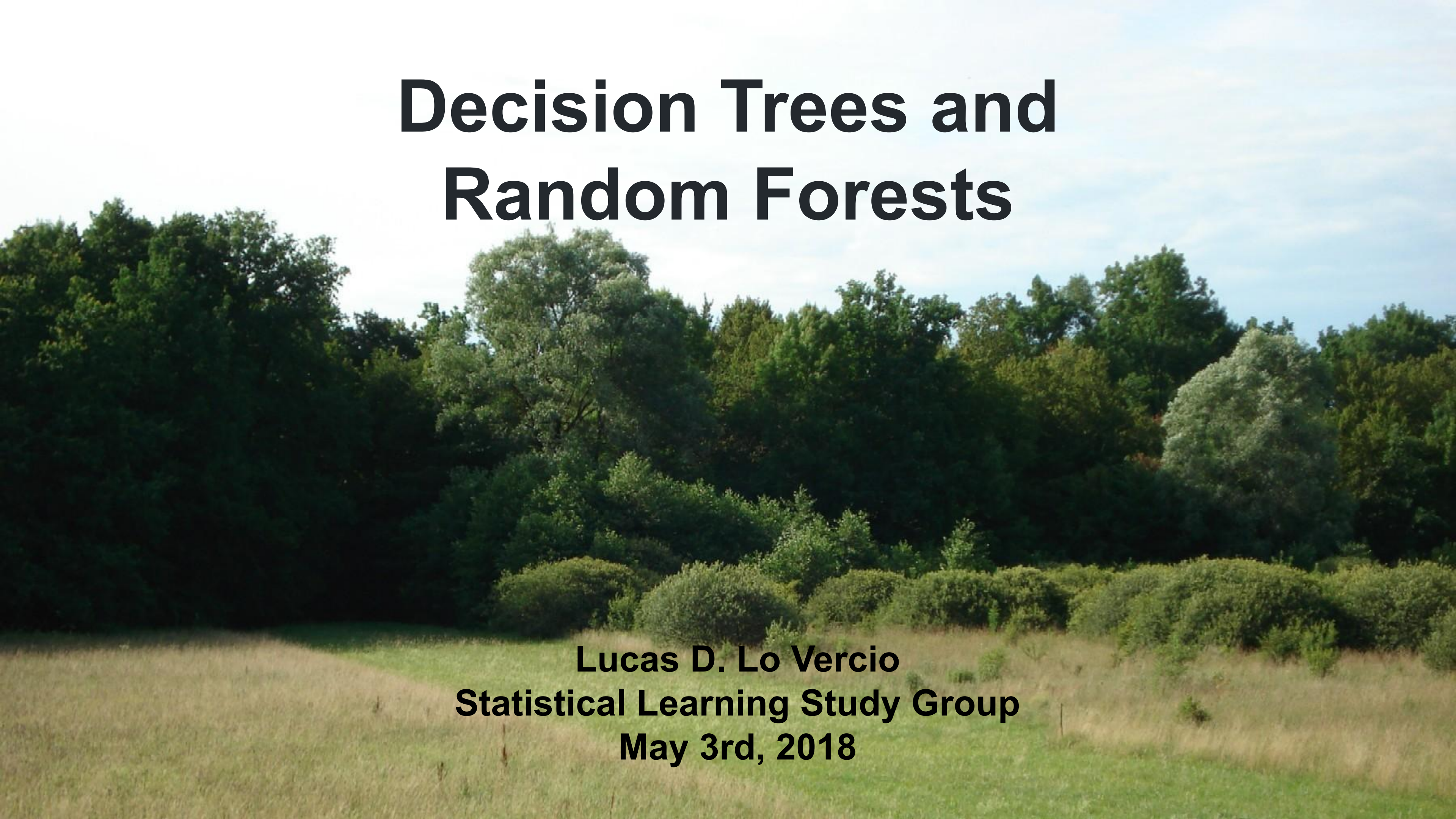
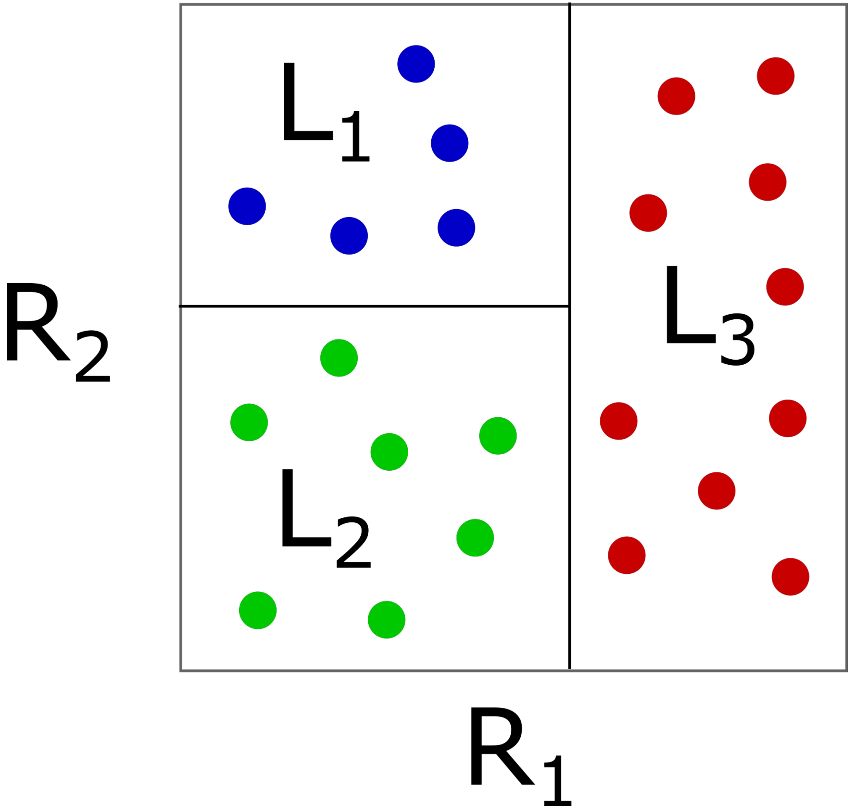
Decision Trees and Random Forests

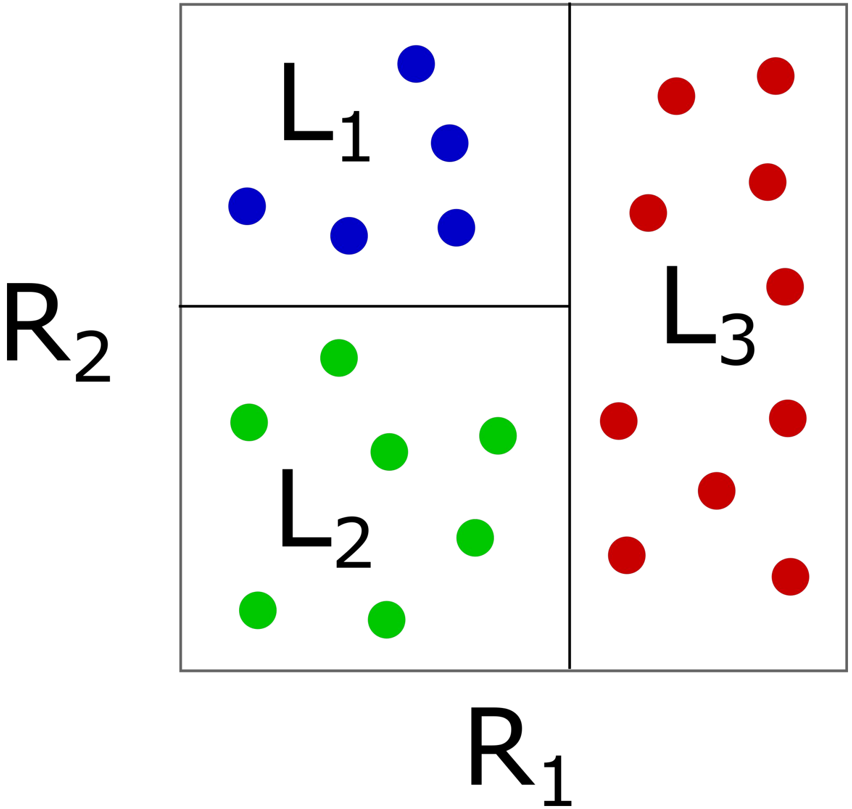


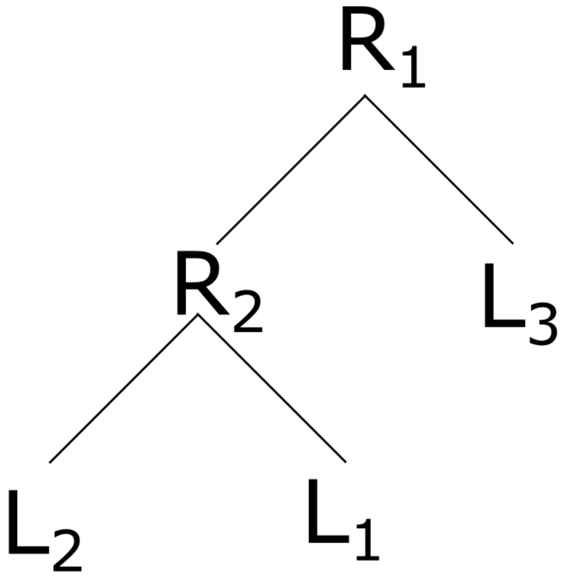
Lucas D. Lo Vercio
Statistical Learning Study Group
May 3rd, 2018

