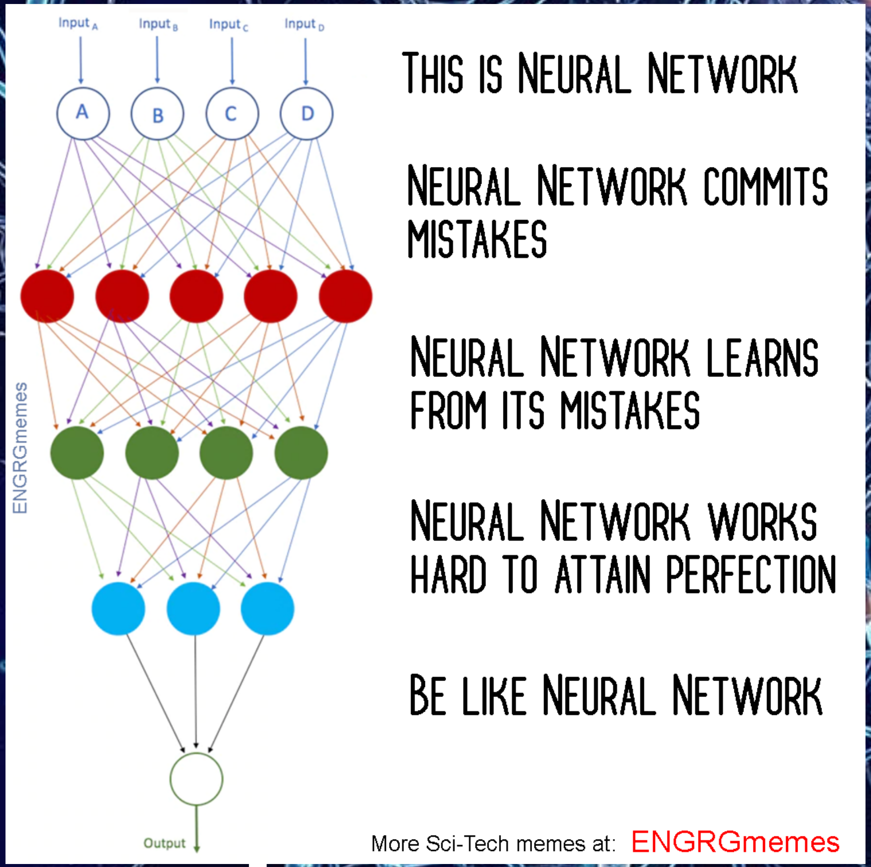
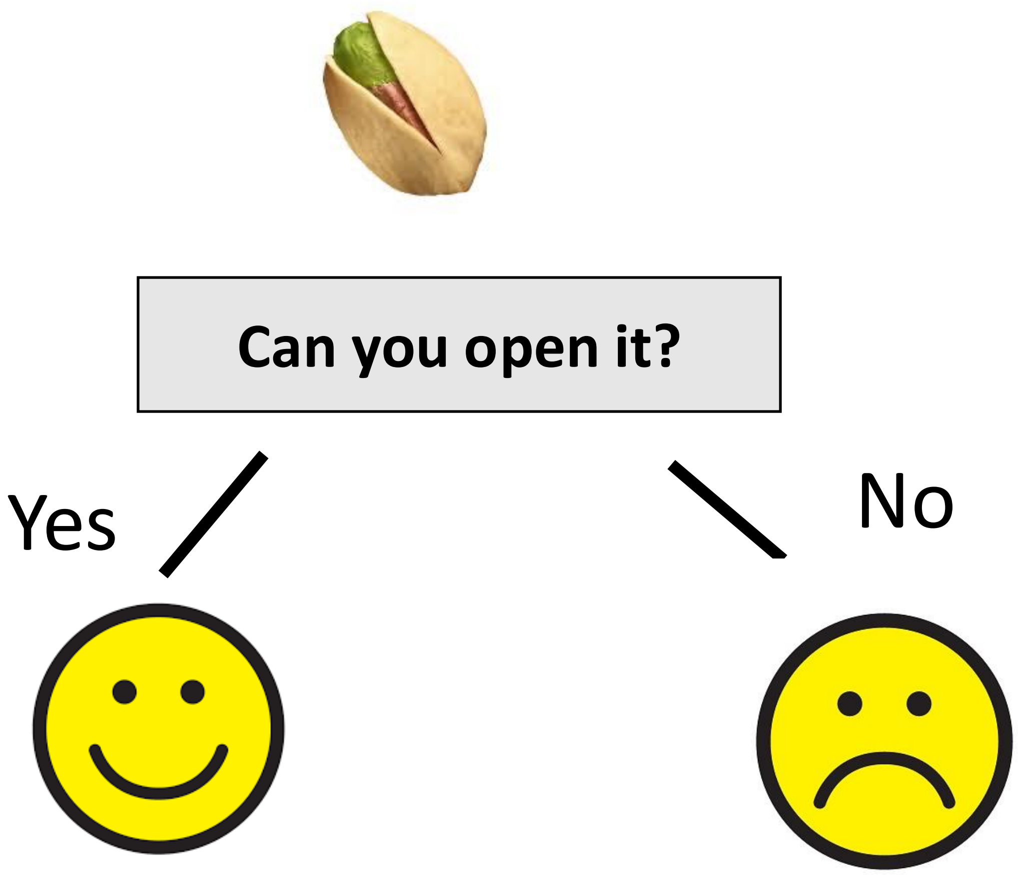
PLS 206 Applied Multivariate Modeling in Agricultural and Environmental Sciences

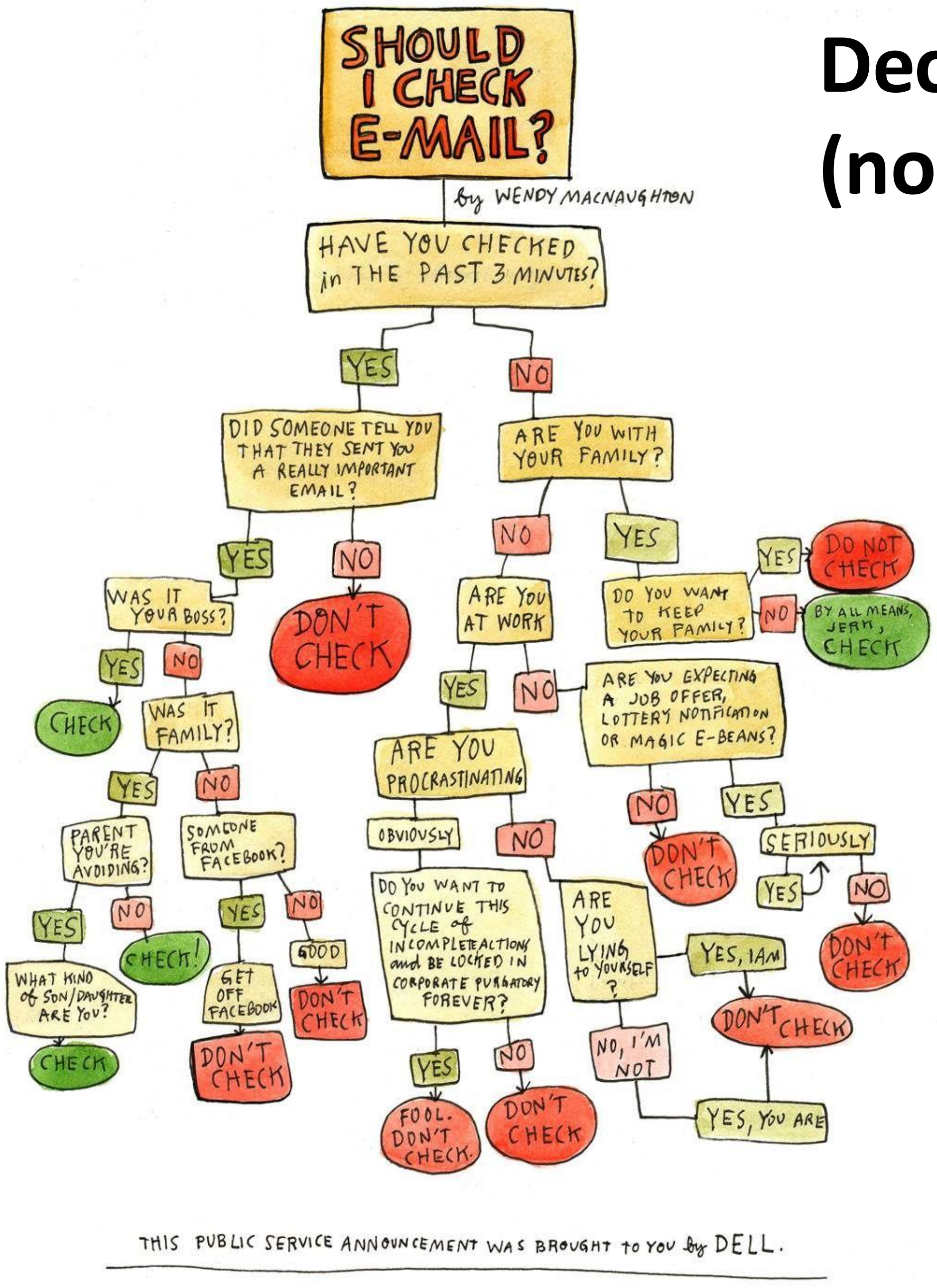
Random forest



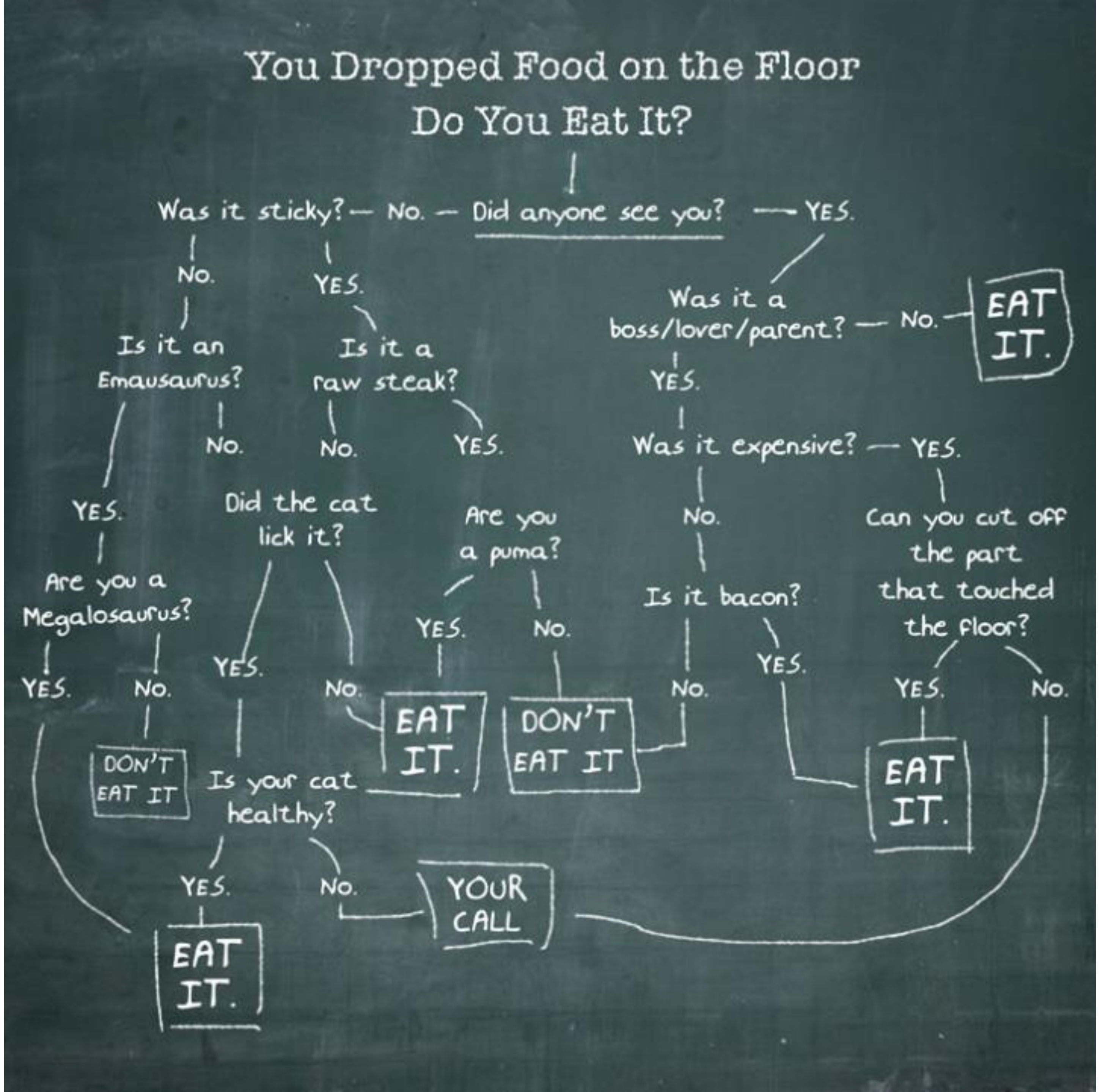
Neural network

Can you open it?

