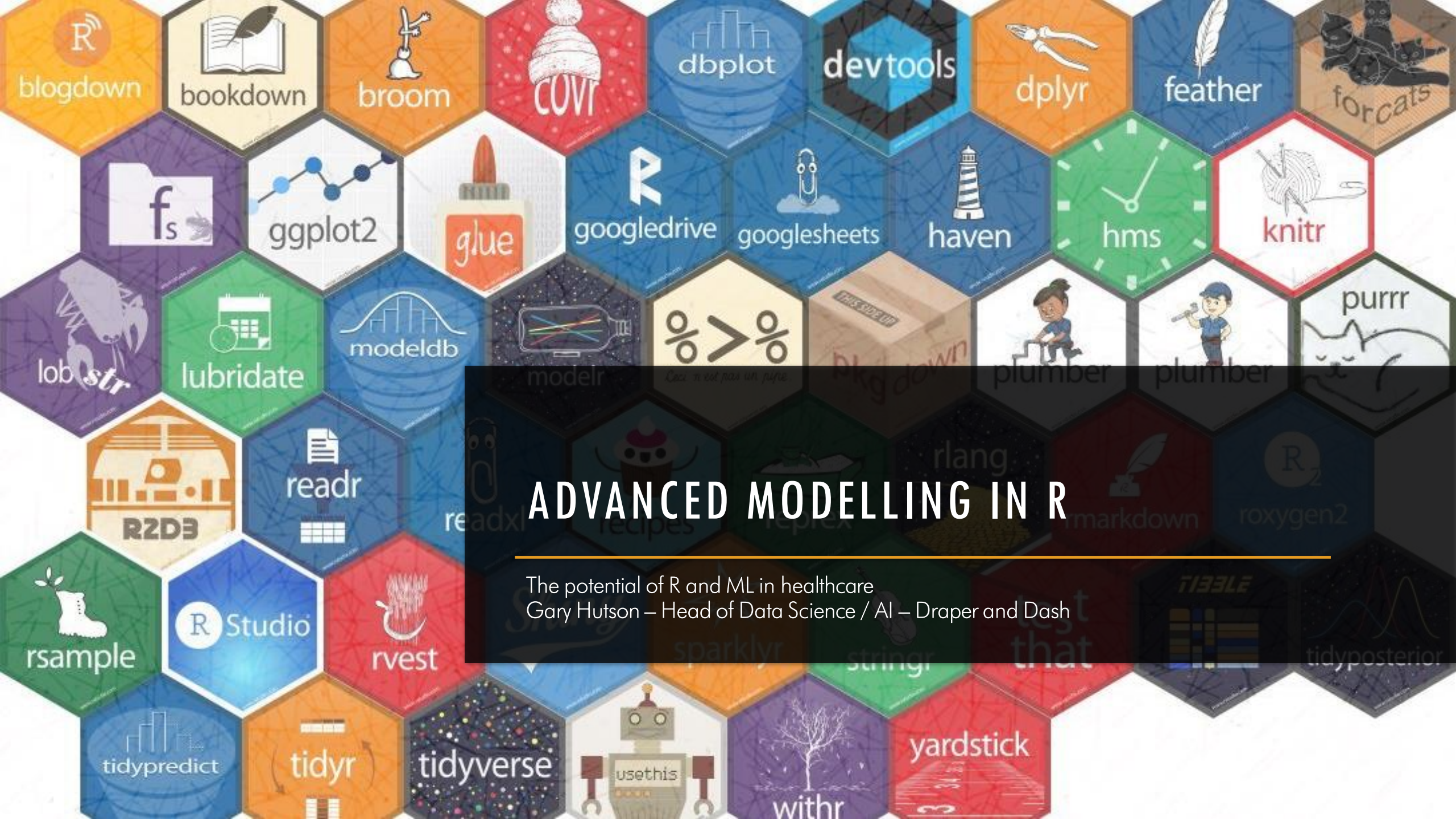


ADVANCED MODELLING IN R

The potential of R and ML in healthcare

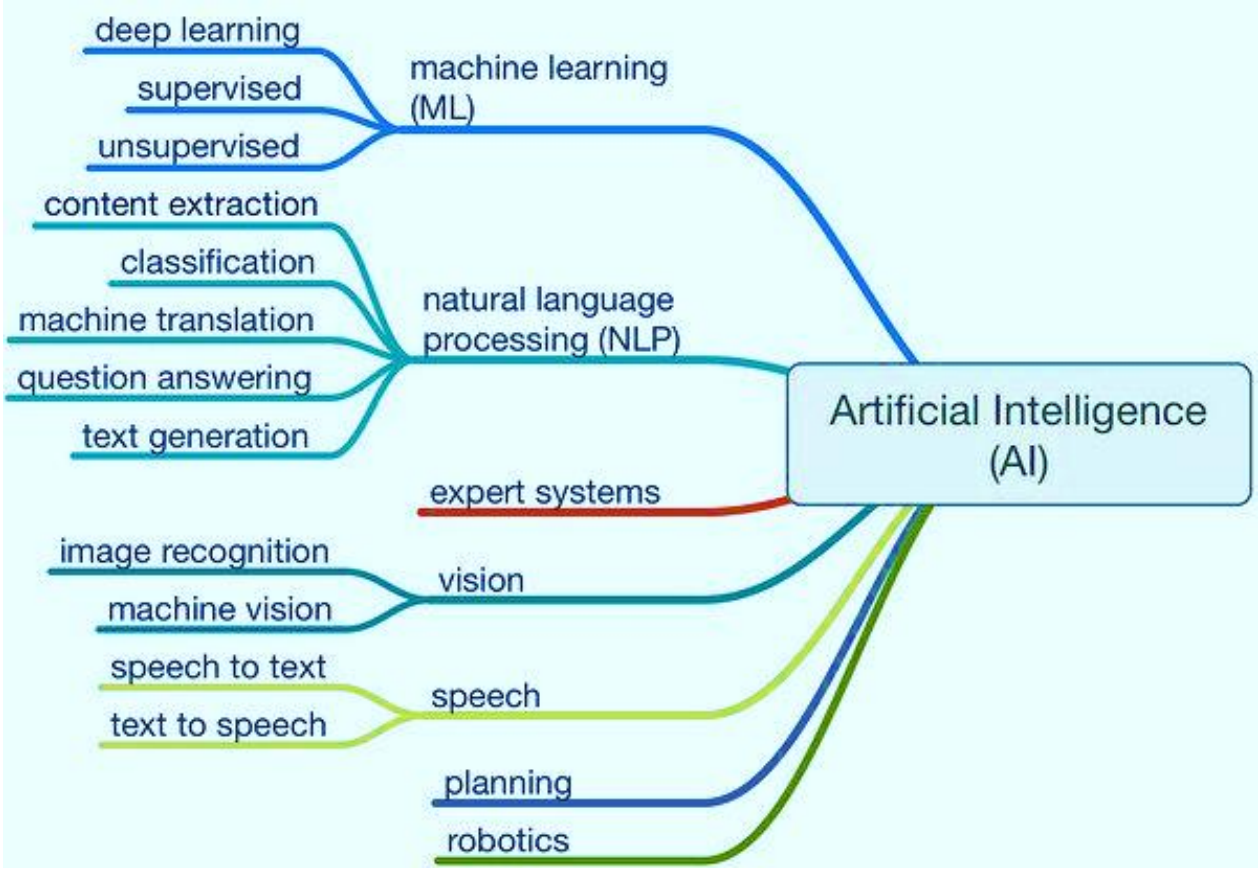
Gary Hutson – Head of Data Science / AI – Draper and Dash

AI OR ML – WHAT IS THE DIFFERENCE

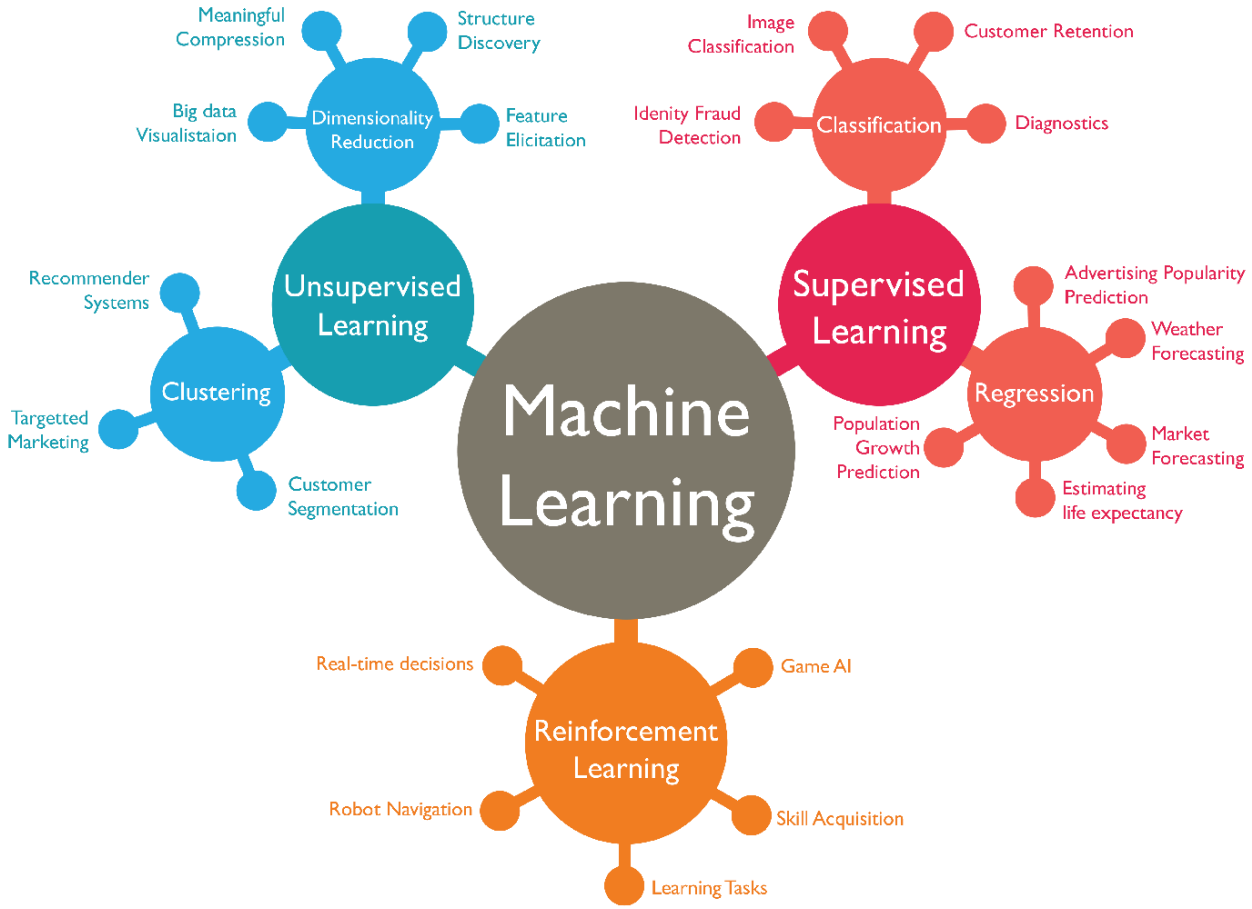
Definition:

Artificial Intelligence (AI) – “It is the study of how to train machines or computers, in order to do things, which at present humans can do better”

Machine Learning (ML) – “the study of computer algorithms that allow computer programs to automatically improve through experience”



WHAT MAKES UP AI?