Motivation

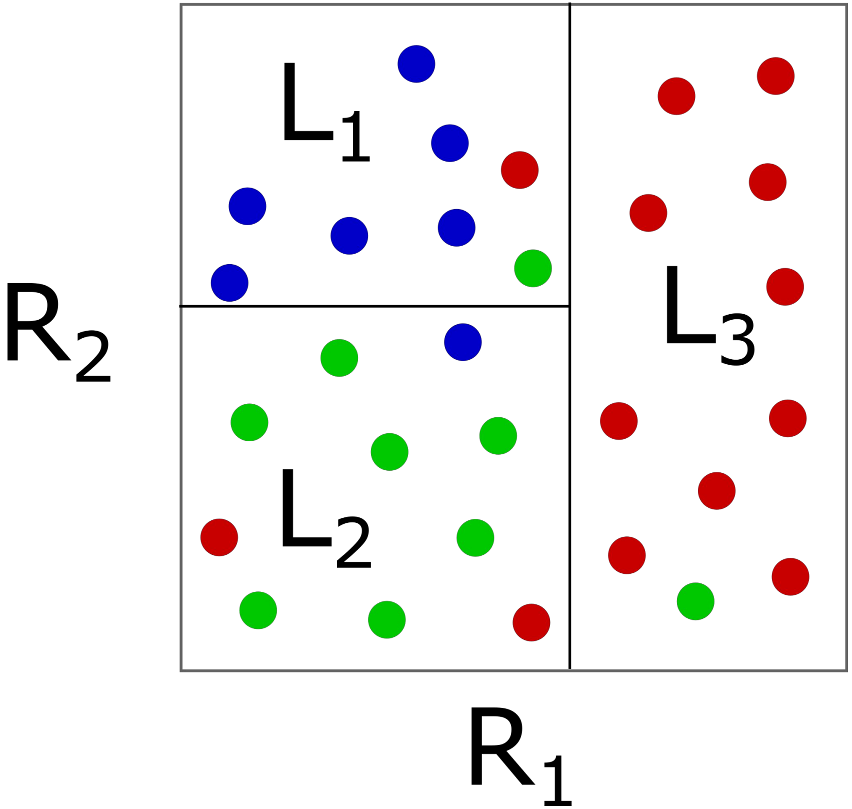


Motivation





Motivation



?

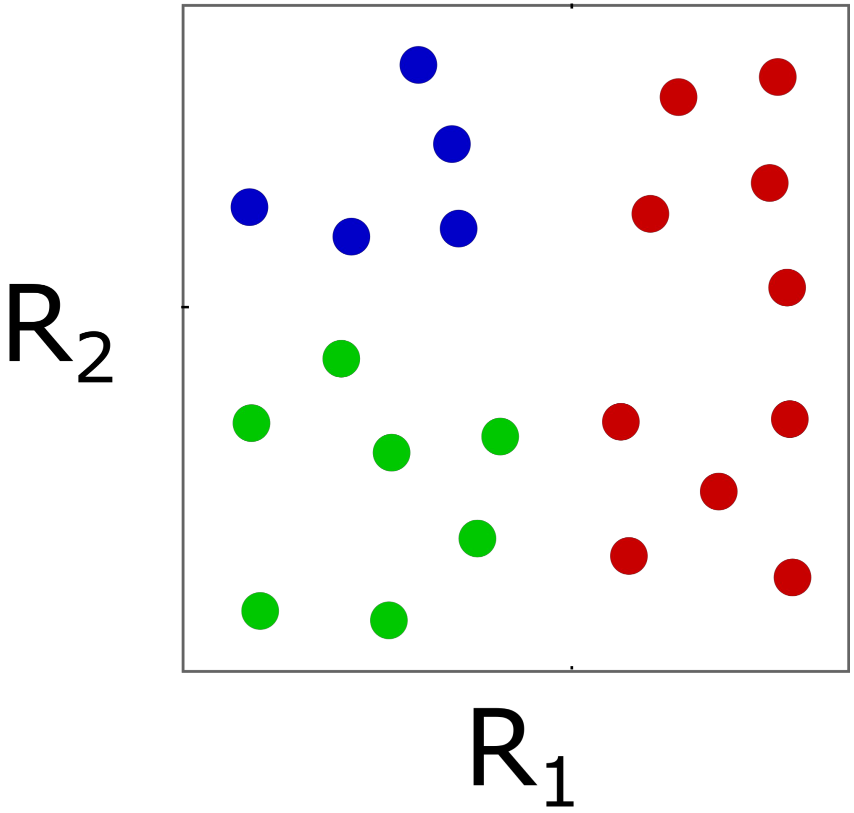
Agenda

- Decision tree
  - Construction
  - Parameters
- Random forest
  - Construction
  - Application examples

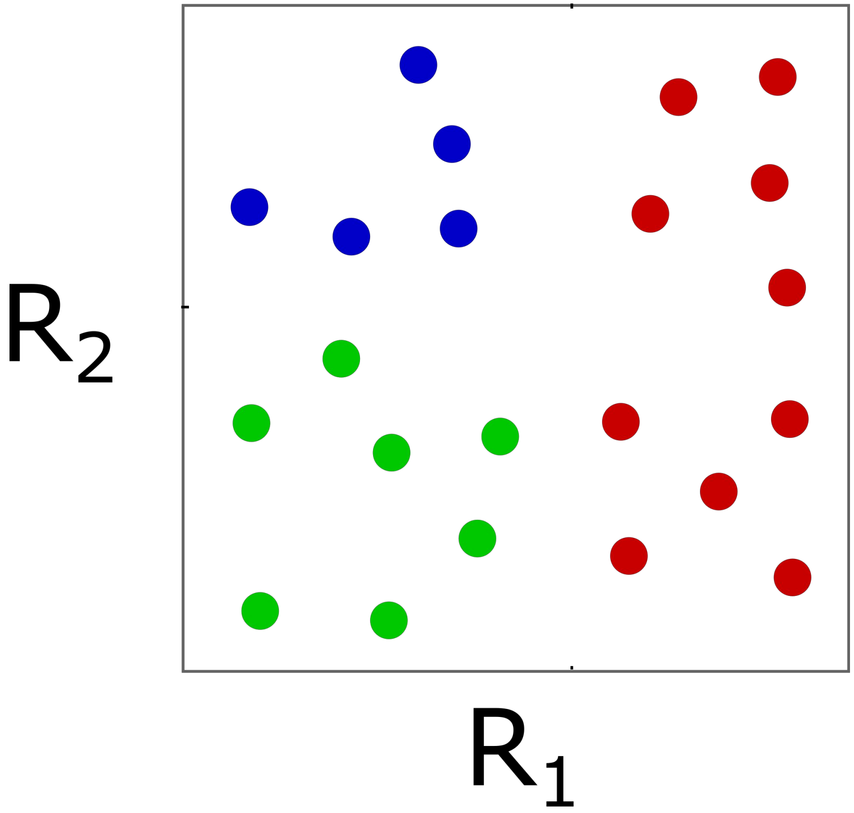
Agenda

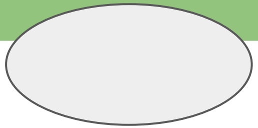
- **Decision tree**
  - **Construction**
  - **Parameters**
- Random forest
  - Construction
  - Application examples