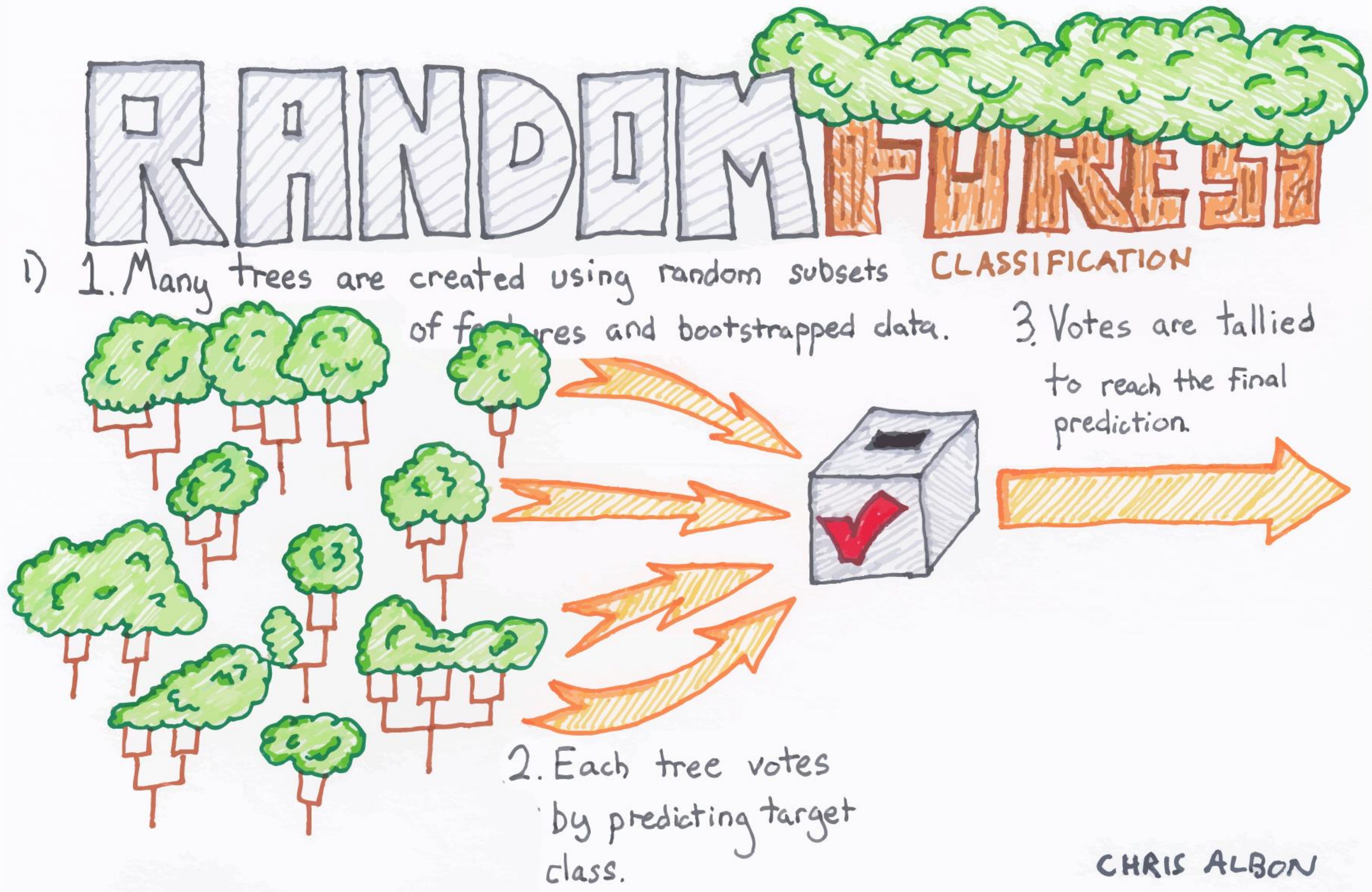

Can you open it?
Yes
No



Decision trees are made of yes/no questions (nodes)

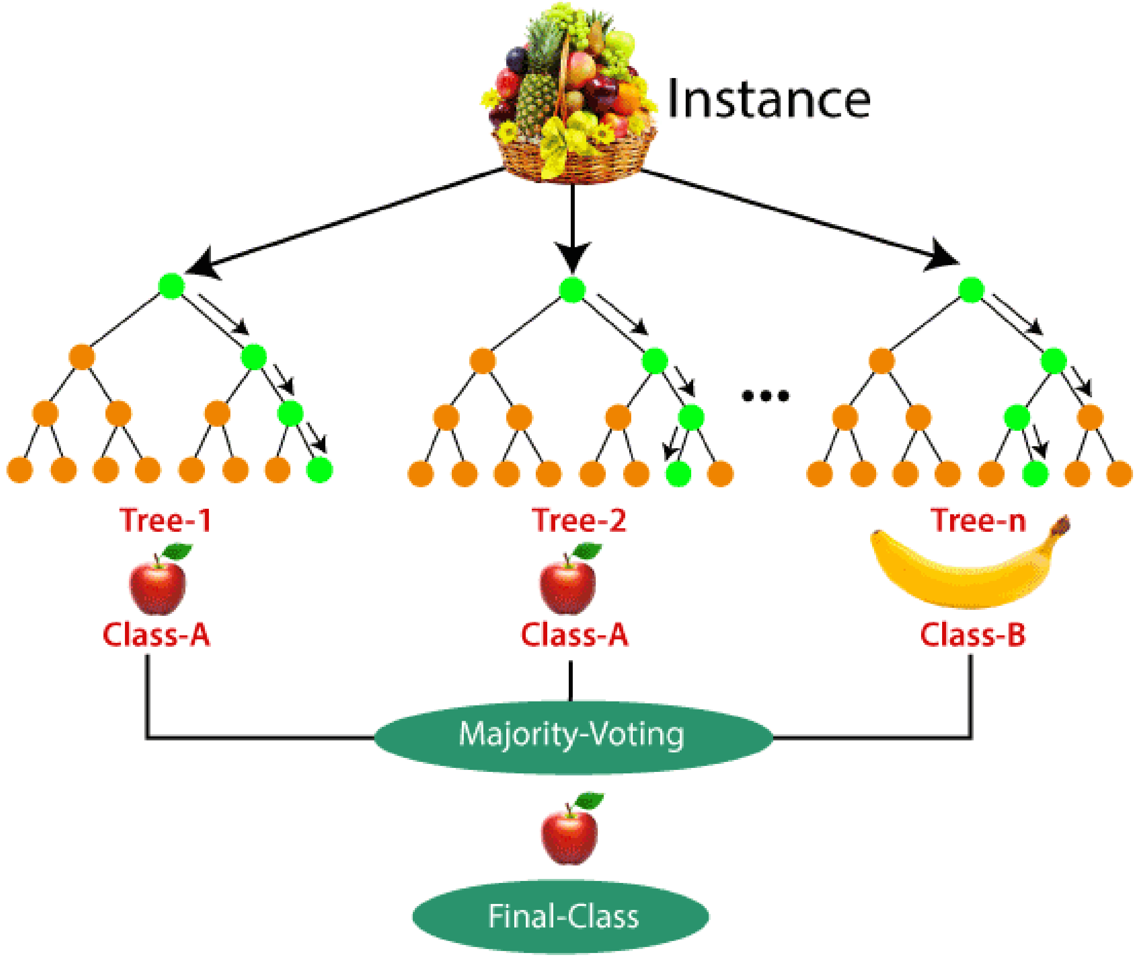


How do we know if this is the best branching structure?

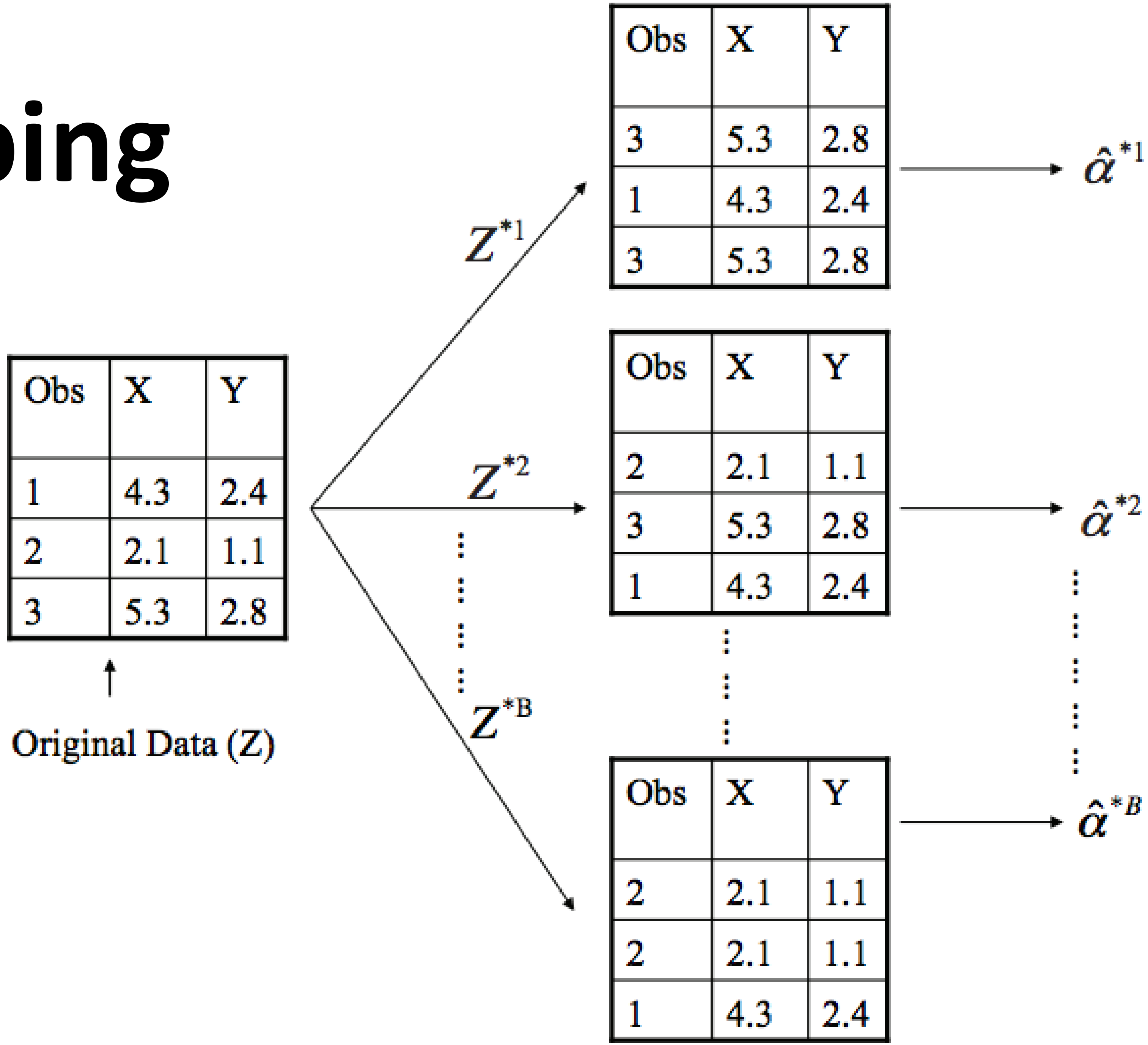

RANDOM FOREST
CLASSIFICATION
1) 1. Many trees are created using random subsets
of fe ures and bootstrapped data.
2. Each tree votes
by predicting target
class.
3. Votes are tallied
to reach the final
prediction.
CHRIS ALBON

Random Forests

- Made of decision trees
- Uses bootstrapping observations and predictors to make a “forest” of decision trees which are aggregated (Bagging)
- Observations left out (due to bootstrapping – OOB – out-of-bag) are used to estimate error
- Can extract importance of variables ”Gini” Score by comparing trees with/without variable included

