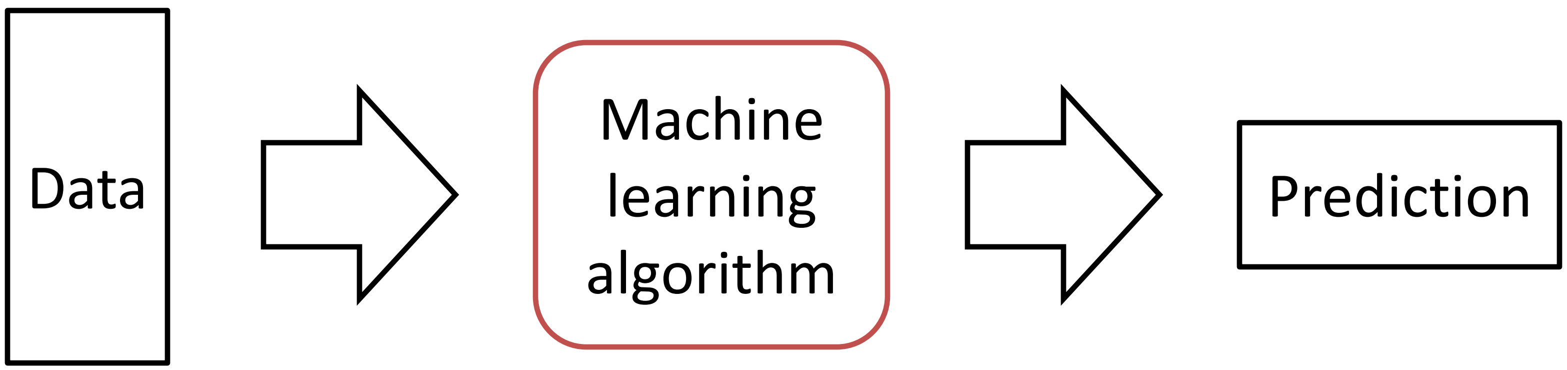
Random Forest Ensemble Visualization

CPSC 547 Project

Ken Lau

Prediction



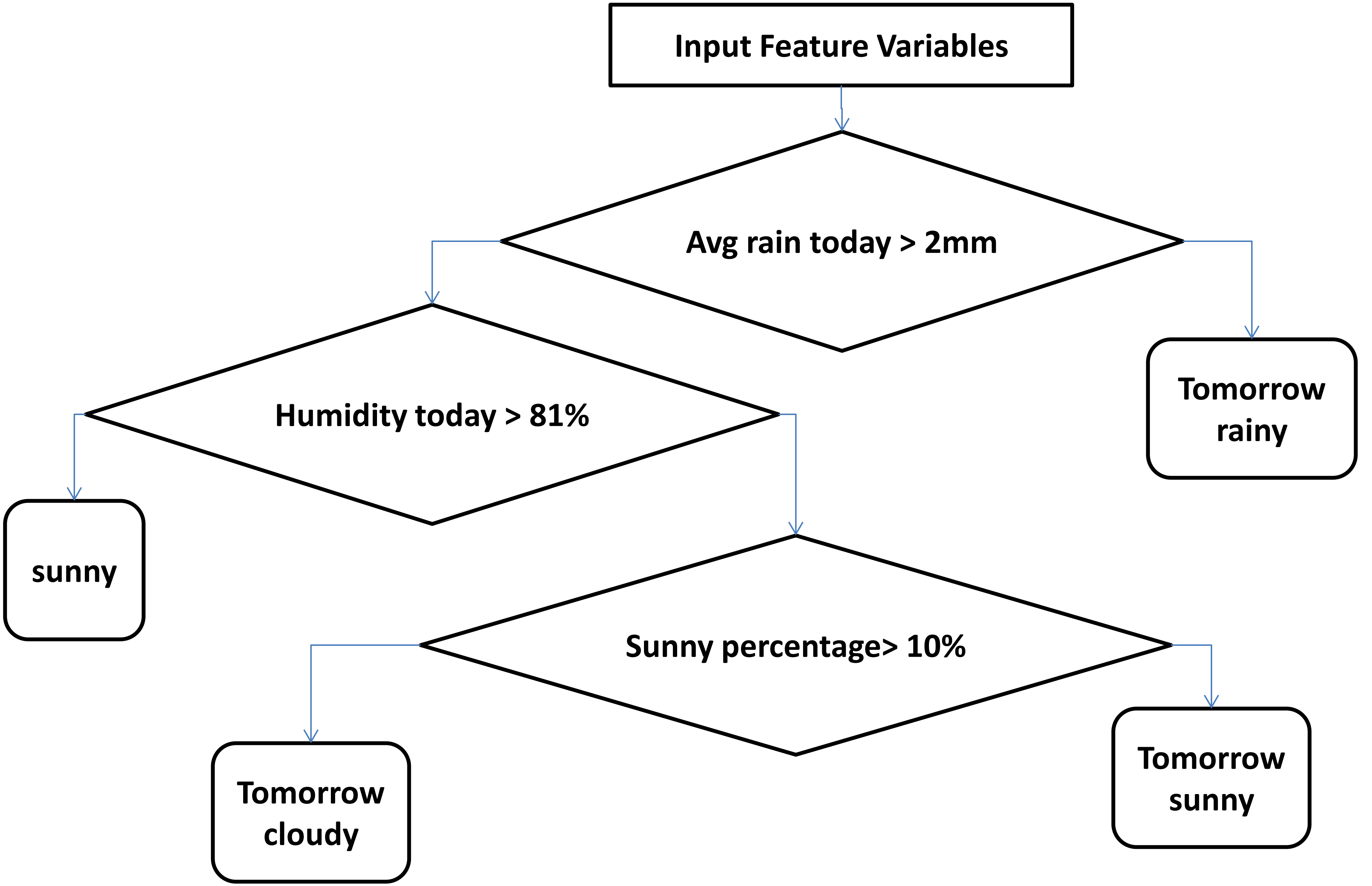
Weather Data Example

Class Prediction Variables