Decision tree - Construction

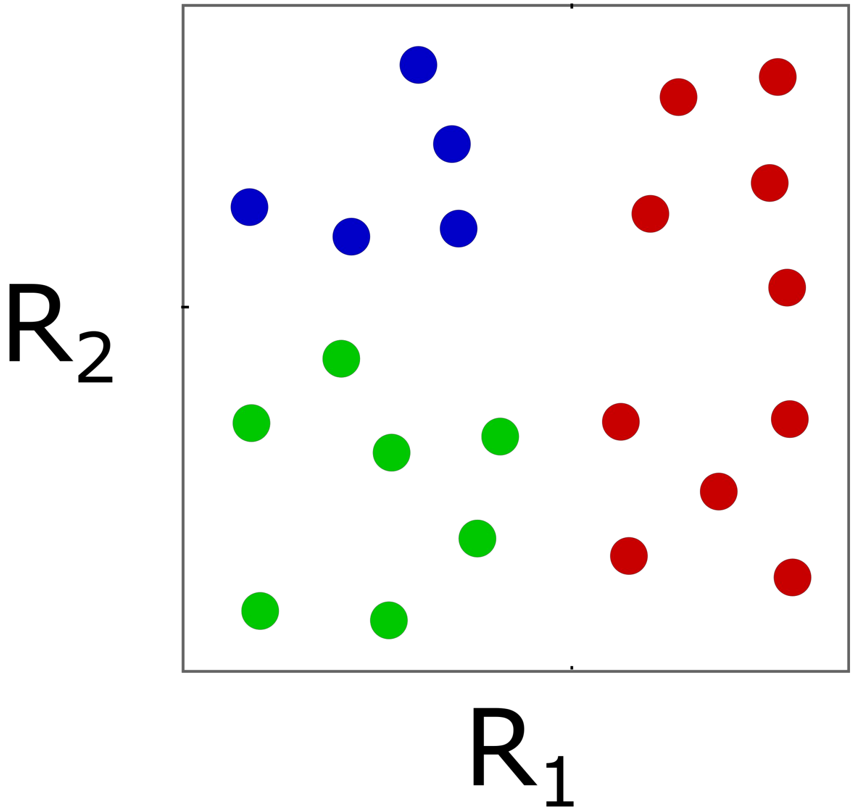


Decision tree - Construction





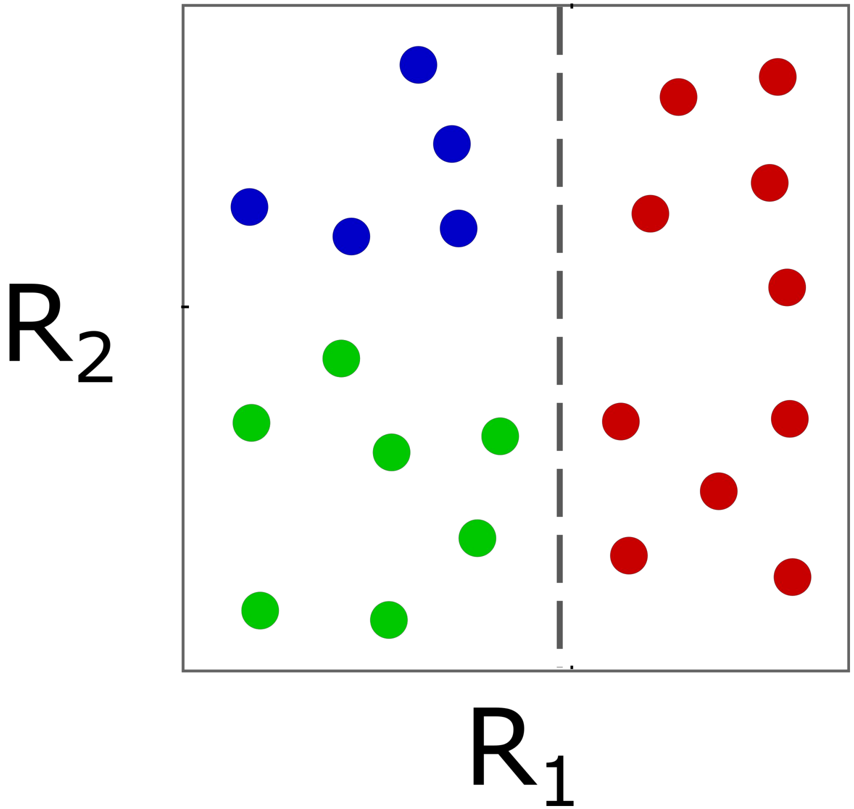
Decision tree - Construction

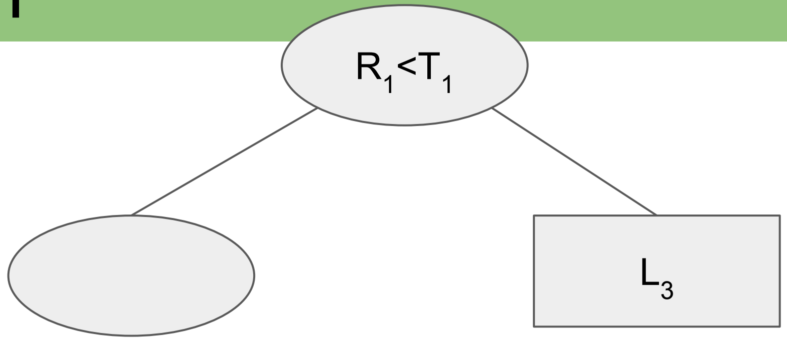


Impurity measures:

- Entropy
- Gini index
- Misclassification error

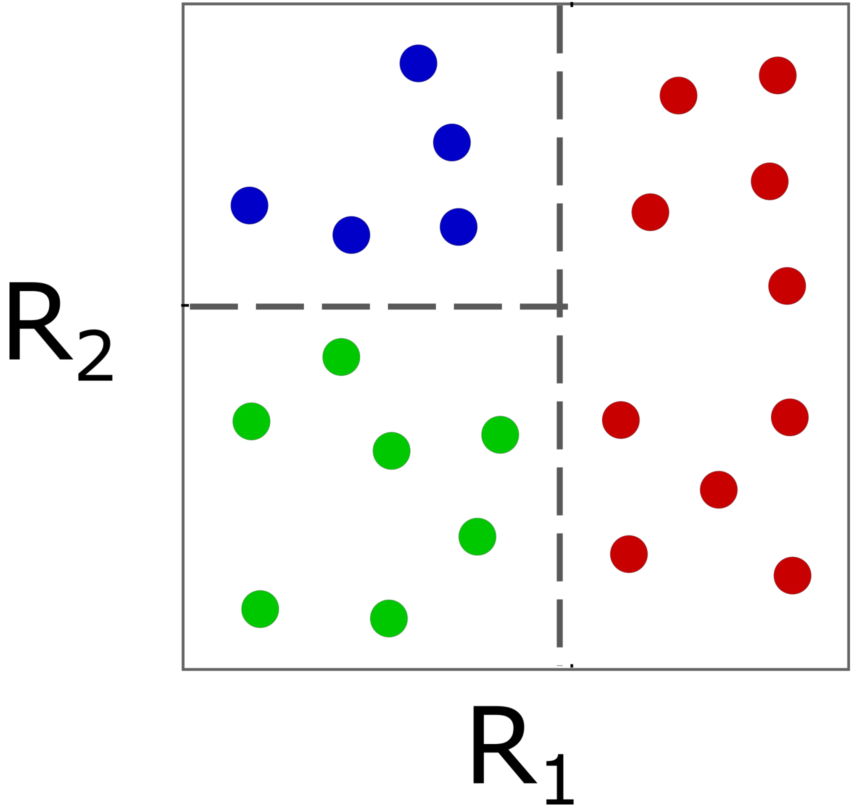
Decision tree - Construction

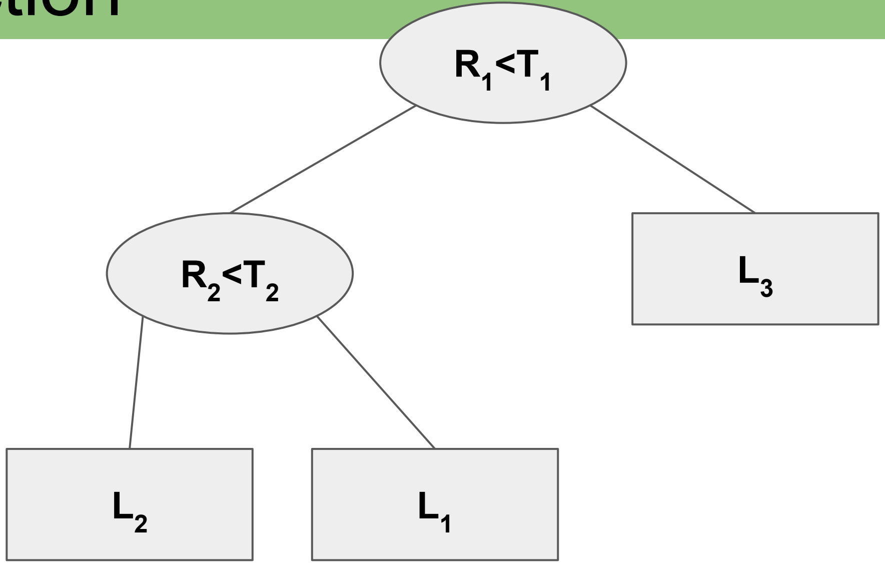



$R_1<T_1$

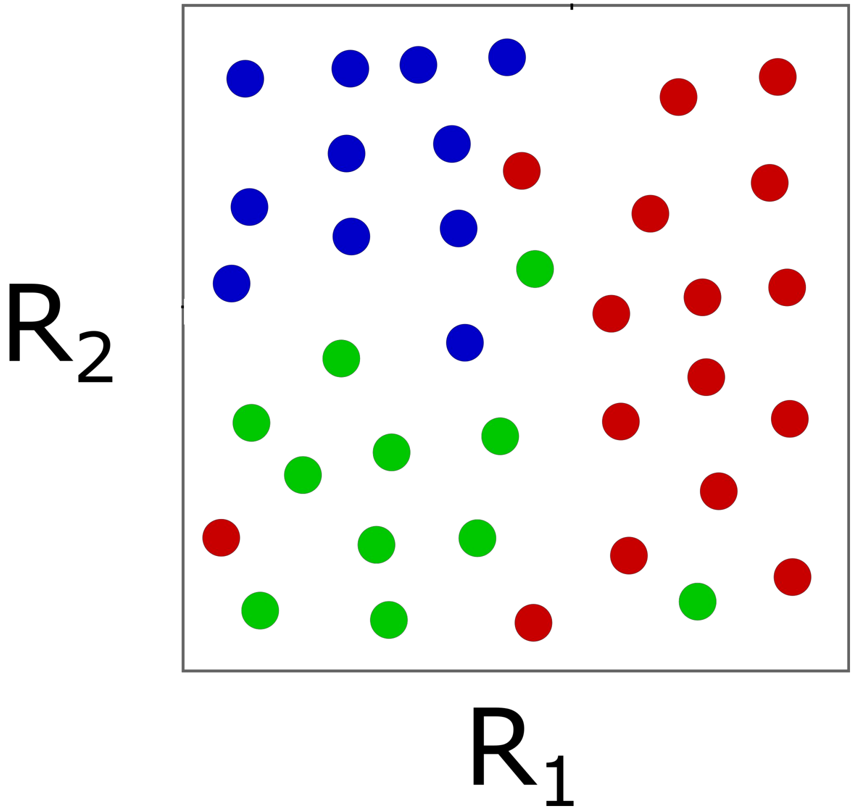
L_3

Decision tree - Construction

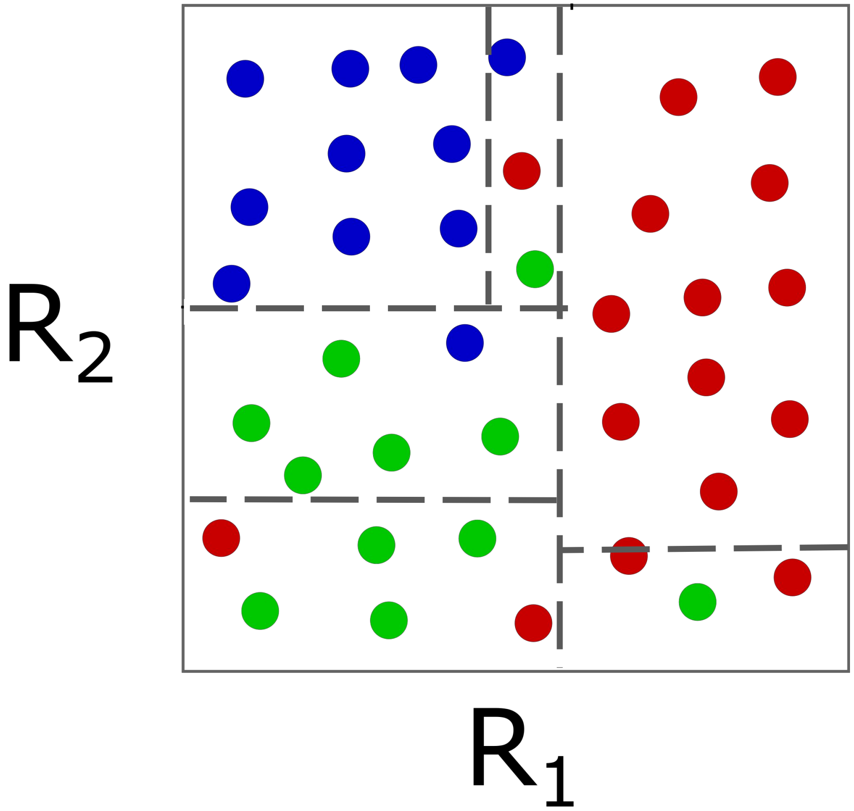


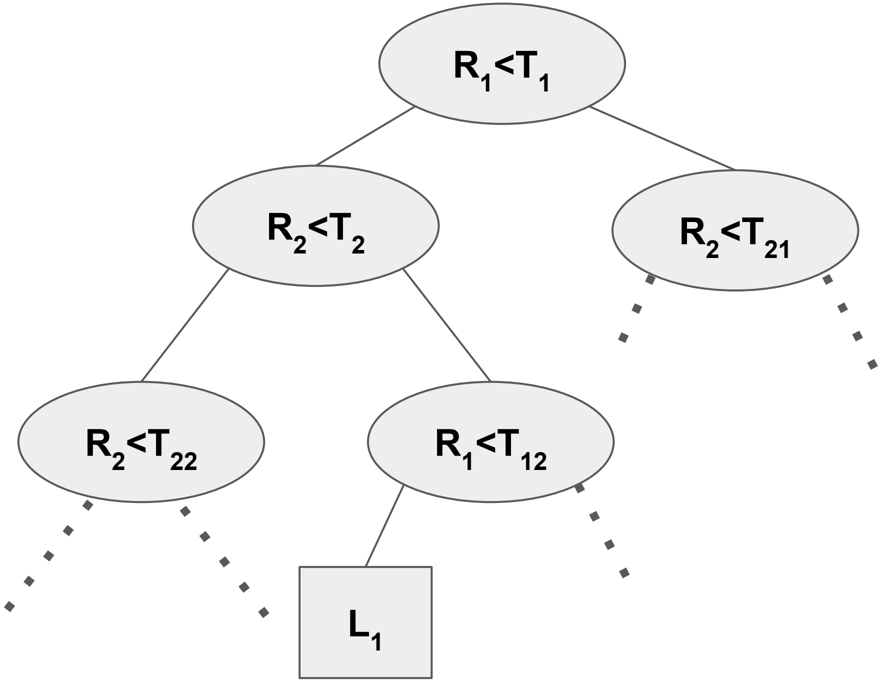


Decision tree - Construction

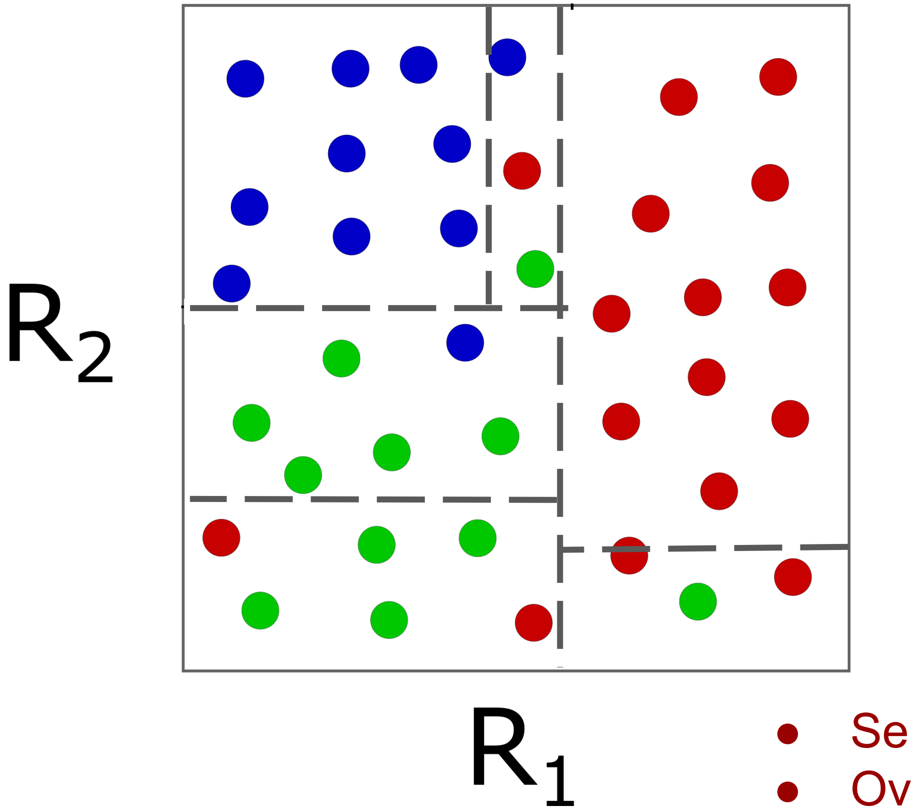


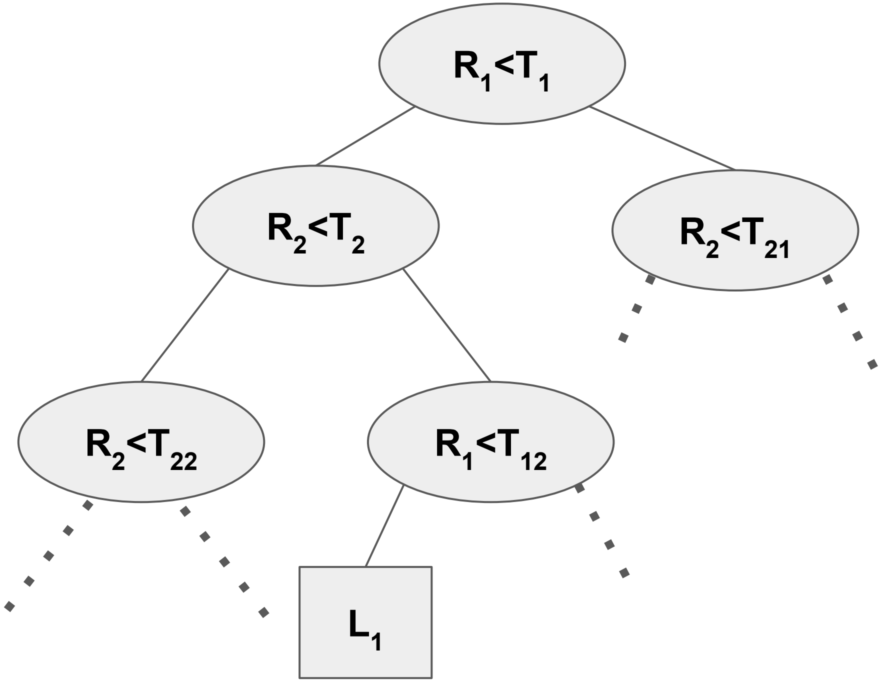
Decision tree - Construction





Decision tree - Construction





- Sensible to noise
- Overfitting

Decision tree - Parameters

How to avoid the overfitting/ensure generalization?

