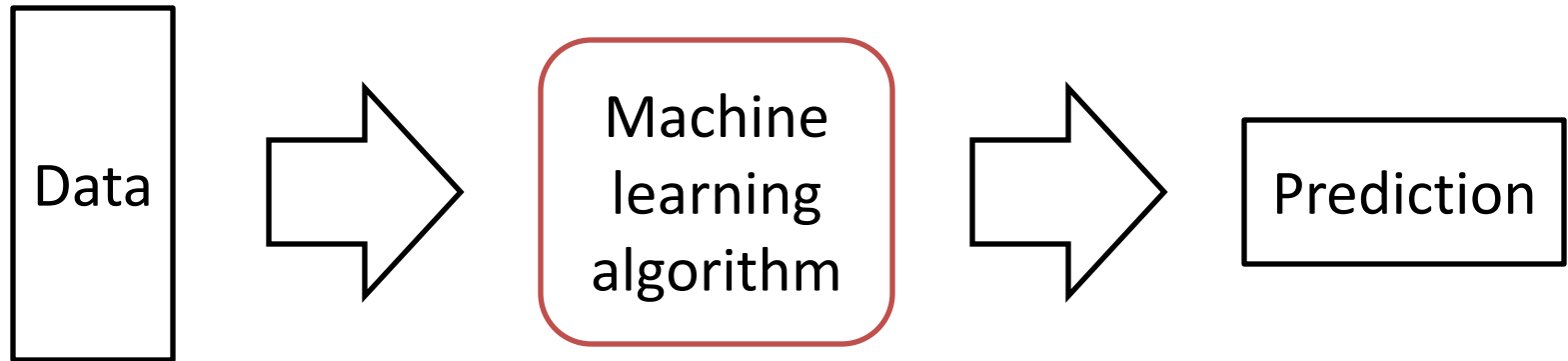


Random Forest Ensemble Visualization

CPSC 547 Project

Ken Lau


Prediction



Weather Data Example

