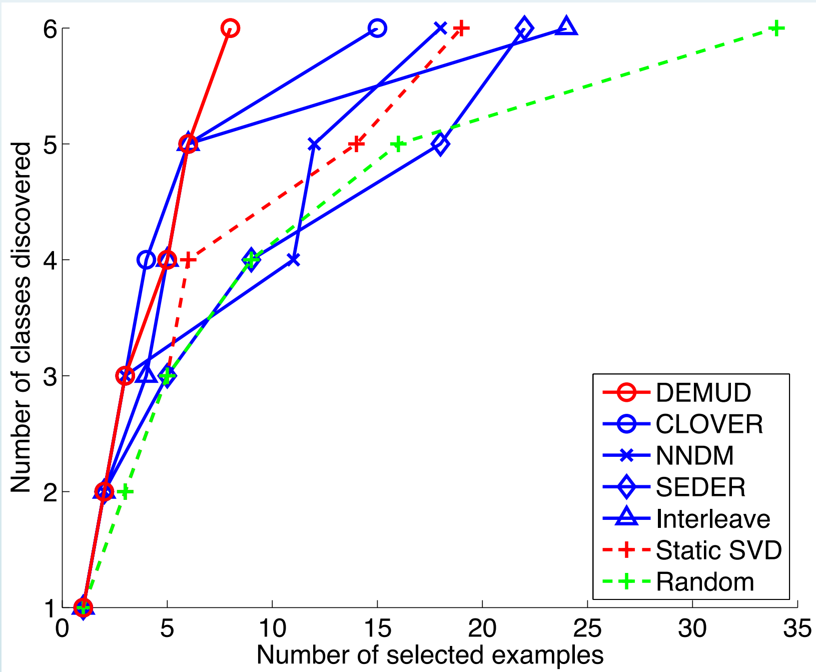
Characteristic evaluation

- Select a diverse set of “case studies” from the data set
- Goal: expose strengths and weaknesses
- Characteristic, not statistical

Characteristic evaluation with a DEMUD traversal

- DEMUD: Iterative discovery [Wagstaff et al., 2013]
  - Incremental SVD model of growing user knowledge
  - Selects most-different item given current model
  - Quickly discovers classes/sub-populations

Glass data set

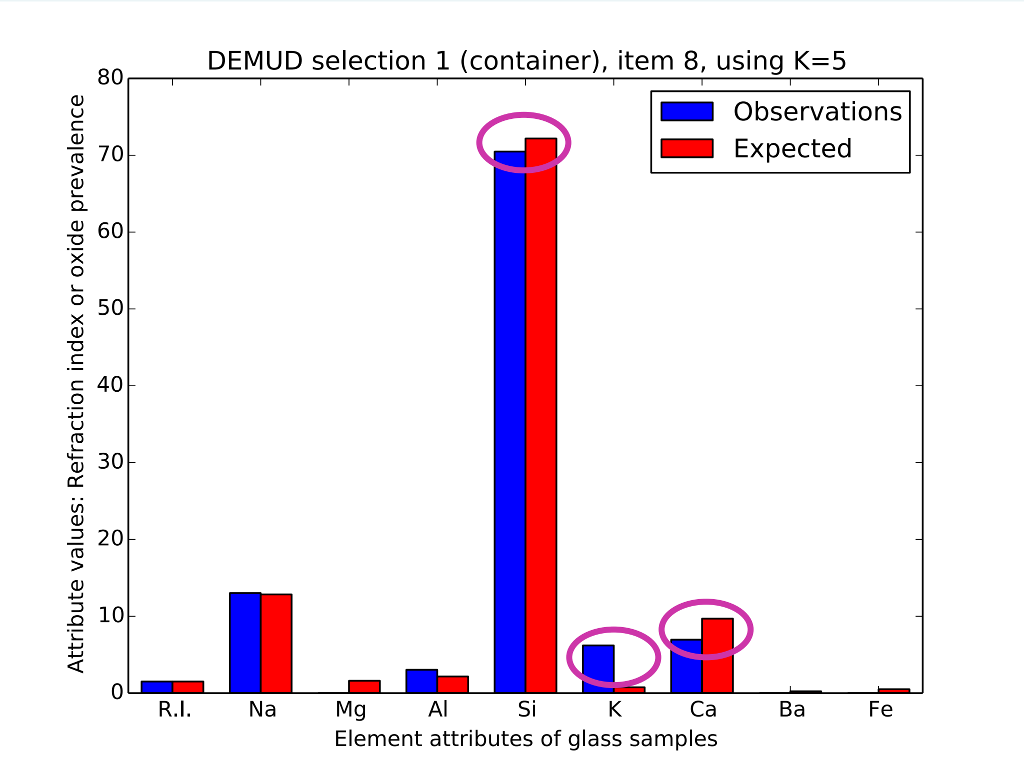


Characteristic evaluation with a DEMUD traversal

- DEMUD: Iterative discovery [Wagstaff et al., 2013]
- Additional benefits
  - Minimizes redundancy
  - Provides per-item explanations
  - Finds sub-populations
    - Unlike clustering, they are prioritized
    - Don’t need to specify how many

DEMUD explanations

- What is surprising about this item?
- DEMUD: Observations vs. expected values (SVD reconstruction)



DEMUD for error understanding

- What is surprising about this item, given its class?