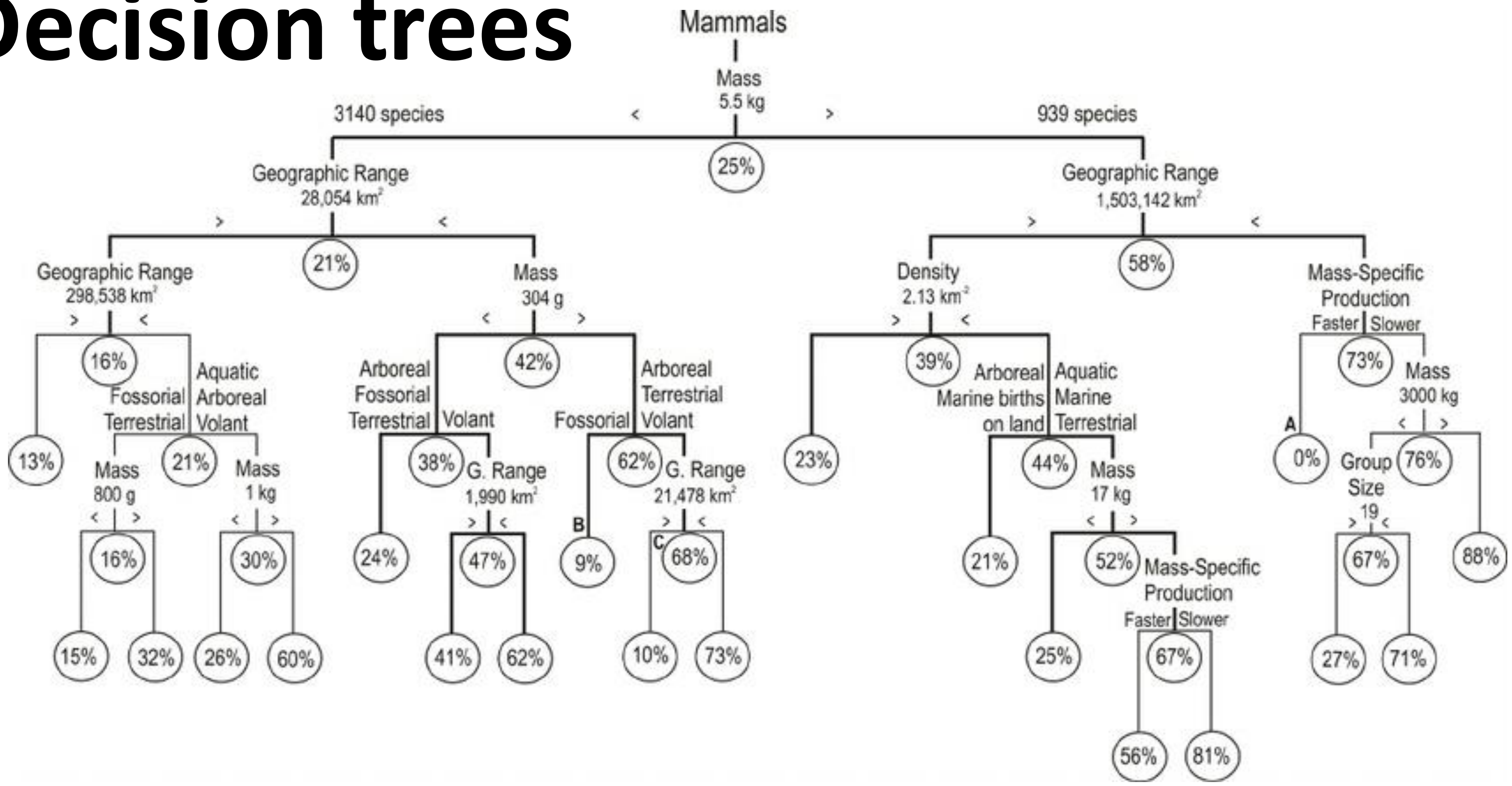


Bootstrapping



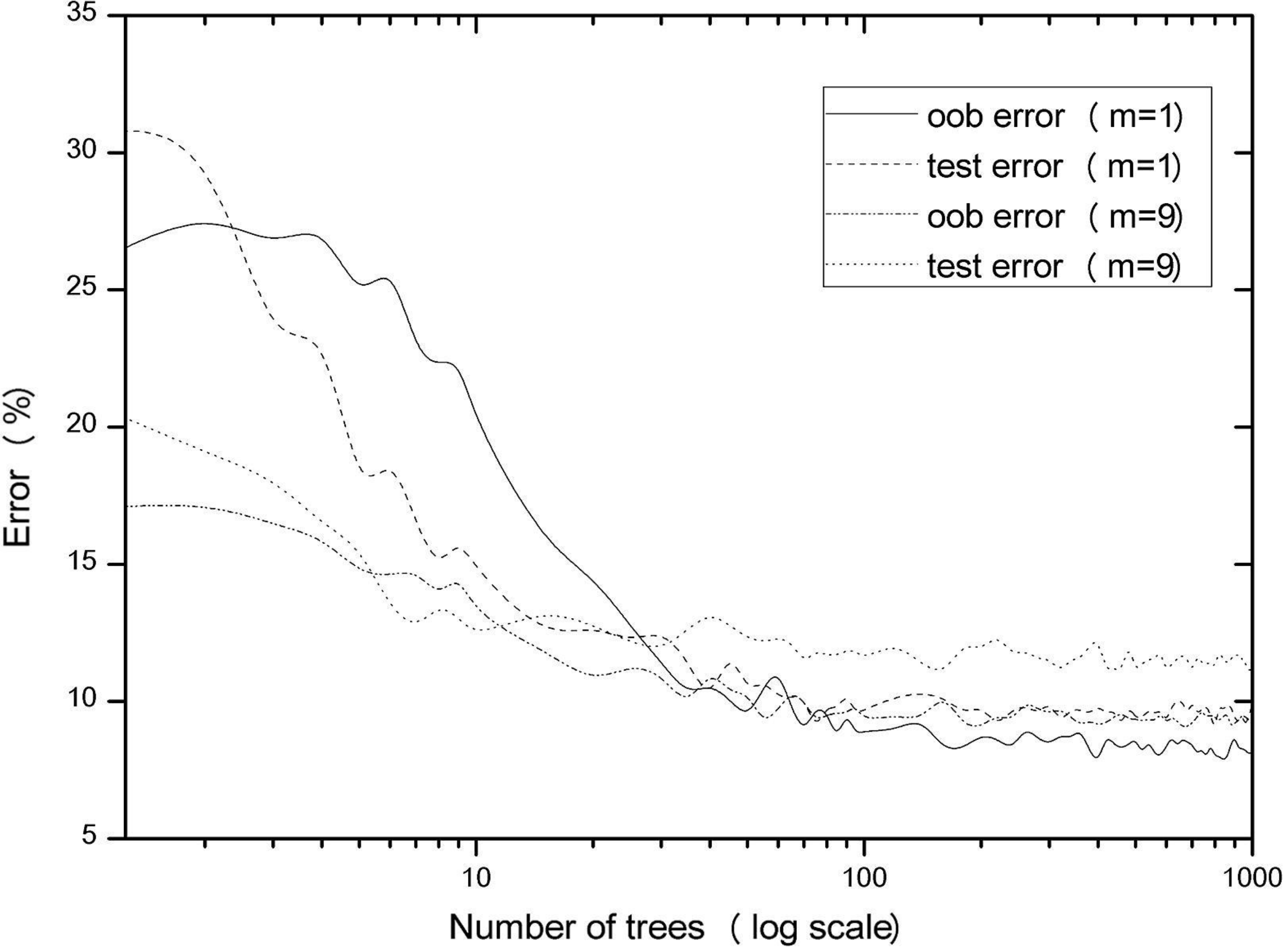
https://timvink.nl/post-balanced-trees/

Decision trees



Decision tree showing extinction risk based on ecological traits (body mass, geographic range size, mass-specific production rate, population density, group size, home range, activity period, type of landmass, habitat mode, sociality, trophic group).

https://www.pnas.org/doi/10.1073/pnas.0901956106



The more trees in the random forest, the more accurate the model (with diminishing returns)

Error is estimated during bootstrapping – called out of bag (OOB) error

An assessment of the effectiveness of a random forest classifier for land-cover classification

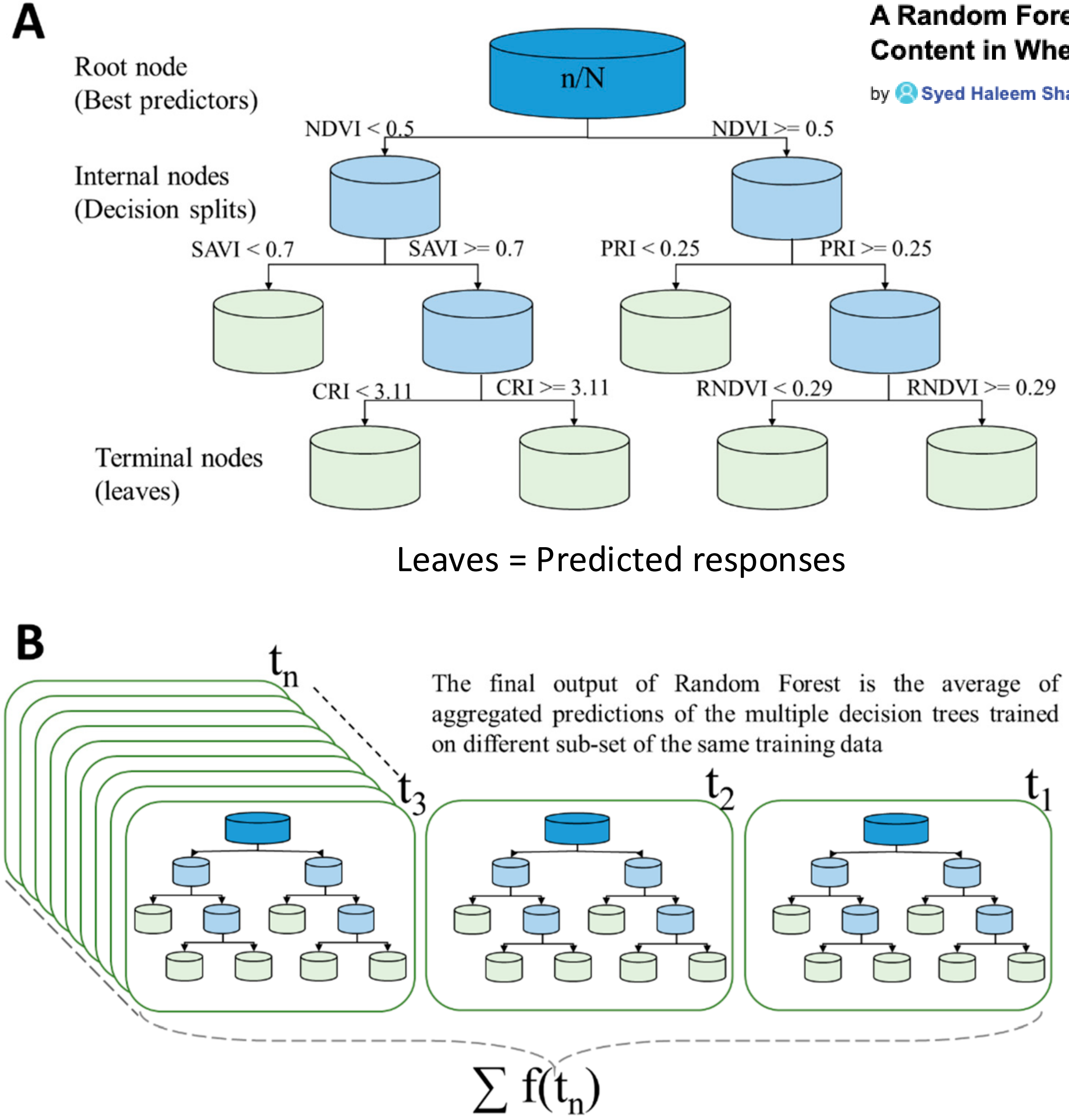
V.F. Rodriguez-Galiano [a], B. Ghimire [b], J. Rogan [b], M. Chica-Olmo [a], J.P. Rigol-Sanchez [c]

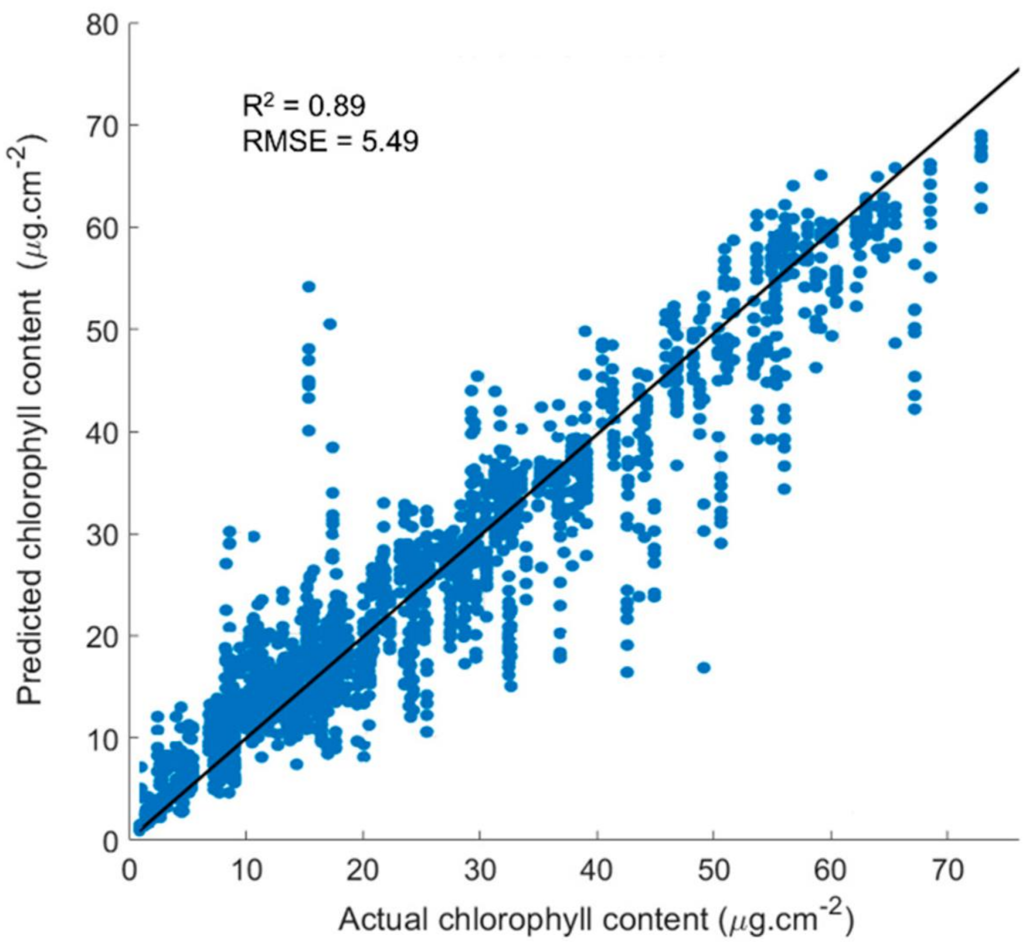
A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat

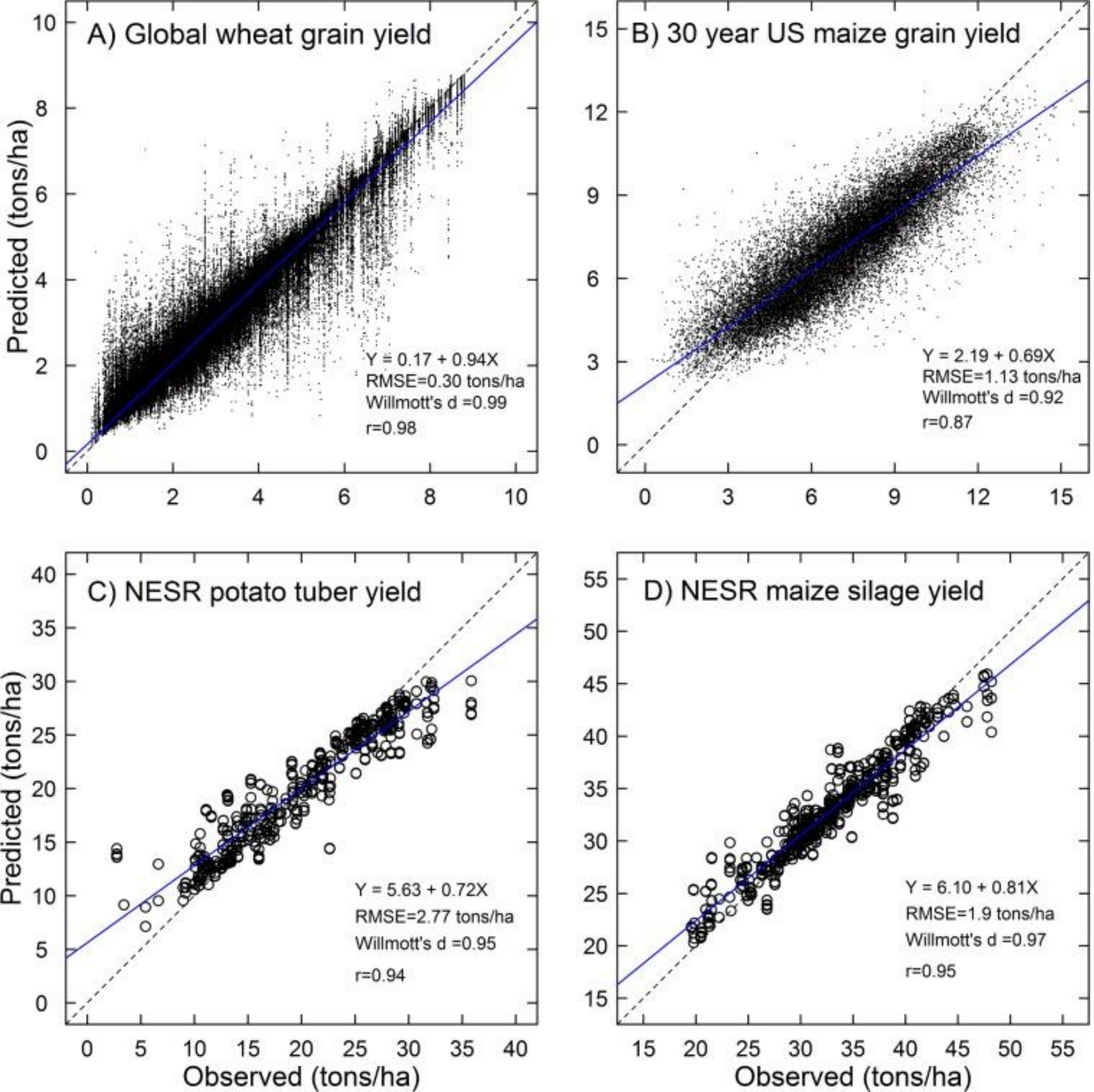
by Syed Haleem Shah [1,*], Yoseline Angel [1], Rasmus Houborg [2], Shawkat Ali [3] and Matthew F. McCabe [1]

A Random Forest Machine Learning Approach for the Retrieval of Leaf Chlorophyll Content in Wheat

by Syed Haleem Shah [1,*], Yoseline Angel [1], Rasmus Houborg [2], Shawkat Ali [3] and Matthew F. McCabe [1]