TYPES OF ML AND USES

R	Python
Supervised and Unsupervised Learning	Supervised and Unsupervised Learning
Natural Language Processing aided by Tidytext and tm packages	Natural Language Processing
Deep Learning – packages such as H2O, Keras and MxNet	Deep Learning and Computer vision (OpenCV was designed for Python and R does not have a suitable
Data mangling, wrangling, encoding, cleaning and visualization	Data mangling, wrangling, feature encoding, cleaning and visualization
	Reinforcement Learning – agent vs reward Q Learning models
	Object orientated platform is easier to use than S3 and S4 in R.

R VS PYTHON – THE GRAND DATA SCIENTIST DEBATE

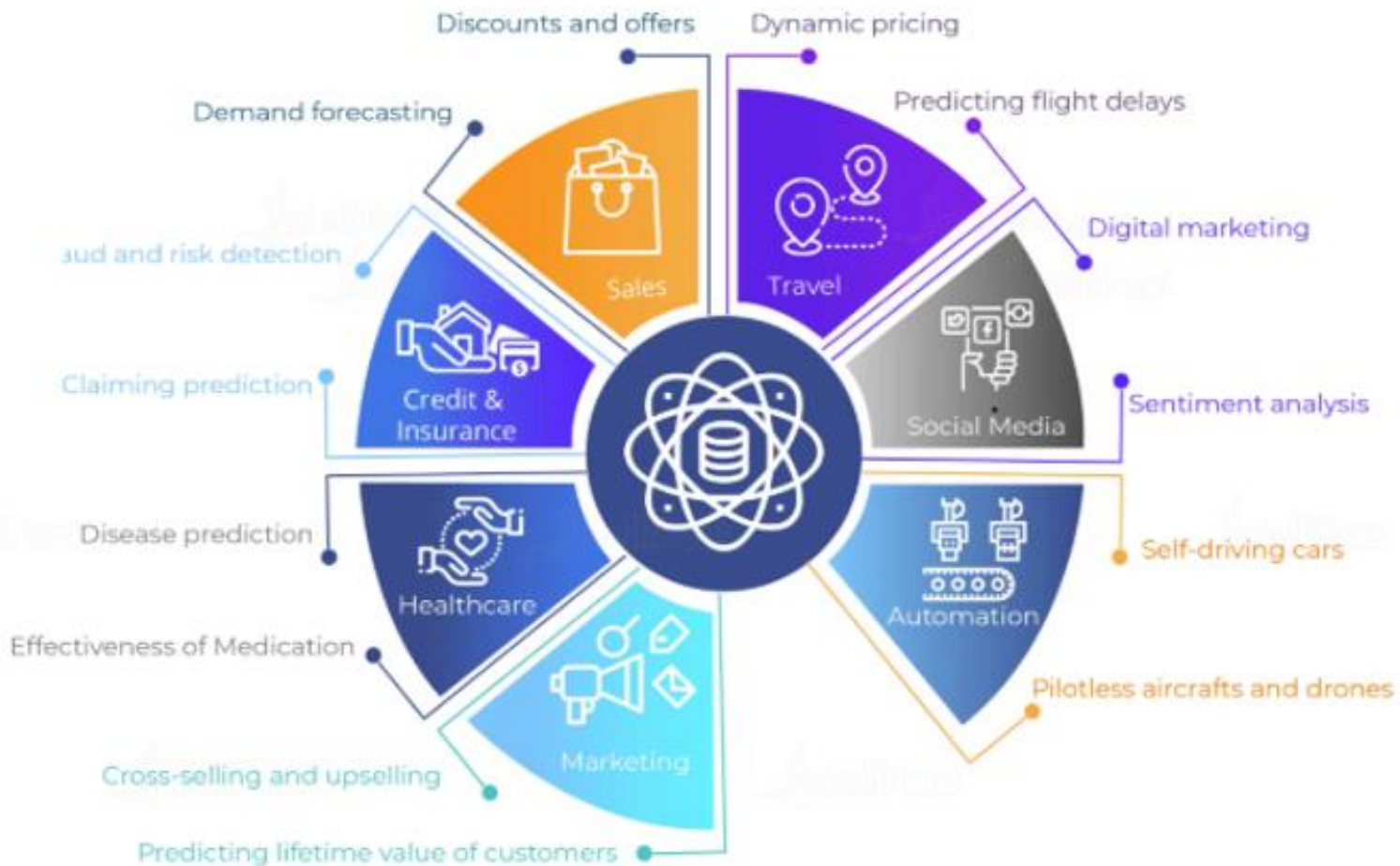
HOW CAN ML BE USED TO AID IN DECISION MAKING AND CHALLENGING ORGANIZATIONAL ISSUES?

APPLICATION IN HEALTHCARE AND BEYOND

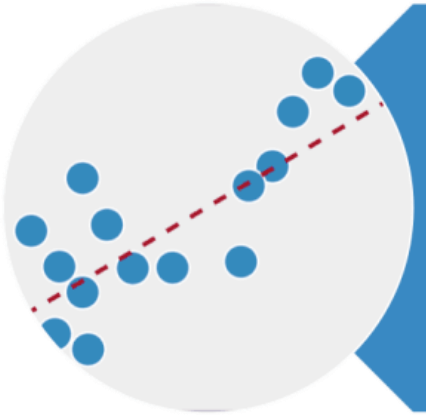
Some example:

- Decision Impact Support Tools – supervised learning models are increasingly being used to look at the relationship multiple variables have on a specific outcome measure or measures – examples are admission prediction tools, readmission avoidance tools, LOS prediction tools, mortality predictors, stranded patient predictors, COVID-19 probabilistic estimators based off similar patient demographics and characteristics, etc.
- Imaging augmentation tools – computer vision algorithms utilized to detect irregular scans i.e. presence of certain conditions using convolutional deep learning
- Deep neural networks will be used to expose hidden relationships and underlying features in all activity, access, diagnostic and other hospital datasets to drive more accurate and precise predictions
- Reinforcement learning will be used to optimize prescription regimes and will be able to work out the reward (effect) associated with every patient in clinical trials

WIDER USAGE

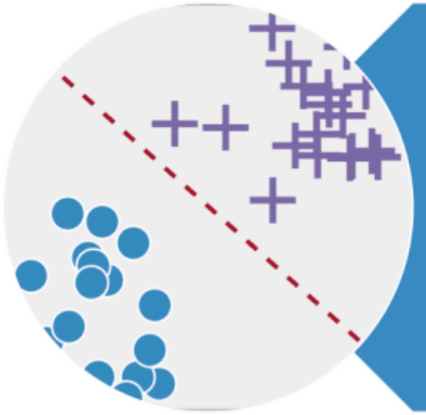

Discounts and offers
Dynamic pricing
Demand forecasting
Predicting flight delays
Digital marketing
Sales
Travel
Credit & Insurance
Social Media
Claiming prediction
Sentiment analysis
Disease prediction
Self-driving cars
Healthcare
Automation
Effectiveness of Medication
Pilotless aircrafts and drones
Marketing
Cross-selling and upselling
Predicting lifetime value of customers