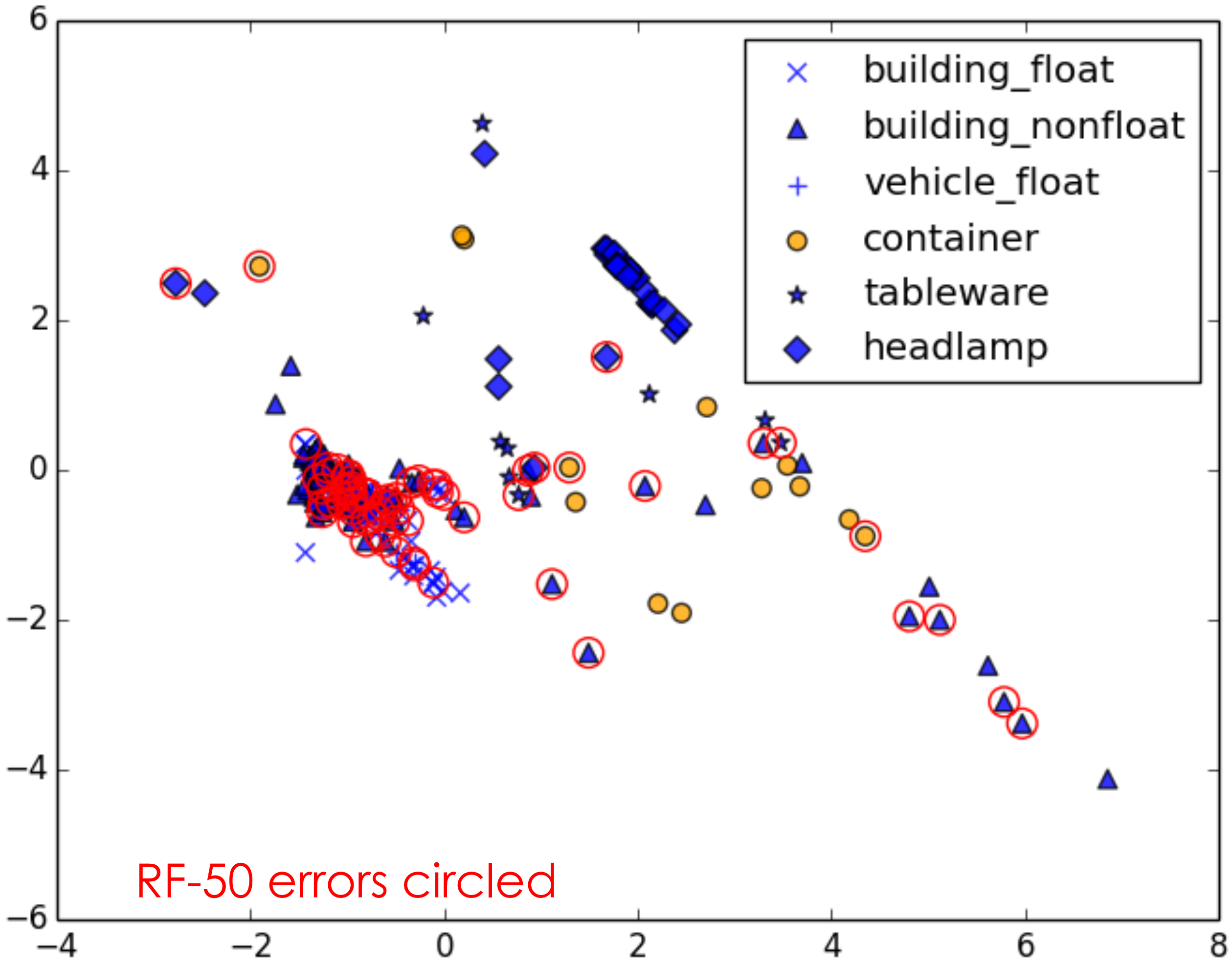
- Top 5 anomaly types in the "Container" class

Item explanation	SVM	5-NN	RF-50	RF-100
Class accuracy	69%	62%	77%	77%
↑ Al,Fe	❌	❌	❌	❌
↓ Ca,Mg,Si; ↑ K,Al	✅	✅	✅	✅
↓ K,Ca; ↑ Mg,Ba	❌	❌	❌	❌
↓ Mg,Fe; ↑ Na,Ca	✅	✅	❌	❌
↓ Mg,Fe; ↑ Si	✅	✅	✅	✅

