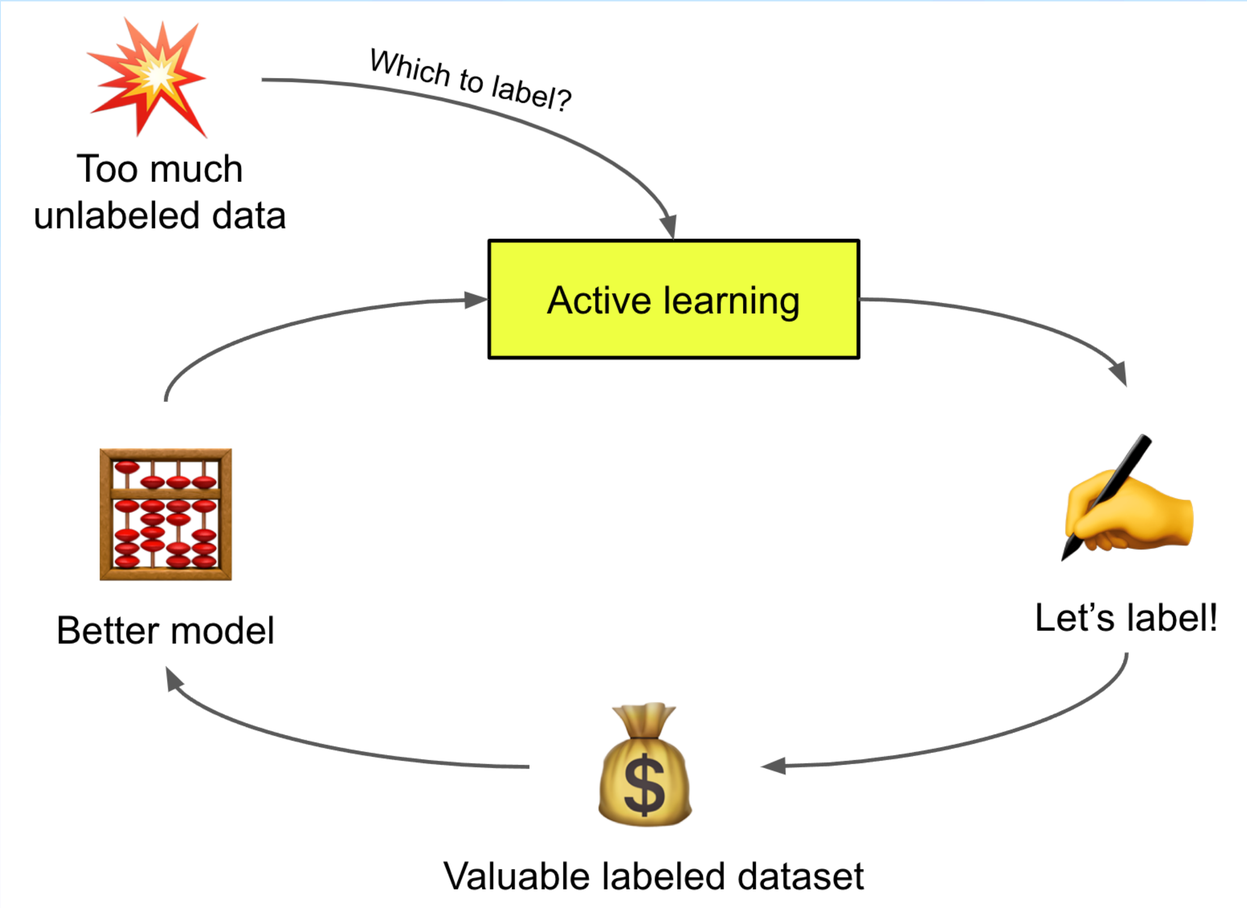
Rashomon Ambiguity Averse Active Learning

December 5 Update

Simon Nguyen

What data point should we label?

- Labeling training data may be expensive!
- Which data points should we label?