SUPERVISED LEARNING



Regression

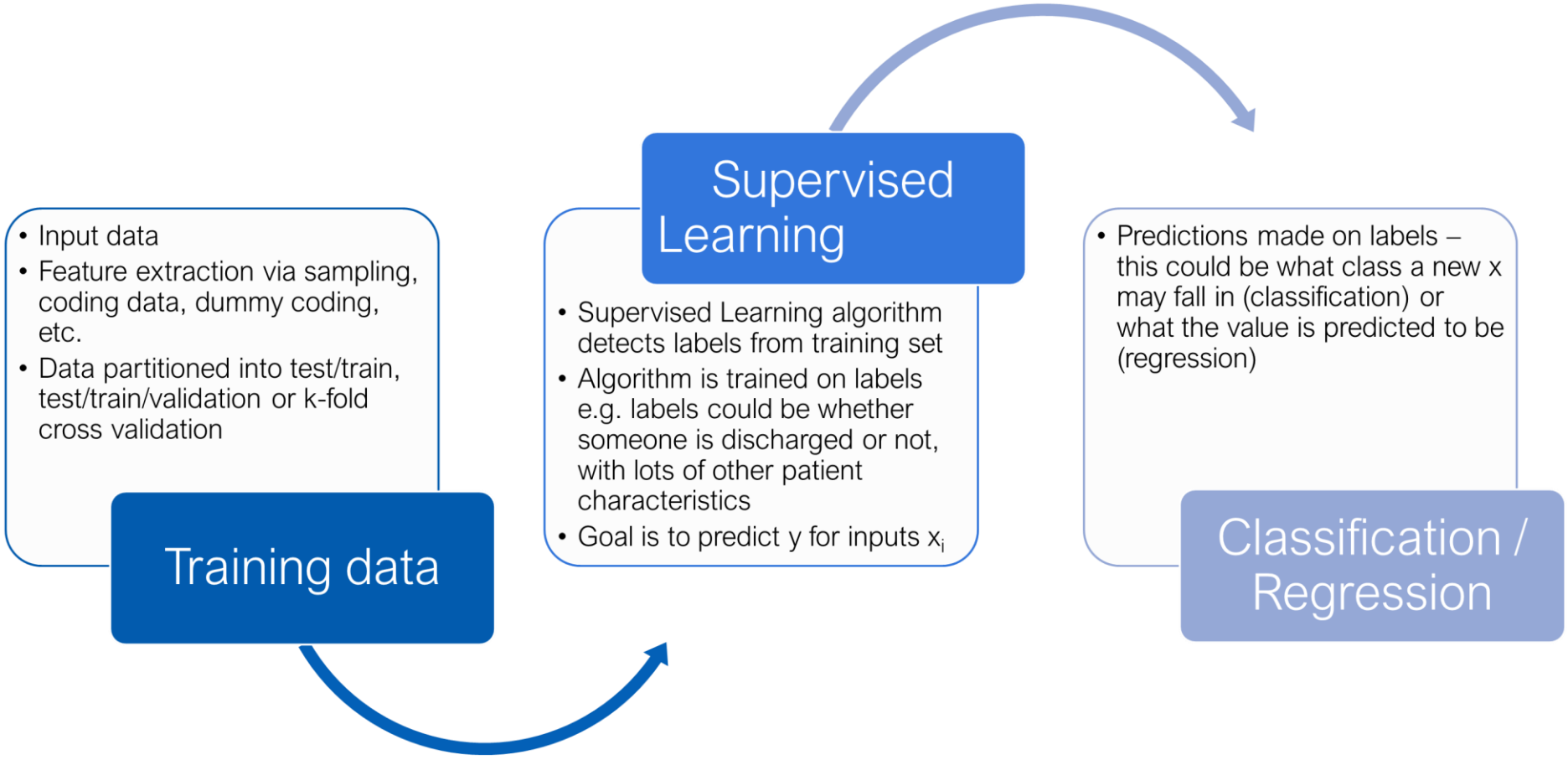
- Linear Regression
- Random Forest
- Multi-layer Perceptron
- AdaBoost
- Gradient Boosting
- Convolutional Neural Networks

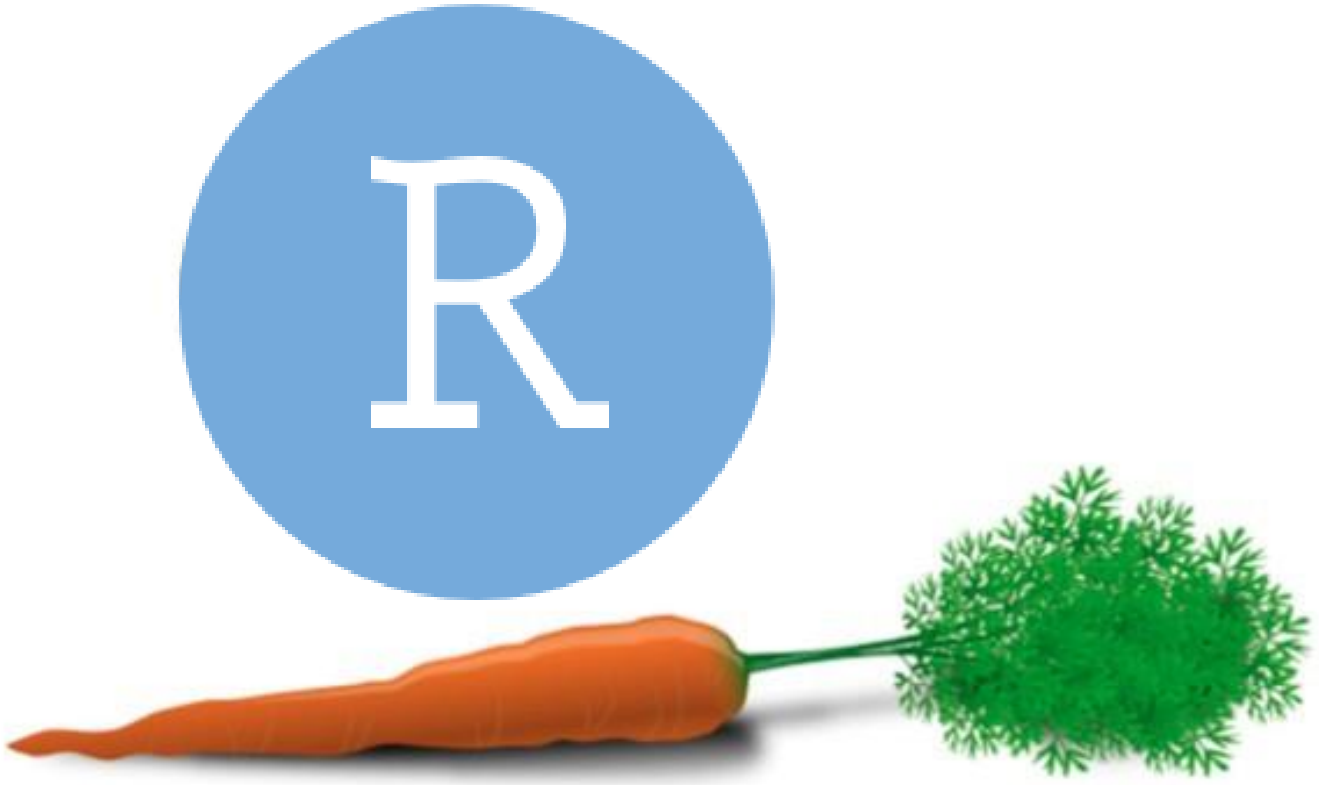


Classification

- Logistic Regression
- Decision Tree
- KNN
- Support vector machines
- Naive Bayes
- Convolutional Neural Networks

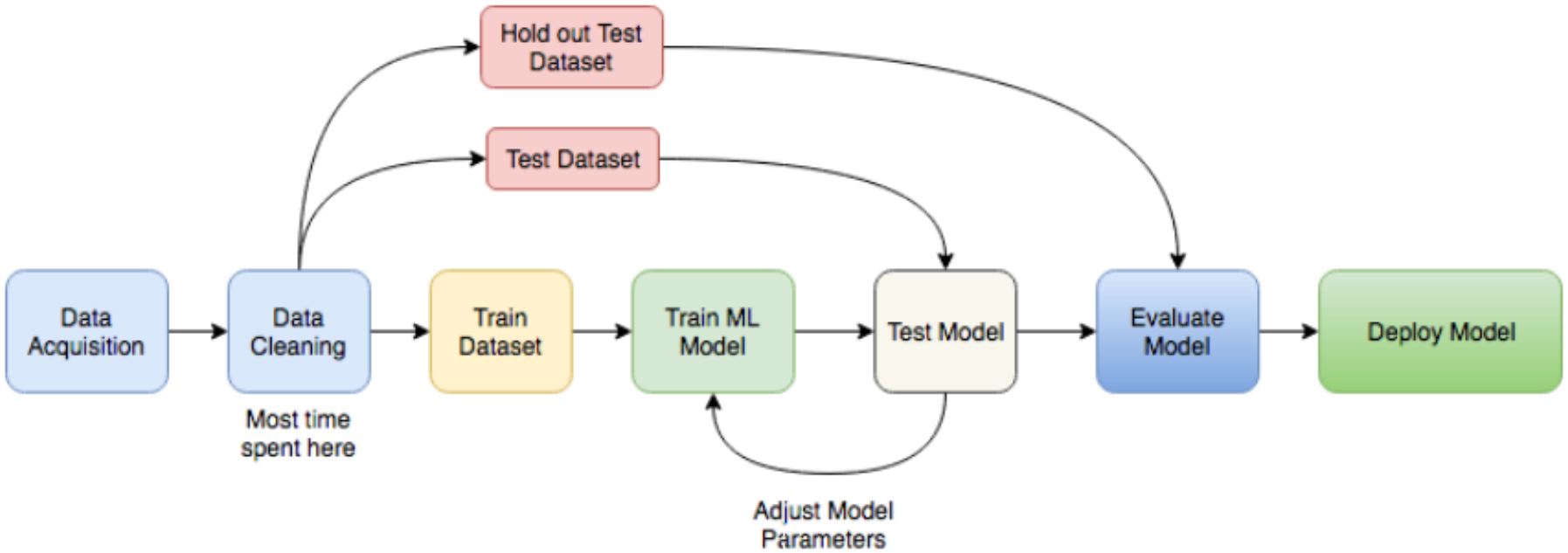
SUPERVISED LEARNING – WHAT IS IT?