Feature Variables

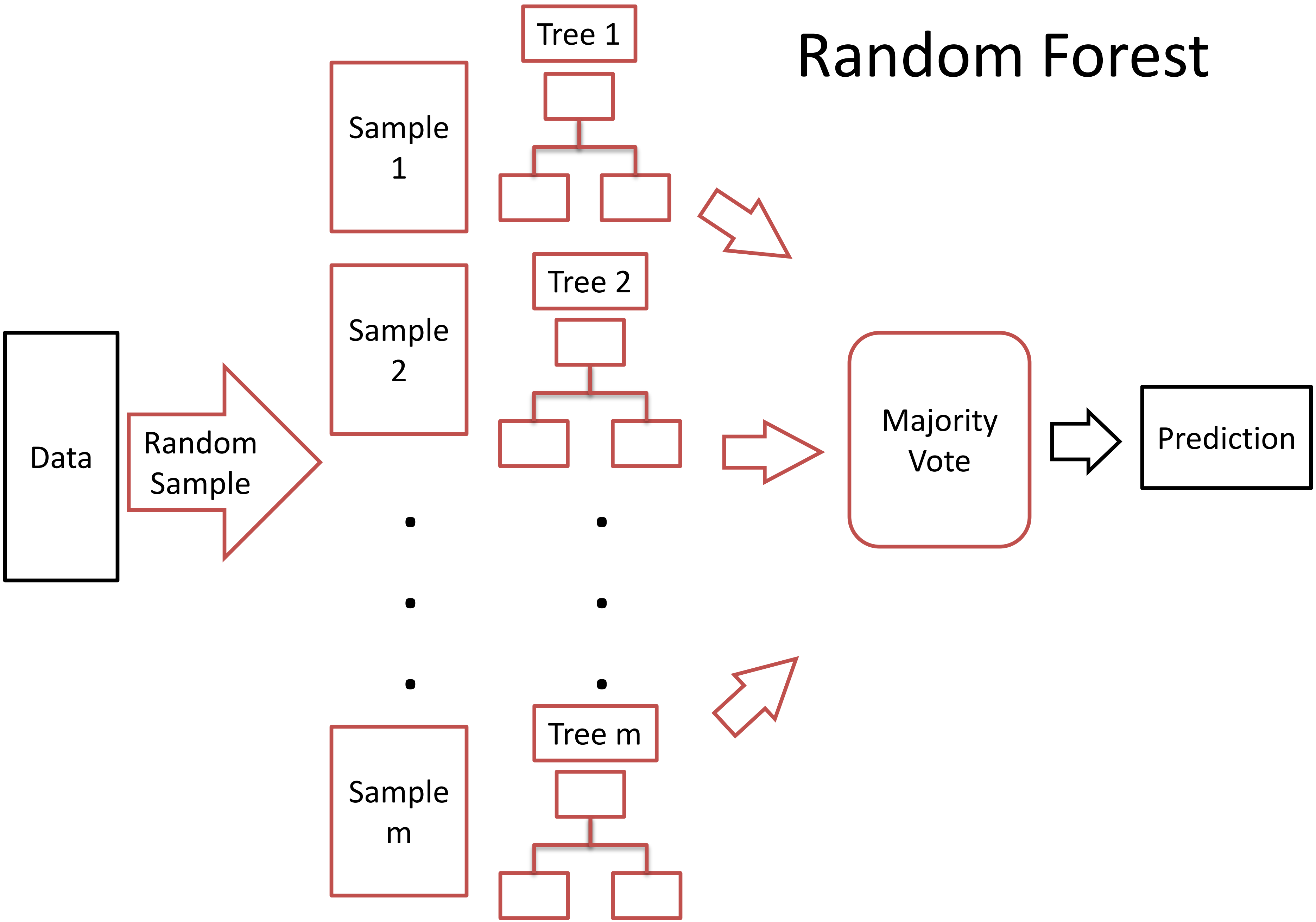
Obs	Avg rain today	Humidity today	Sunny percentage today	Weather Tomorrow
1	1mm	81 %	10%	Rainly
2	2mm	83 %	40%	Cloudy
3	0 mm	80 %	80%	Sunny
4	1 mm	81 %	20 %	Cloudy

