Caret and zoon: machine learning, ecology and domain specific package systems

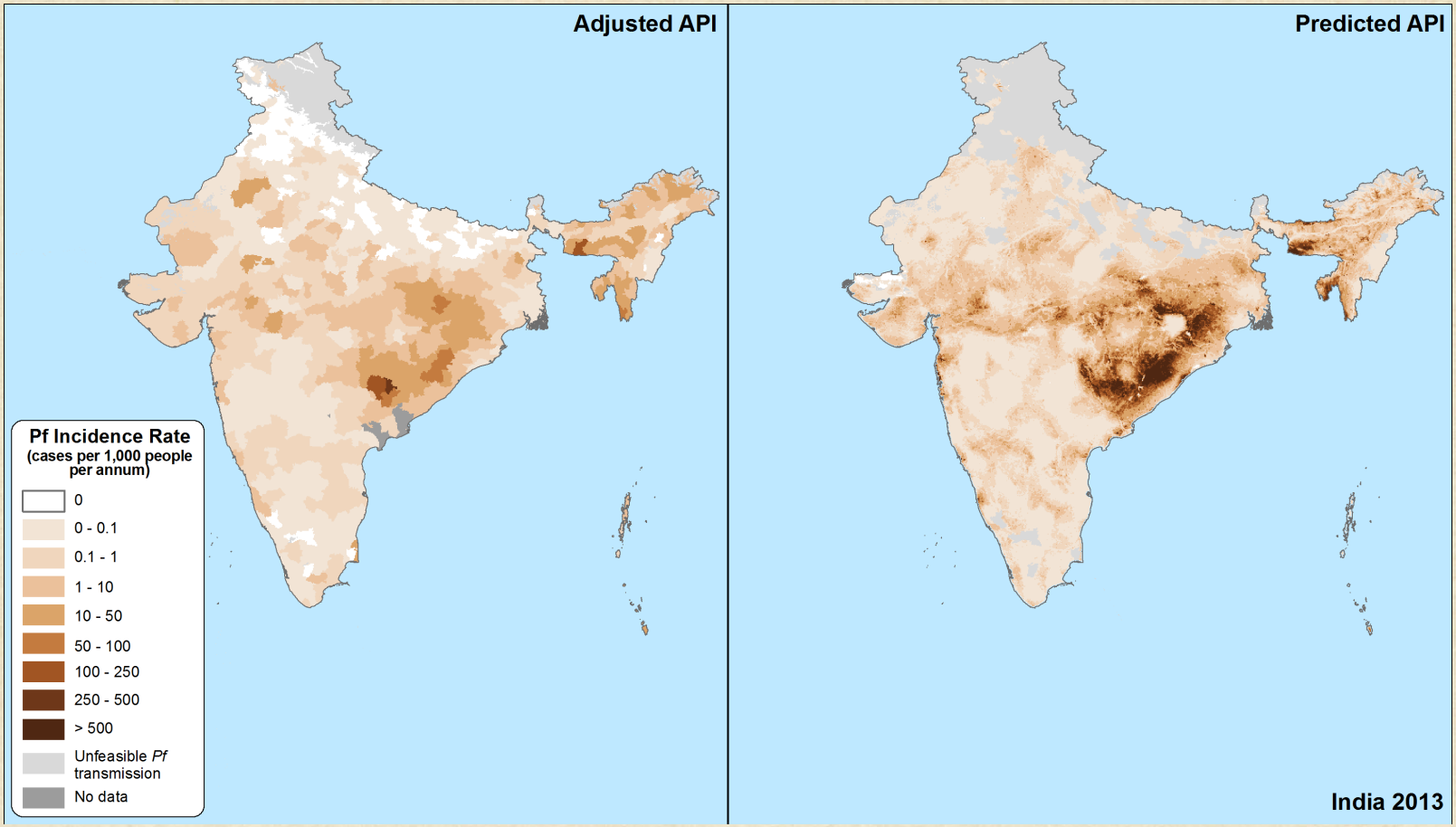
Tim C.D. Lucas
Malaria Atlas Project, BDI, Oxford

@timcdlucas @statsforbios
timcdlucas@gmail.com

Who am I?

Malaria Atlas Project at BDI

Malaria, maps, geostatistics



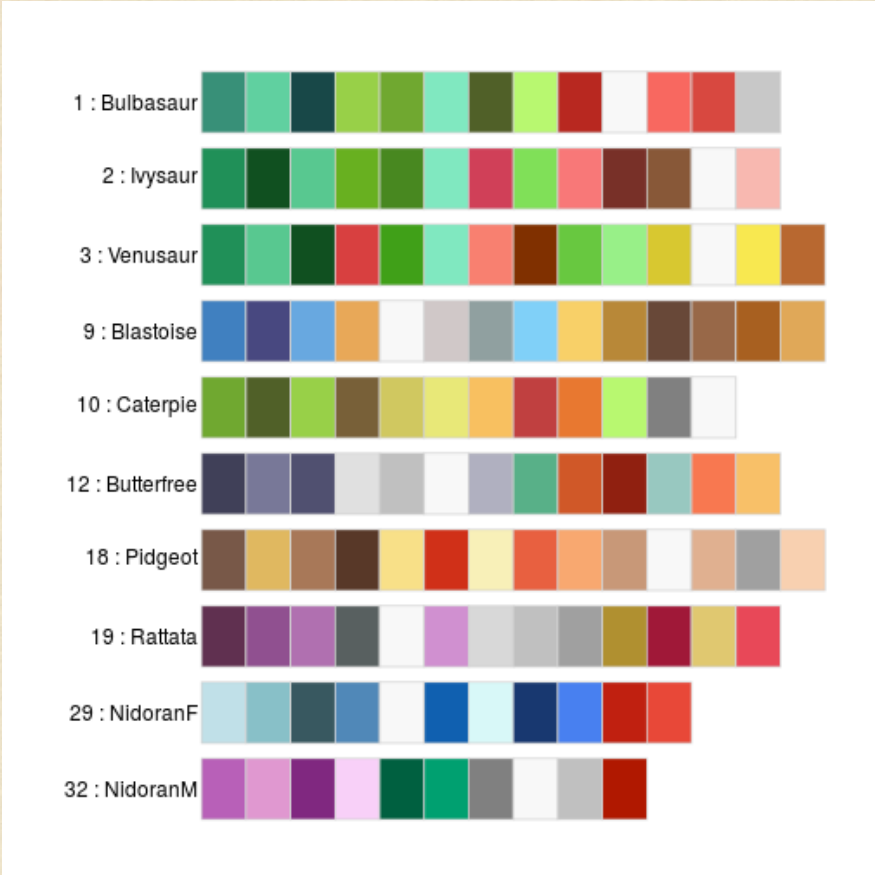
Who am I?

R packages

Zoon

INLAutils

palettetown - my greatest ever achievement



Talk overview

caret

General package for machine learning.

Introduction to the package.

A domain specific package ecosystem?

zoon

General package for species distribution modelling.

What are SDMs?