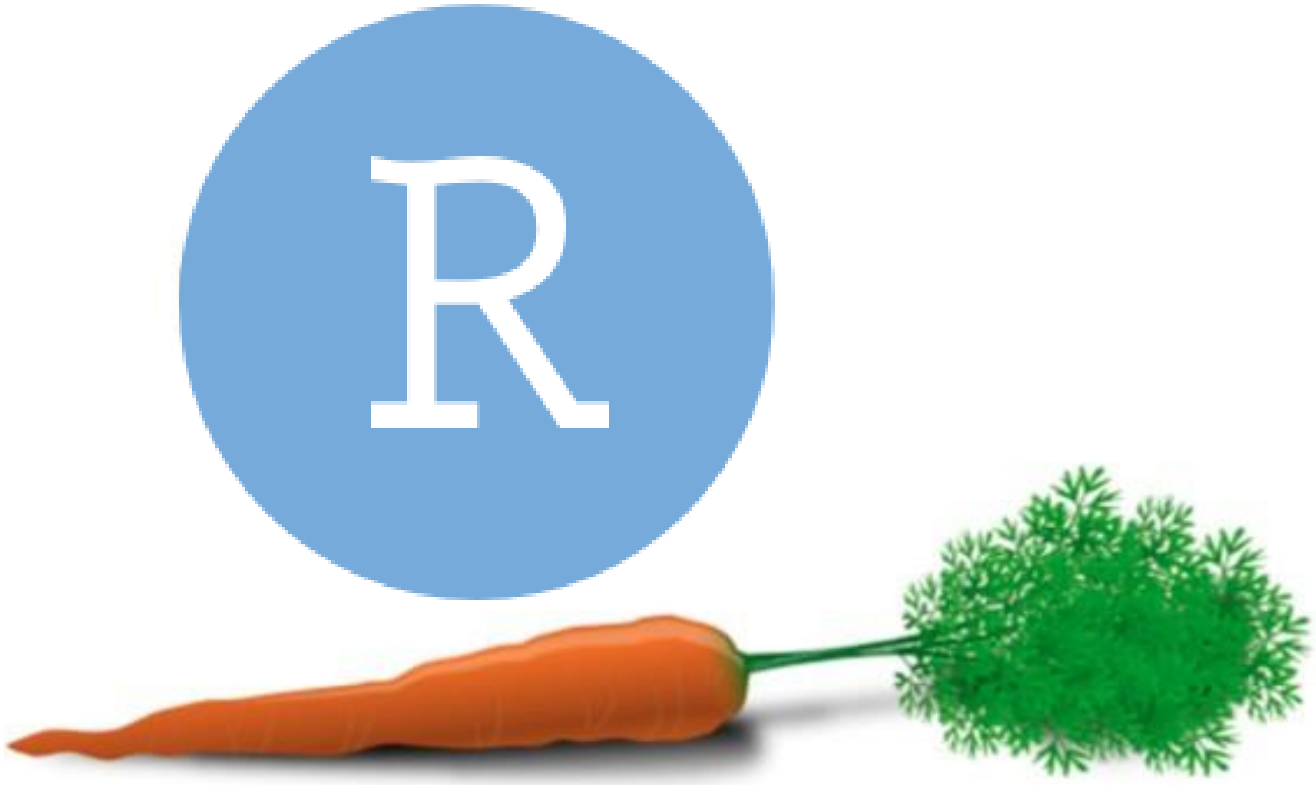




LET'S BUILD A ML MODEL WITH CARET

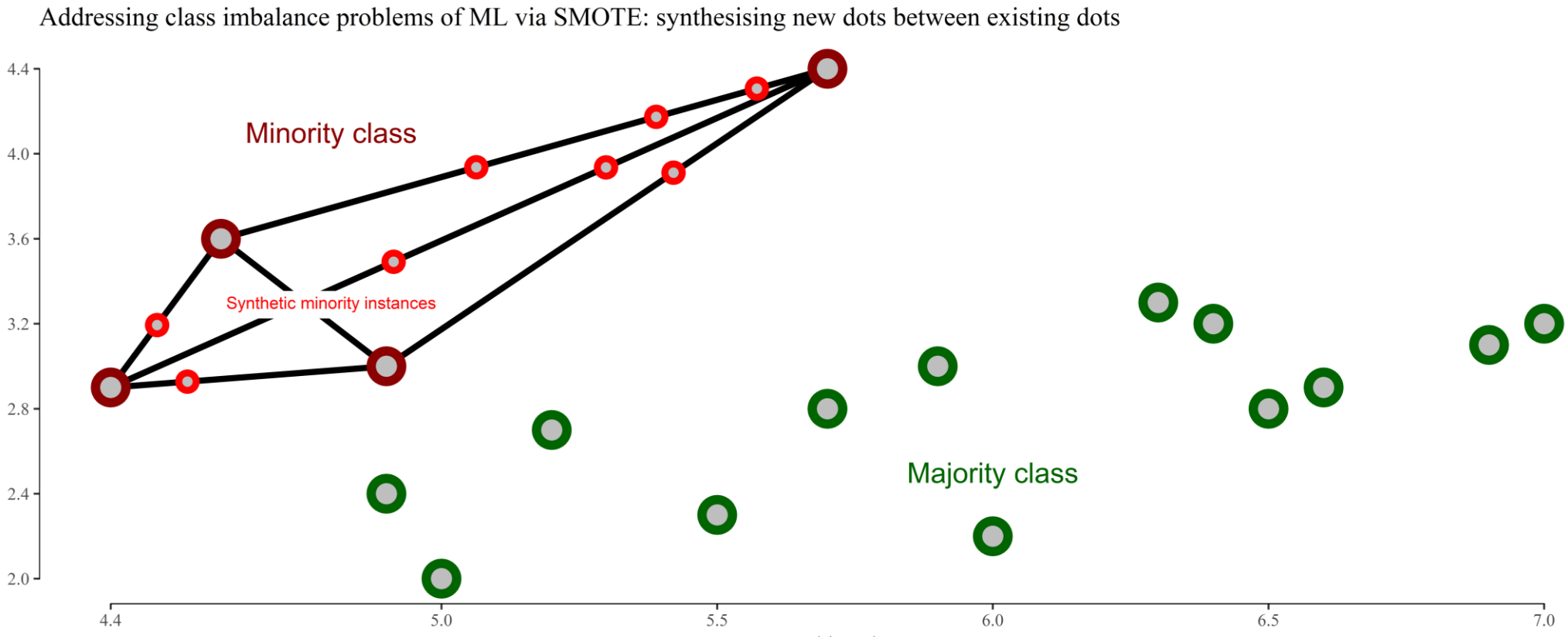
ML PROCESS


Hold out Test Dataset
Test Dataset
Data Acquisition
Data Cleaning
Most time spent here
Train Dataset
Train ML Model
Test Model
Evaluate Model
Deploy Model
Adjust Model Parameters