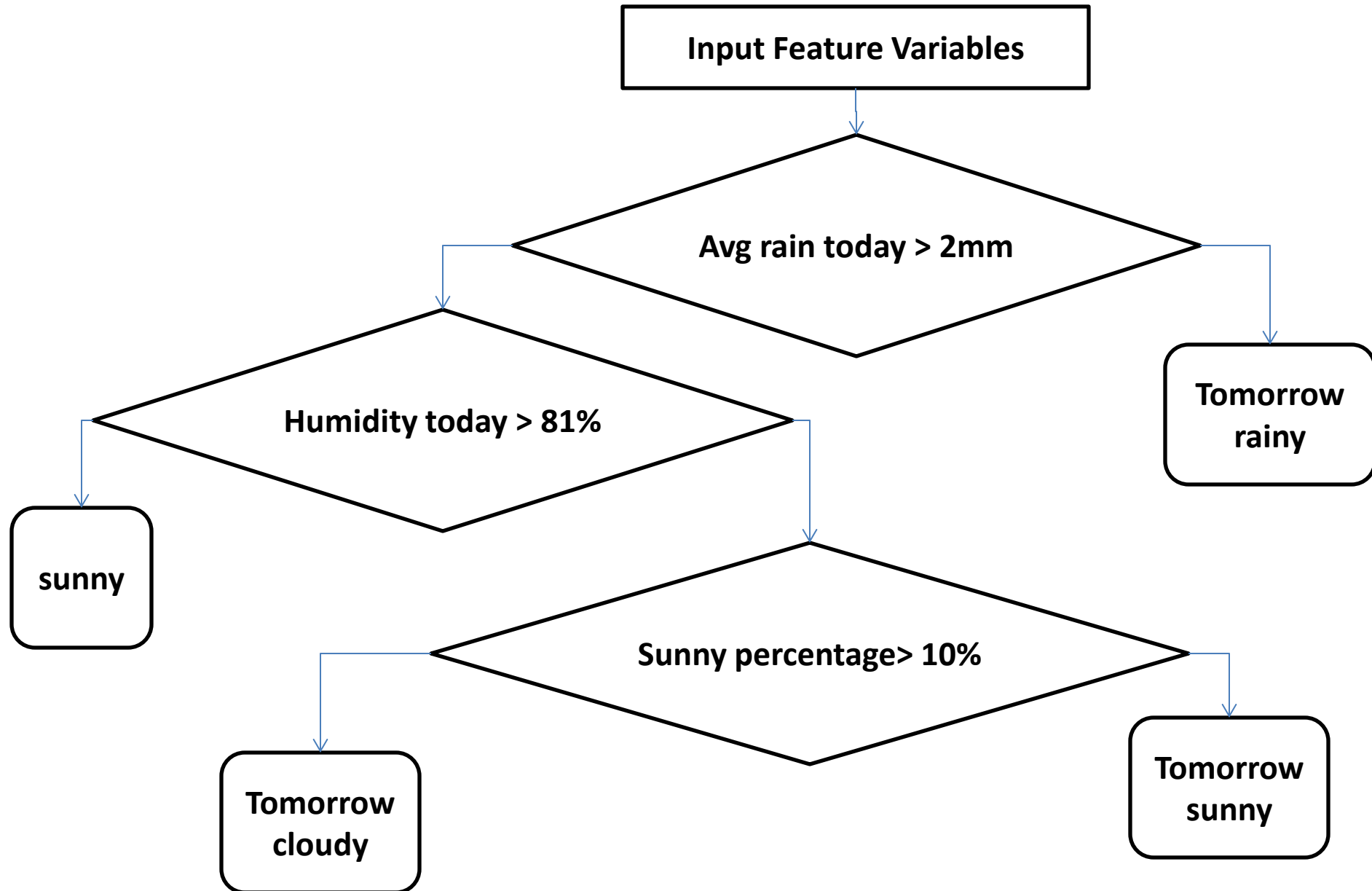
Class Prediction
Variables

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

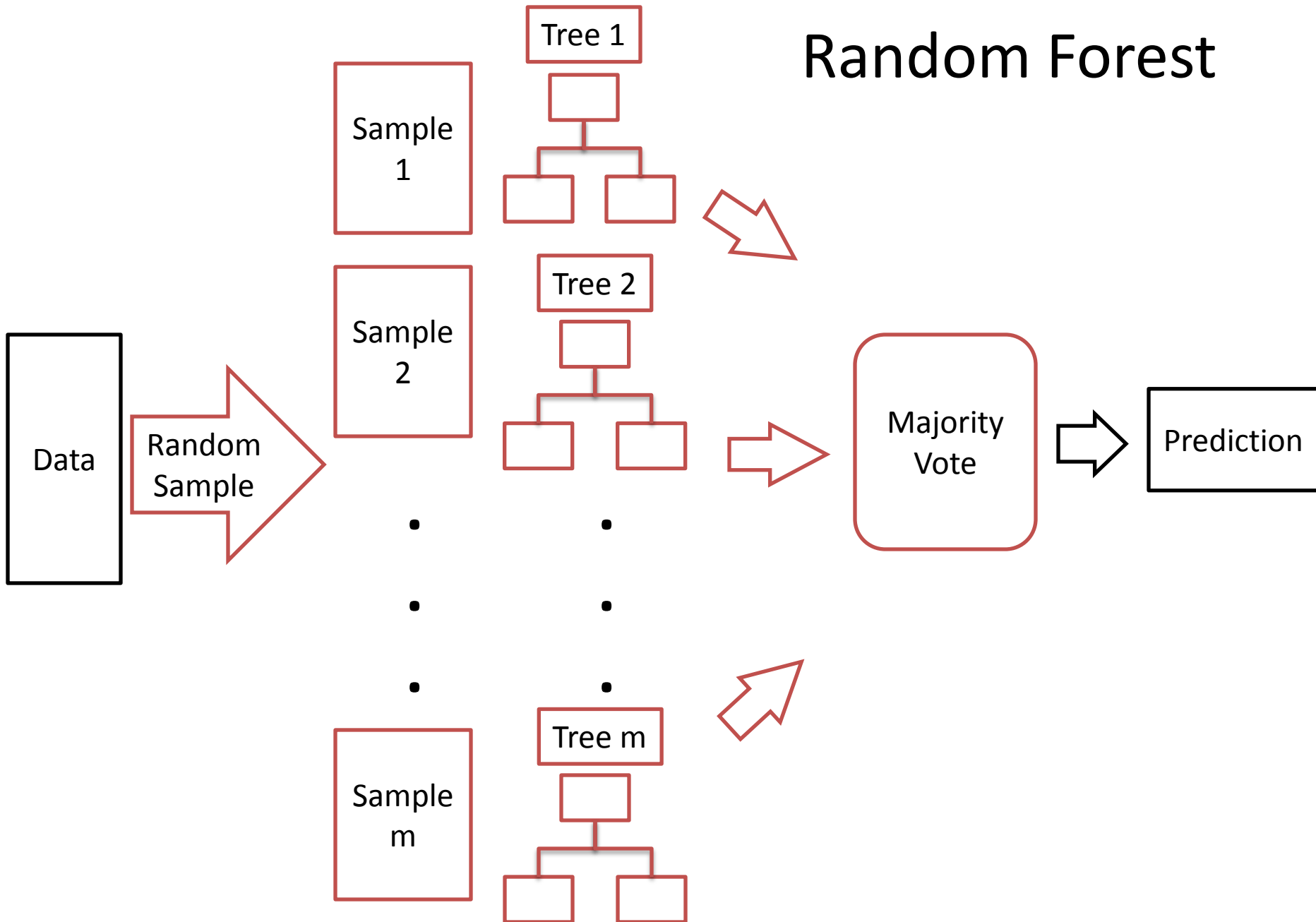