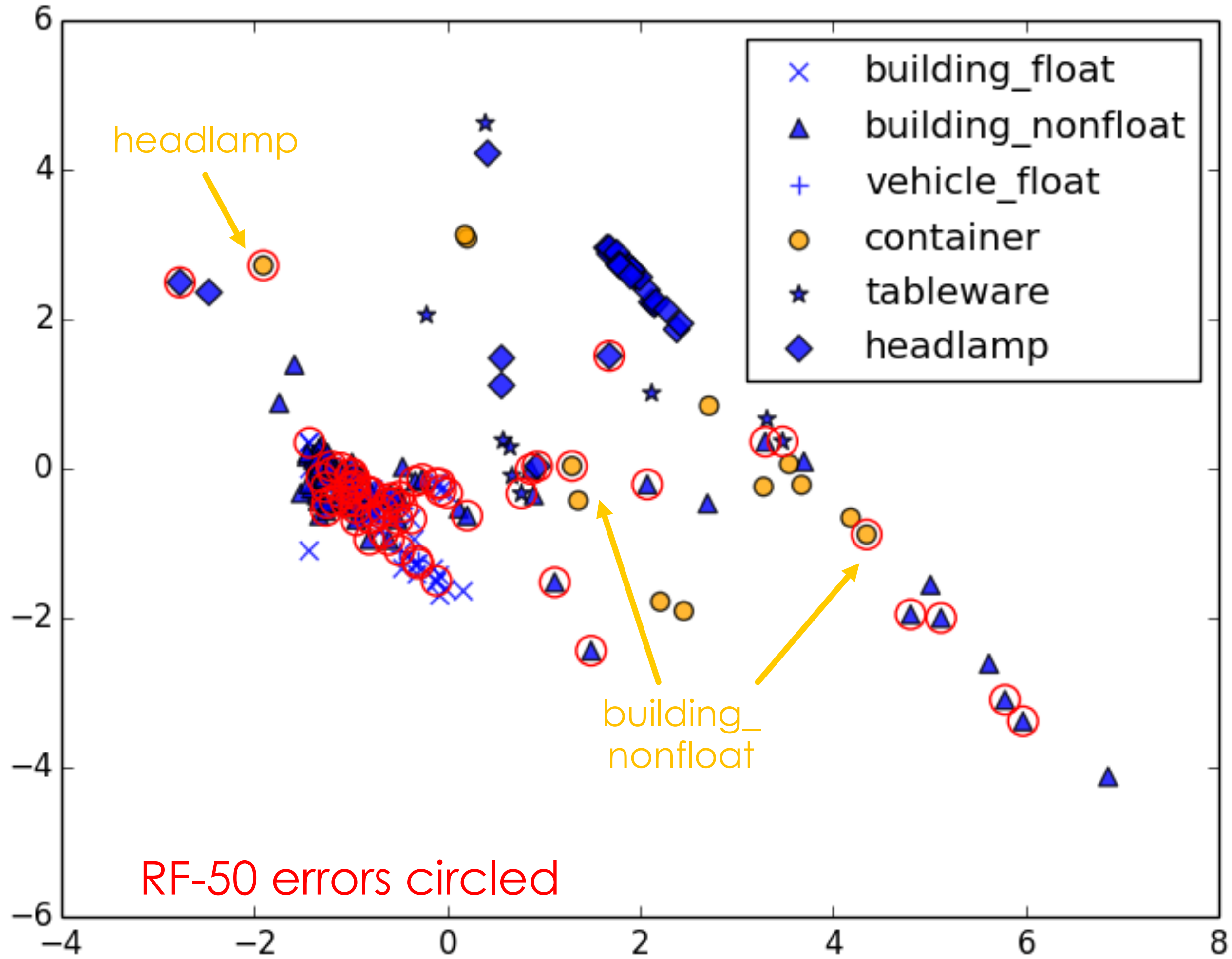
✅ Model is tolerant to this kind of deviation

❌ Model cannot handle this kind of deviation

Container class



Container class



Dark Energy Survey

- Data set: galaxies classified by redshift (distance) bin
  - Low dimension (RGIZ magnitudes)
  - Large N (>3 million galaxies)
  - Wrong redshift bin = distorted model of dark matter distributio
- Goal: find:
  - Potentially mislabeled galaxies
  - Possible errors made by SkyNet (neural network) [Bonnett et al., 2

Summary

- Real-world problems require that we go beyond standard evaluation measures
  - Goal: understand behavior, strengths, weaknesses so we know when to trust a model (or the data)
- One idea: traverse and inspect behavior on subpopulations
  - DEMUD: prioritized traversal by outlier-ness and per-item explanations
  - Could help identify mislabeled items

Thank you!