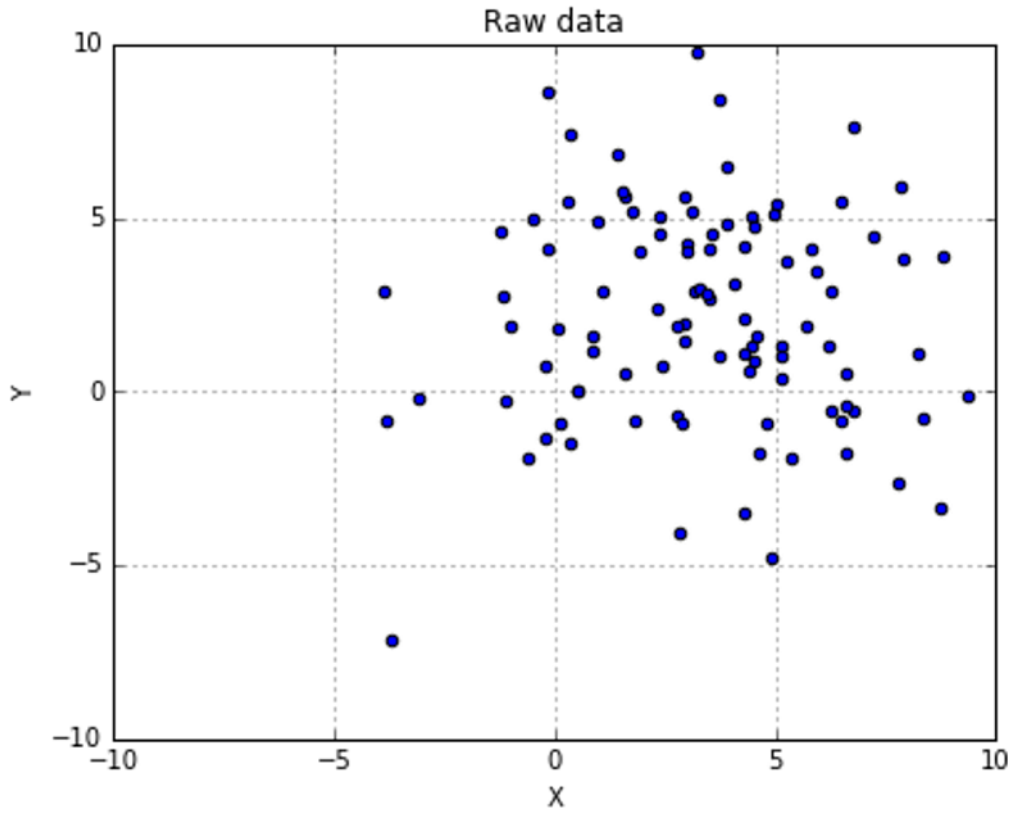


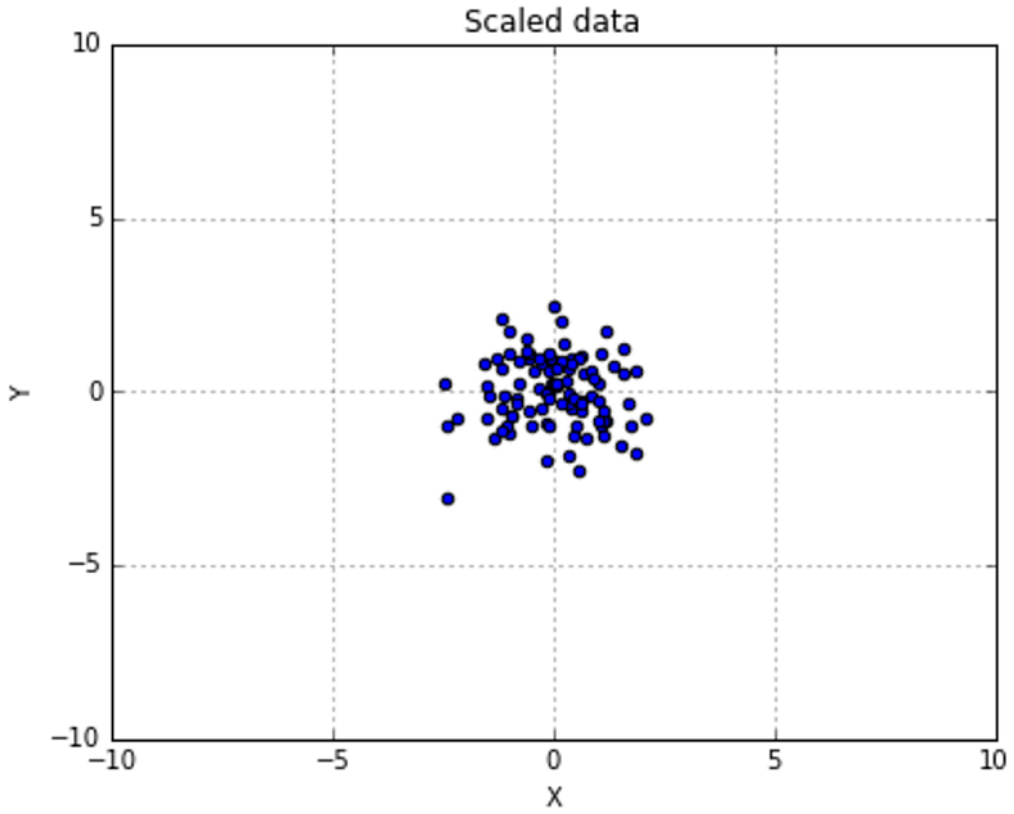
CLASSIFICATION MODELLING

CLASSIFICATION IMBALANCE

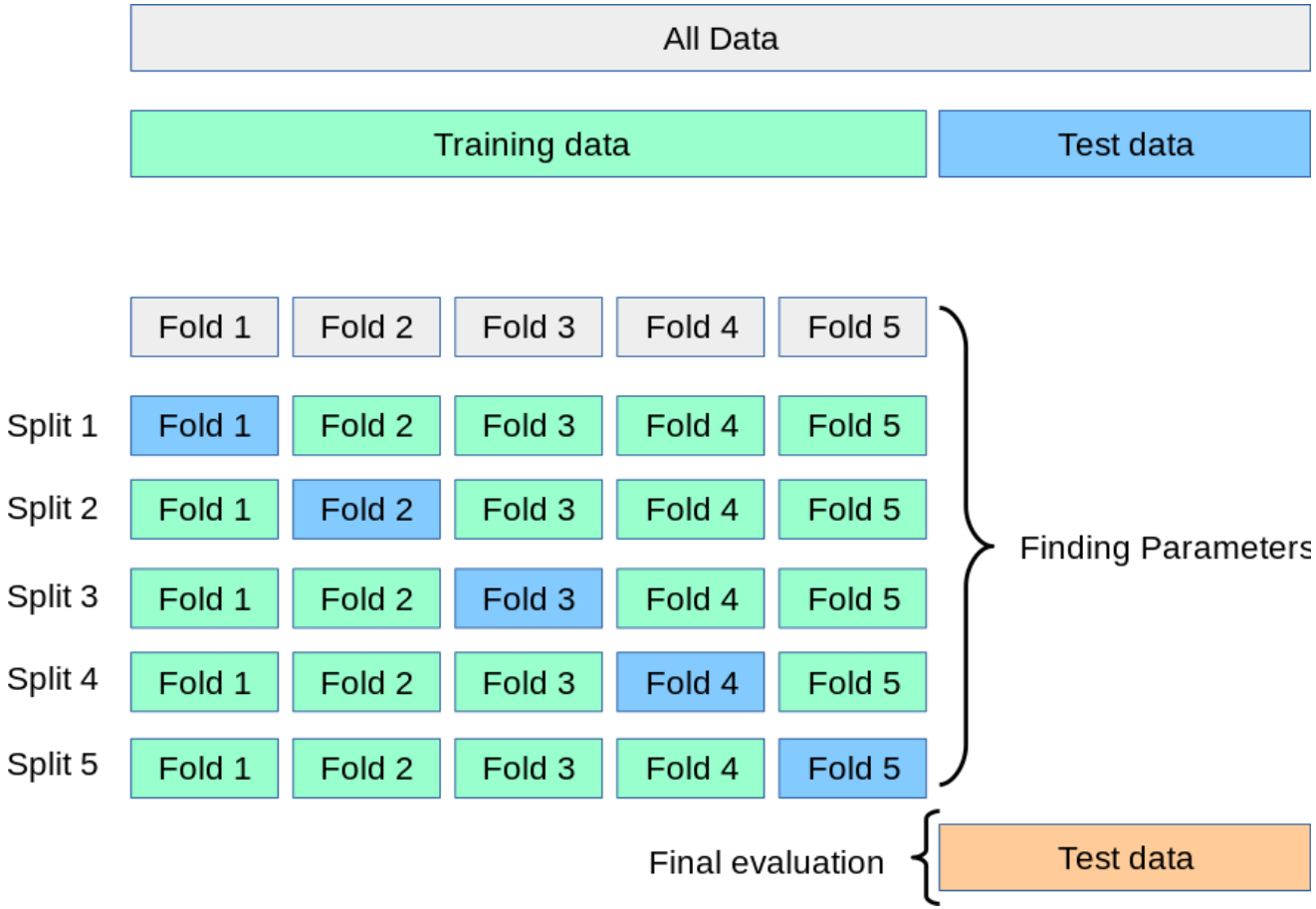


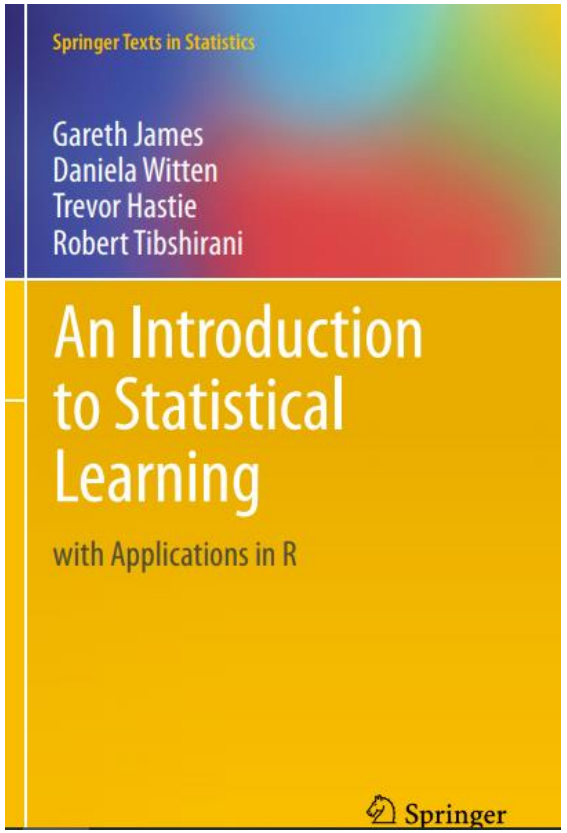
DATA SCALING



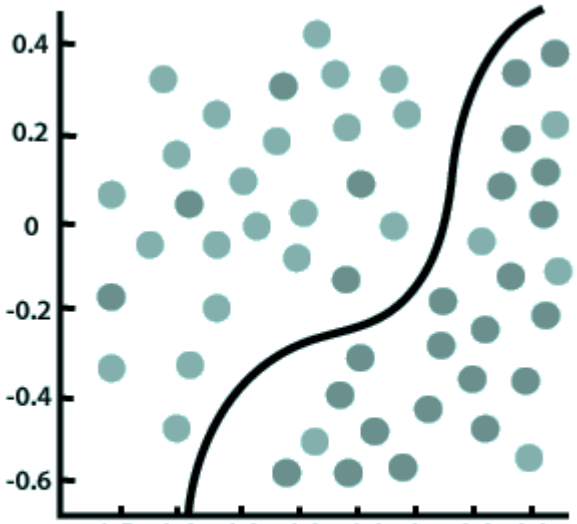


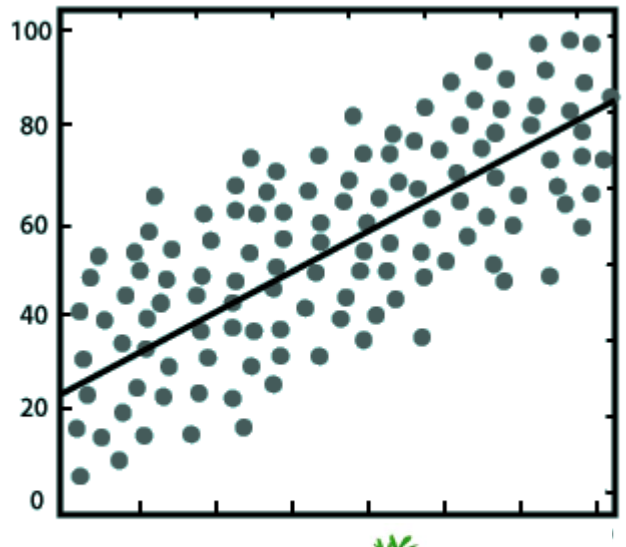
DATA PARTITIONING AND RESAMPLING

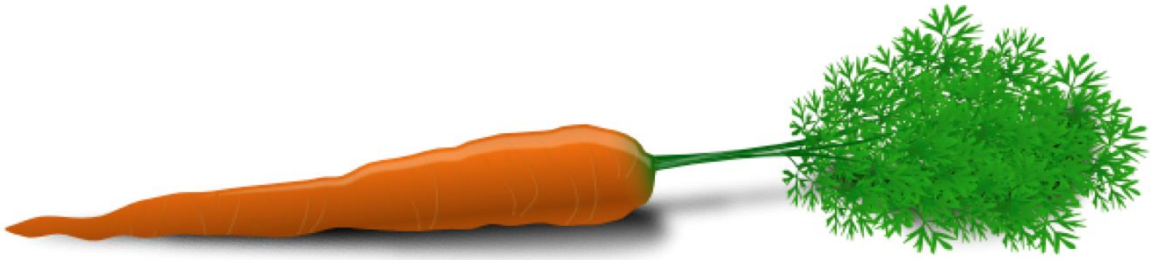




ALGORITHM LOGIC AND INTUITION







CARET

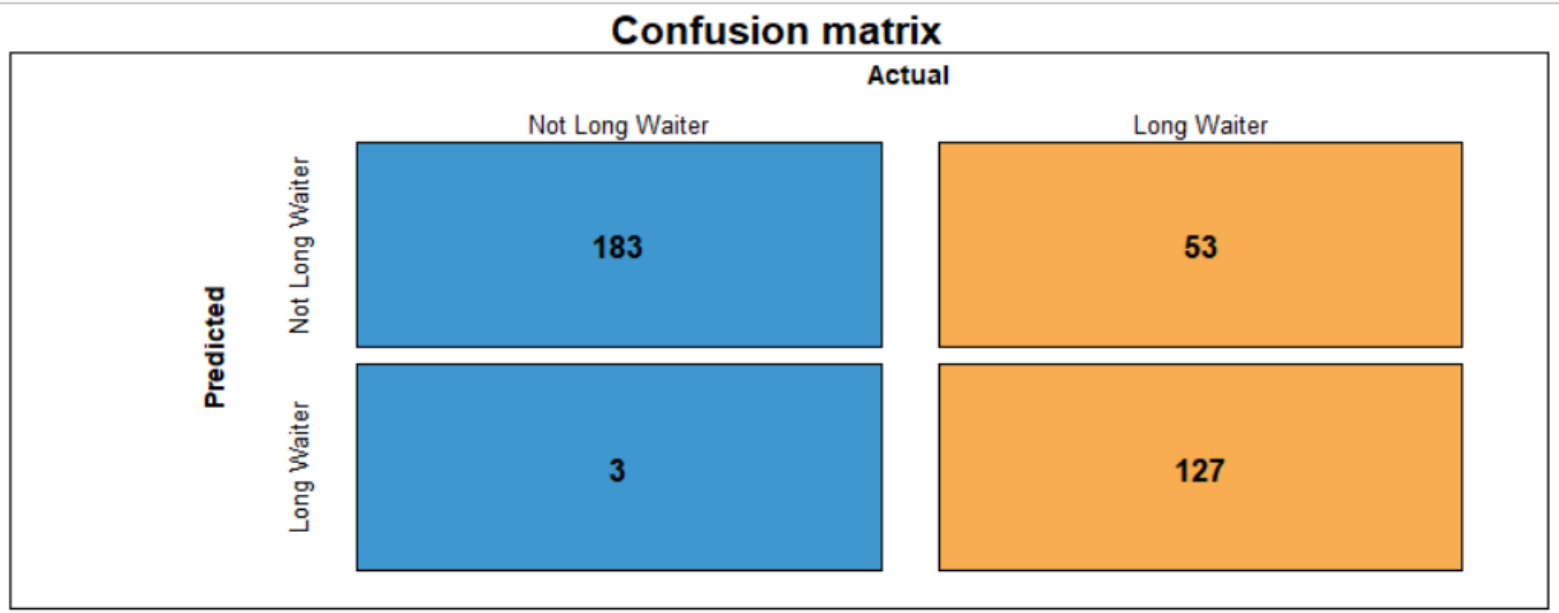
ALGORITHMS

```
                 Reference
Prediction        Long.Waiter Not.Long.Waiter
  Long.Waiter             183              53
  Not.Long.Waiter           3             127

               Accuracy : 0.847
                 95% CI : (0.806, 0.8823)