Bad Trees in Random Forests

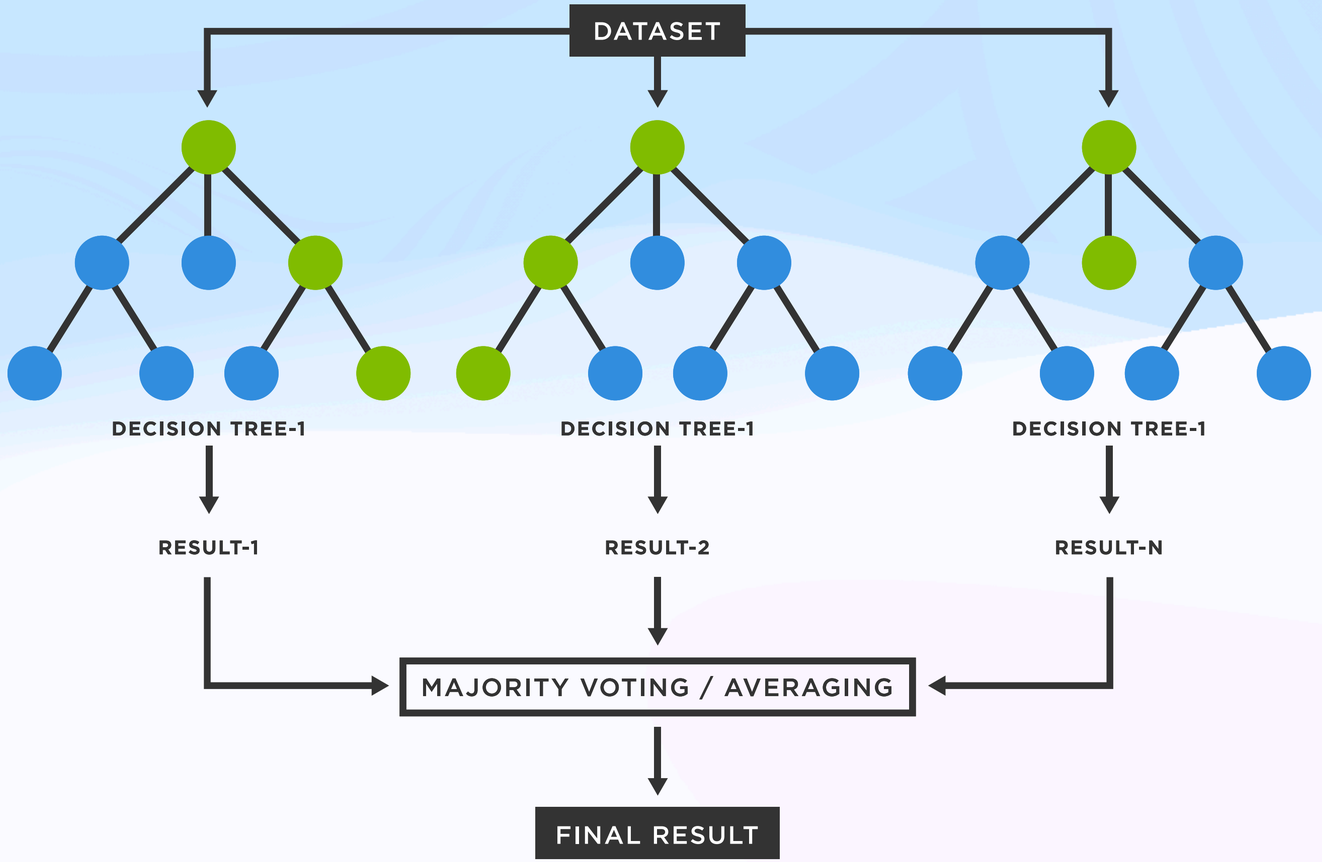
1. Random Forests of often used in active learning classification.