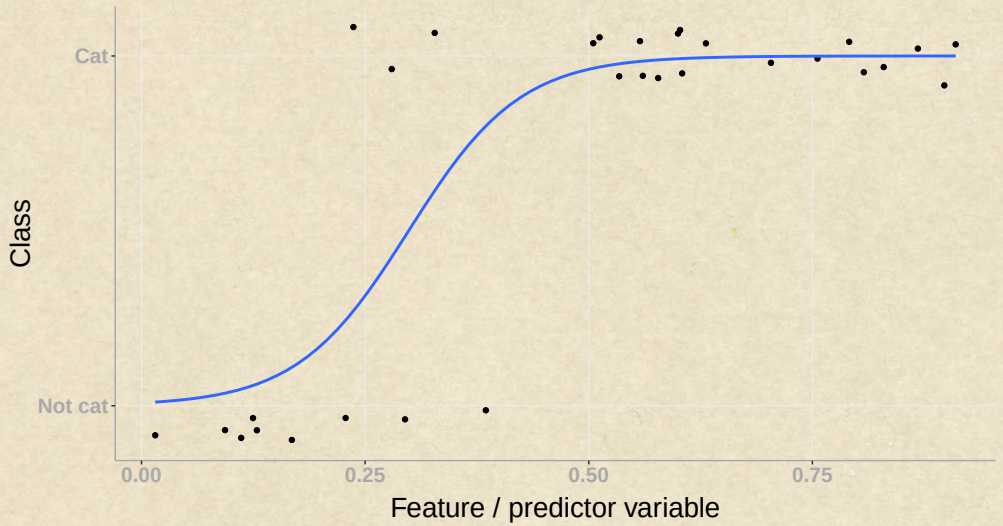
Package overview.

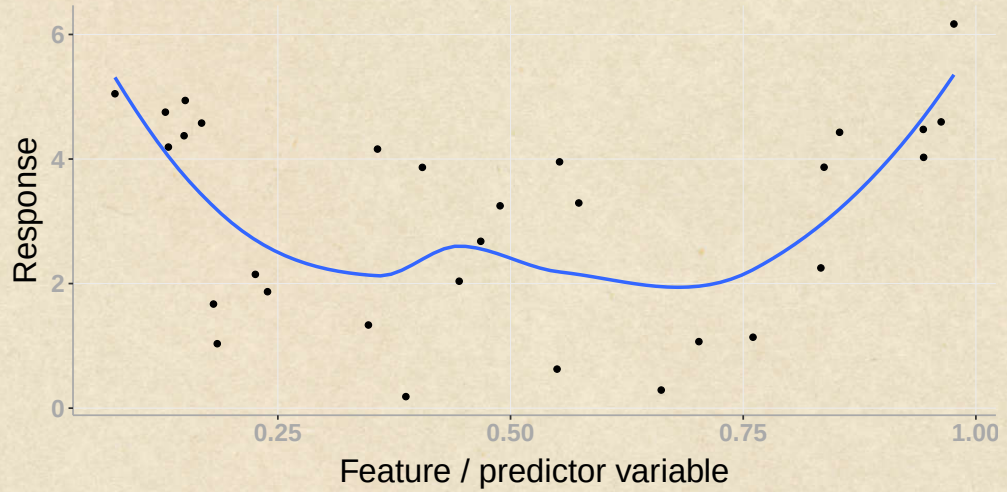
Domain specific ecosystems

Other examples.

Are they a good thing?

What is machine learning?





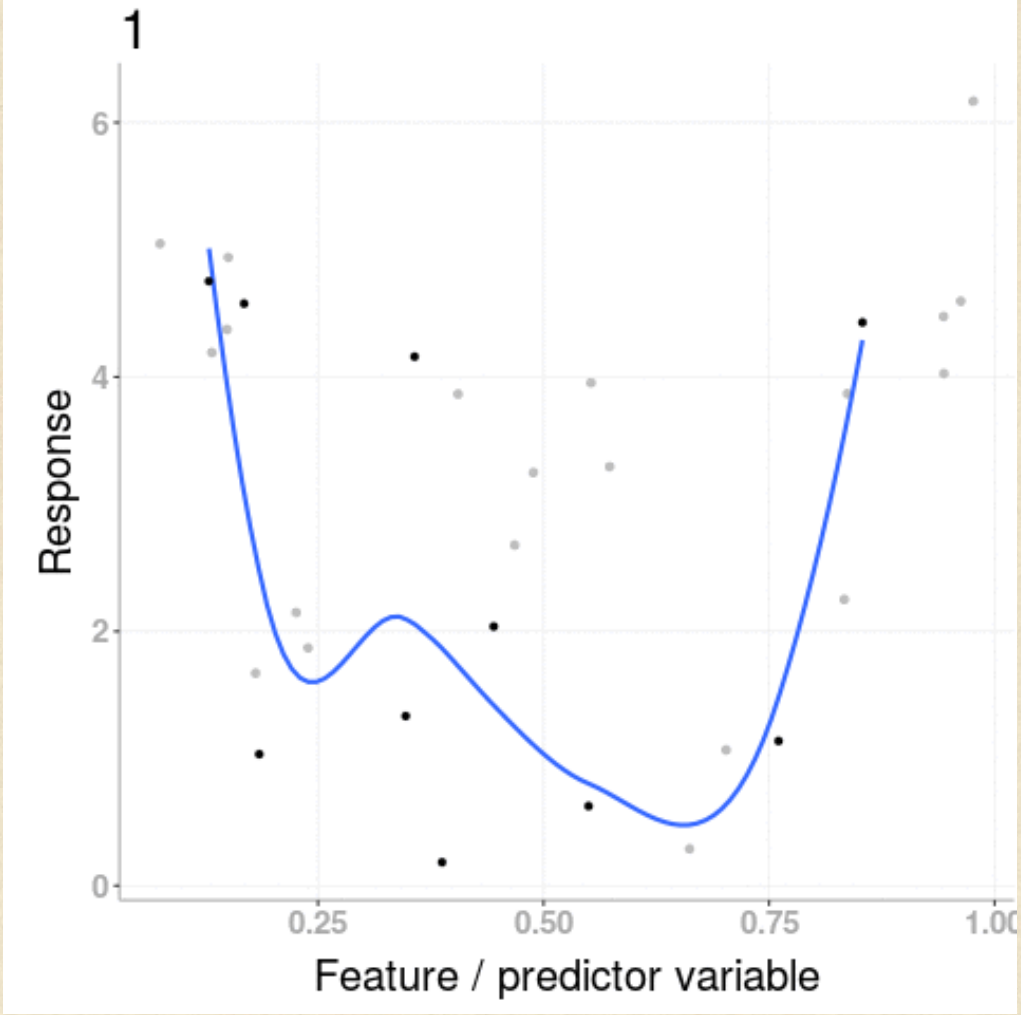
What is machine learning?

Only care about prediction

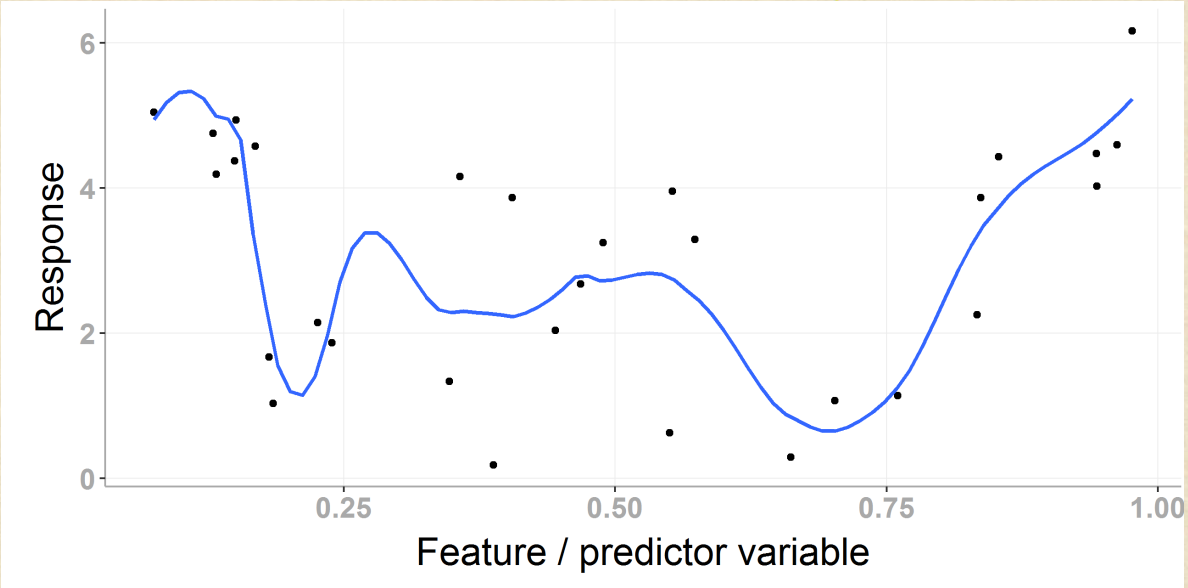
Not mechanistic/process based models

Not inference

Cross-validation



Hyperparameters



Hyperparameters

Number of PCA coordinates

Cut-offs for variable selection

$x + x^2 + x^3 + x^4 + ...$

No free lunch

No such thing as a universal, ‘best’ machine learning model.