- (Pre-,Post-) Pruning
- Impurity tolerance
  - Entropy
  - Gini index
  - Misclassification error

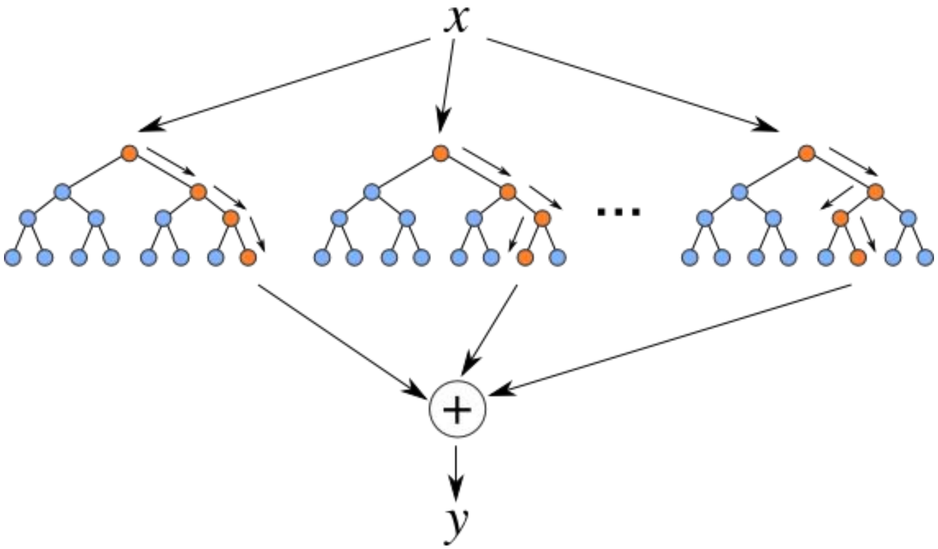
Agenda

- Decision tree
  - Construction
  - Parameters
- **Random forest**
  - **Construction**
  - **Application examples**

Random Forest

A Random Forest (RF) consist in train N **uncorrelated** trees.

A new sample is labelled using the most frequent labelling by the N trees.



Random Forest - Construction

	Feature 1	Feature 2	Feature 3	...	Label
Sample 1					1
Sample 2					2
Sample 3					1
Sample 4					3
Sample 5					1
...					
Sample p					3

