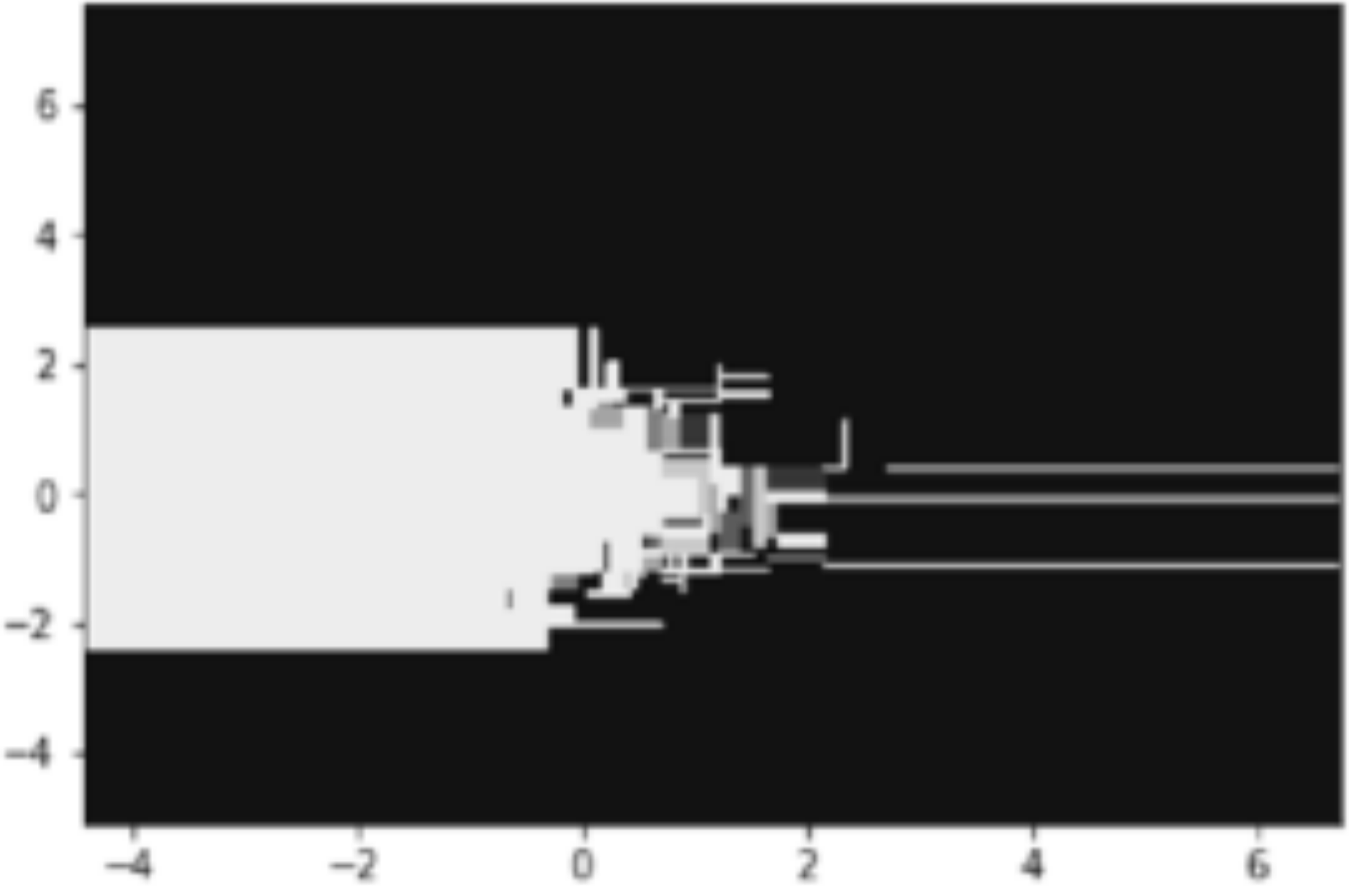

H2O.ai

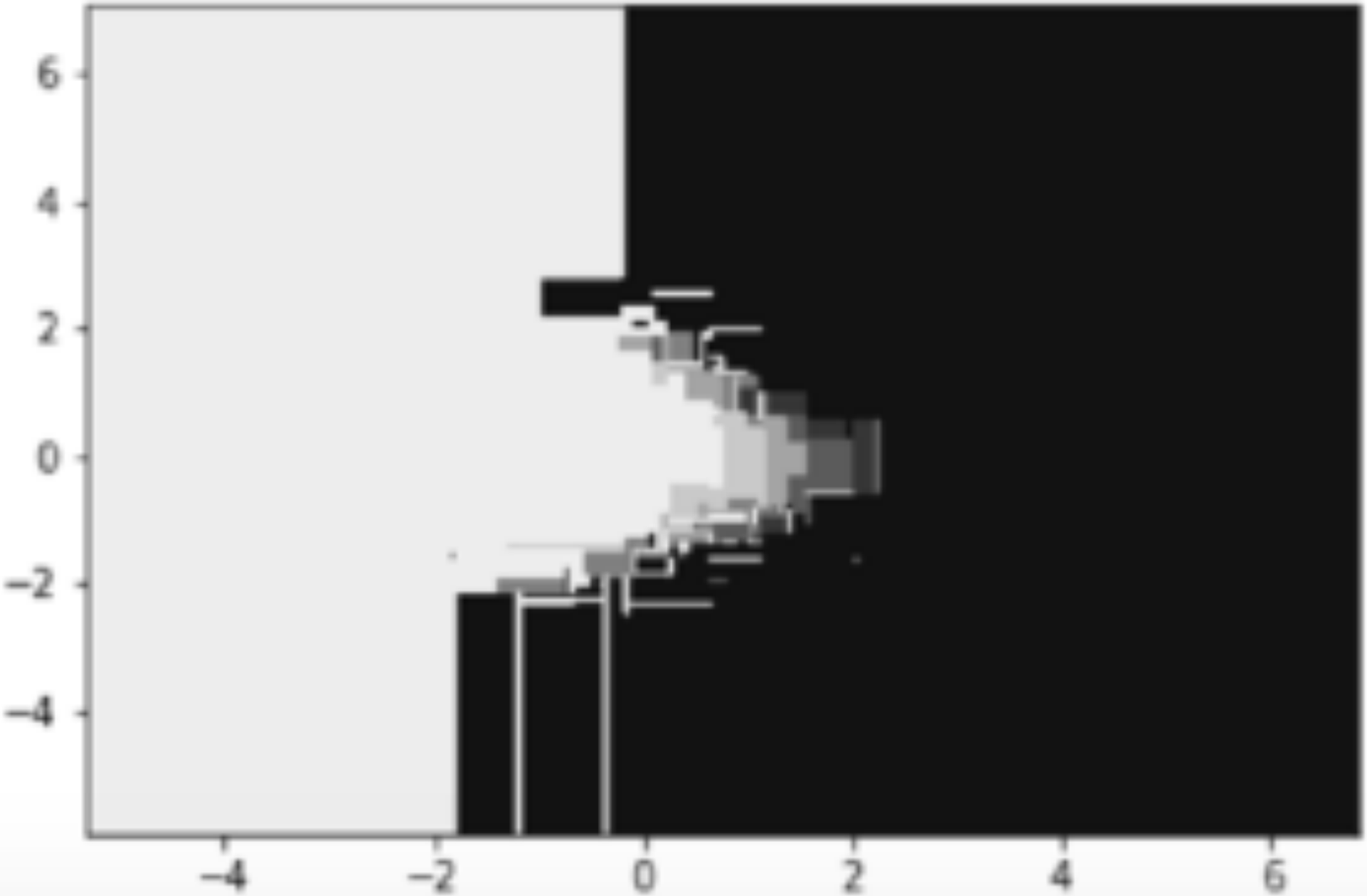
Pros/Cons of Decision Trees

✔ simple to understand/interpret
✔ little data prep
- natural handling of "mixed" data types
- handling of missing values
- handing of multi-class outputs
✔ robustness to outliers in input space
✔ insensitive to monotonic transformations of inputs
✔ computational scalability (large N)

✔ automatically ignores irrelevant inputs

✗ weak predictors
✗ can be unstable to small variations in the data
✗ poor ability to extract linear relationships
✗ can create biased trees if classes unbalanced


6
4
2
0
−2
−4
−4
−2
0
2
4
6

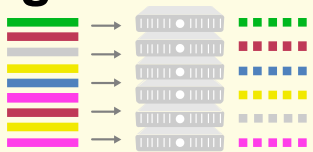

6
4
2
0
−2
−4
−4
−2
0
2
4
6

Exercise 1: Creating Decision Tree in sklearn

$\mathcal{O}(n \log n)$

Scalable Implementation in H2O

1 Parallel Data Ingest



Data is stored in-memory on all cluster compute nodes

- **Rows are evenly distributed across the cluster**
- **Columns are stored separately and compressed**