Classification Tree



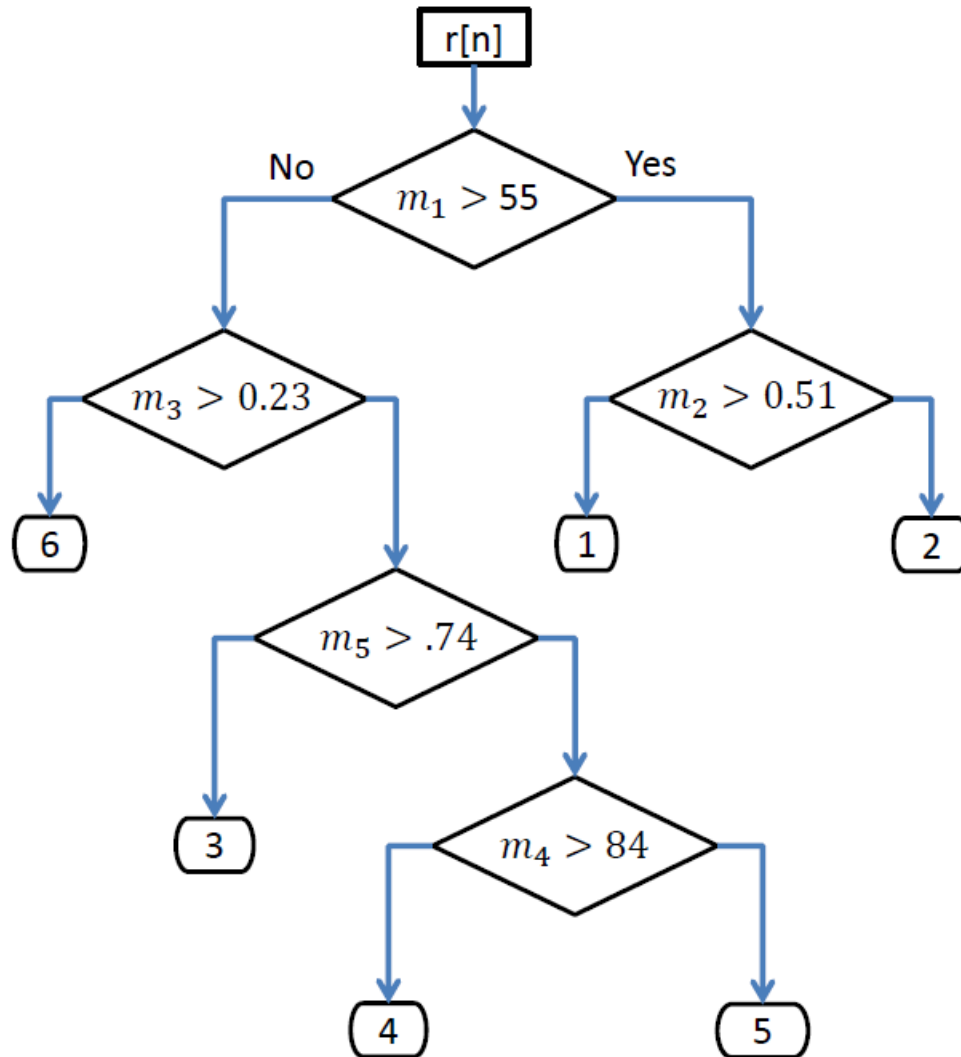
Random Forest

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

Random Forest



My Data