2. However, Random Forests ensemble a random selection of data and covariates

3. This potentially incorporates bad decision trees.

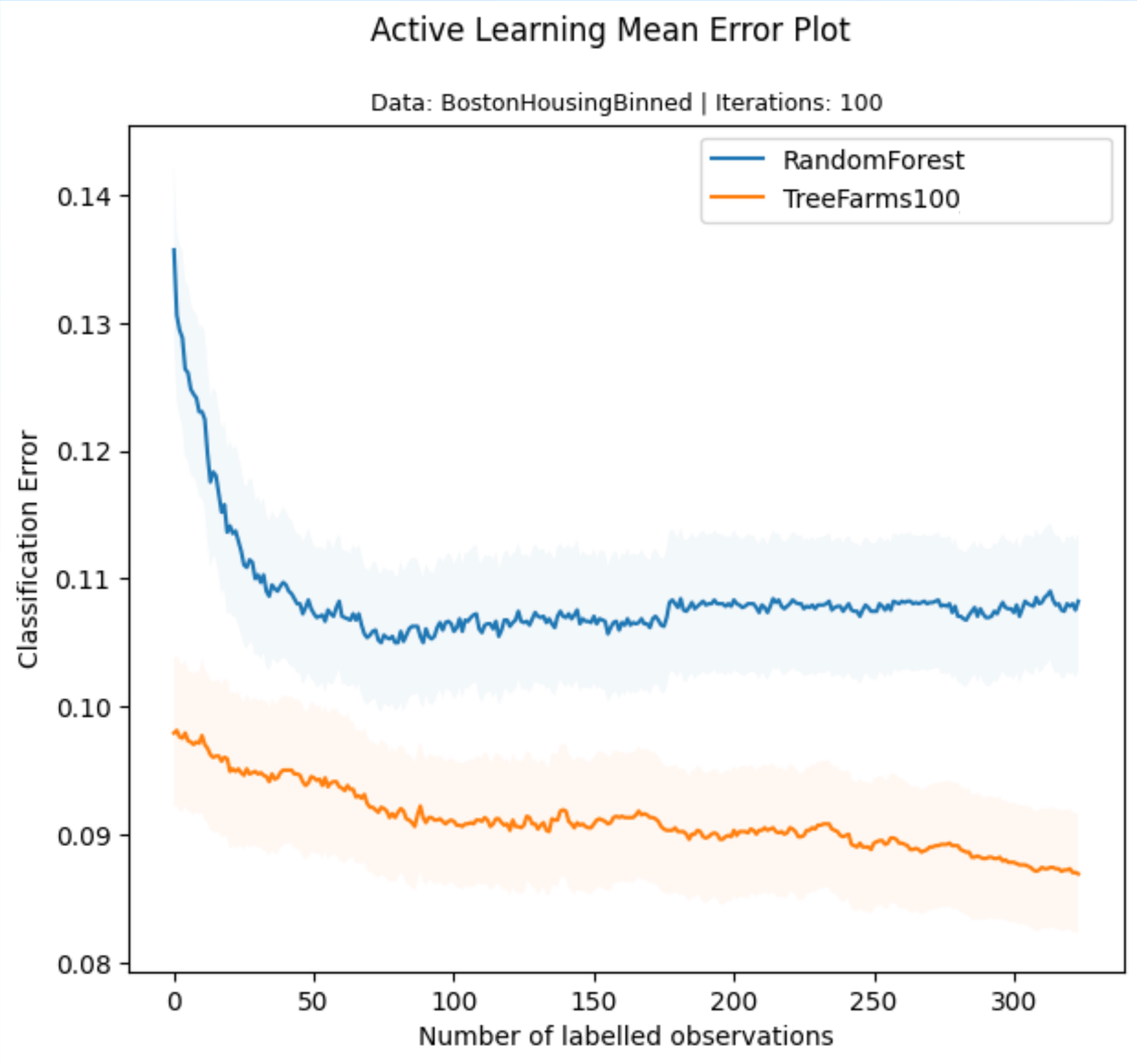
4. This motivates the use of only good decision trees in ensemble methods.