Classification Tree

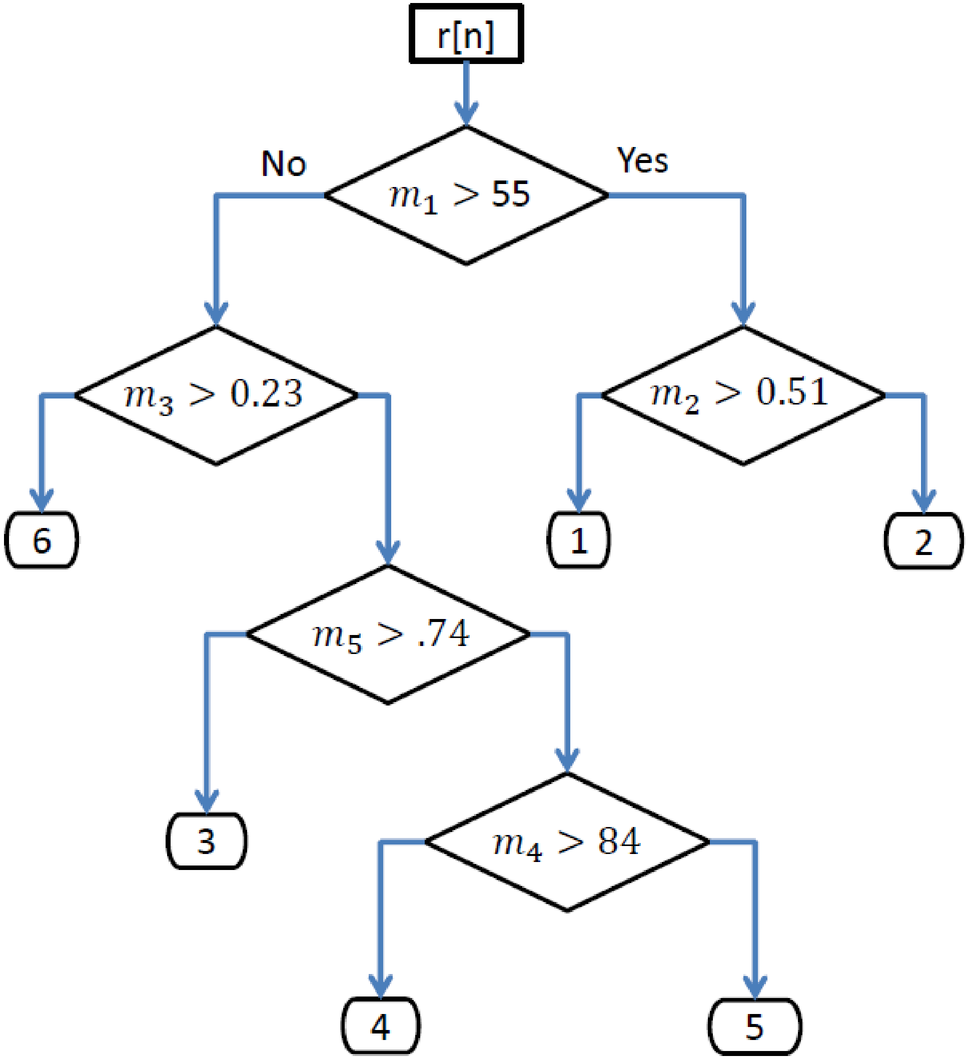


Random Forest

- Collection of classification trees
  - Usually 500-1000
- Popular
- Black box


Random Forest
Data
Random Sample
Sample 1
Sample 2
Sample m
Tree 1
Tree 2
Tree m
Majority Vote
Prediction

My Data

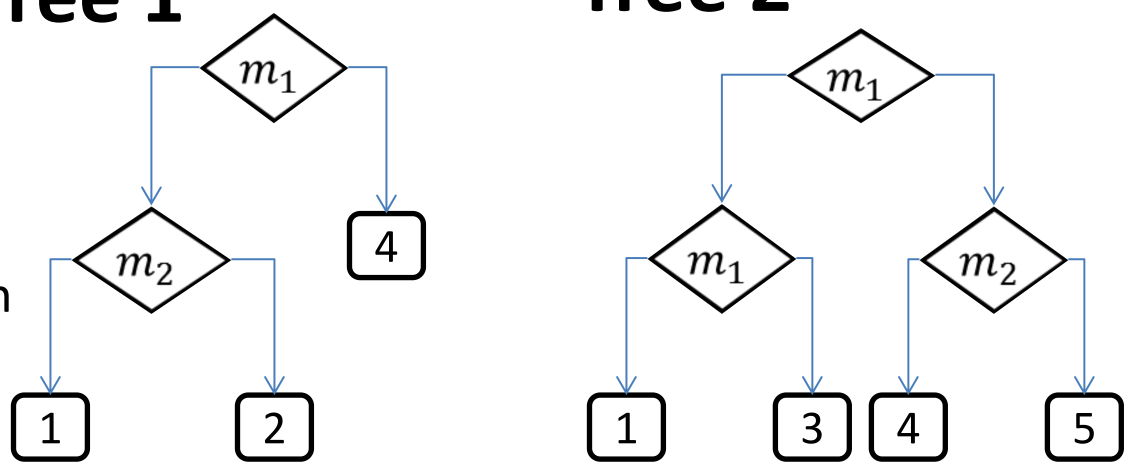


Problem:

How to visualize the collection of Classification trees

Aggregate: Features Variables

Tree 1 **Tree 2**



Encode
Colour Saturation

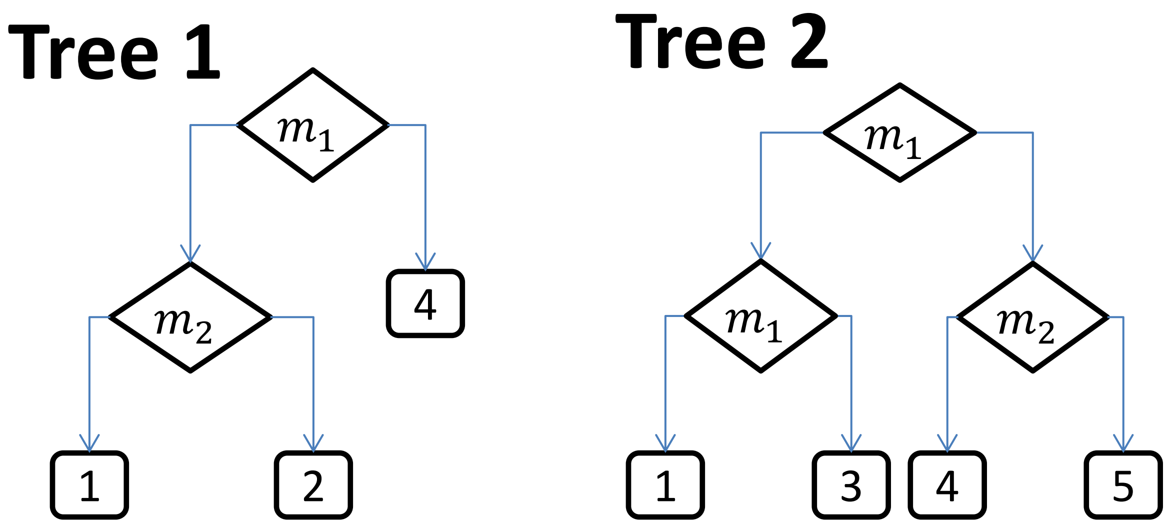
Depth	Feature	Appearance
root	m1	2

Data Derive

Depth	Parent	Feature	Left Split	Right Split
2	m1	m2	1	1

Depth	Parent	Feature	Left Split	Right Split
2	m1	m1	1	0

Aggregate: Class Prediction Variables

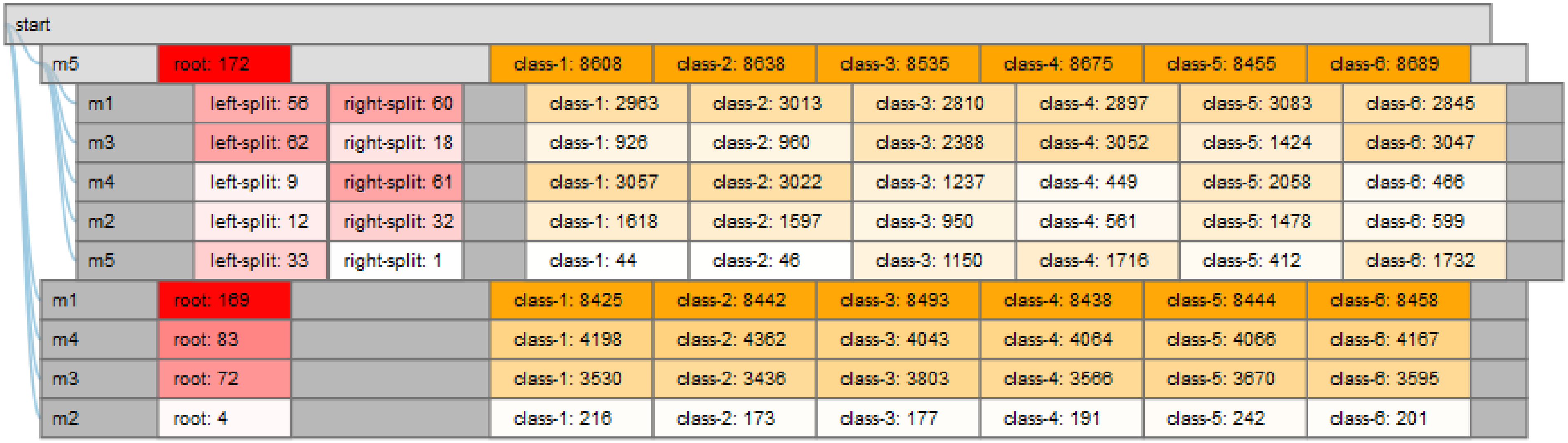


Depth	Feature	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6
root	m1	2	1	1	2	1	0

Depth	Parent	Feature	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6
2	m1	m2	1	1	0	1	1	0

Depth	Parent	Feature	Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	Cl 6
2	m1	m1	1	0	1	0	0	0

Visualization



- So What?
  - Feature importance and interaction
  - Tree pruning when non-uniform class count distribution occurs
  - Class count predictions given nodes traversed so far

Software

- Python
  - Model fitting
  - Information retrieval
  - Aggregation

- D3
  - Encoding
  - Based on Indented Tree (Mike Bostock, 2011)

  http://bl.ocks.org/mbostock/1093025

Demo

Visualization link:
http://kenlau177.github.io/Indented-Agg-Tree/

Scale

- Manageable up to trees of depth 8 with 5 feature variables.
  - Out of memory issue
  - There is a step that generates all possible permutations of features variables
  - Instead keep only variables that appear at least once in the collection of trees

- Handles more than 1000 trees fast with depth less than 7

Number of Trees	Depth	Time
200	7	59 sec
800	3	10 sec
1500	3	15 sec
1500	6	22 sec

Quantify the Tree Ensembles

- Measure diversity among trees based on class predictions
- Unrelated members are the reason for high accuracy
- Hamann Similarity Measure
  - Multivariate version