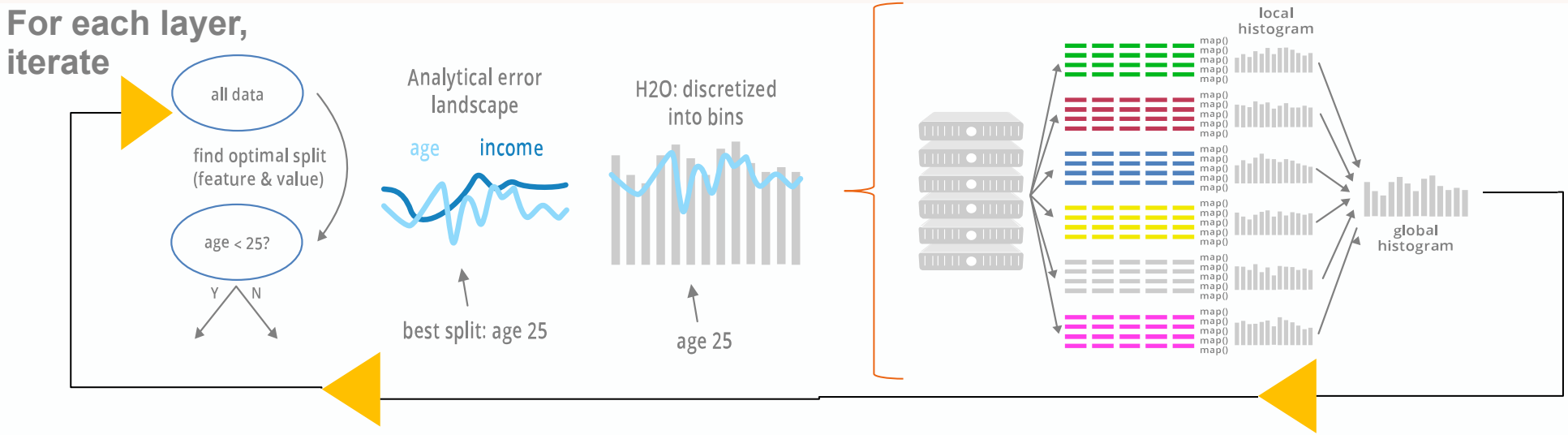
Basis for fine-grain Map/Reduce for histogram calculation

2 Distributed Tree Building via Fine-Grain Map/Reduce to find optimal split points of data layer by layer

Start with root node and build layers of tree nodes **[ILLUSTRATION BELOW]**

For each layer, repeat the following:

- **For a set of features, split the data at every possible split point**
- **Find the split that leads to best model improvement**
- **Use discretization to limit the number of potential splits**
  - **To find the split, local histograms are calculated on each node and then aggregated into a global histogram**
  - **From the global histogram, the best split column is chosen**



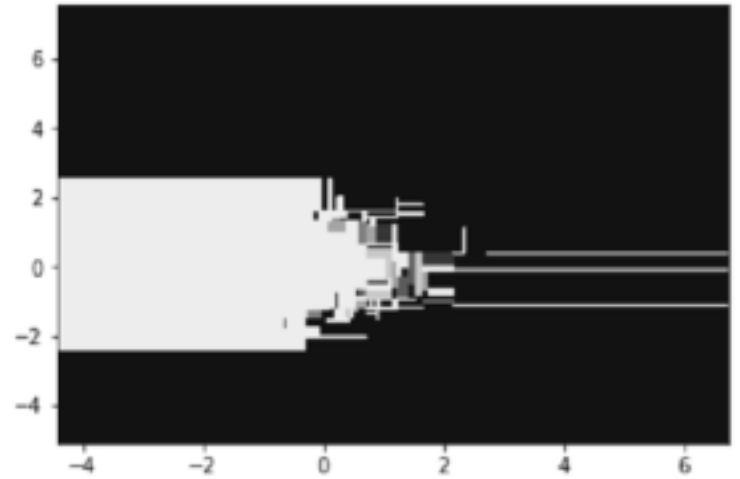
H2O.ai

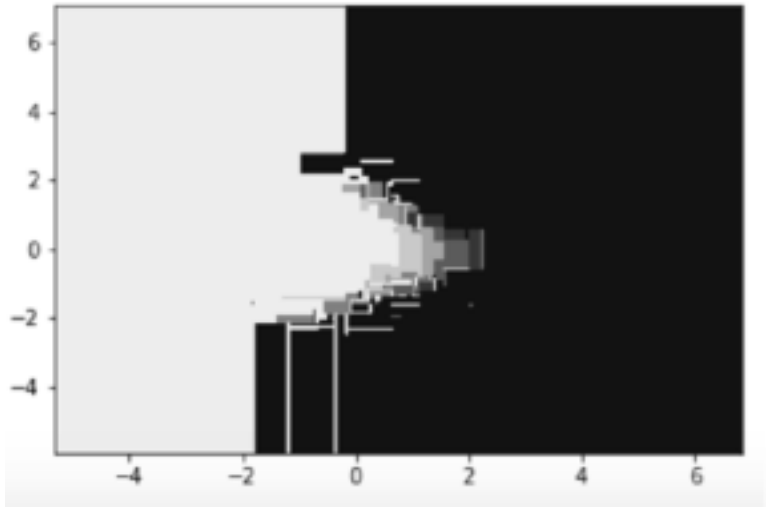
Pros/Cons of Decision Trees

✔ simple to understand/interpret
✔ little data prep
  - natural handling of "mixed" data types
  - handling of missing values
  - handing of multi-class outputs
✔ robustness to outliers in input space
✔ insensitive to monotonic transformations of inputs
✔ computational scalability (large N) $\mathcal{O}(n \log n)$

✔ automatically ignores irrelevant inputs

✗ weak predictors
✗ can be unstable to small variations in the data
✗ poor ability to extract linear relationships
✗ can create biased trees if classes unbalanced





Exercise 1: Creating Decision Tree in sklearn