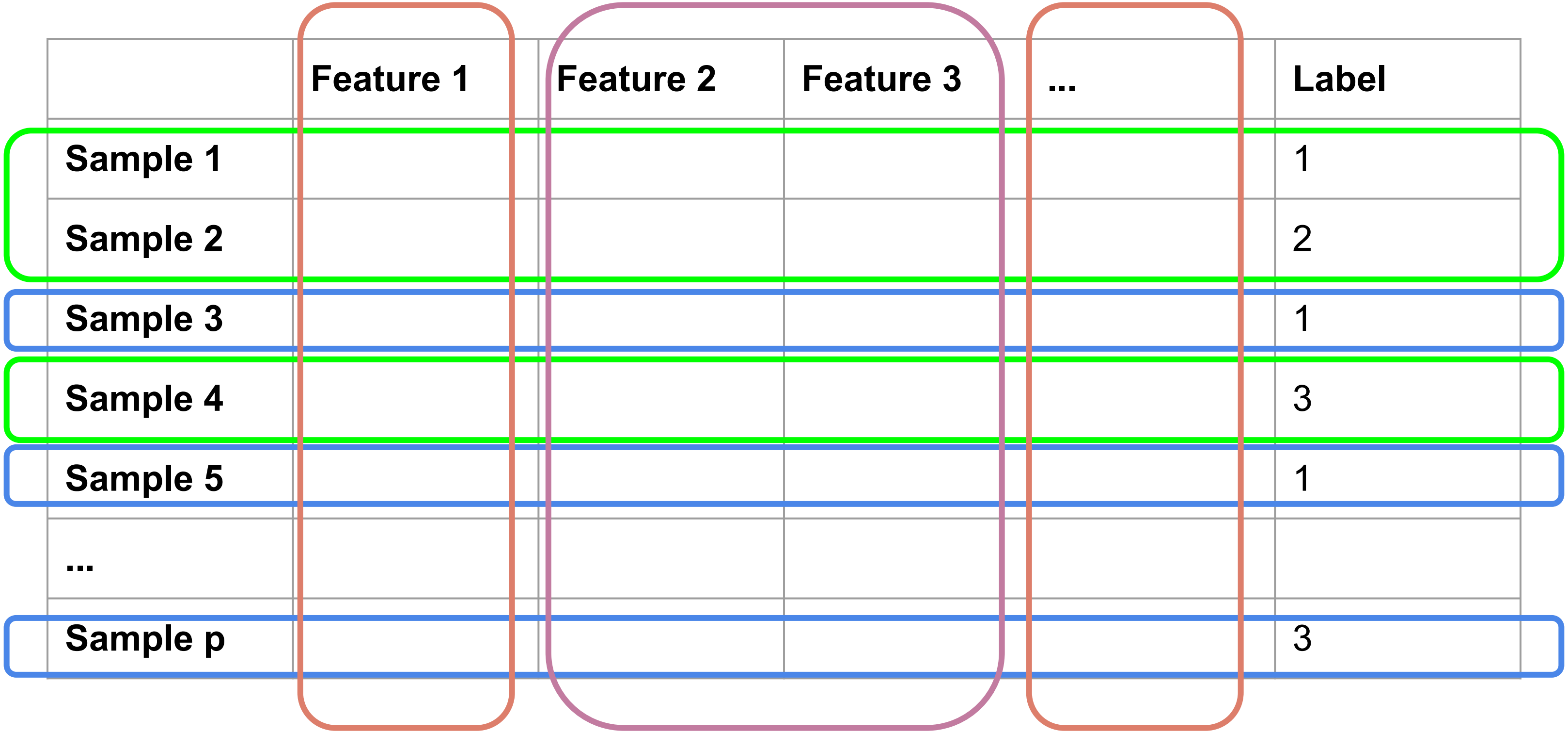
Random Forest - Construction

Bagging

	Feature 1	Feature 2	Feature 3	...	Label
Sample 1					1
Sample 2					2
Sample 3					1
Sample 4					3
Sample 5					1
...					
Sample p					3

Random Forest - Construction

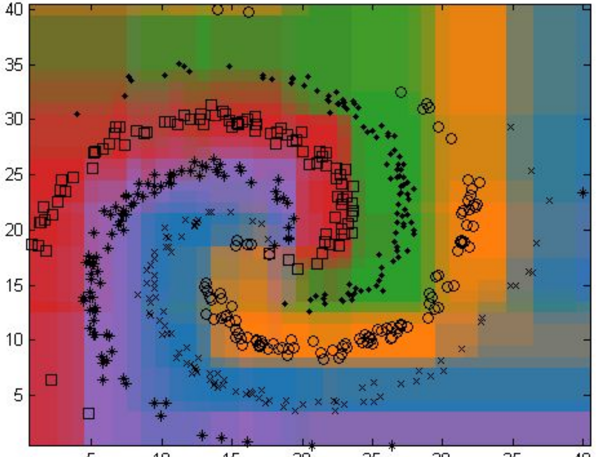
Subset of features to evaluate



	Feature 1	Feature 2	Feature 3	...	Label
Sample 1					1
Sample 2					2
Sample 3					1
Sample 4					3
Sample 5					1
...					
Sample p					3

Random Forest - Parameters

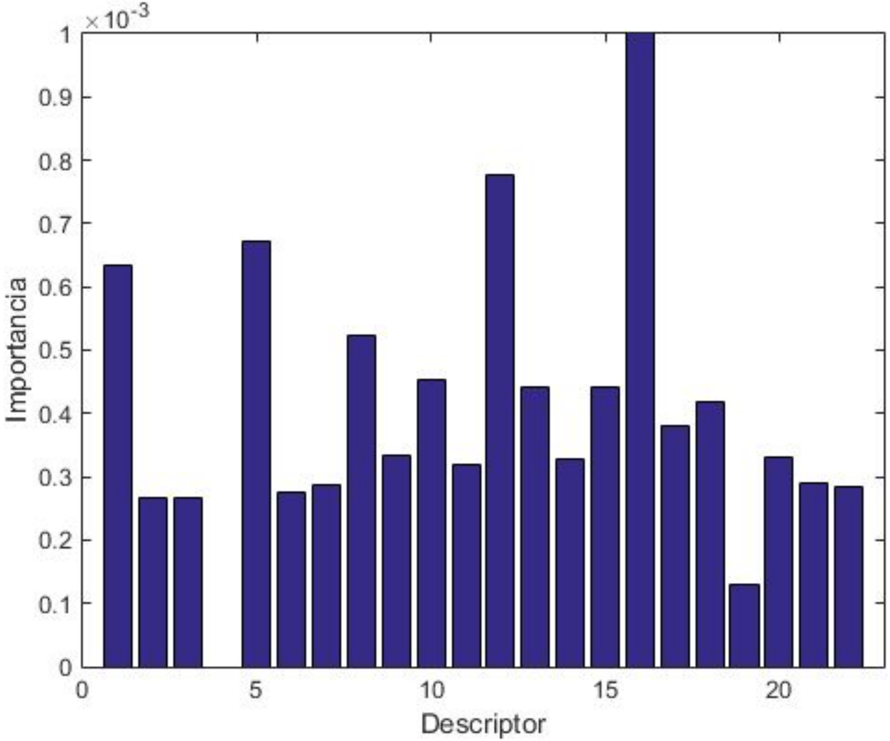
- ~~(Pre-,Post-) Pruning~~
- ~~Impurity tolerance~~
- Number of features to evaluate in each split
- Number of trees



Randomforest Example
by Wasit Limprasert
(Mathworks)

Random Forest - Feature importance

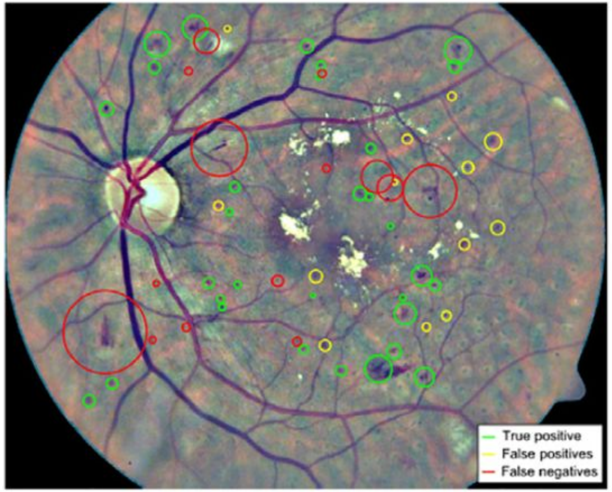
- Number of times a feature is selected for splitting
- Distance to the root when is selected



Application examples

An ensemble deep learning based approach for red lesion detection in fundus images

José Ignacio Orlando[a,b,*], Elena Prokofyeva[d,e], Mariana del Fresno[a,c], Matthew B. Blaschko[f]