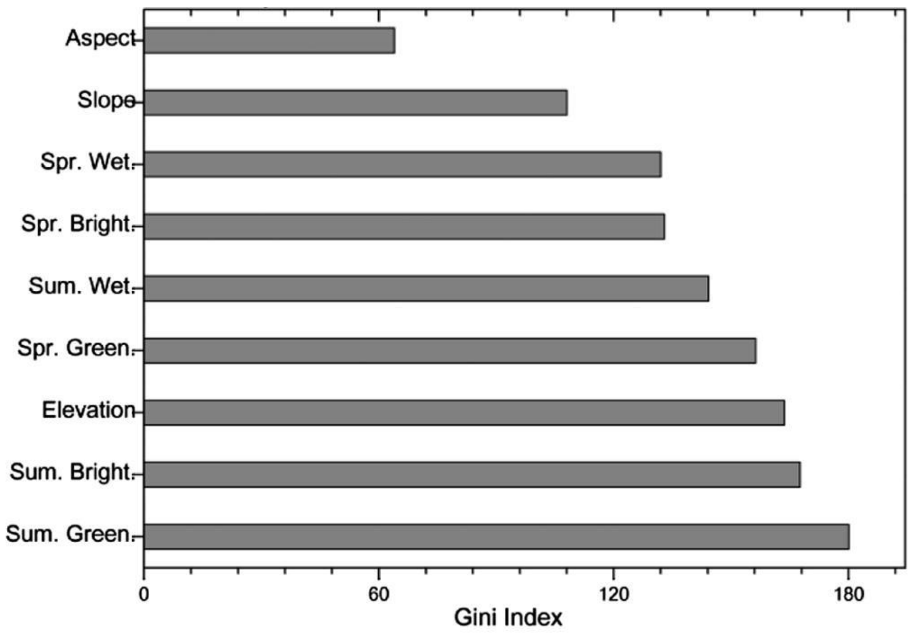


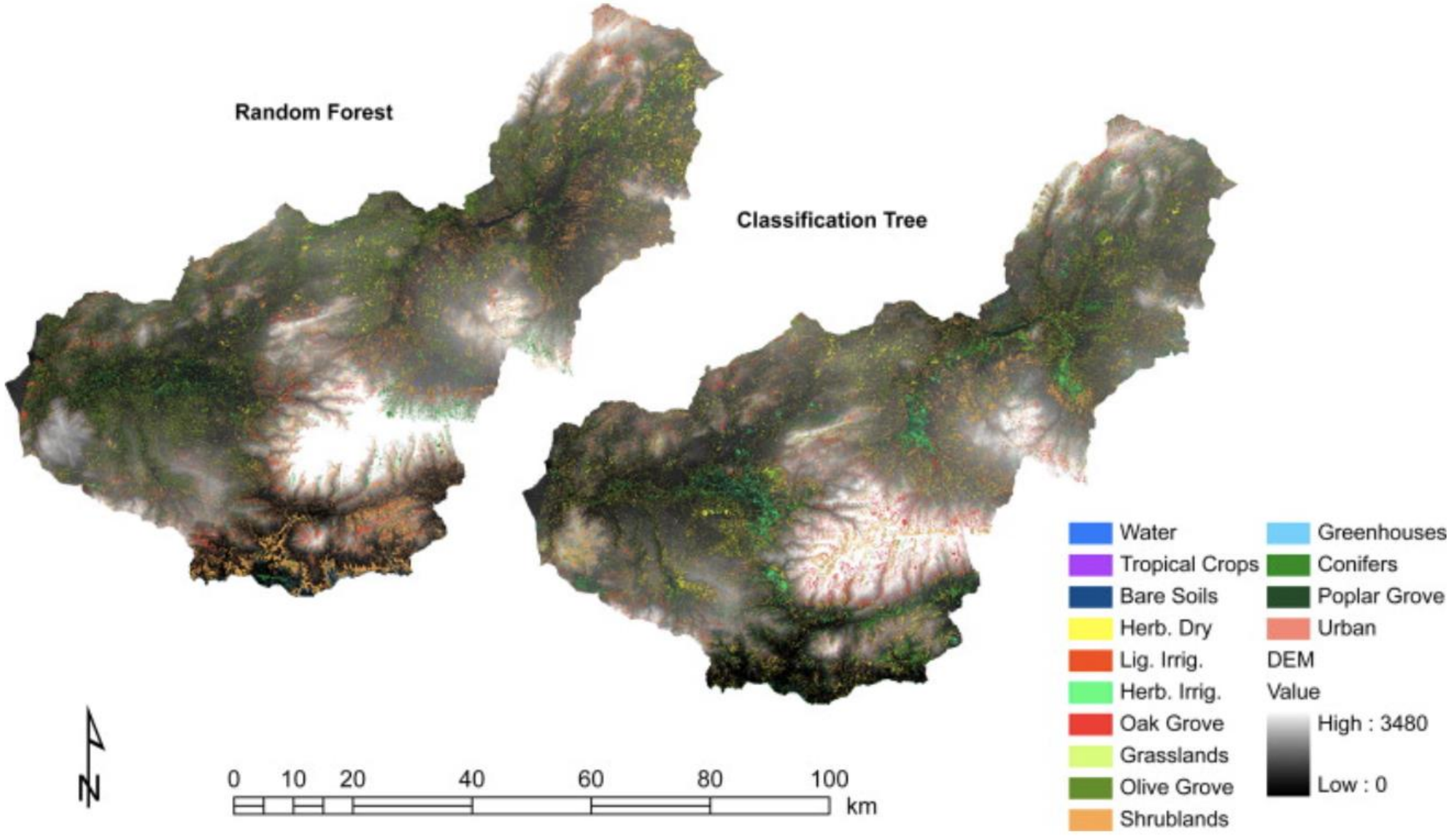
Random Forests for Global and Regional Crop Yield Predictions

Jig Han Jeong,[1] Jonathan P. Resop,[2,3] Nathaniel D. Mueller,[4,5] David H. Fleisher,[3] Kyungdahm Yun,[1] Ethan E. Butler,[6] Dennis J. Timlin,[3] Kyo-Moon Shim,[7] James S. Gerber,[8] Vangimalla R. Reddy,[3] and Soo-Hyung Kim[1,*]

Global wheat and US maize grain yields			Importance Rank	
Variable	Abbreviation	Unit	wheat	maize
Averaged monthly temperature	AVT	°C	8	9
Annual evapotranspiration	EVA	mm	2	6
Summer solstice day length	DAYL	hour	4	7
Maximum monthly temperature	MAX	°C	7	3
Mean coldest quarter Temperature	MCQ	°C	9	11
Minimum monthly temperature	MIN	°C	10	8
Mean warmest quarter temperature	MWQ	°C	6	10
Nitrogen fertilizer application rate	NFERT	kg/ha	1	2
Growing season precipitation	PRE49	mm	3	4
Annual precipitation	PRECI	mm	5	5
Year (US maize only)	YR		-	1

Gini score helps interpretability
Provides measure of variable importance

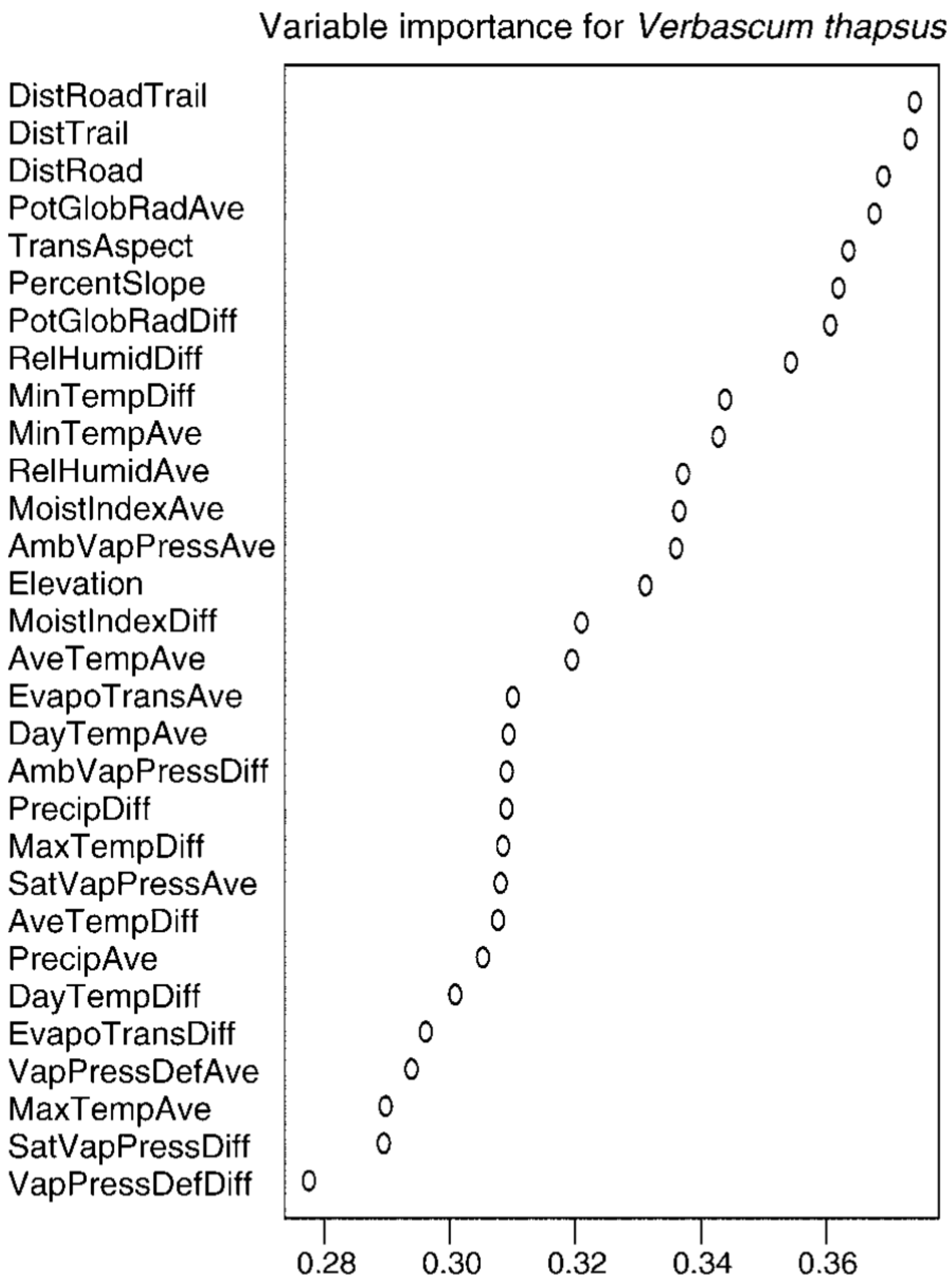




Download : Download high-res image (1MB) Download : Download full-size image

An assessment of the effectiveness of a random forest classifier for land-cover classification

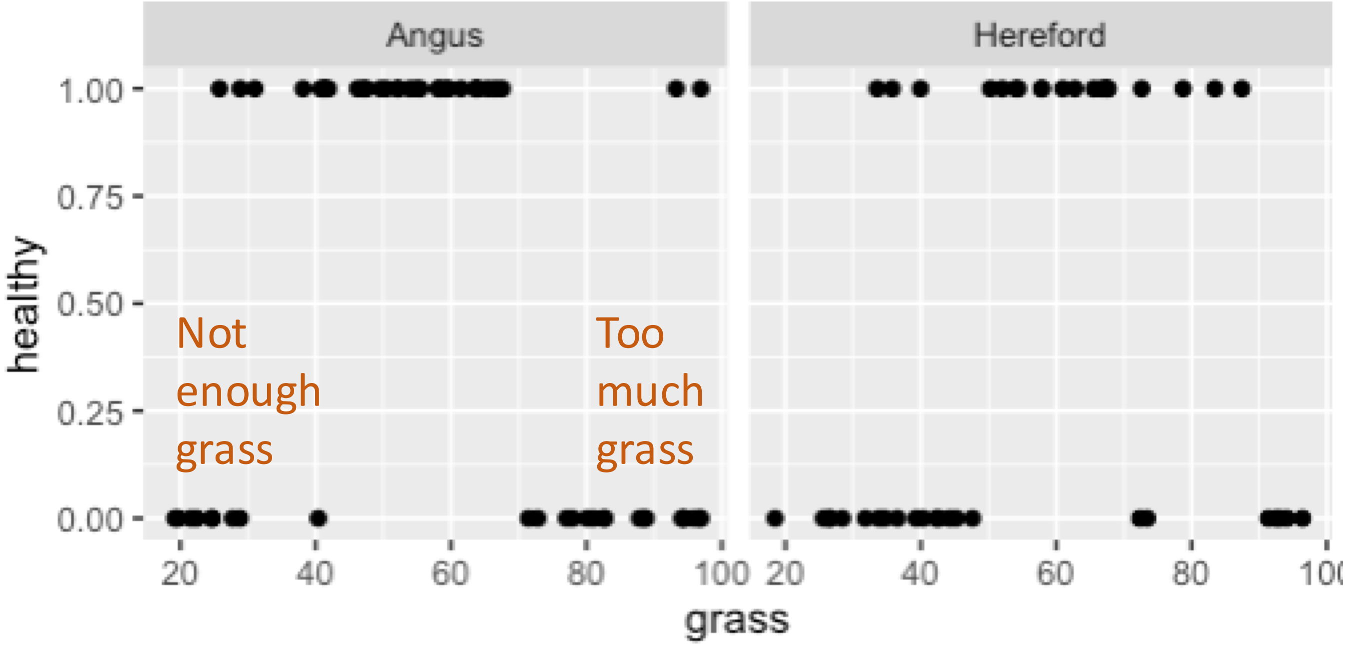
V.F. Rodriguez-Galiano [a], B. Ghimire [b], J. Rogan [b], M. Chica-Olmo [a], J.P. Rigol-Sanchez [c]



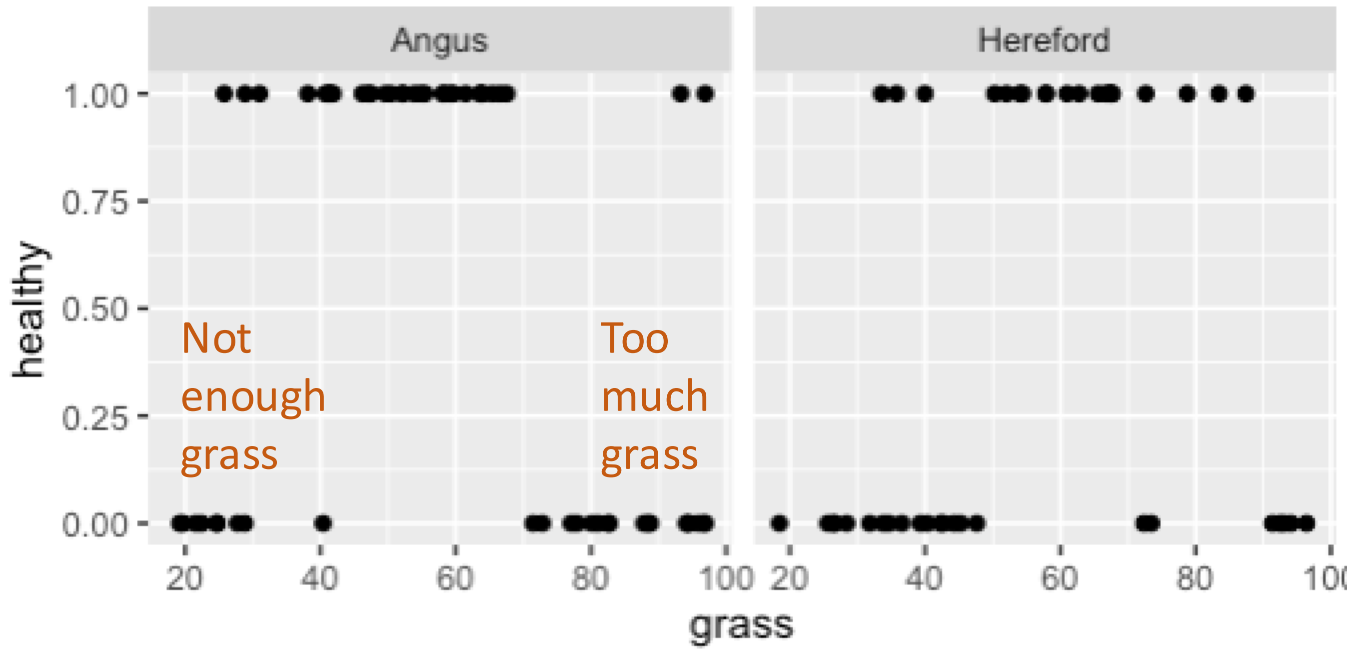
RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY