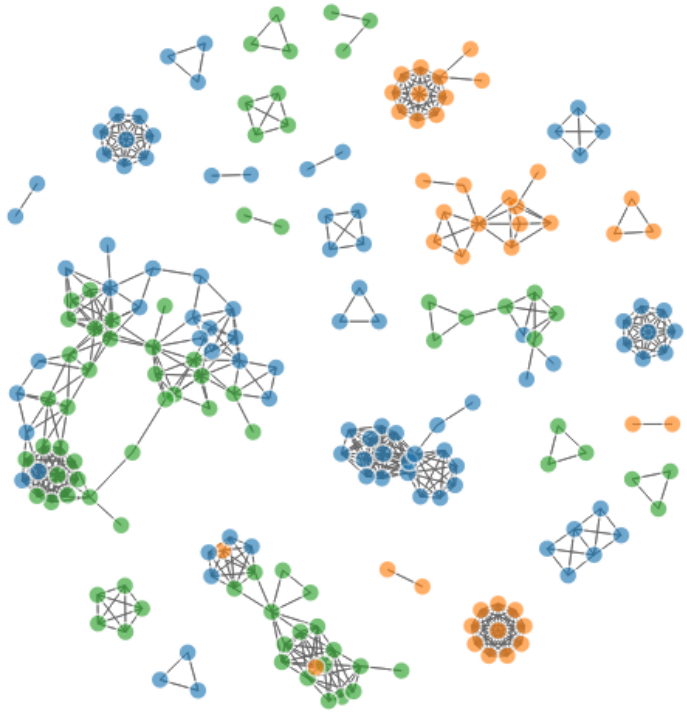
caret

https://topepo.github.io/caret/model-training-and-tuning.html

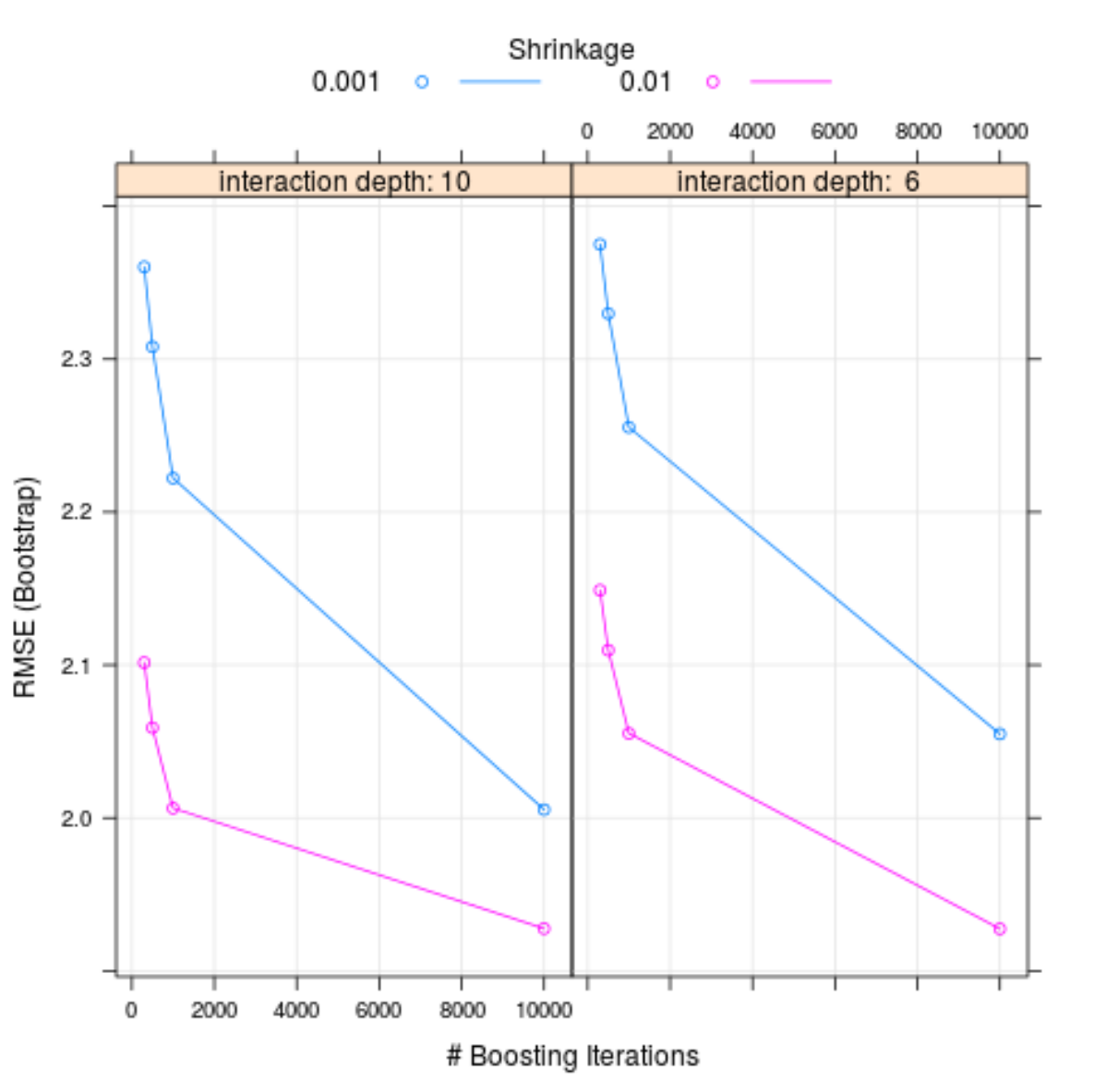
What does caret do?

8 Models Clustered by Tag Similarity

This page shows a network diagram of all the models that can be accessed by `train`. See the Revolutions blog for details about how this visualization was made (and this page has updated code using the `networkD3` package). In summary, the package annotates each model by a set of tags (e.g. "Bagging", "L1 Regularization" etc.). Using this information we can cluster models that are similar to each other.



What does caret do?