5. The Rashomon Set of good decision trees: TreeFarms!



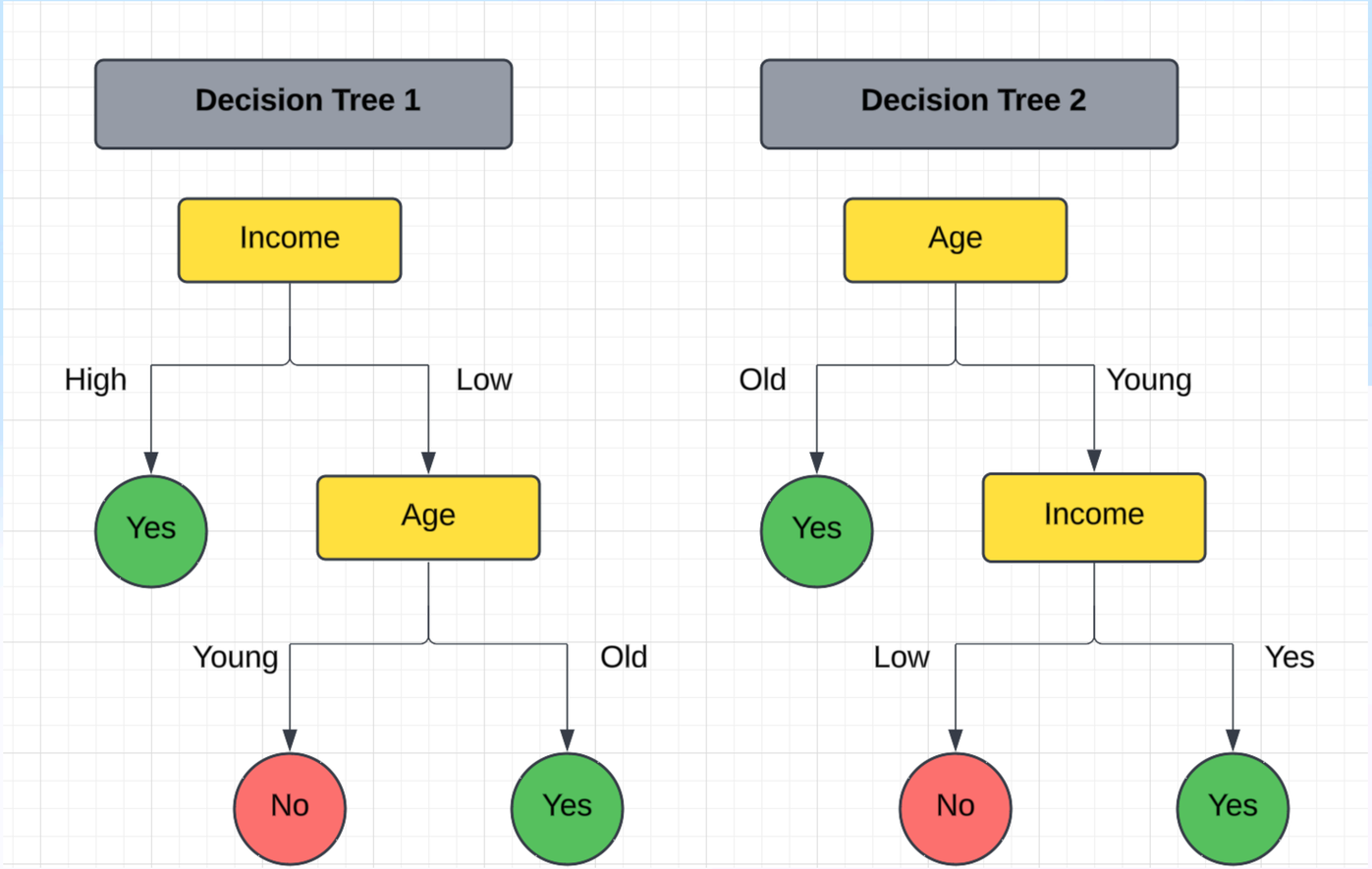
Active learning

- Simulation:
  - Yellow line: random forests.
  - Blue line TreeFarms with the best 100 decision trees.
- Clearly, using the best 100 decision trees is much better than ensembling all the decision trees in random forests.
- Problem solved, right?



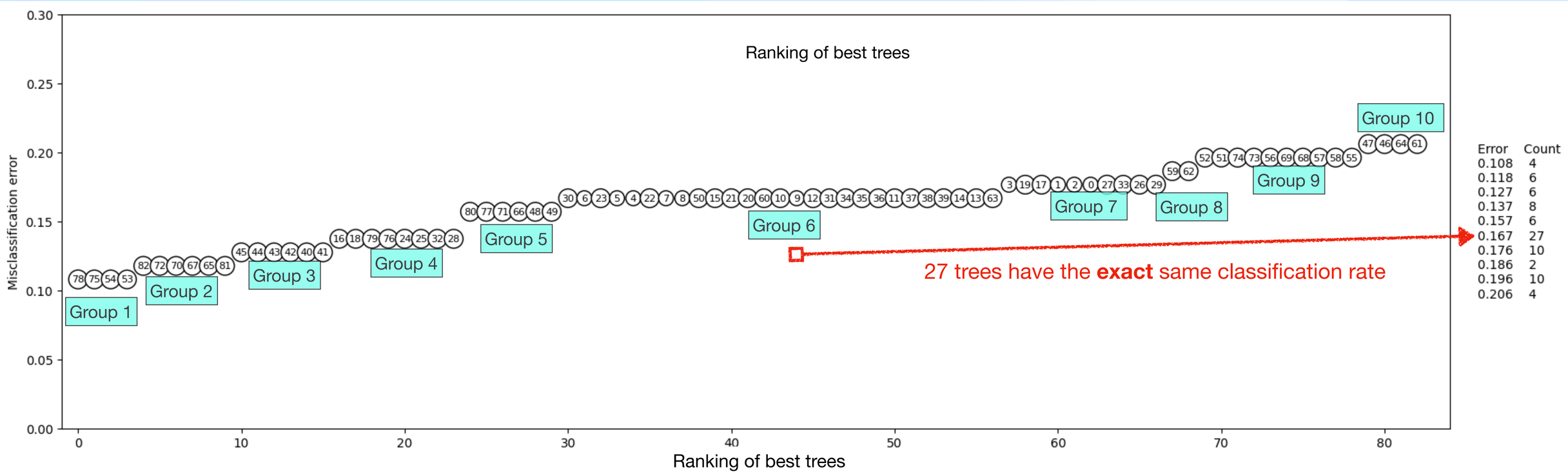
Multiplicity of Explanations in Tree Farms

1. However, TreeFarms suffers from multiplicity of explanations.
    - ie. many trees repeat the same explanation!
2. Redundancy in explanations leads redundancy in predictions.
3. This redundancy skews our notion of uncertainty.
4. Let’s take a look at what this means!