D. Richard Cutler,[1,7] Thomas C. Edwards, Jr.,[2] Karen H. Beard,[3] Adele Cutler,[4] Kyle T. Hess,[4] Jacob Gibson,[5] and Joshua J. Lawler[6]

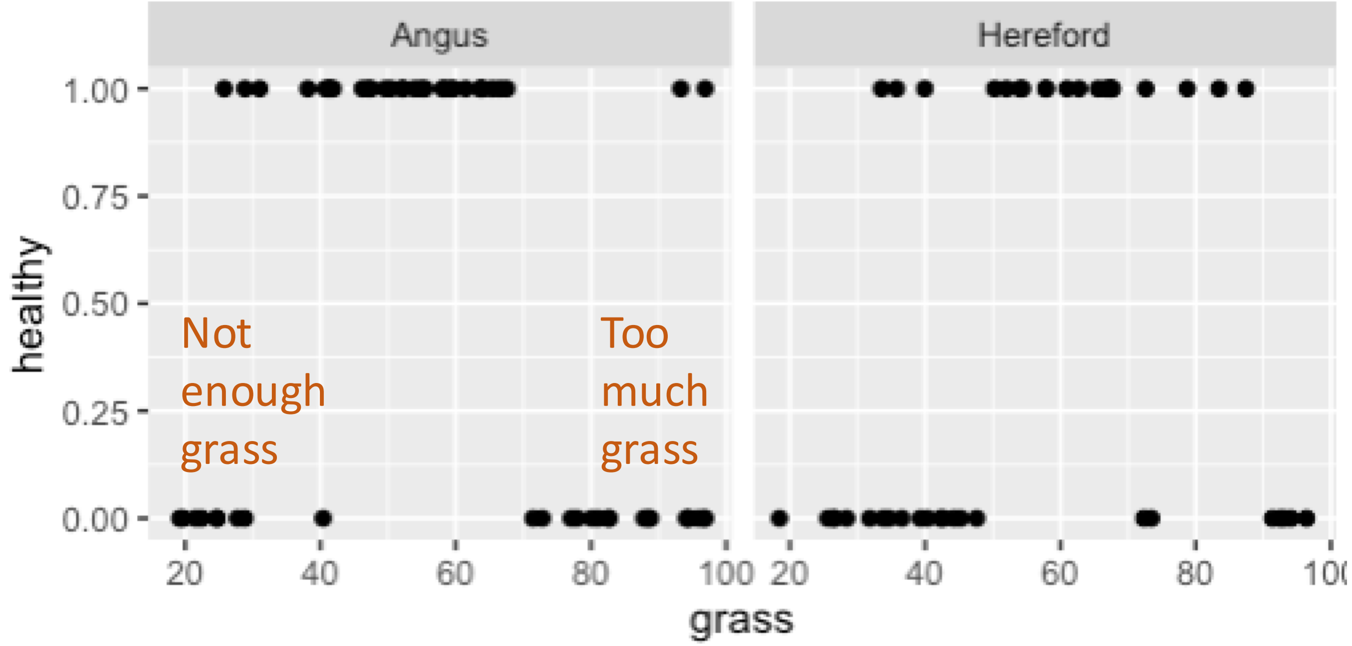
Common Mullen

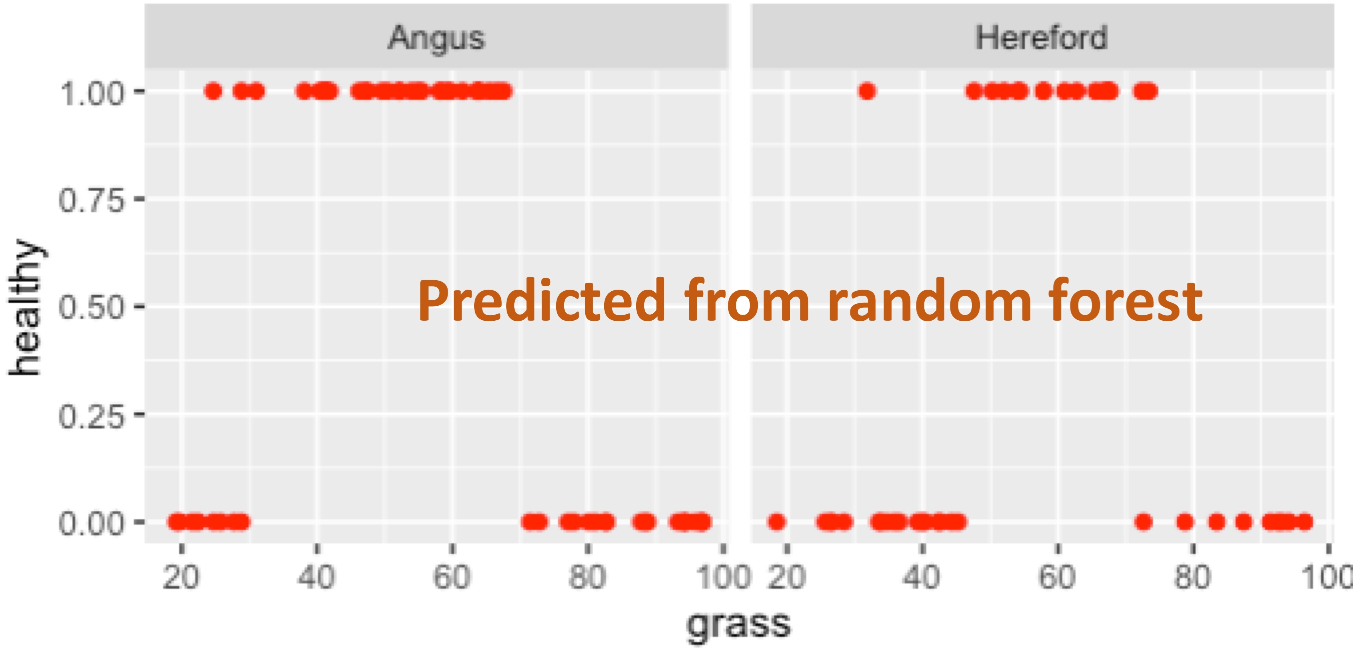


Non-linear relationships are difficult to model with standard linear models



```
model<-randomForest(factor(healthy)~grass+breed, data=data, ntree=100)
```





```
      MeanDecreaseGini
grass        23.778664
breed         2.030991
```

Grass is important predictor

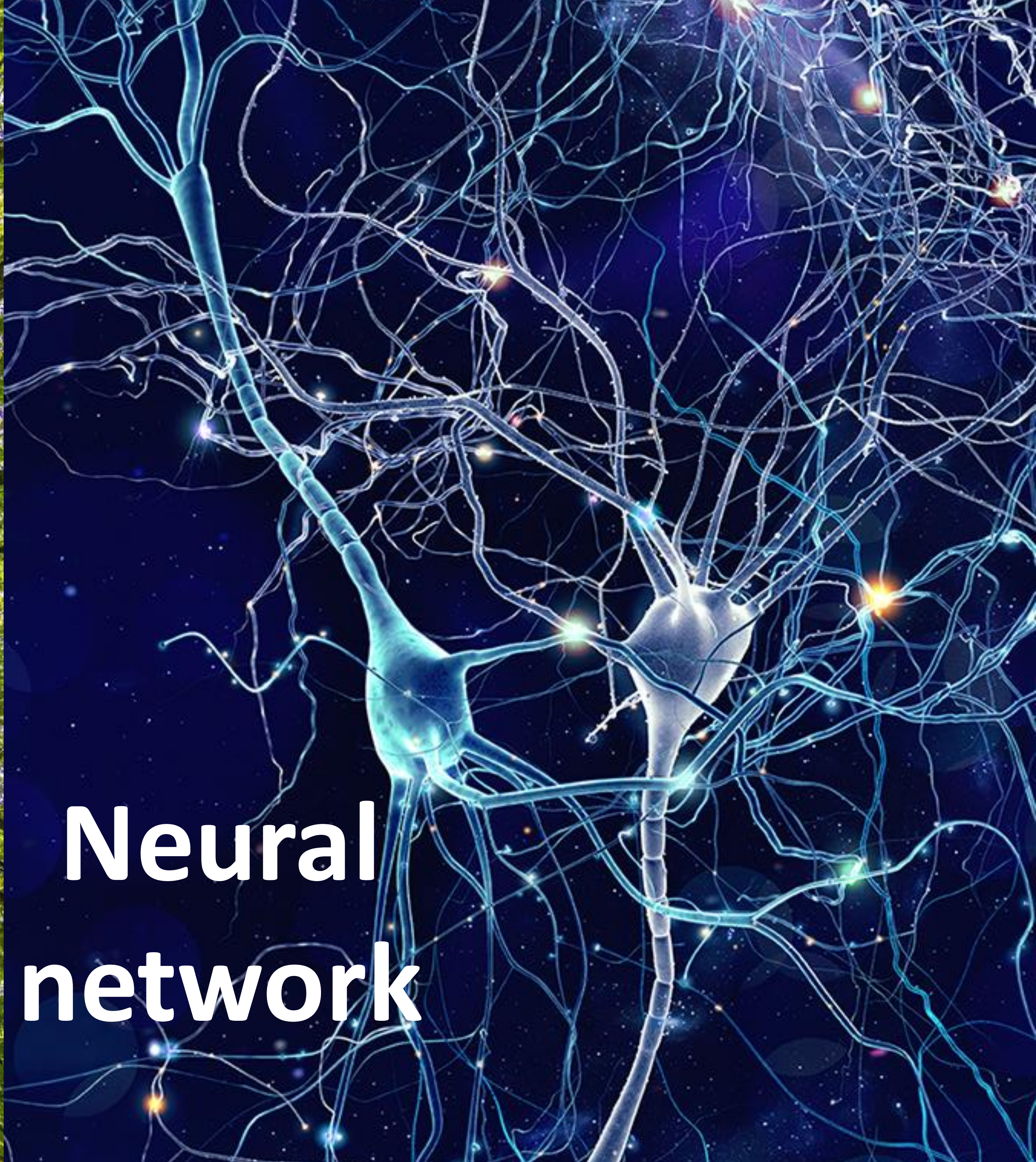
Random Forest

Advantages

1.Accuracy under complexity: Random Forest is known for its high accuracy when relationships between variables are not straightforward (interactions and non-linearities).
2.Variable Importance: It provides insights into which variables are most important in prediction, aiding in hypothesis testing and feature selection (though multicollinearity is an issue)
3.Versatility: Suitable for both classification and regression tasks, making it a versatile tool for various types of data.

Disadvantages

1.Model Interpretability: Less interpretable compared to simpler models like linear regression.
2.Computationally Intensive: It can be computationally expensive, especially with large datasets and a large number of trees.
3.Bias in Multiclass Problems: Tends to be biased towards classes with more instances in classification problems (address by sub-setting data for balanced classes to train model)


Random
forest

Neural
network

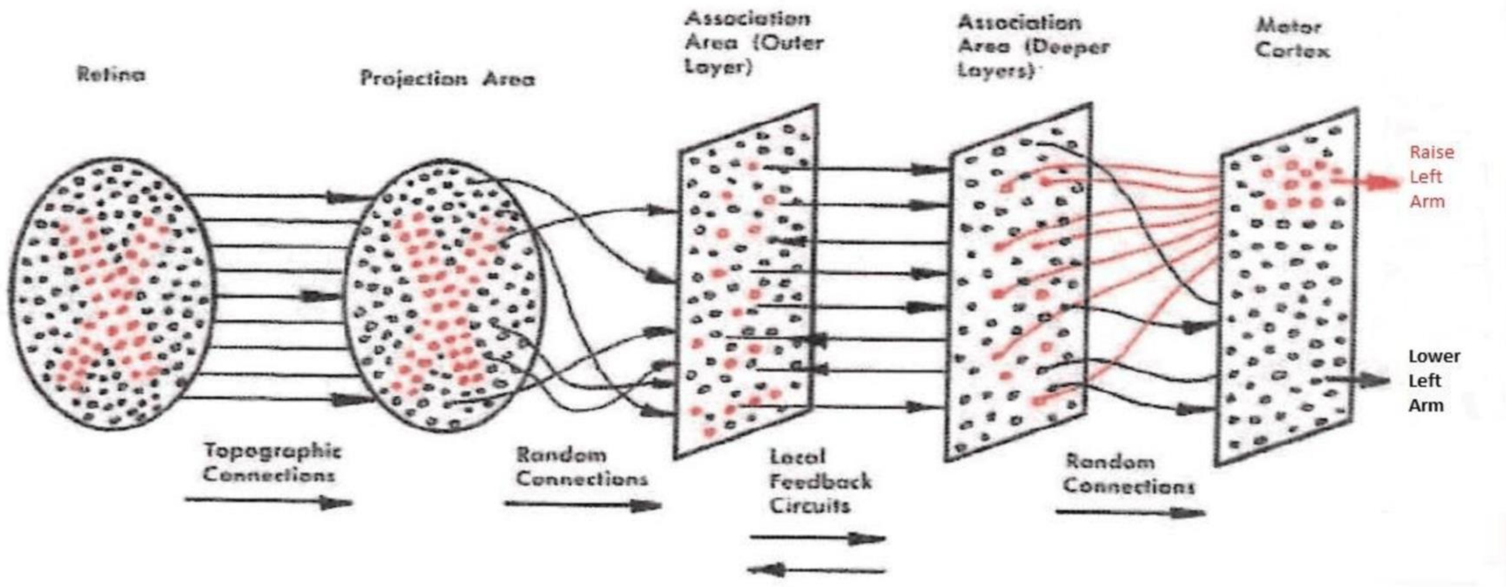


FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)