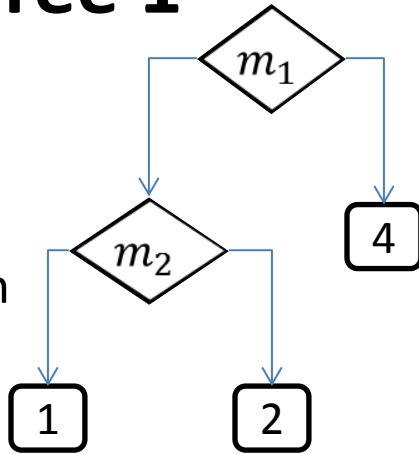


Problem:

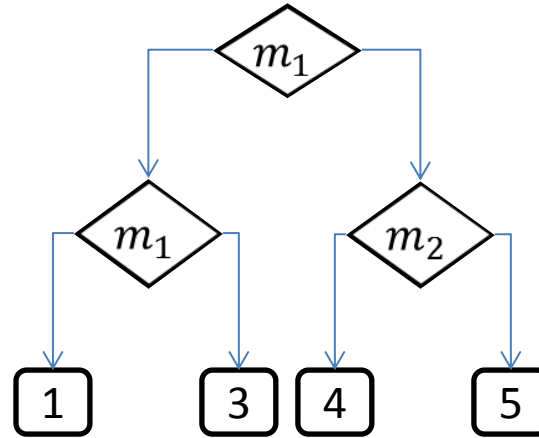
How to visualize the collection of
Classification trees