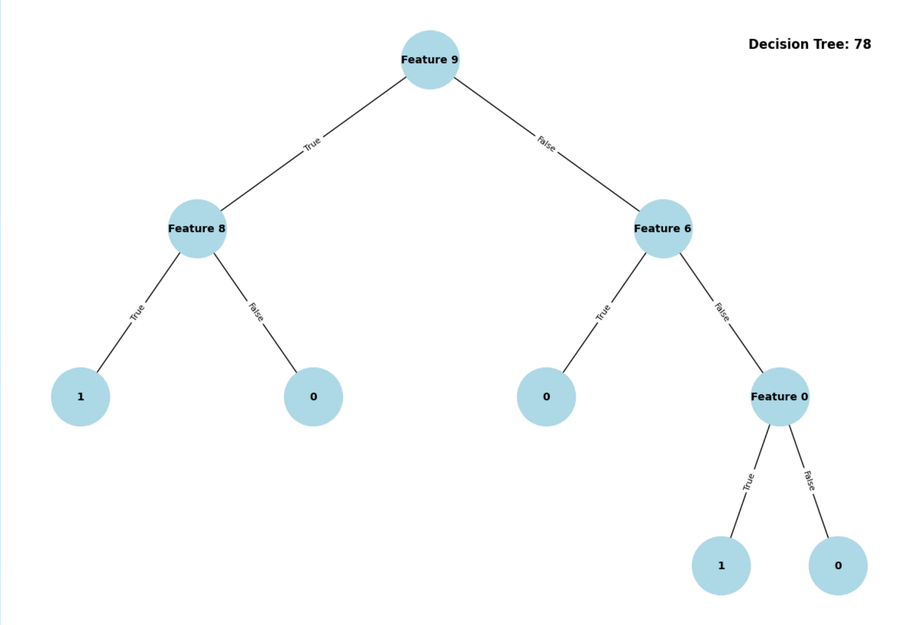
Multiplicity of Explanations in TreeFarms (Grouped)

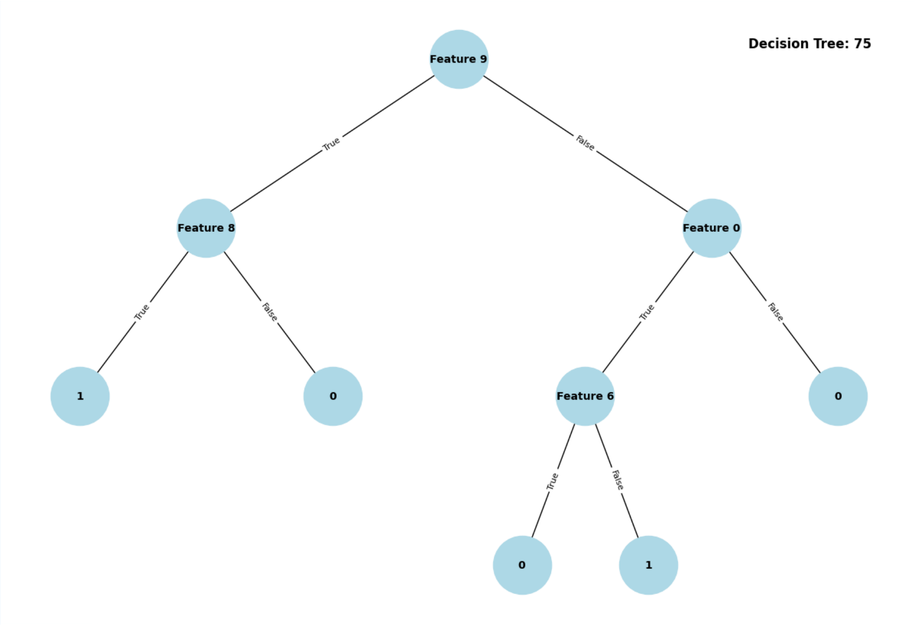
- I group the trees by their misclassification error (y-axis).
- Note how many trees have the same **exact** misclassification rate!
- This suggests many trees share the same explanation of the data, but order covariates differently.