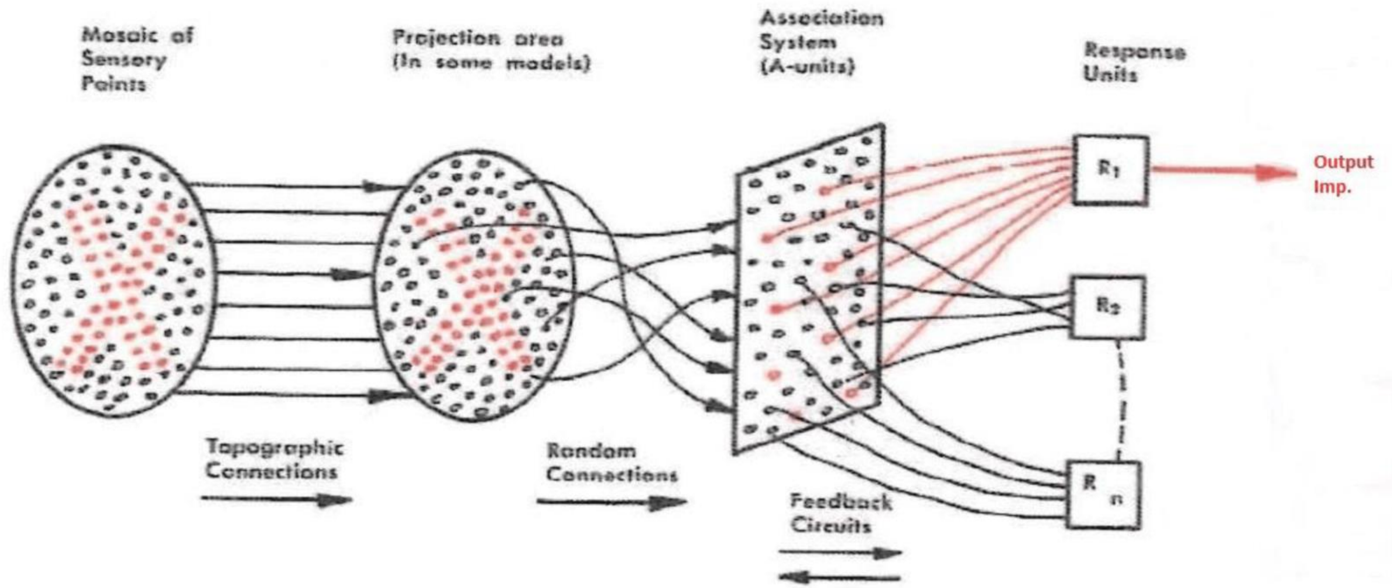
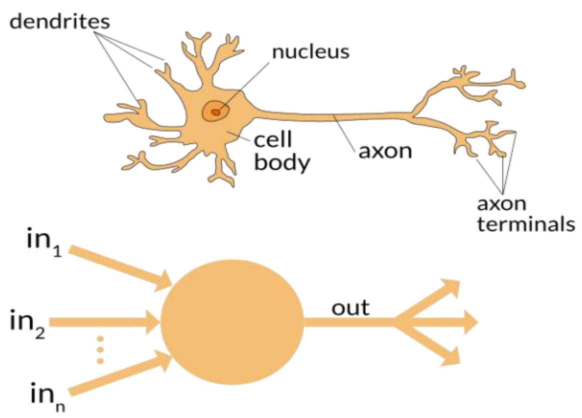
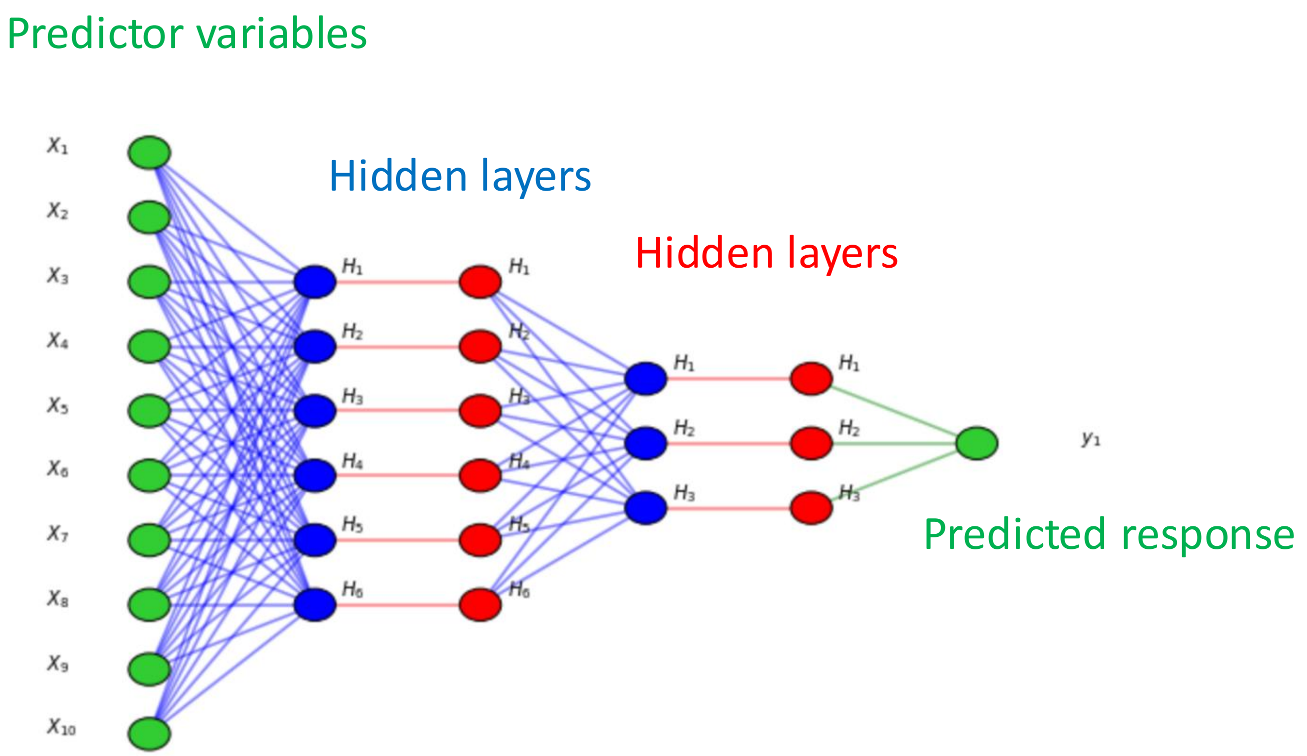


FIG. 2 — Organization of a perceptron.

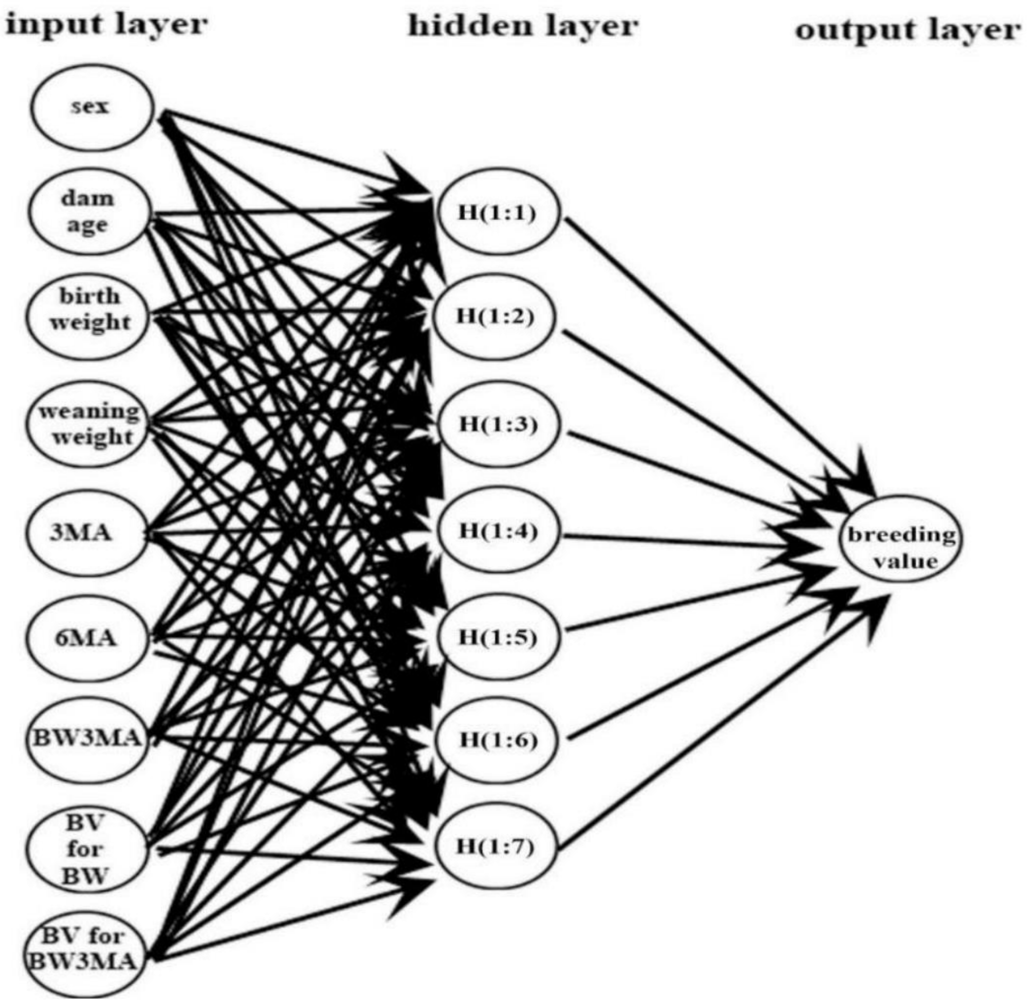
Neural network

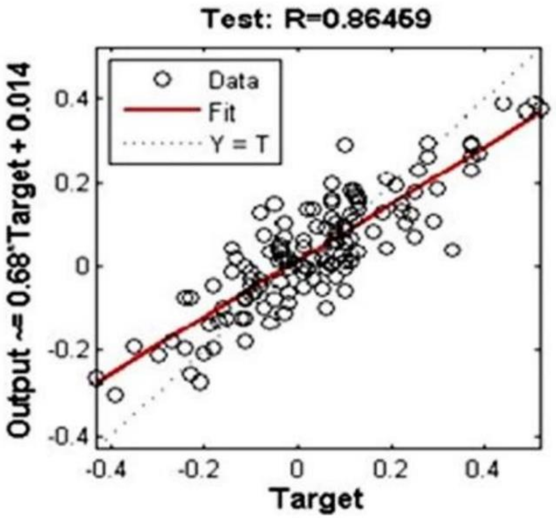
- Internal nodes are functions
- Edges between nodes influence outcome
- Outcome is product of effect of all input and transformations in internal nodes
- Values of weights and edges are those that fit the data best





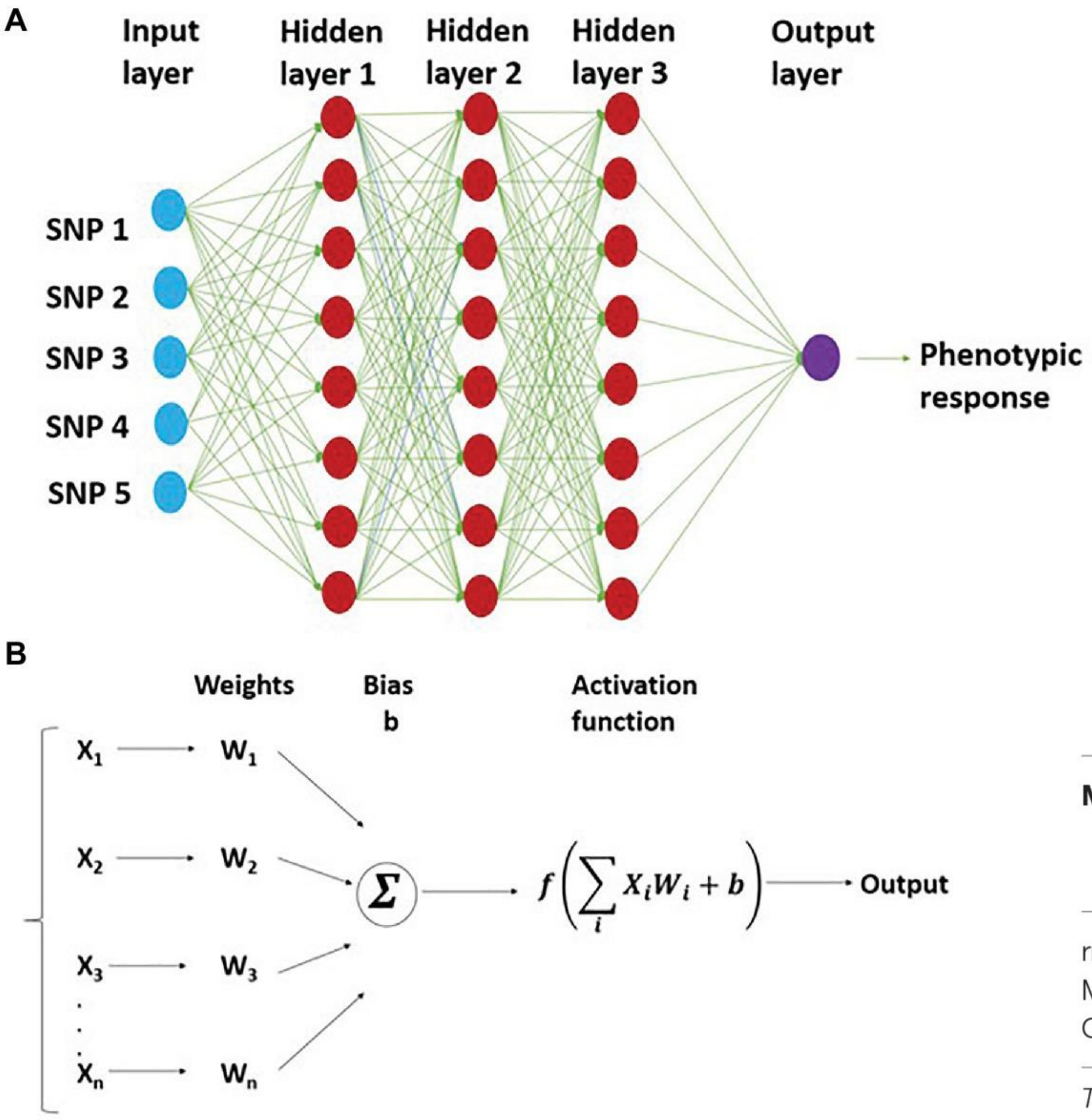
"Deep learning" refers to having multiple hidden layers





Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed

Hamidreza Ghotbaldini[1], Mohammadreza Mohammadabadi[1*], Hossein Nezamabadi-pour[1], Olena Ivanivna Babenko[2], Maryna Vitaliivna Bushtruk[1] and Serhii Vasyliovych Tkachenko[1]