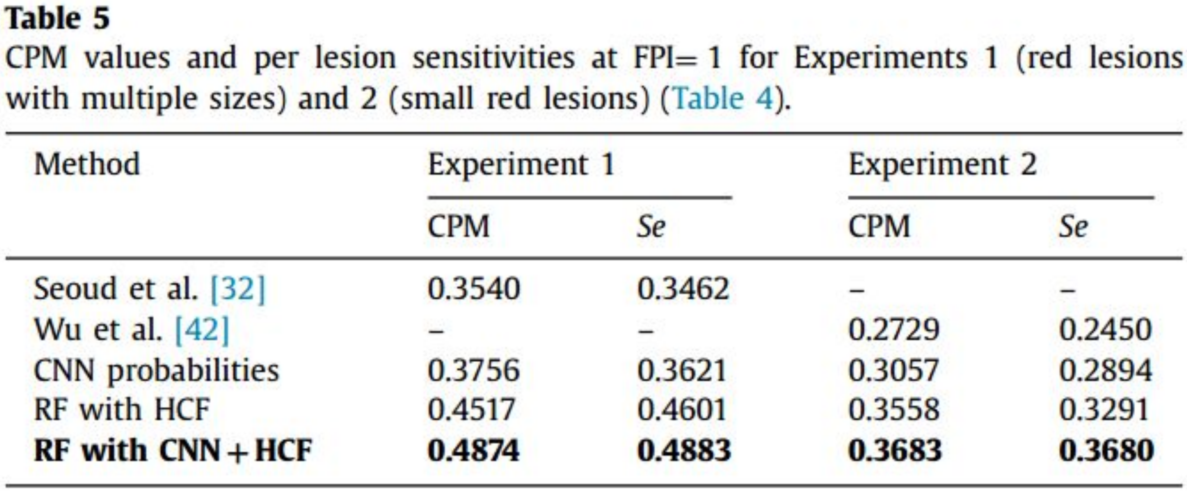
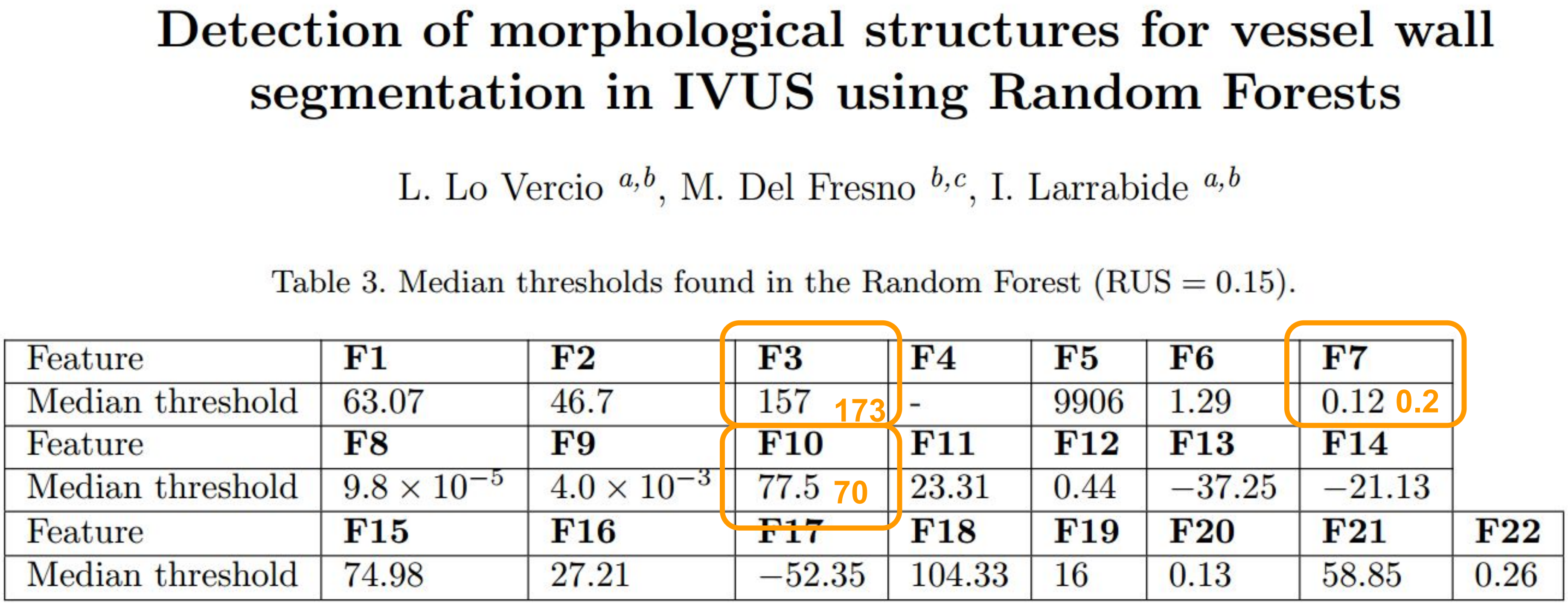


Table 5
CPM values and per lesion sensitivities at FPI= 1 for Experiments 1 (red lesions with multiple sizes) and 2 (small red lesions) (Table 4).

Method	Experiment 1		Experiment 2	
	CPM	*Se*	CPM	*Se*
Seoud et al. [32]	0.3540	0.3462	–	–
Wu et al. [42]	–	–	0.2729	0.2450
CNN probabilities	0.3756	0.3621	0.3057	0.2894
RF with HCF	0.4517	0.4601	0.3558	0.3291
RF with CNN + HCF	**0.4874**	**0.4883**	**0.3683**	**0.3680**

Application examples



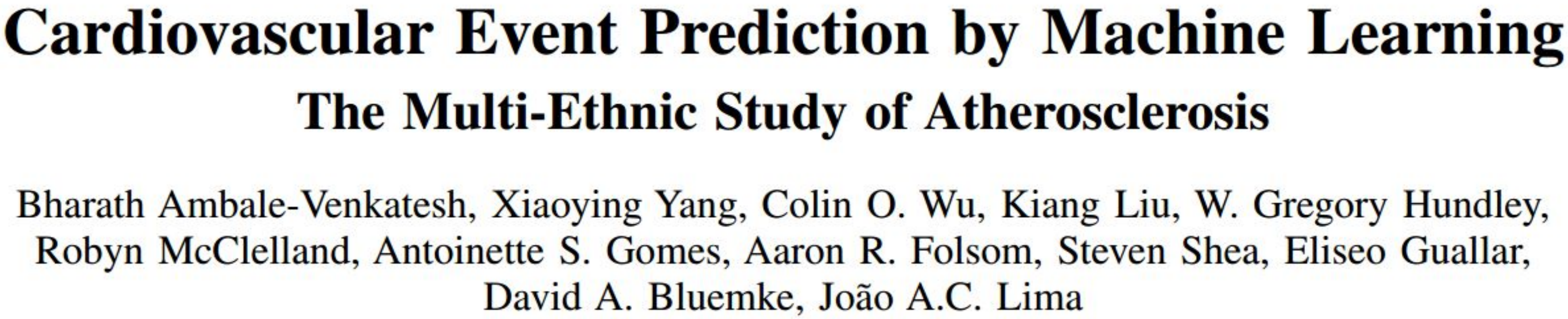
Detection of morphological structures for vessel wall segmentation in IVUS using Random Forests

L. Lo Vercio [a,b], M. Del Fresno [b,c], I. Larrabide [a,b]

Table 3. Median thresholds found in the Random Forest (RUS = 0.15).

Feature	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	
Median threshold	63.07	46.7	157 173	-	9906	1.29	0.12 0.2	
Feature	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	
Median threshold	9.8×10^{-5}	4.0×10^{-3}	77.5 70	23.31	0.44	-37.25	-21.13	
Feature	**F15**	**F16**	**F17**	**F18**	**F19**	**F20**	**F21**	**F22**
Median threshold	74.98	27.21	-52.35	104.33	16	0.13	58.85	0.26

Application examples

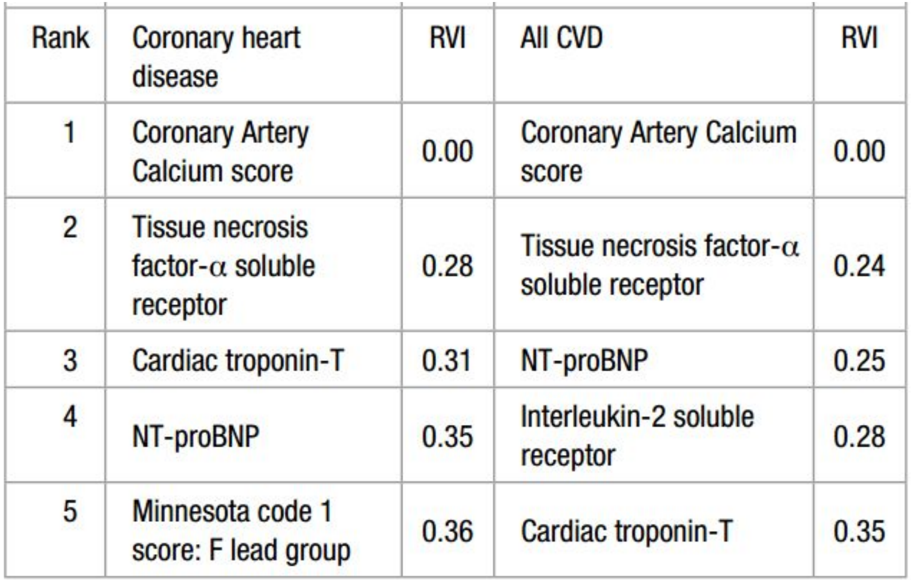


Cardiovascular Event Prediction by Machine Learning

The Multi-Ethnic Study of Atherosclerosis

Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O. Wu, Kiang Liu, W. Gregory Hundley, Robyn McClelland, Antoinette S. Gomes, Aaron R. Folsom, Steven Shea, Eliseo Guallar, David A. Bluemke, João A.C. Lima