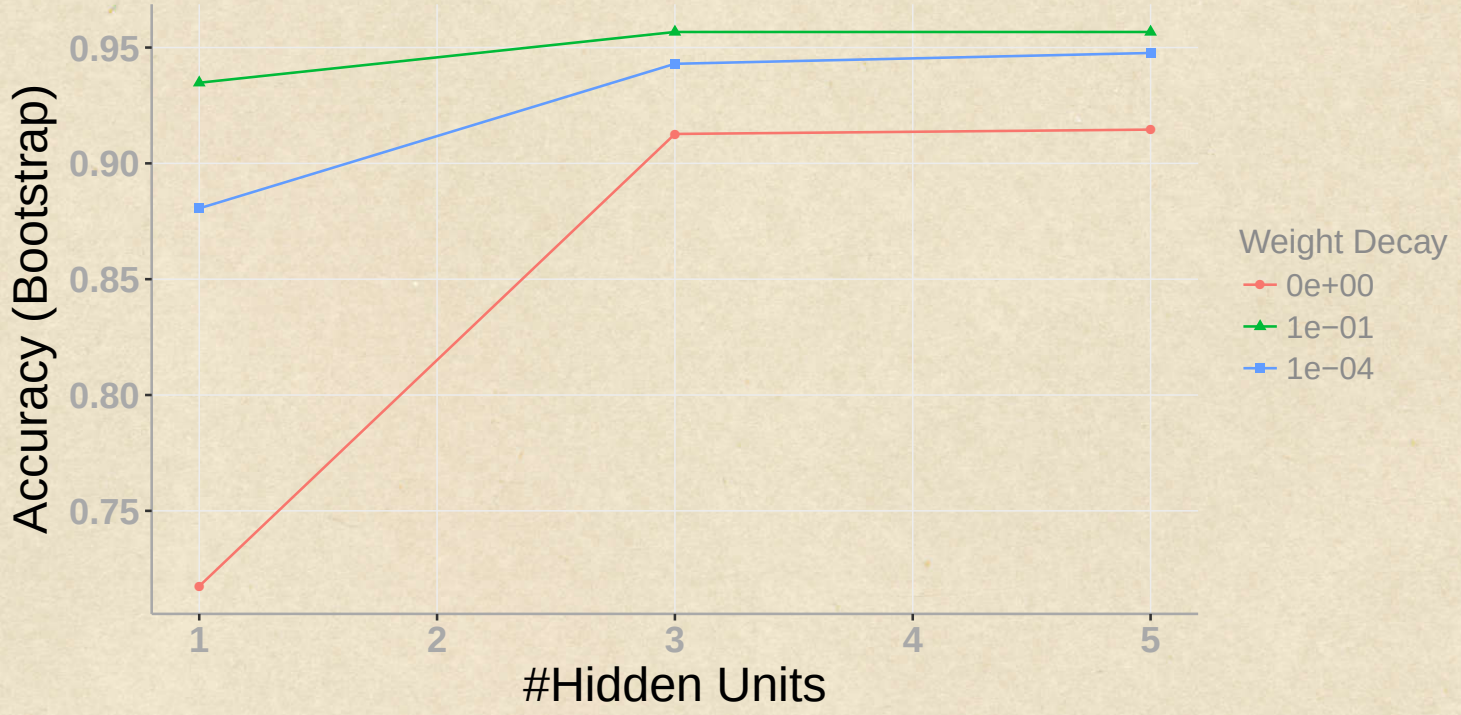


Training a model

```
m1 <- train(Species ~ .,
            iris,
            method = 'gbm')
```

Training a different model

```
m2 <- train(Species ~ .,
            iris,
            method = 'nnet')
```



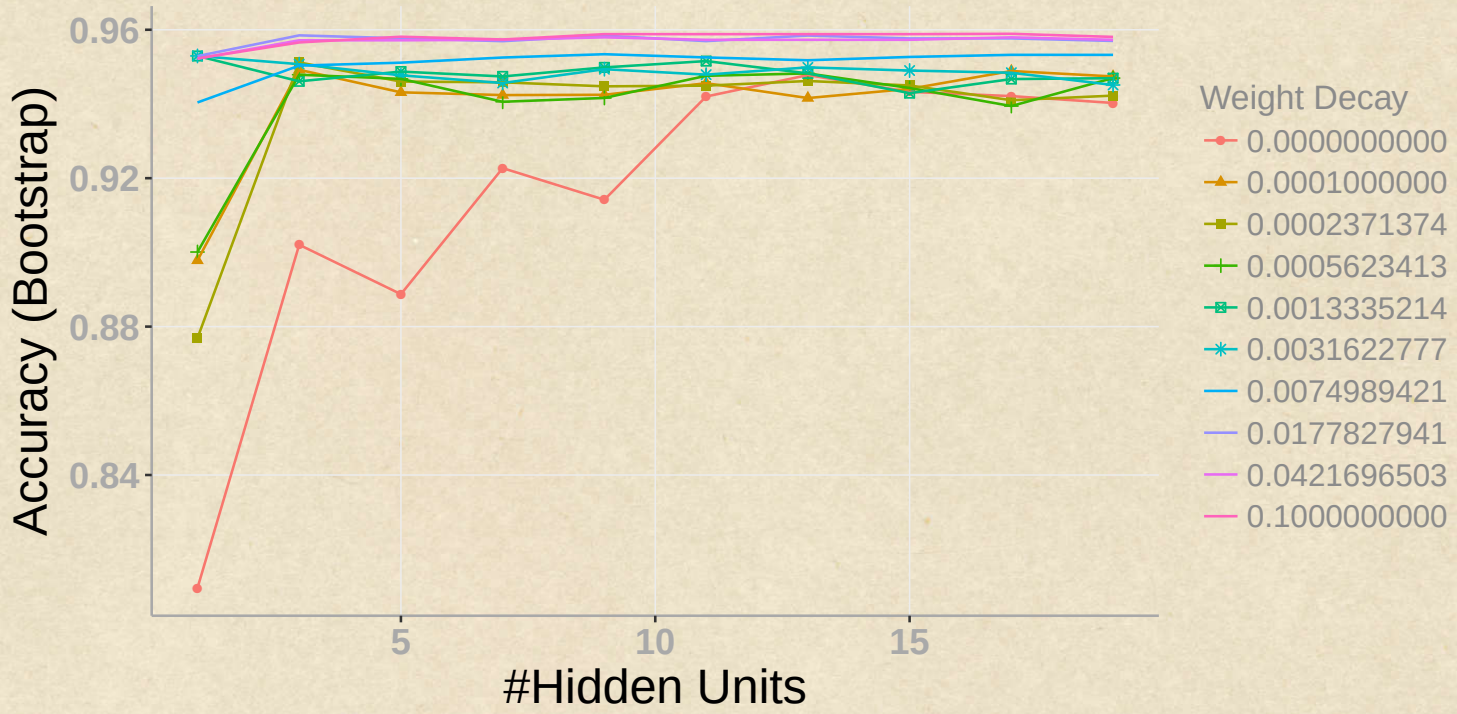
Controlling Crossvalidation

```
tr <- trainControl(method = 'cv', number = 5)

m3 <- train(Species ~ .,
            iris,
            trControl = tr,
            method = 'nnet')
```