Predicted **Same** Class

Tree 2

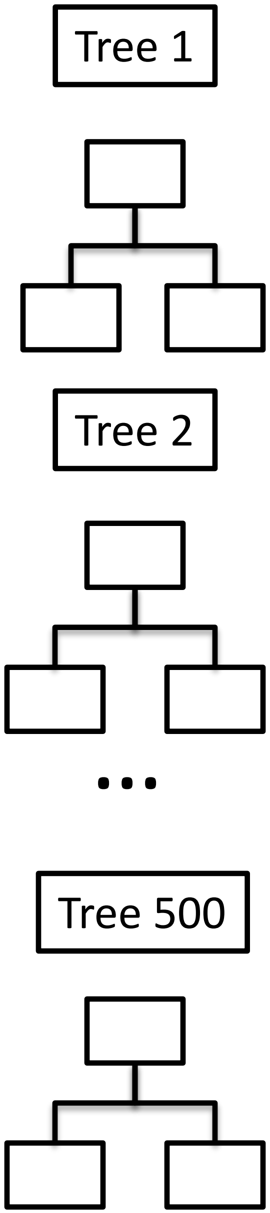
Tree 1		Correct	Incorrect
	Correct	a1	0
	Incorrect	0	d1

Predicted **Different** Class

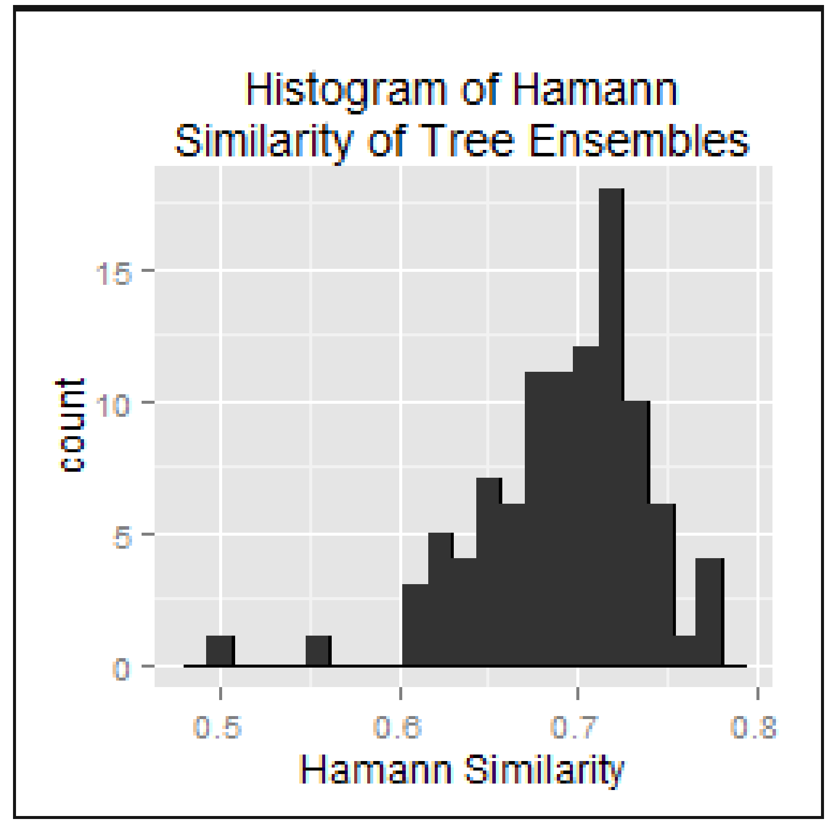
Tree 2

Tree 1		Correct	Incorrect
	Correct	0	b2
	Incorrect	c2	d2

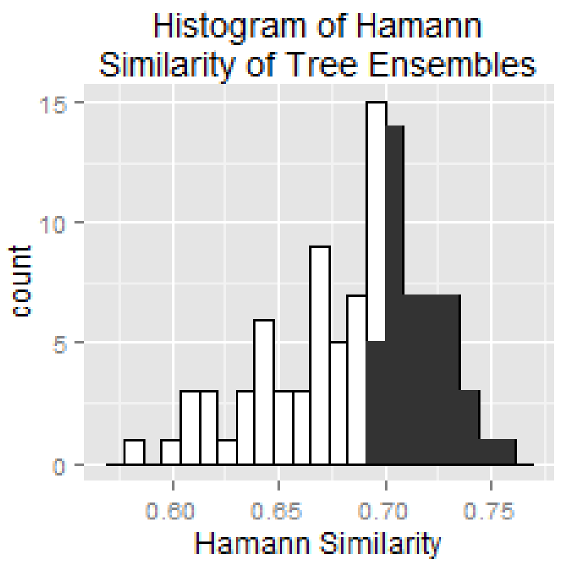
$$H = \frac{(a_1 + d_1) - (b_2 + c_2 + d_2)}{a_1 + d_1 + b_2 + c_2 + d_2}$$