Aggregate: Features Variables

Tree 1



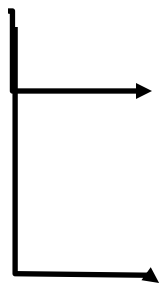
Tree 2



Encode
Colour Saturation

Depth	Feature	Appearance
root	m1	2

Data Derive

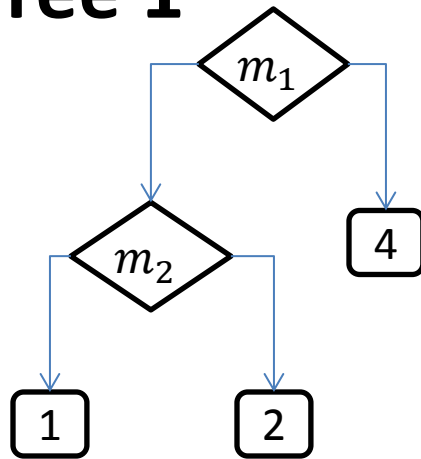


Depth	Parent	Feature	Left Split	Right Split
2	m1	m2	1	1

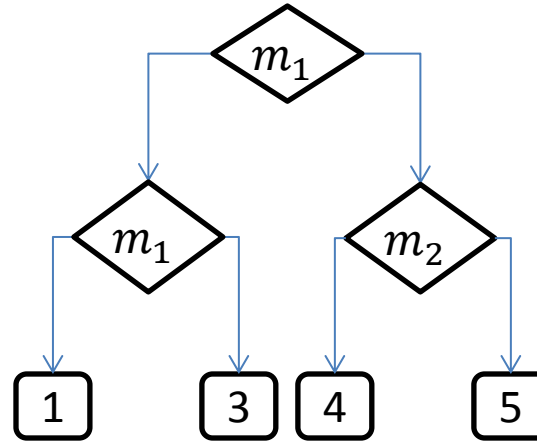
Depth	Parent	Feature	Left Split	Right Split
2	m1	m1	1	0

Aggregate: Class Prediction Variables

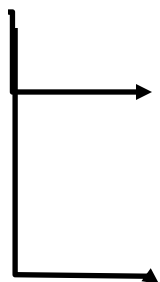
Tree 1



Tree 2



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

start										
m5	root: 172		class-1: 8608	class-2: 8638	class-3: 8535	class-4: 8675	class-5: 8455	class-6: 8689		
m1	left-split: 56	right-split: 60	class-1: 2983	class-2: 3013	class-3: 2810	class-4: 2897	class-5: 3083	class-6: 2845		
m3	left-split: 62	right-split: 18	class-1: 926	class-2: 960	class-3: 2388	class-4: 3052	class-5: 1424	class-6: 3047		
m4	left-split: 9	right-split: 61	class-1: 3057	class-2: 3022	class-3: 1237	class-4: 449	class-5: 2058	class-6: 466		
m2	left-split: 12	right-split: 32	class-1: 1618	class-2: 1597	class-3: 950	class-4: 561	class-5: 1478	class-6: 599		
m5	left-split: 33	right-split: 1	class-1: 44	class-2: 46	class-3: 1150	class-4: 1716	class-5: 412	class-6: 1732		
m1	root: 169		class-1: 8425	class-2: 8442	class-3: 8493	class-4: 8438	class-5: 8444	class-6: 8458		
m4	root: 83		class-1: 4198	class-2: 4362	class-3: 4043	class-4: 4064	class-5: 4066	class-6: 4167		
m3	root: 72		class-1: 3530	class-2: 3436	class-3: 3803	class-4: 3566	class-5: 3670	class-6: 3595		
m2	root: 4		class-1: 216	class-2: 173	class-3: 177	class-4: 191	class-5: 242	class-6: 201		

- 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

Predicted **Different** Class

Tree 2

Tree 1

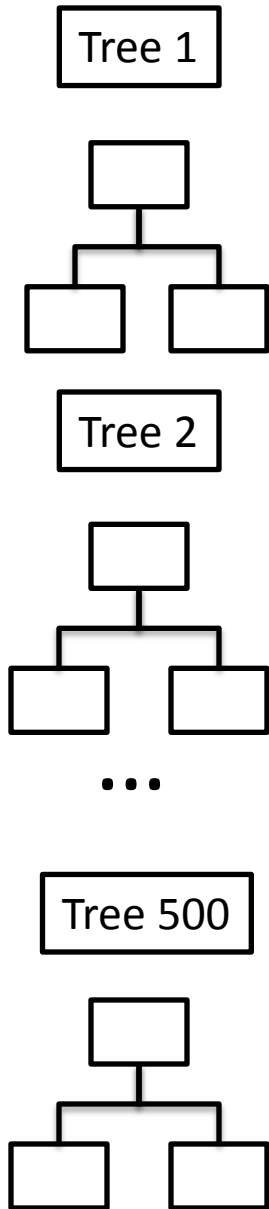
		Correct	Incorrect
Correct	a1	0	
Incorrect	0	d1	