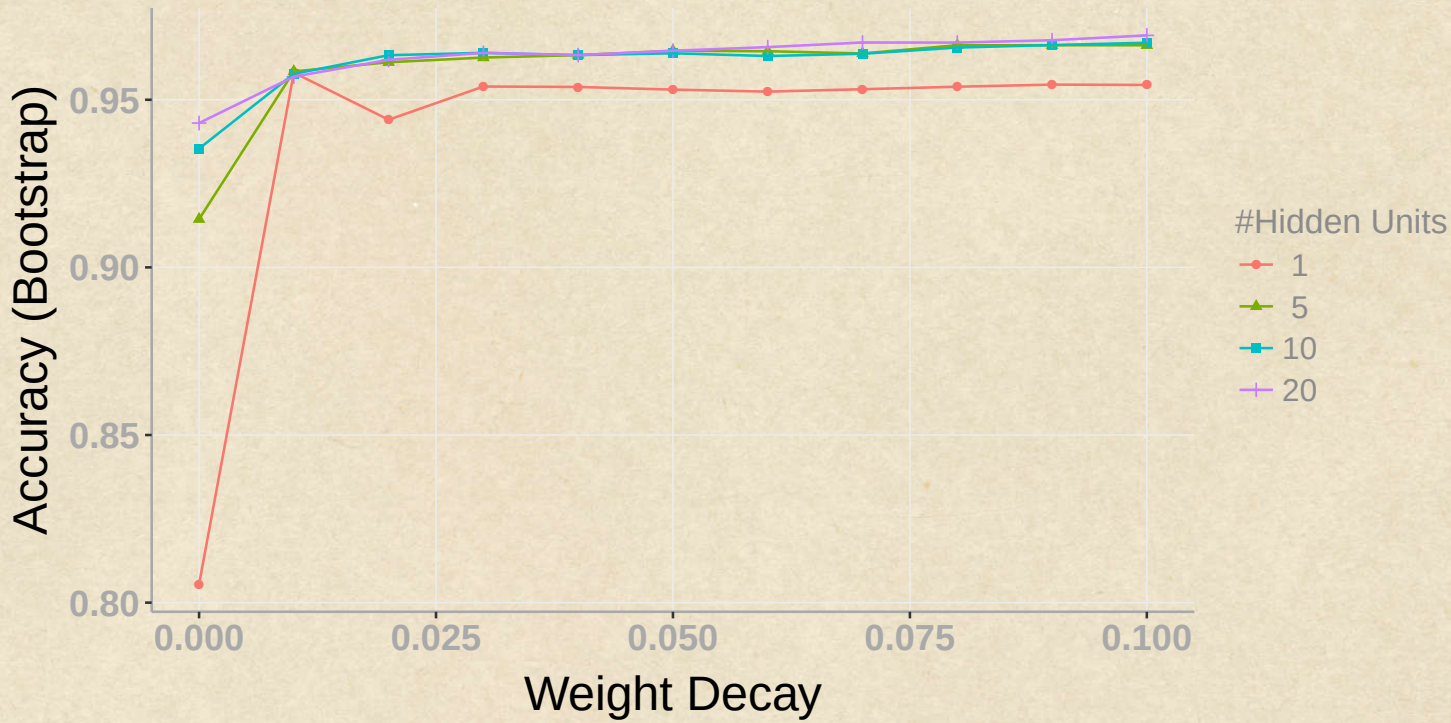
Try more hyperparameter values

```
m4 <- train(Species ~ .,
            iris,
            tuneLength = 10,
            method = 'nnet')
```



Use chosen hyperparameter values

```
m5 <- train(Species ~ .,
            iris,
            tuneGrid = expand.grid(size=c(1,5,10,20),
                                   decay=seq(0,0.1,0.01)),
            method = 'nnet')
```



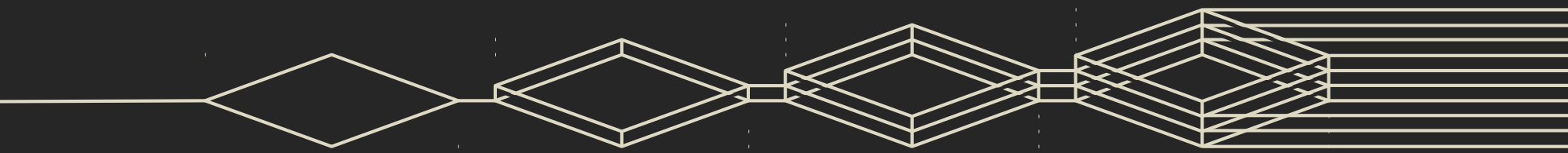
Contributions

Add your own models.

Share by github pull request.

But aim is for devs to keep package up to date.

ZOÖN



2020Science

Who develops zoon?