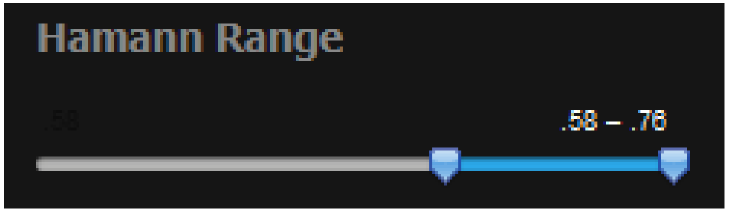


Derived Data

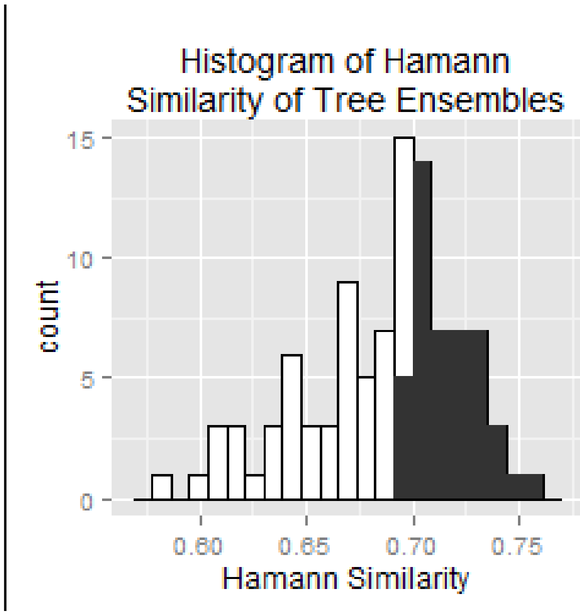


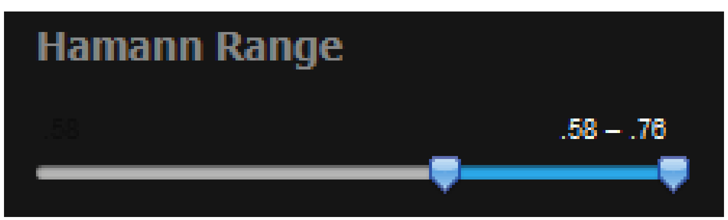
Filter Trees based on Hamann Similarity



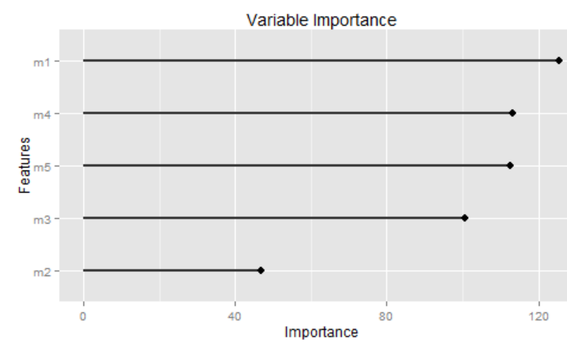


Filter Trees based on Hamann Similarity

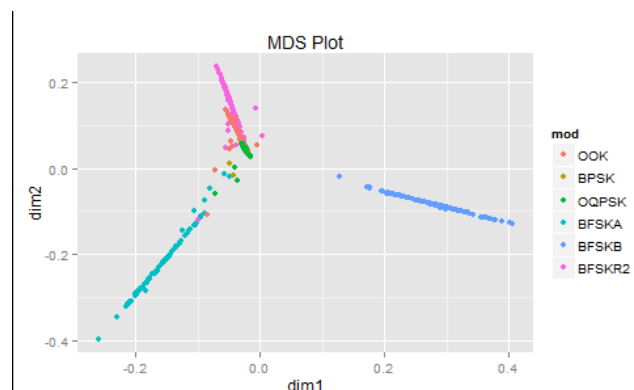




Variable Importance



Multi-dimensional scaling



Thank you

R Shiny App:
https://kenlau177.shinyapps.io/randomForestApp/