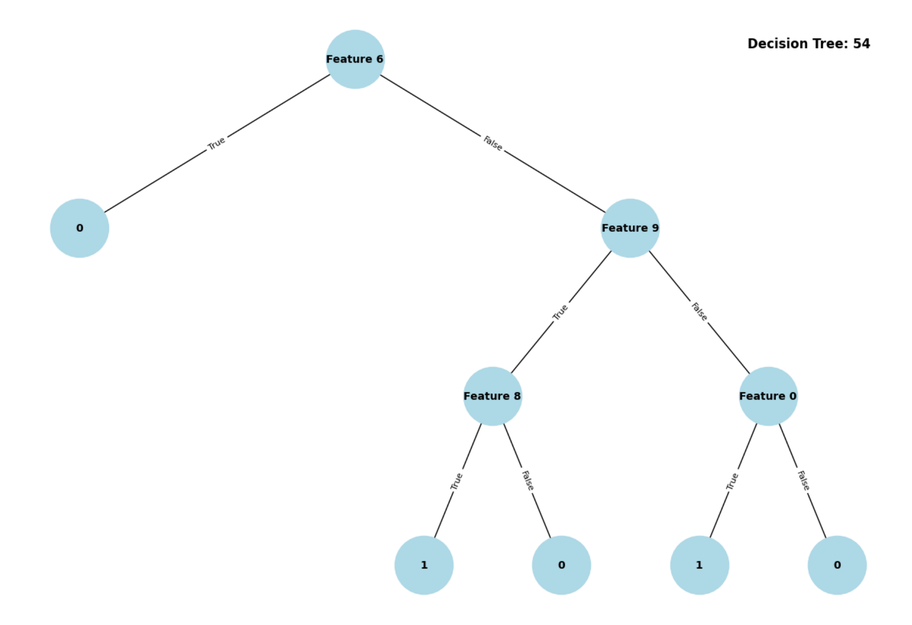


Group 1

Apologies on the lack of image quality/size

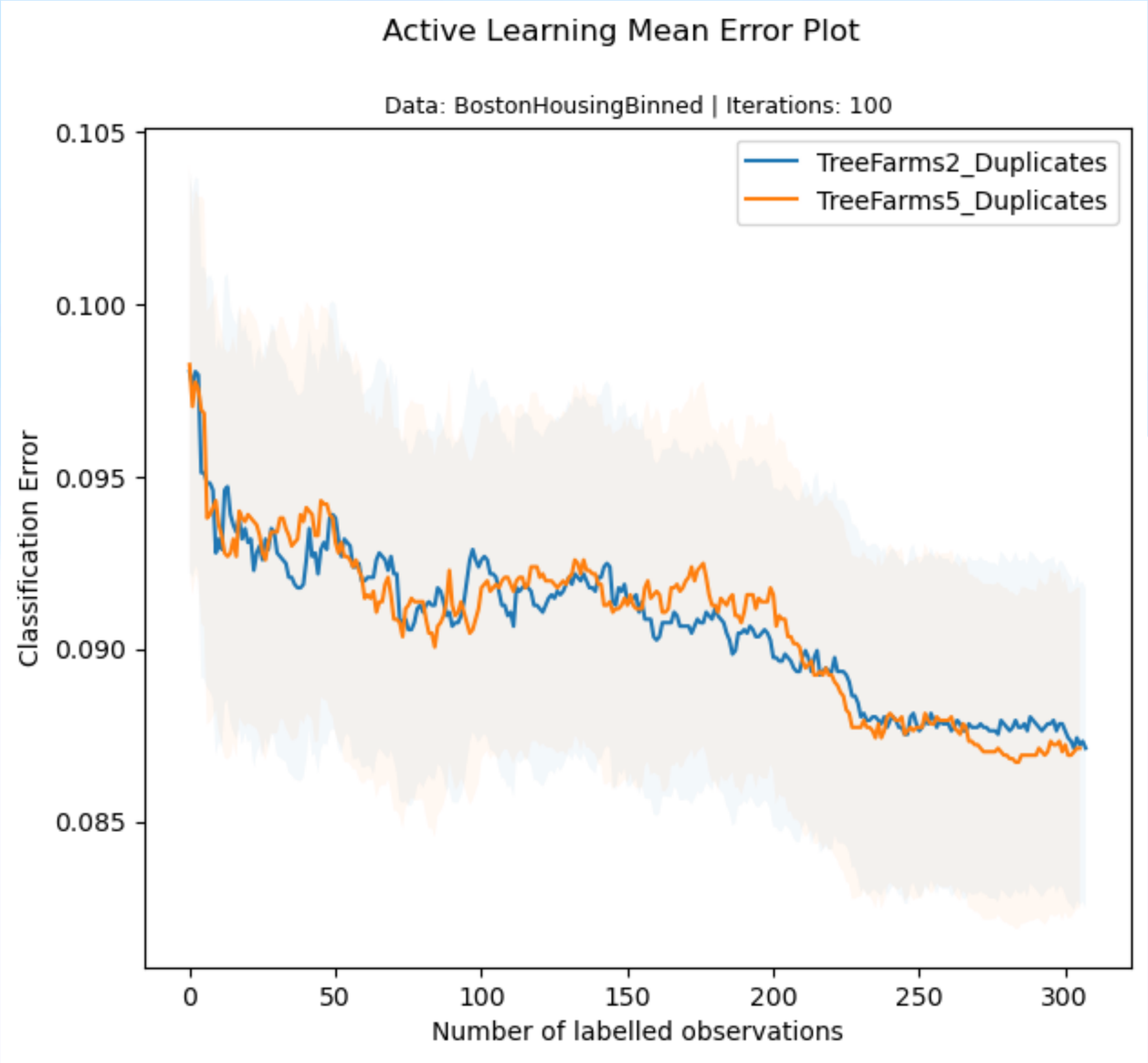






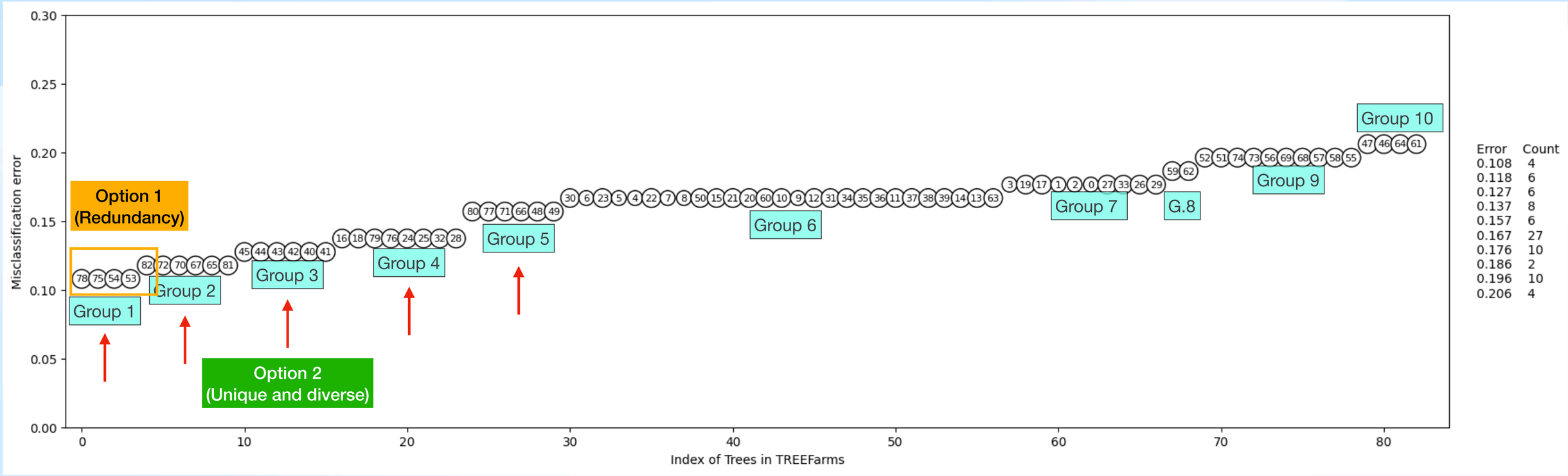
Active learning with Redundant Trees

- In the simulation to the right, active learning is performed with only the best 2 tree!
- Note how similar it is to the active learning method using the top 5 trees!
- The duplication of trees affects query-selection in active learning.