Tom August

Me

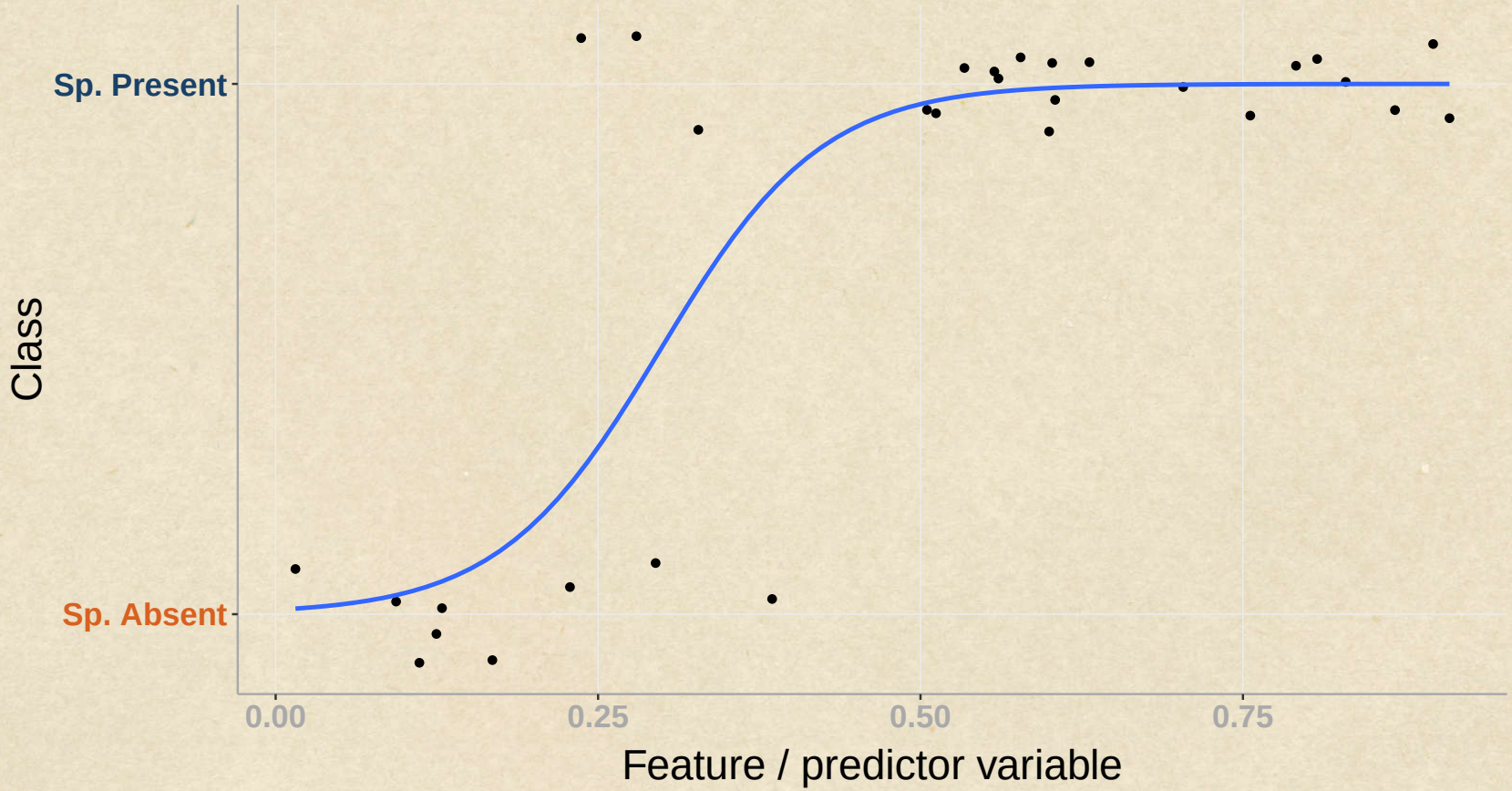
Nick Golding

Emiel van Loon

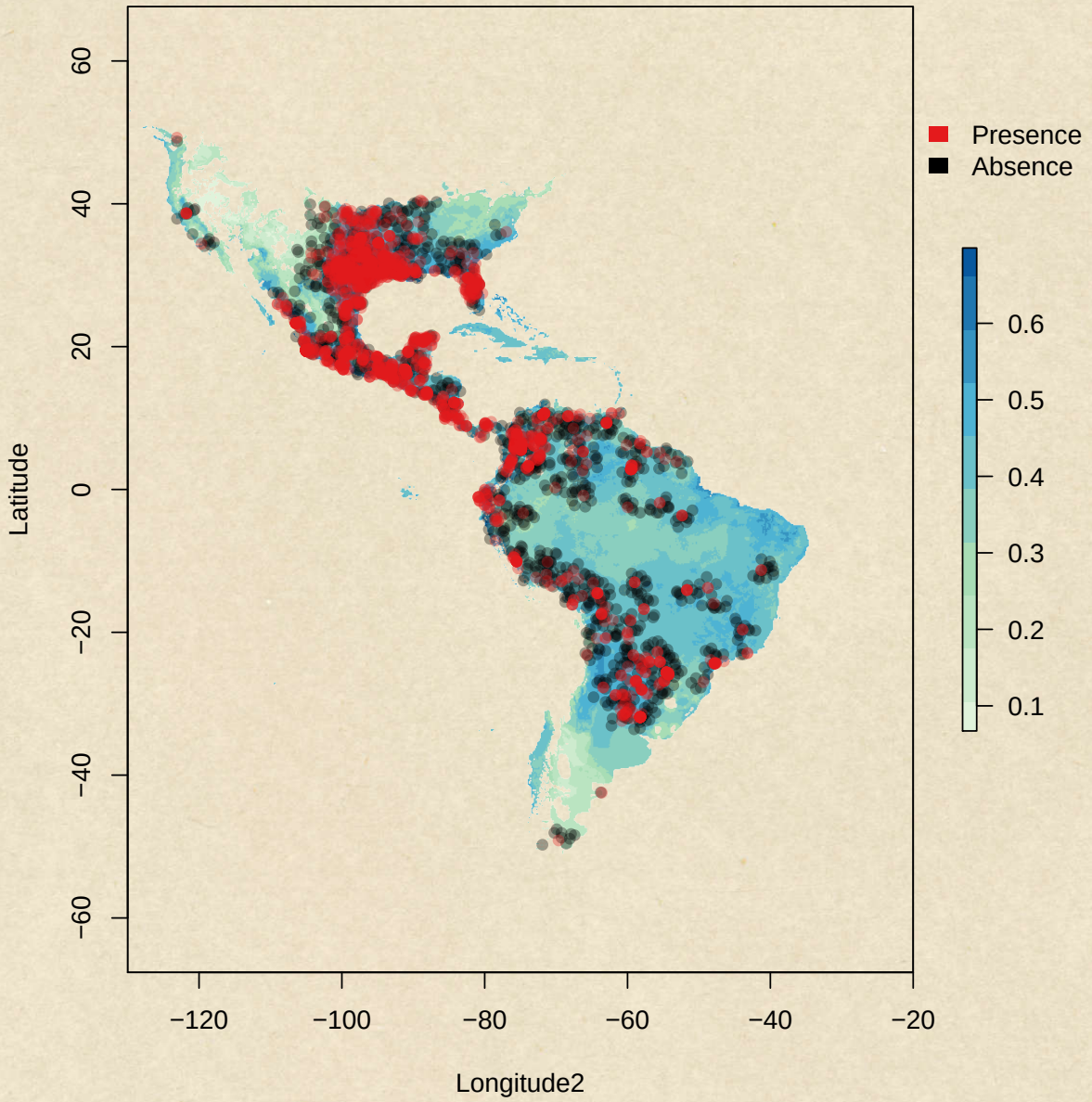
David Gavaghan

Greg McInerny

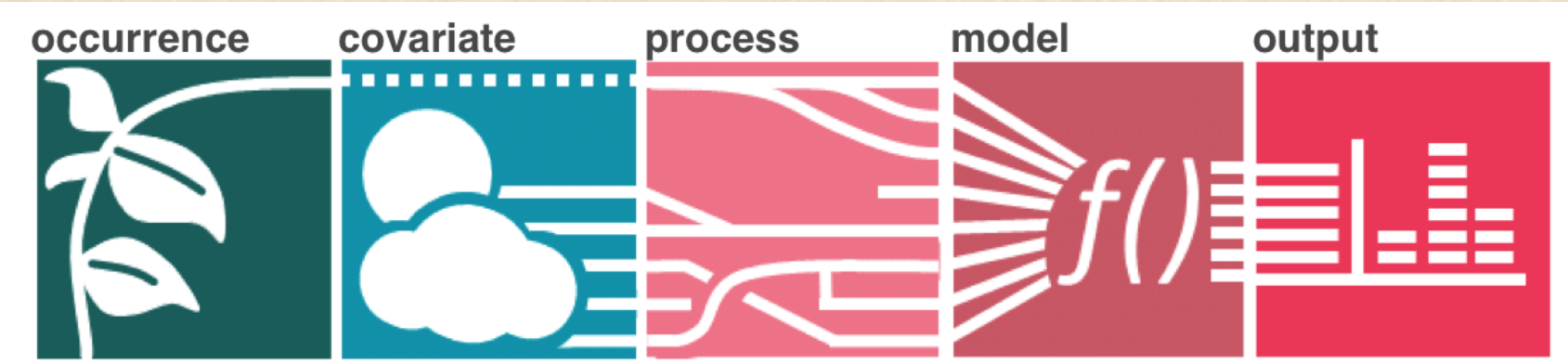
What is species distribution modelling?



What is species distribution modelling?