    No Information Rate : 0.5082
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6925

 Mcnemar's Test P-Value : 5.835e-11

            Sensitivity : 0.9839
            Specificity : 0.7056
         Pos Pred Value : 0.7754
         Neg Pred Value : 0.9769
             Prevalence : 0.5082
         Detection Rate : 0.5000
   Detection Prevalence : 0.6448
      Balanced Accuracy : 0.8447
```

https://draperanddash.com/machinelearning/confusion-matrices-evaluating-your-classification-models/

CONFUSION MATRICES

CONFUSION MATRIX

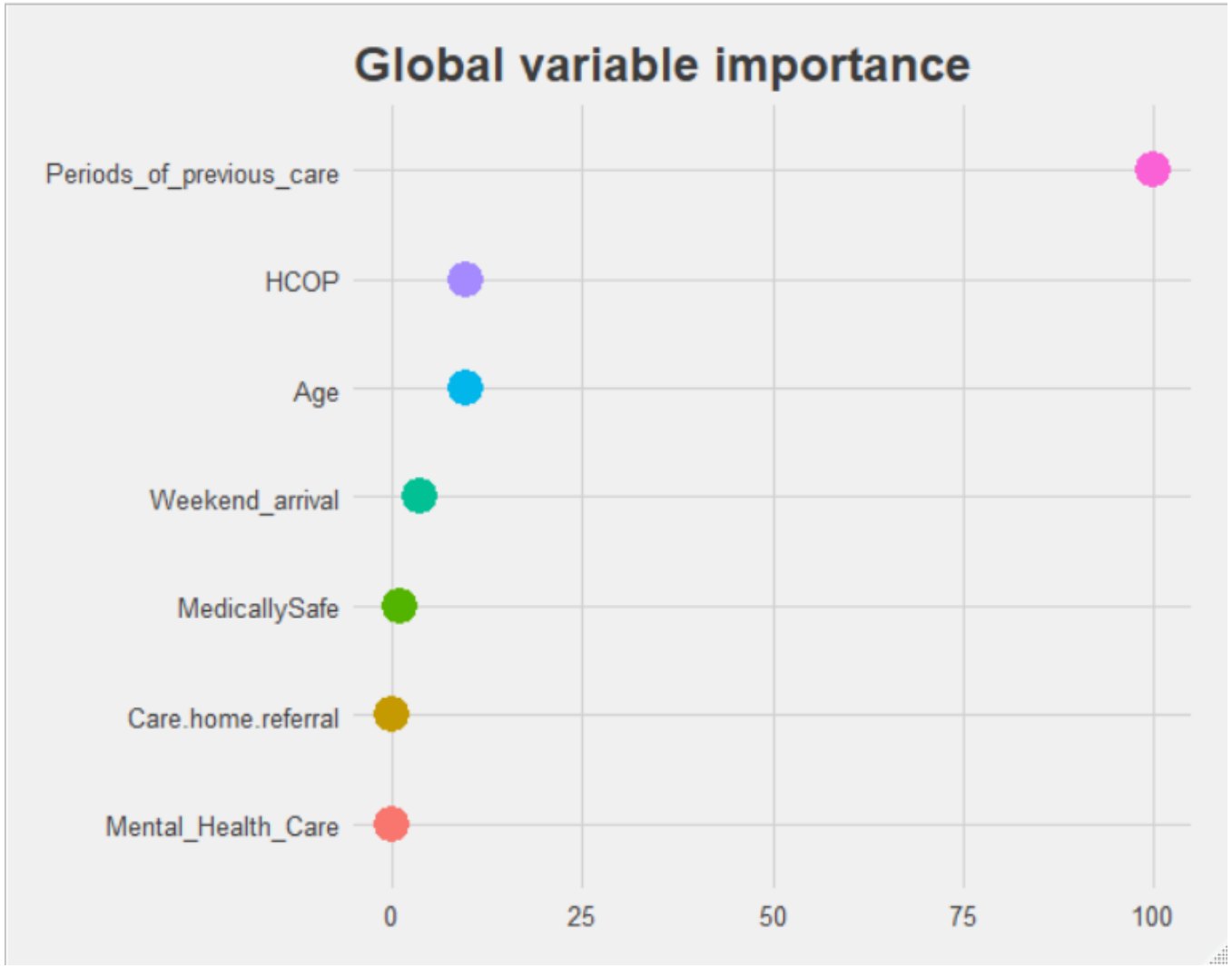


Confusion matrix statistics

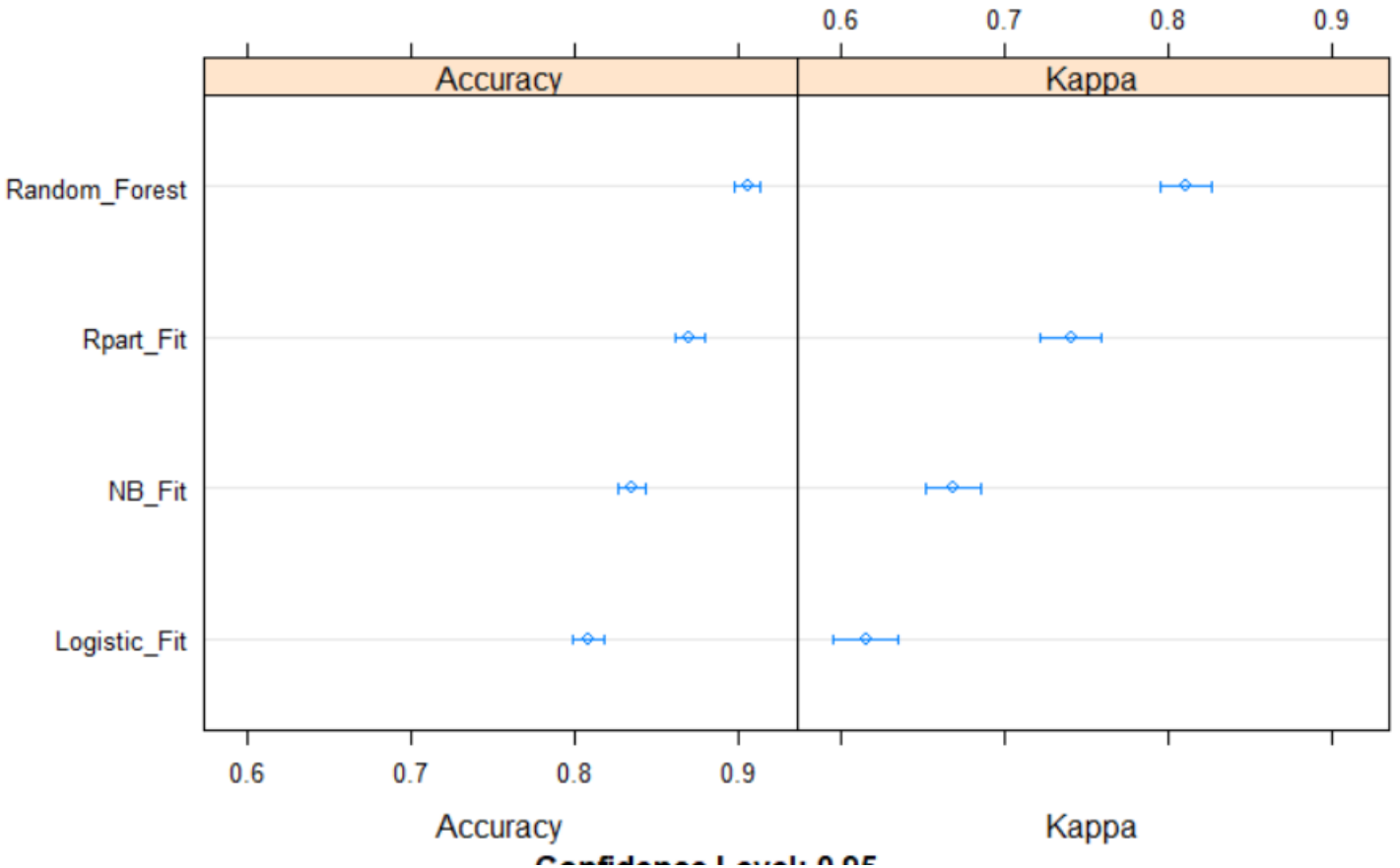
Sensitivity	Specificity	Precision	Recall	Balanced Accuracy
0.98	0.71	0.78	0.98	0.84