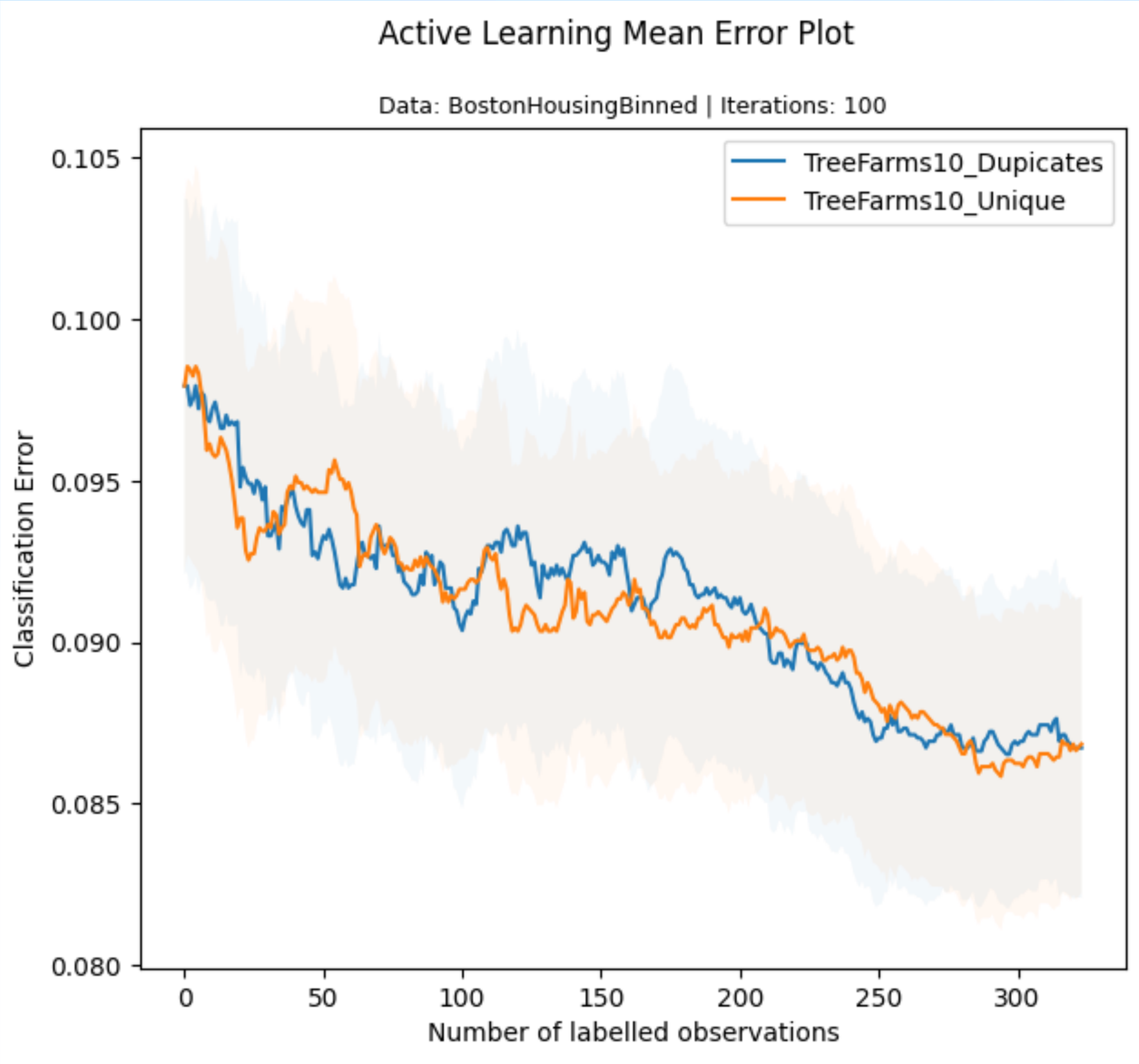
How do we fix this?

- **Possible solution:** We only extract one tree from each group.
- Let's say we measure uncertainty by ensembling the top 4 models.
- Approaches:
  1. Due to redundancy in TreeFarms, we would choose trees [78, 75, 54, 53] (this is what we have been doing - ignoring the redundancy).
  2. Accounting for this redundancy, we would instead choose trees [78, 82, 45, 16, 80] (or any arbitrary tree across the top 5 groups).
- Rather than relying on redundant trees, I attempt to ensure the ensemble reflects multiple explanations of the data across different groups.



How well does selecting unique decision trees from TreeFarms work?

- The drawing of unique trees **does not work that well**!
- The active learning methods are still very similar!



Next Steps

- Suggests we should partition the dataset differently!
- Use of Rashomon Partition Sets

ROBUSTLY ESTIMATING HETEROGENEITY IN FACTORIAL DATA USING RASHOMON PARTITIONS

APARAJITHAN VENKATESWARAN[§], ANIRUDH SANKAR[‡], ARUN G. CHANDRASEKHAR[‡,⋆], AND TYLER H. MCCORMICK[§,¶]