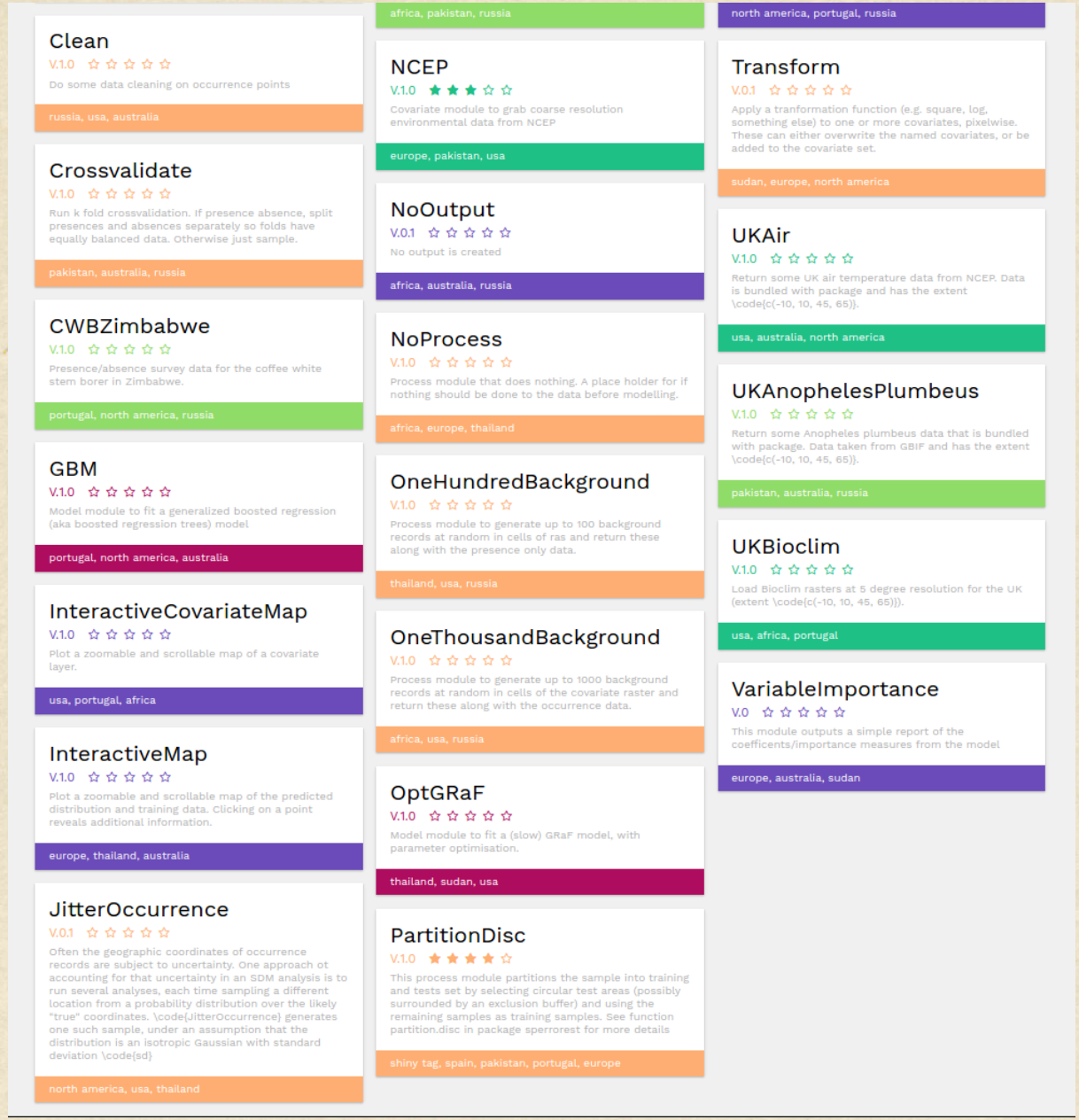


A basic workflow




```
work1 <- workflow(
  occurrence = UKAnophelesPlumbeus,
  covariate  = UKAir,
  process    = OneHundredBackground,
  model      = RandomForest,
  output     = PrintMap
)
```

What does zoon do?



africa, pakistan, russia

north america, portugal, russia

Clean
V.1.0 ☆ ☆ ☆ ☆ ☆

Do some data cleaning on occurrence points

russia, usa, australia

Crossvalidate
V.1.0 ☆ ☆ ☆ ☆ ☆

Run k fold crossvalidation. If presence absence, split presences and absences separately so folds have equally balanced data. Otherwise just sample.

pakistan, australia, russia

CWBZimbabwe
V.1.0 ☆ ☆ ☆ ☆ ☆

Presence/absence survey data for the coffee white stem borer in Zimbabwe.

portugal, north america, russia

GBM
V.1.0 ☆ ☆ ☆ ☆ ☆

Model module to fit a generalized boosted regression (aka boosted regression trees) model

portugal, north america, australia

InteractiveCovariateMap
V.1.0 ☆ ☆ ☆ ☆ ☆

Plot a zoomable and scrollable map of a covariate layer.

usa, portugal, africa

InteractiveMap
V.1.0 ☆ ☆ ☆ ☆ ☆

Plot a zoomable and scrollable map of the predicted distribution and training data. Clicking on a point reveals additional information.

europe, thailand, australia

JitterOccurrence
V.0.1 ☆ ☆ ☆ ☆ ☆

Often the geographic coordinates of occurrence records are subject to uncertainty. One approach ot accounting for that uncertainty in an SDM analysis is to run several analyses, each time sampling a different location from a probability distribution over the likely "true" coordinates. \code{JitterOccurrence} generates one such sample, under an assumption that the distribution is an isotropic Gaussian with standard deviation \code{sd}

north america, usa, thailand

NCEP
V.1.0 ★ ★ ★ ☆ ☆

Covariate module to grab coarse resolution environmental data from NCEP

europe, pakistan, usa

NoOutput
V.0.1 ☆ ☆ ☆ ☆ ☆

No output is created

africa, australia, russia

NoProcess
V.1.0 ☆ ☆ ☆ ☆ ☆

Process module that does nothing. A place holder for if nothing should be done to the data before modelling.

africa, europe, thailand

OneHundredBackground
V.1.0 ☆ ☆ ☆ ☆ ☆

Process module to generate up to 100 background records at random in cells of ras and return these along with the presence only data.

thailand, usa, russia

OneThousandBackground
V.1.0 ☆ ☆ ☆ ☆ ☆

Process module to generate up to 1000 background records at random in cells of the covariate raster and return these along with the occurrence data.

africa, usa, russia

OptGRaF
V.1.0 ☆ ☆ ☆ ☆ ☆

Model module to fit a (slow) GRaF model, with parameter optimisation.

thailand, sudan, usa

PartitionDisc
V.1.0 ★ ★ ★ ★ ☆

This process module partitions the sample into training and tests set by selecting circular test areas (possibly surrounded by an exclusion buffer) and using the remaining samples as training samples. See function partition.disc in package sperrorest for more details

shiny tag, spain, pakistan, portugal, europe

Transform
V.0.1 ☆ ☆ ☆ ☆ ☆

Apply a tranformation function (e.g. square, log, something else) to one or more covariates, pixelwise. These can either overwrite the named covariates, or be added to the covariate set.

sudan, europe, north america

UKAir
V.1.0 ☆ ☆ ☆ ☆ ☆

Return some UK air temperature data from NCEP. Data is bundled with package and has the extent \code{c(-10, 10, 45, 65)}.

usa, australia, north america

UKAnophelesPlumbeus
V.1.0 ☆ ☆ ☆ ☆ ☆

Return some Anopheles plumbeus data that is bundled with package. Data taken from GBIF and has the extent \code{c(-10, 10, 45, 65)}.

pakistan, australia, russia

UKBioclim
V.1.0 ☆ ☆ ☆ ☆ ☆

Load Bioclim rasters at 5 degree resolution for the UK (extent \code{c(-10, 10, 45, 65)}).

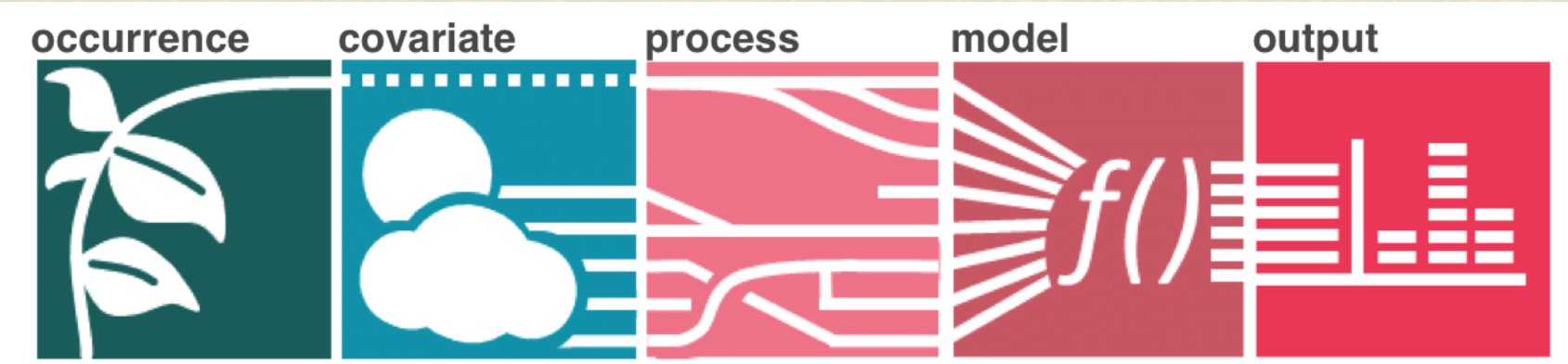
usa, africa, portugal

VariableImportance
V.0 ☆ ☆ ☆ ☆ ☆

This module outputs a simple report of the coefficents/importance measures from the model