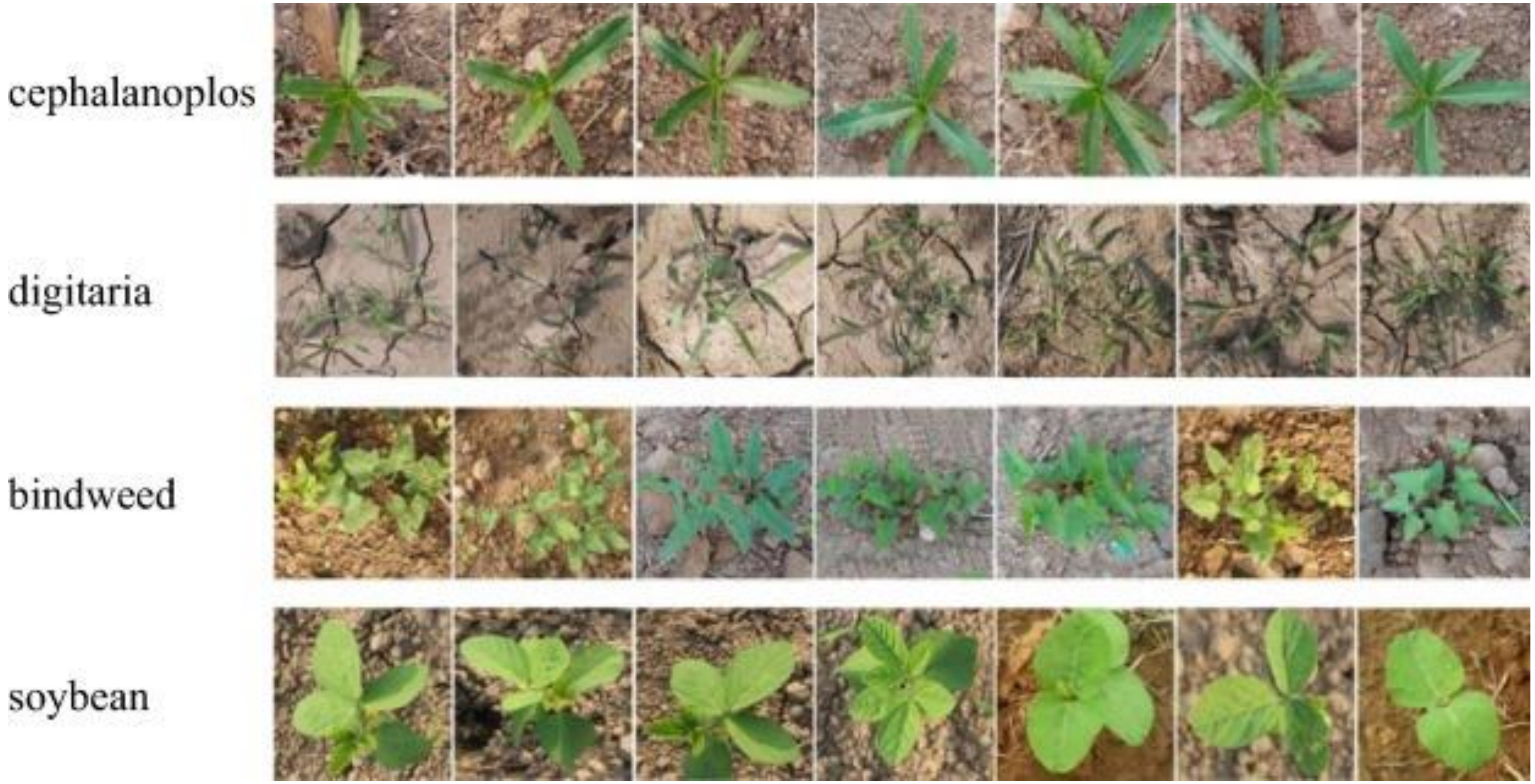


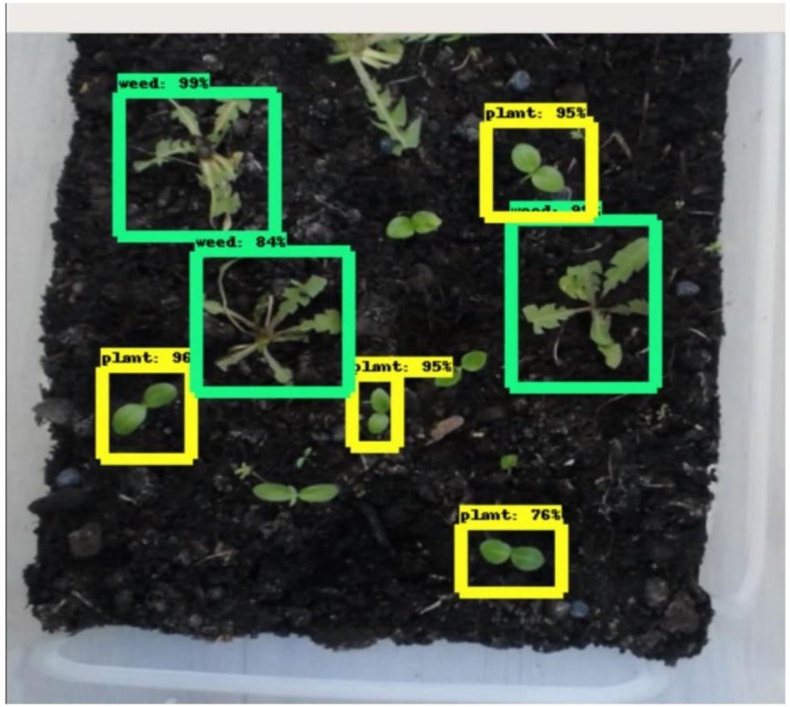
Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program

Karansher S. Sandhu[1], Dennis N. Lozada[2], Zhiwu Zhang[1], Michael O. Pumphrey[1] and Arron H. Carter[1*]

Model	Grain yield	Grain protein content	Test weight	Plant height	Heading date
rrBLUP	0.39	0.48	0.45	0.52	0.46
MLP	**0.44**	**0.53**	**0.48**	**0.57**	**0.51**
CNN	0.39	0.48	0.47	0.55	0.49

The highest prediction accuracy is bolded for each trait under each model scenario.





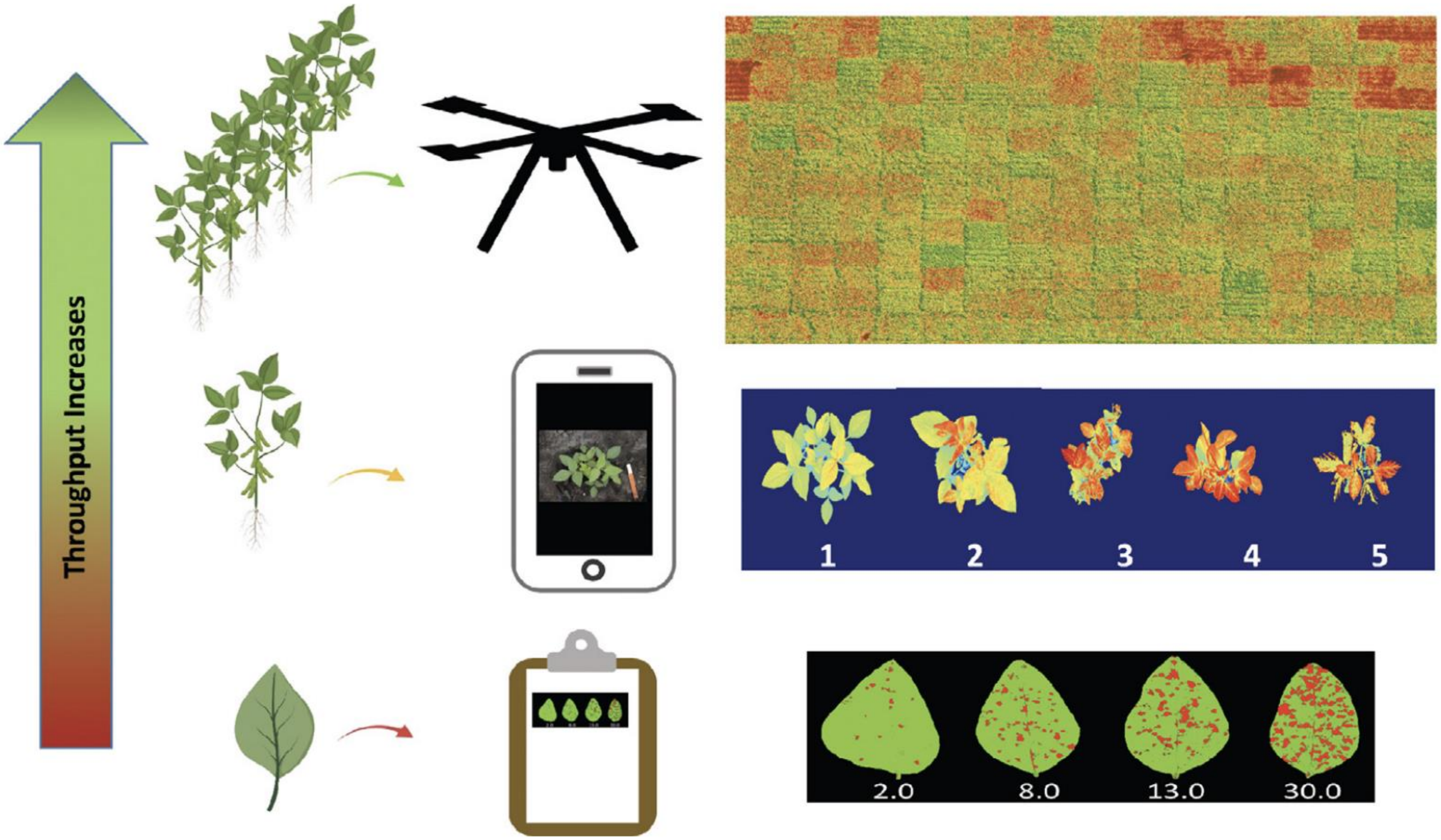
Plant and Weed Identifier Robot as an Agroecological Tool Using Artificial Neural Networks for Image Identification

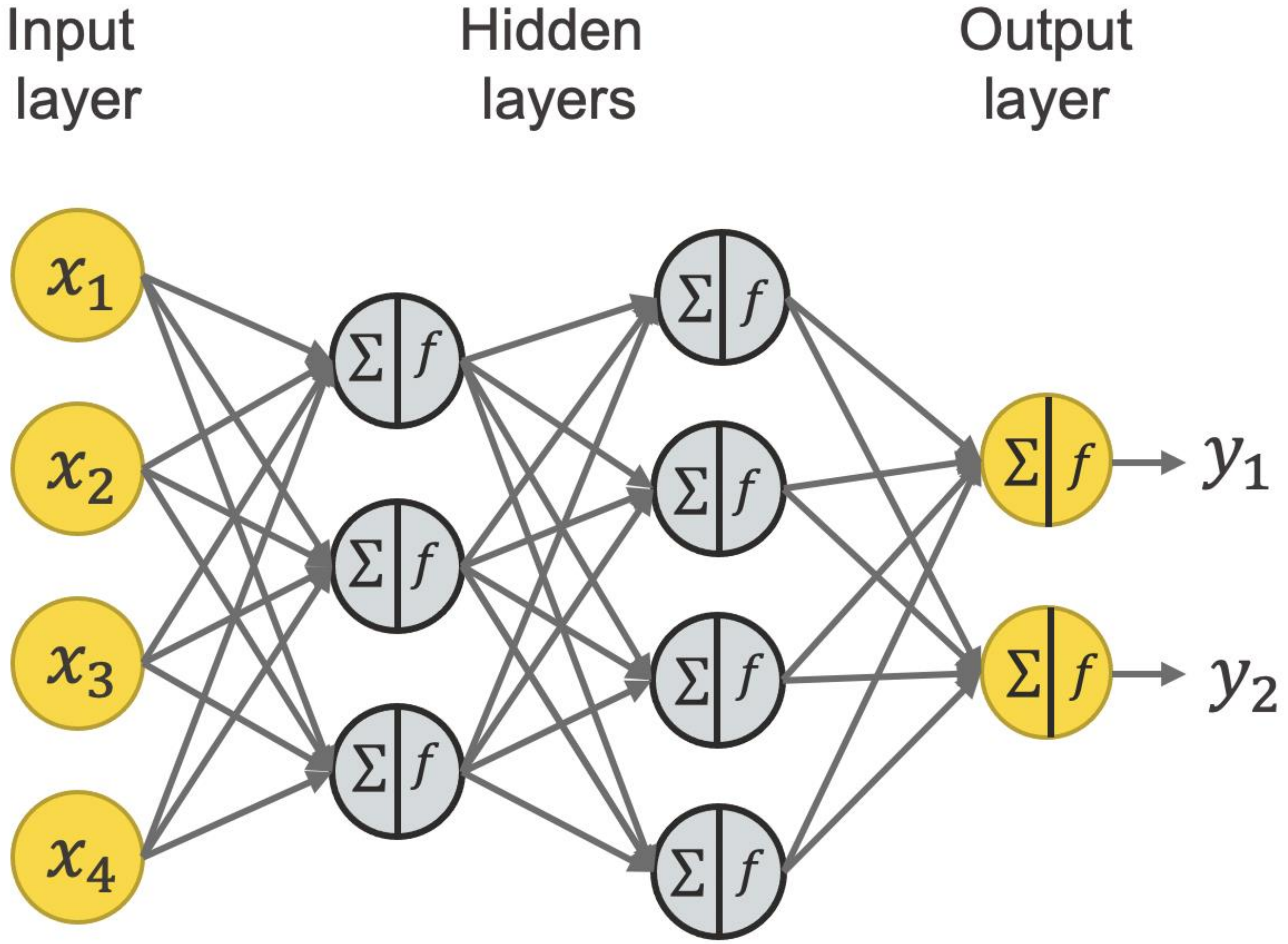
by Tavseef Mairaj Shah [*,†], Durga Prasad Babu Nasika [†] and Ralf Otterpohl