Application examples



Rank	Coronary heart disease	RVI	All CVD	RVI
1	Coronary Artery Calcium score	0.00	Coronary Artery Calcium score	0.00
2	Tissue necrosis factor-α soluble receptor	0.28	Tissue necrosis factor-α soluble receptor	0.24
3	Cardiac troponin-T	0.31	NT-proBNP	0.25
4	NT-proBNP	0.35	Interleukin-2 soluble receptor	0.28
5	Minnesota code 1 score: F lead group	0.36	Cardiac troponin-T	0.35

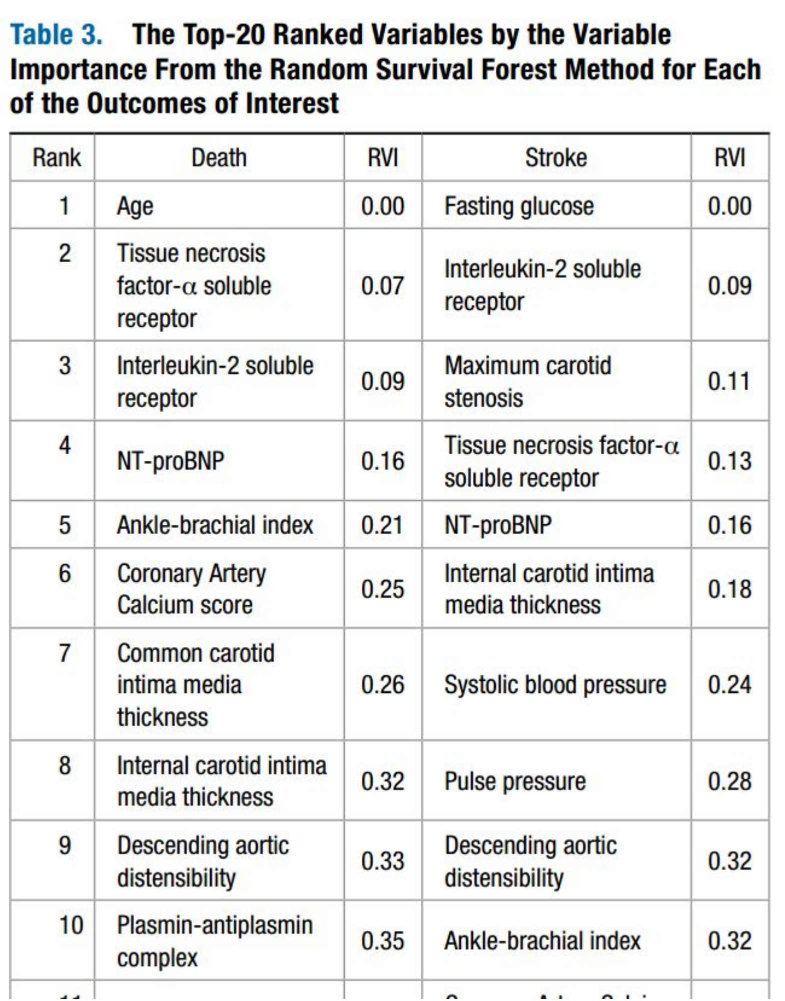
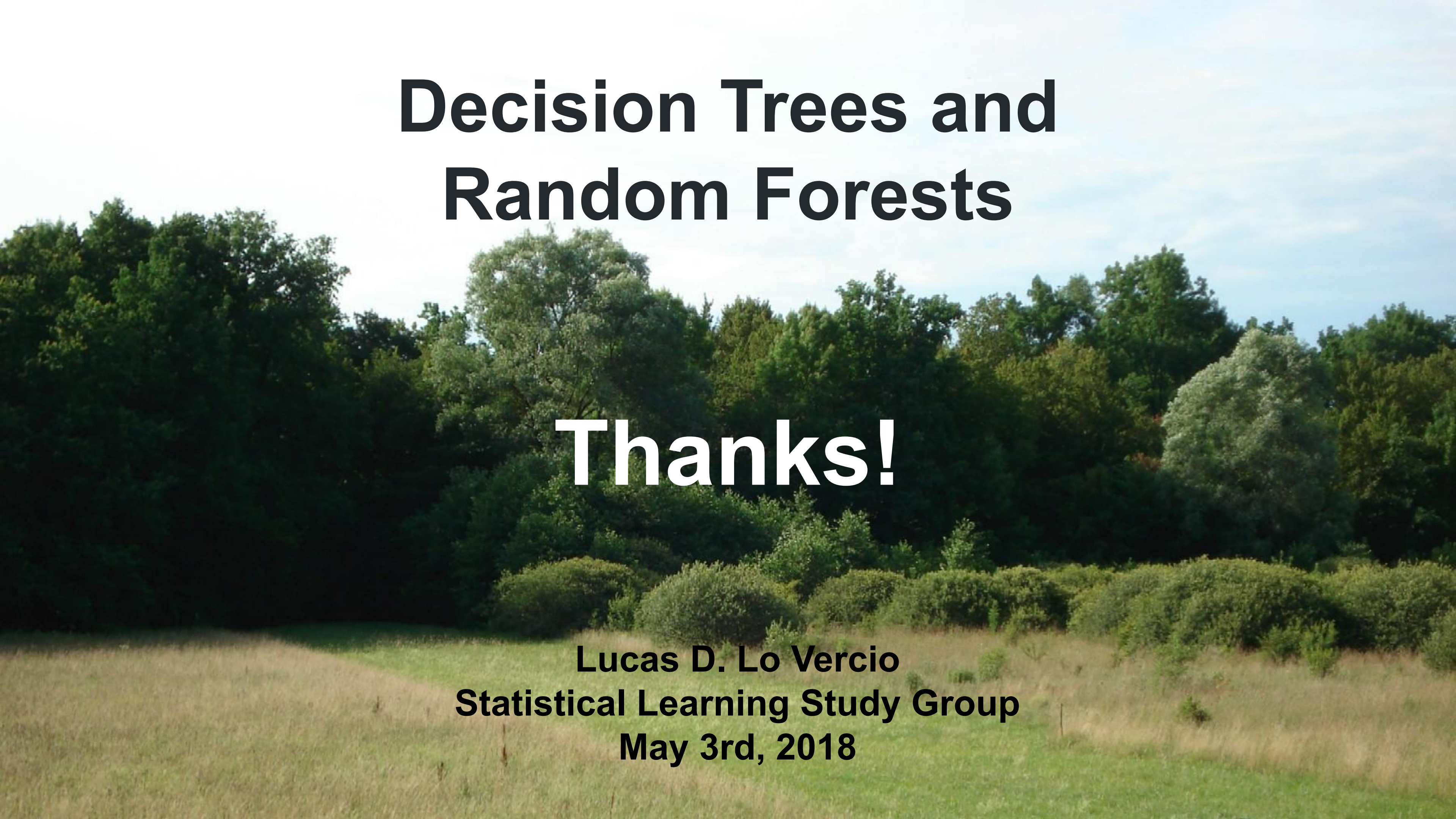


Table 3. The Top-20 Ranked Variables by the Variable Importance From the Random Survival Forest Method for Each of the Outcomes of Interest

Rank	Death	RVI	Stroke	RVI
1	Age	0.00	Fasting glucose	0.00
2	Tissue necrosis factor-α soluble receptor	0.07	Interleukin-2 soluble receptor	0.09
3	Interleukin-2 soluble receptor	0.09	Maximum carotid stenosis	0.11
4	NT-proBNP	0.16	Tissue necrosis factor-α soluble receptor	0.13
5	Ankle-brachial index	0.21	NT-proBNP	0.16
6	Coronary Artery Calcium score	0.25	Internal carotid intima media thickness	0.18
7	Common carotid intima media thickness	0.26	Systolic blood pressure	0.24
8	Internal carotid intima media thickness	0.32	Pulse pressure	0.28
9	Descending aortic distensibility	0.33	Descending aortic distensibility	0.32
10	Plasmin-antiplasmin complex	0.35	Ankle-brachial index	0.32



Decision Trees and Random Forests

Thanks!

Lucas D. Lo Vercio
Statistical Learning Study Group
May 3rd, 2018