Tree 2

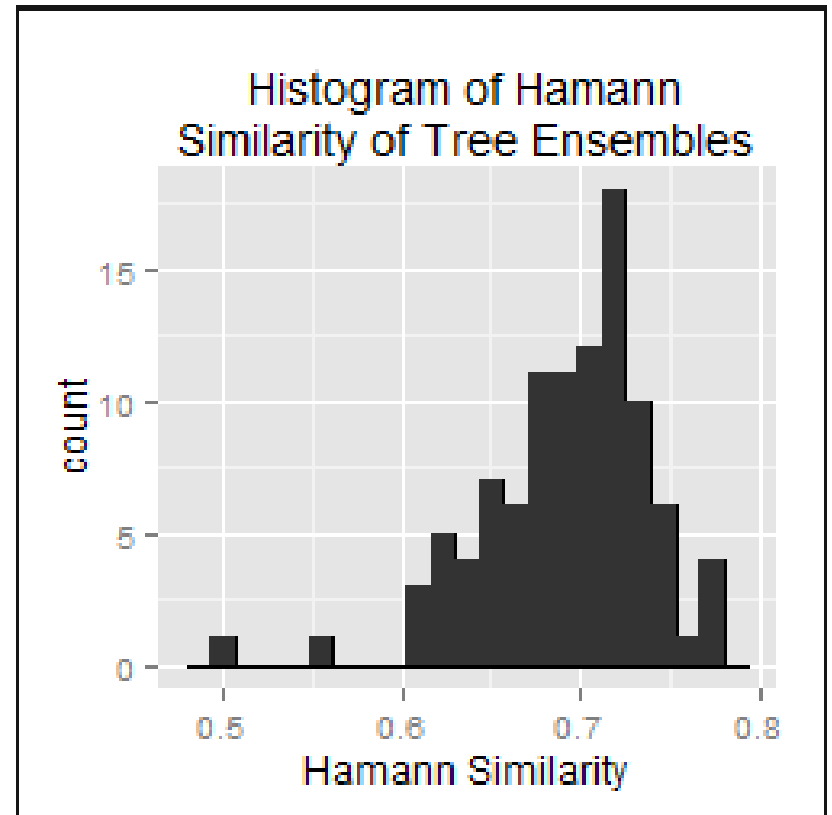
Tree 1

		Correct	Incorrect
Correct	0	b2	
Incorrect	c2	d2	

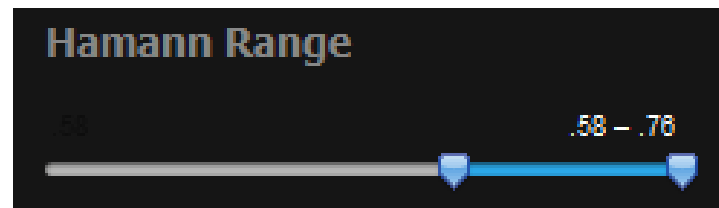
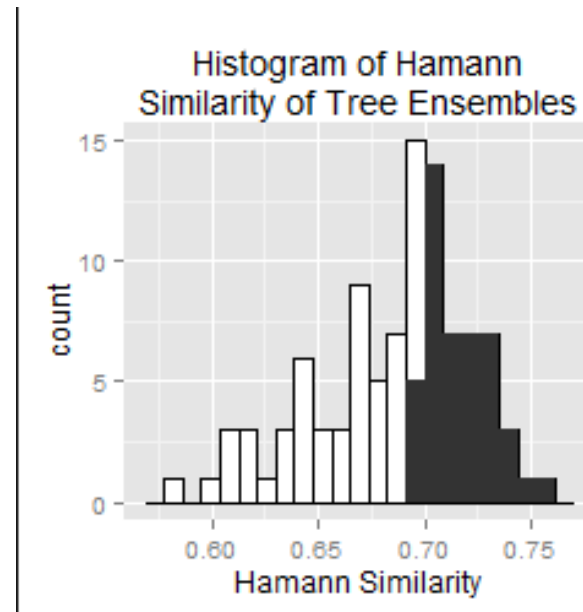
$$H = \frac{(a_1 + d_1) - (b_2 + c_2 + d_2)}{a_1 + d_1 + b_2 + c_2 + d_2}$$