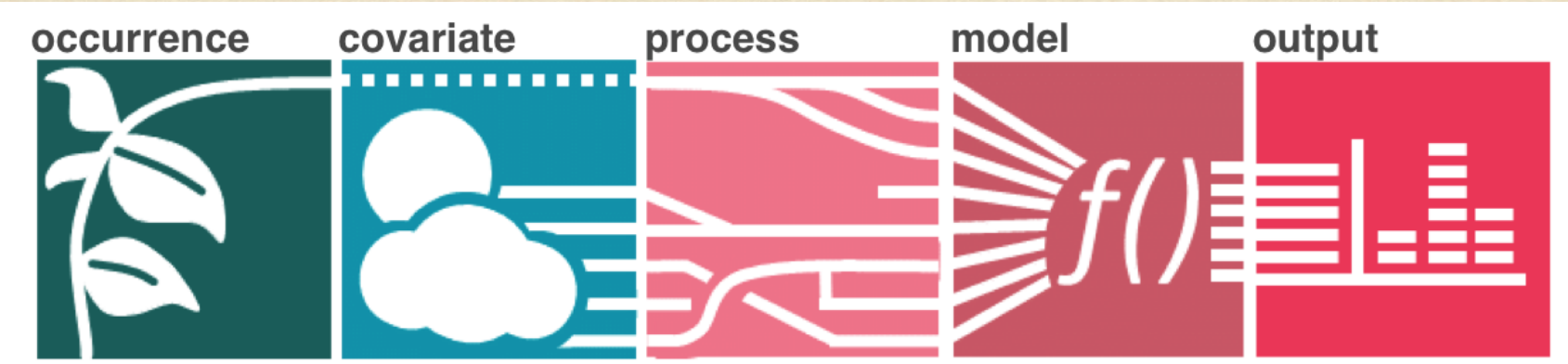
europe, australia, sudan

A different model



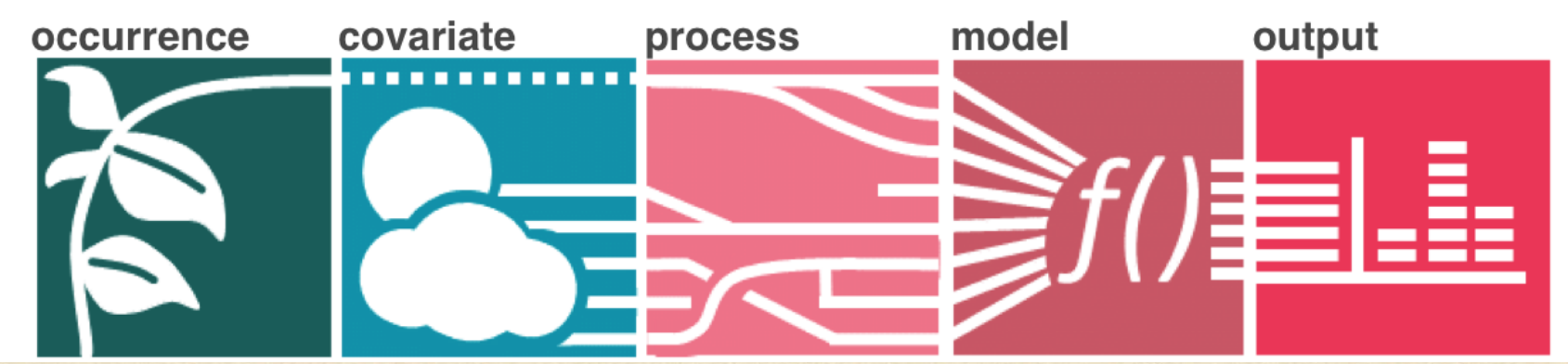

```
work2 <- workflow(
  occurrence = UKAnophelesPlumbeus,
  covariate  = UKAir,
  process    = OneHundredBackground,
  model      = MaxEnt,
  output     = PrintMap
)
```

A different workflow



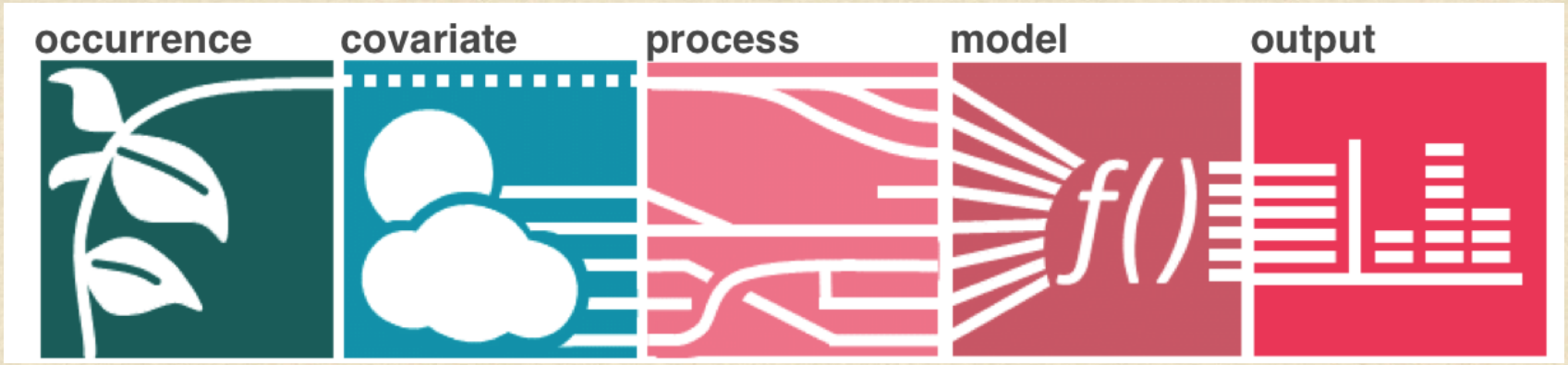

```
work3 <- workflow(
  occurrence = UKAnophelesPlumbeus,
  covariate  = UKBioclim,
  process    = Background(n = 500),
  model      = RandomForest,
  output     = Appify
)
```

caret in zoon




```
work4 <- workflow(
  occurrence = UKAnophelesPlumbeus,
  covariate  = UKAir,
  process    = OneHundredBackground,
  model      = MachineLearn(method = 'nnet',
                            tuneLength = 8),
  output     = PrintMap
)
```

A more complicated workflow




```
work5 <- workflow(
  occurrence = SpOcc(species = 'Eresus kollari',
                     extent = c(-10, 10, 45, 65)),
  covariate  = UKBioclim,
  process    = BackgroundAndCrossvalid(k = 5),
  model      = list(LogisticRegression,
                    RandomForest),
  output     = Chain(PrintMap(plot = FALSE),
                     PerformanceMeasures)
)
```

Contributions

Add your own methods.

Share by web form or github.

Not the aim for devs to keep package up to date.

Cue hand-waving

Package ecosystems

CRAN zoon dismo

Bioconductor caret

Package ecosystems

CRAN zoon dismo

Bioconductor caret

User contribution

Extendability

Any Questions ?

Tim C.D. Lucas