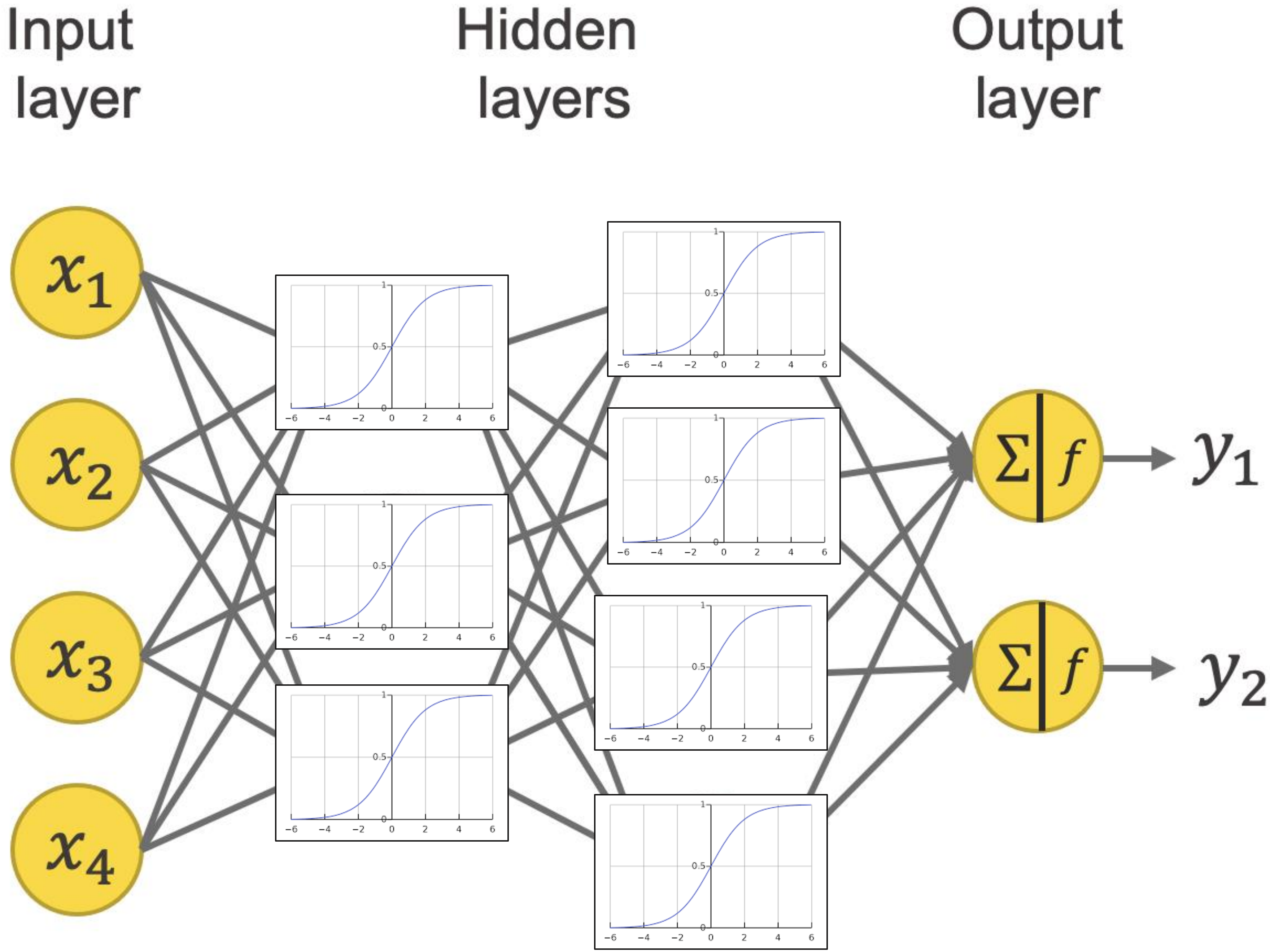
Feature Review

Challenges and Opportunities in Machine-Augmented Plant Stress Phenotyping

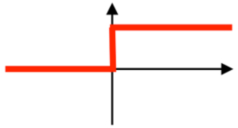
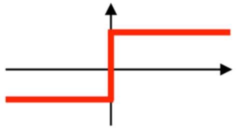
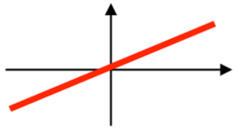
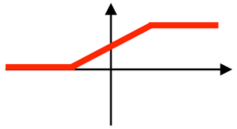
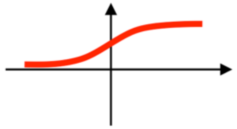
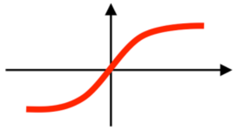
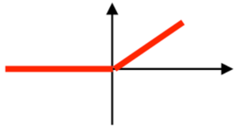
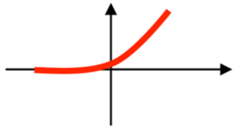
Arti Singh,[1,*] Sarah Jones,[1] Baskar Ganapathysubramanian,[2] Soumik Sarkar,[2] Daren Mueller,[3]
Kulbir Sandhu,[1] and Koushik Nagasubramanian[4]

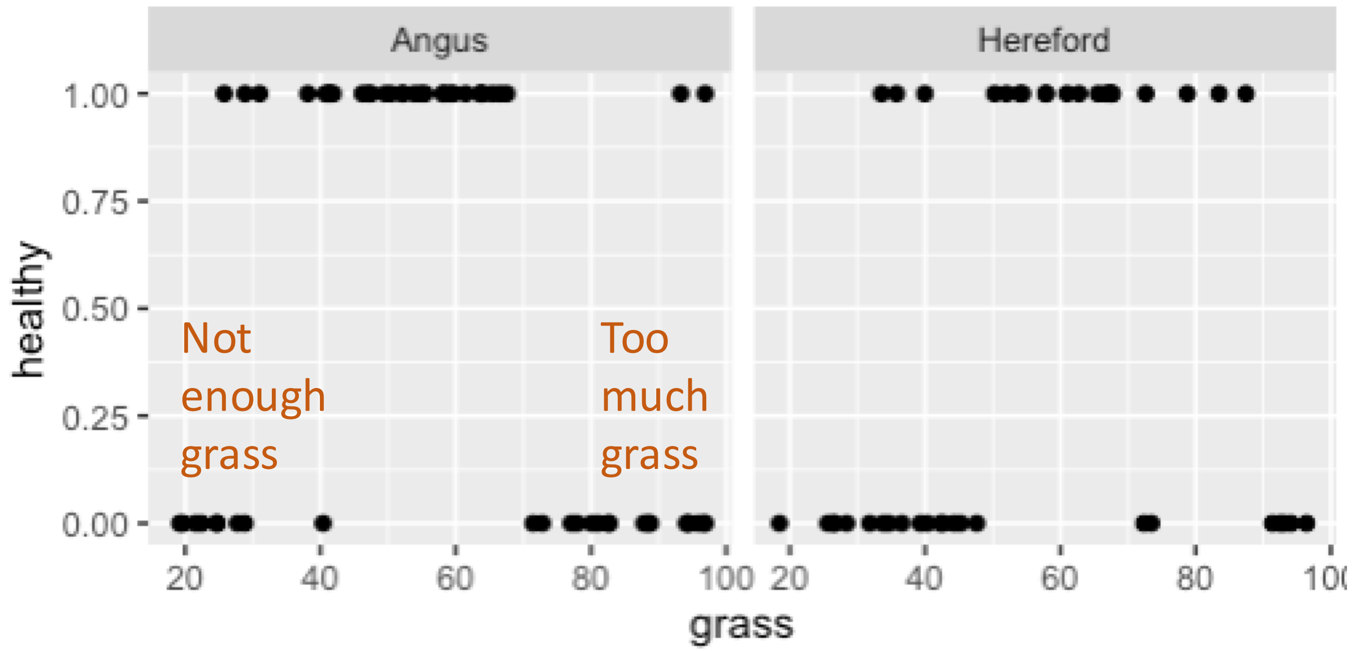



Input layer
Hidden layers
Output layer
x_1
x_2
x_3
x_4
Σ | f
y_1
y_2

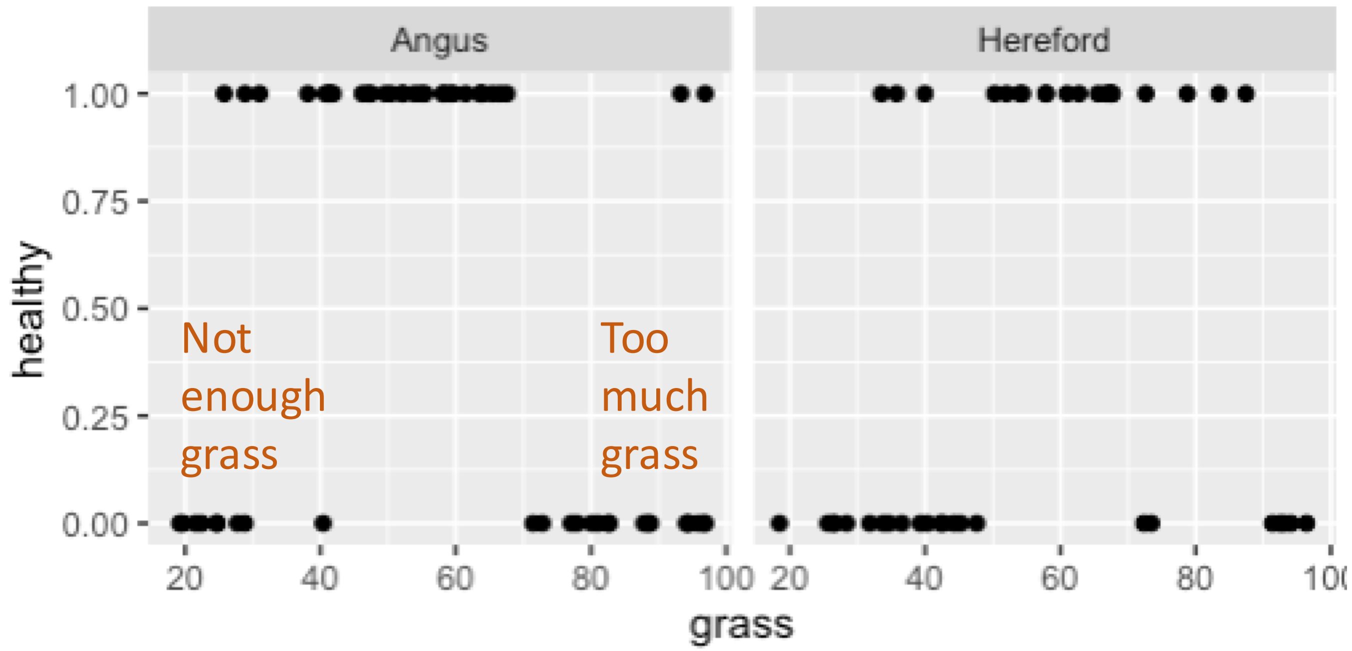

Input layer
Hidden layers
Output layer
x_1
x_2
x_3
x_4
Σ | f
Σ | f
y_1
y_2

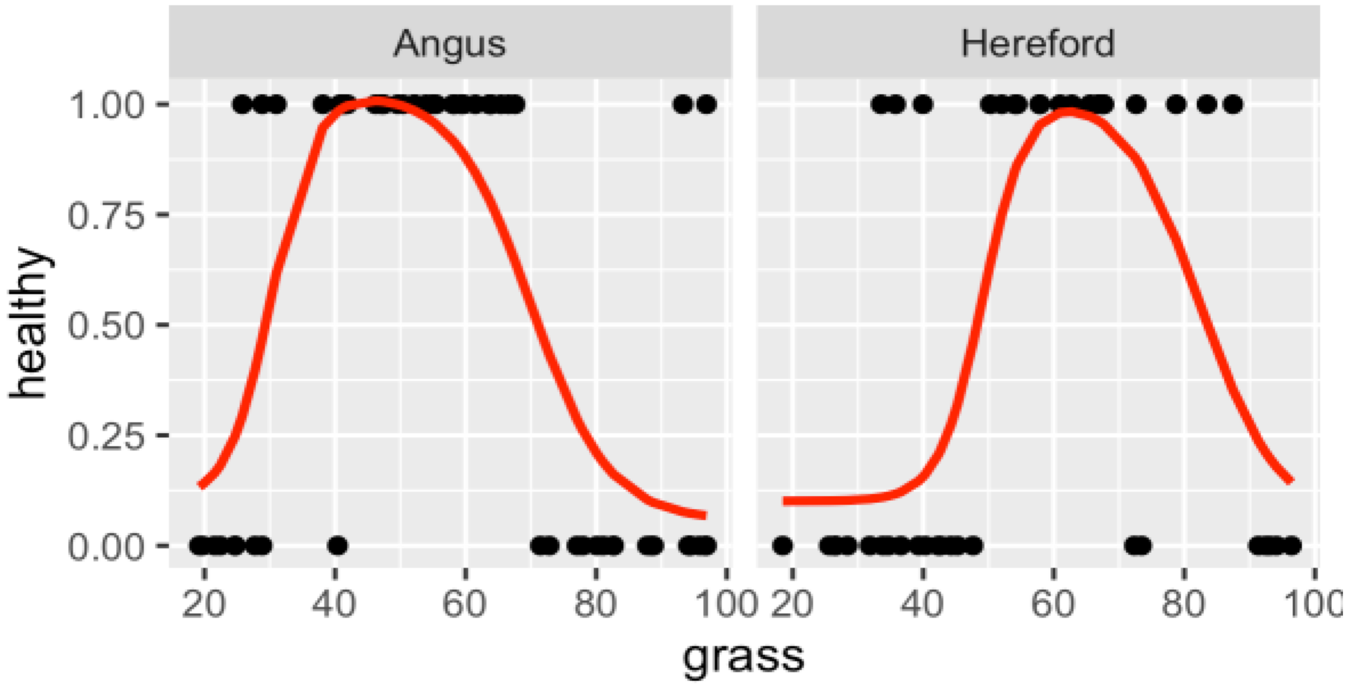
Activation functions

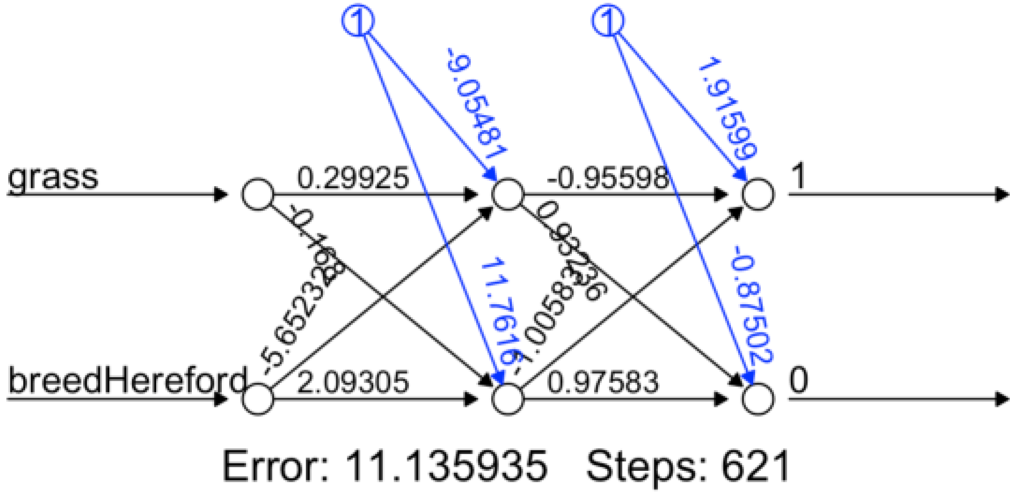
Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^{z})$	Multi-layer Neural Networks	



```
model<-neuralnet(factor(healthy)~grass+breedHereford, data = mod_mat, hidden = c(2), rep=100,  threshold=.2, act.fct="logistic")
```


Angus
Hereford
healthy
1.00
0.75
0.50
0.25
0.00
20
40
60
80
100
grass
Not enough grass
Too much grass


Angus
Hereford
healthy
1.00
0.75
0.50
0.25
0.00
20
40
60
80
100
grass


1
1
-9.05481
1.91599
grass
0.29925
-0.95598
1
11.7616
-0.87502
breedHereford
2.09305
0.97583
0
Error: 11.135935 Steps: 621