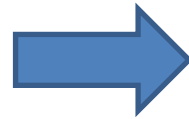
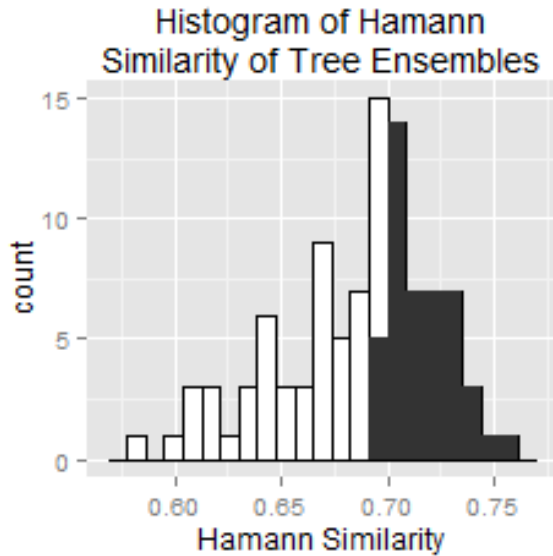
Derived Data



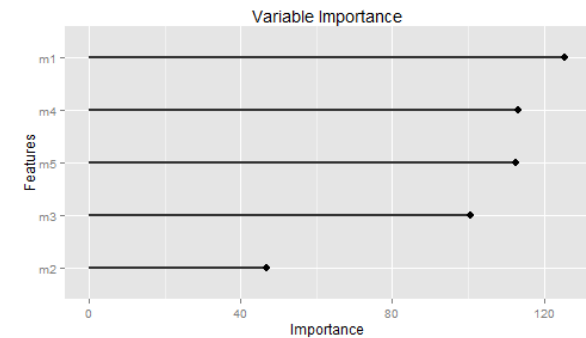
Filter Trees based on Hamann Similarity



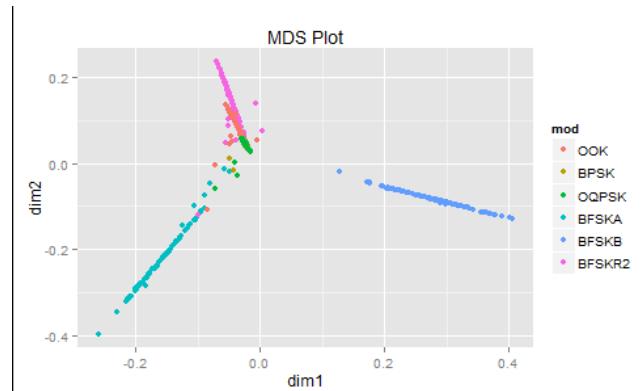
Filter Trees based on Hamann Similarity



Variable Importance

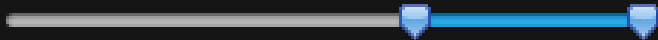


Multi-dimensional scaling



Hamann Range

.58 - .76



Thank you

R Shiny App:

<https://kenlau177.shinyapps.io/randomForestApp/>