Accuracy	Kappa
0.847	0.69

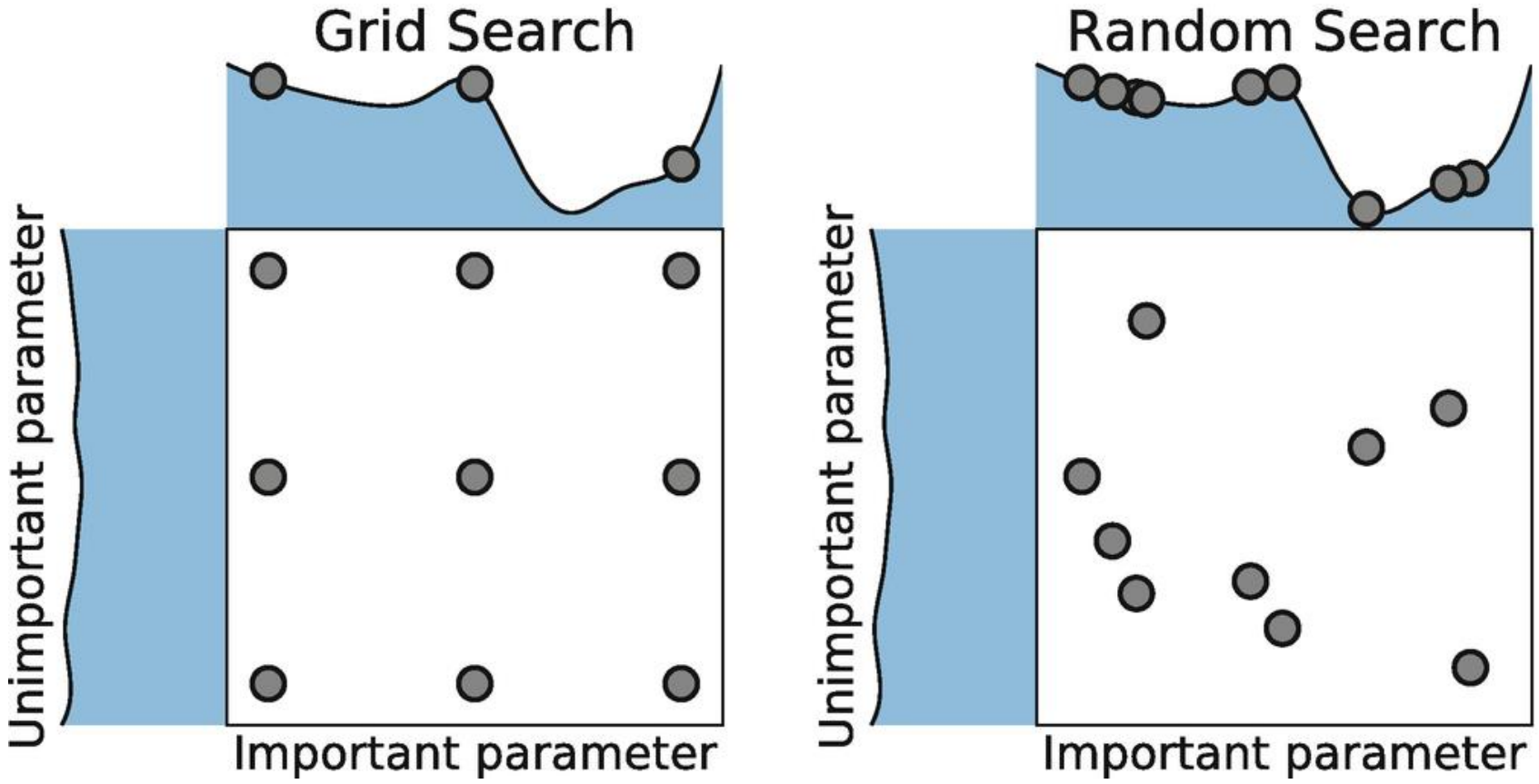
GLOBAL VARIABLE IMPORTANCE



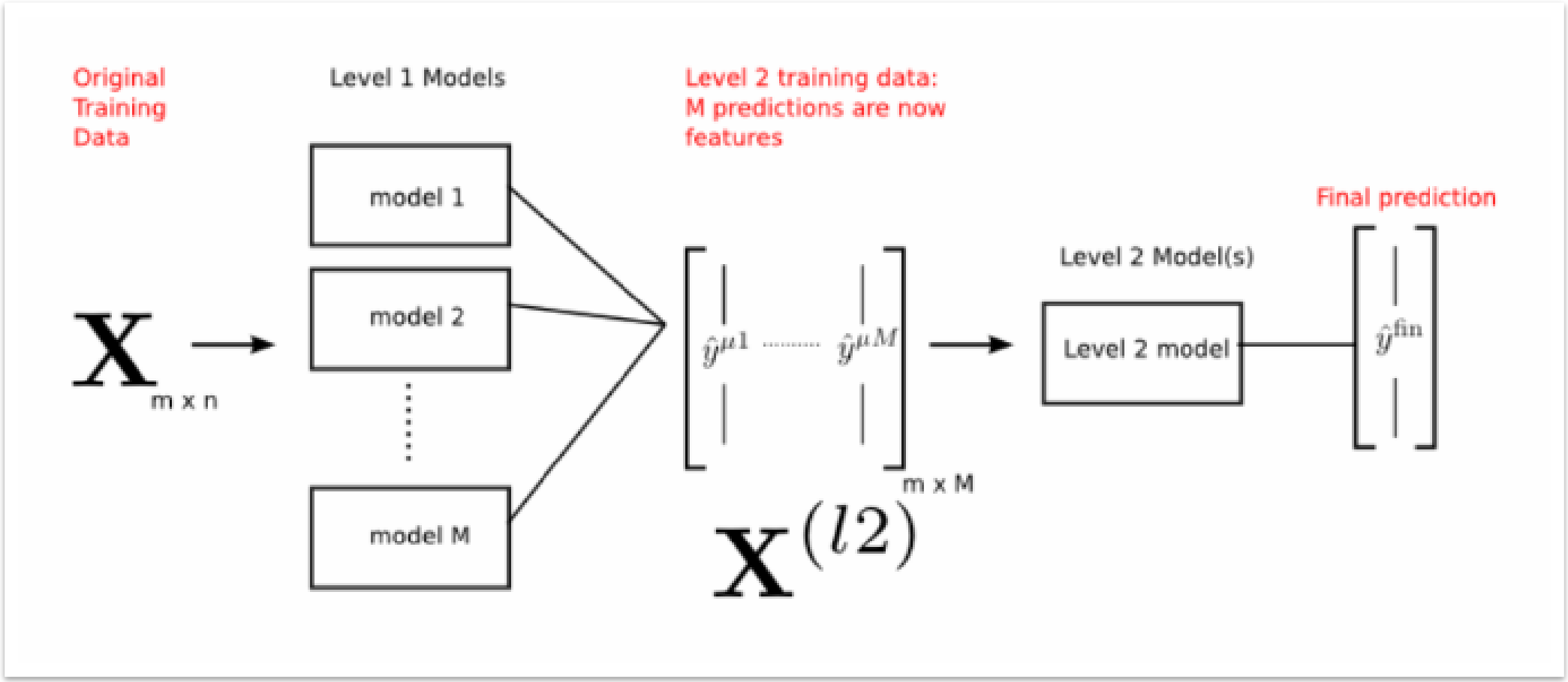
ALGORITHM BENCHMARKING

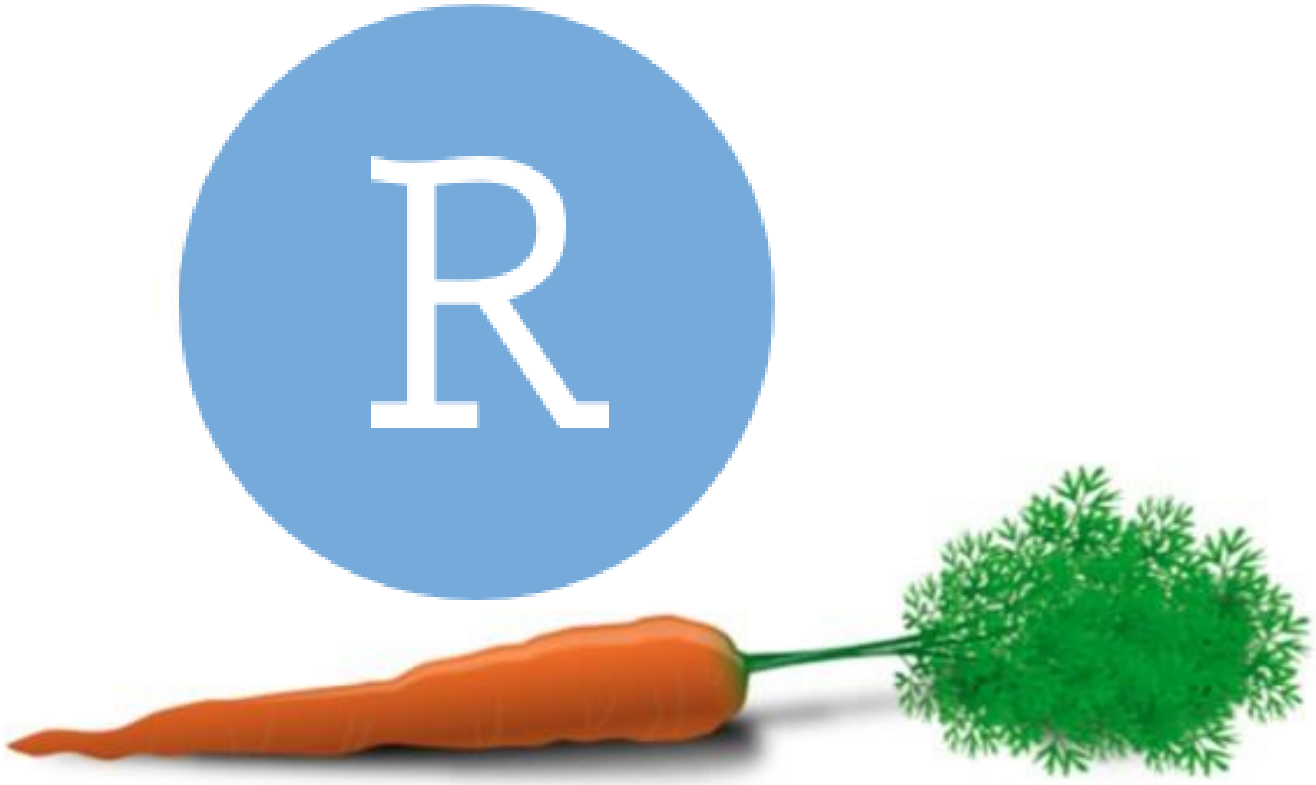


HYPERPARAMETER TUNING – RANDOM AND GRID SEARCH



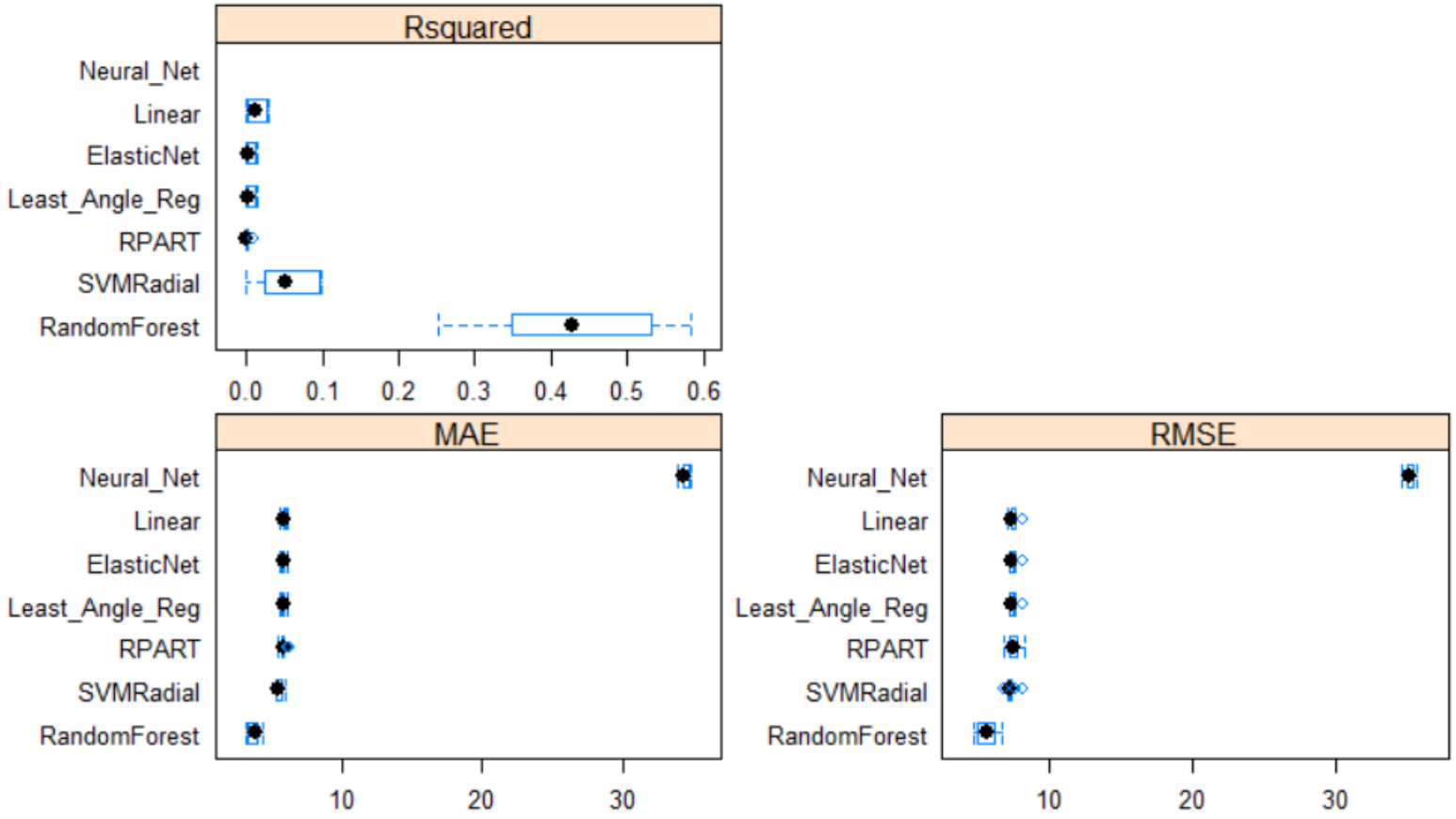
STACKING ENSEMBLE ALGORITHM





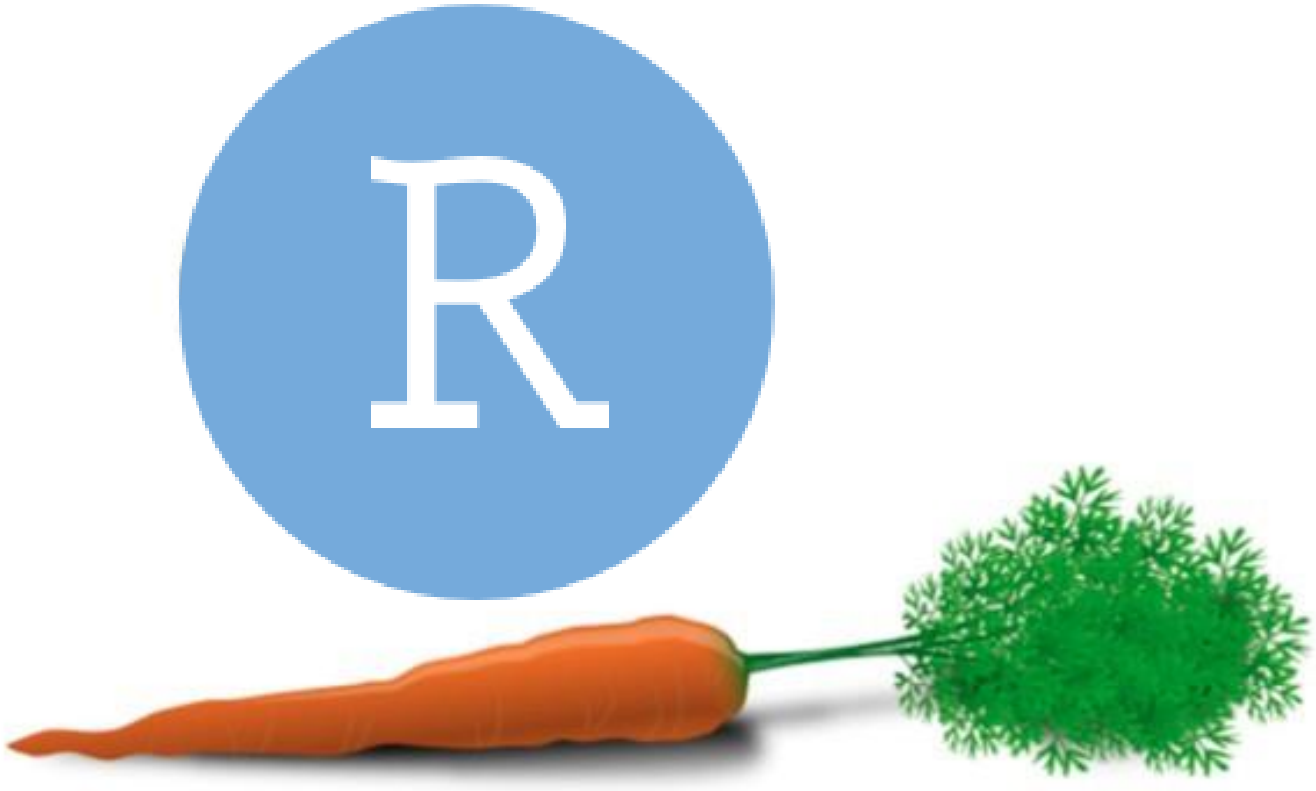
REGRESSION MODELLING

REGRESSION BENCHMARKING



REGRESSION ENSEMBLING FUNCTION

```r
ensemble_function <- function(Y.label, df, k_folds, meta_model_name){
  ensemble_control <- caret::trainControl(method="cv", number = k_folds,
                                          savePredictions = 'final', classProbs = TRUE)

  ensemble_alg_list <- c("rf", "qrf")
  ensemble_models <- caretEnsemble::caretList(
    as.formula(paste(Y.label, "~ .")),
    data=df, trControl=ensemble_control, methodList=ensemble_alg_list
  )

  meta_ensemble <- caretEnsemble::caretStack(ensemble_models,
                                             method = as.character(meta_model_name))
}
```



DEPLOYING R MODEL INTO PRODUCTION AND MAKING PREDICTIONS

WHY CARET?

- Deep learning can be more effective – packages such as Keras / Tensorflow, H2O and MxNet being the market leaders. However, most NHS systems do not have sufficient enough servers to take the processing load of a 3 layer model consisting of 128, 256, 128 node model
- CARET is the right fit between serious ML and applications that can be more useful on smaller datasets for supervised ML
- TidyModels vs CARET – whilst Max Kuhn and his team are developing this out recipes and parsnip – these packages have only 25% of all the algorithms listed under CARET. Max admits himself that this is not a replacement to CARET and it is still worth using CARET for ensembling and ML, as this will not be deprecated for a long time

LINKS TO OTHER COOL WORK

- Feature importance – I have identified global feature importance in the previous slides, but for local feature importance – LIME is a great summary tool to look at a local (patient) level: https://cran.r-project.org/web/packages/lime/vignettes/Understanding_lime.html
- Parsnip and recipes: let Max explain as he creates these awesome tools for R:https://www.youtube.com/watch?v=ZFTjroC8bTg
- Speeding up ML models in CARET: https://hutsons-hacks.info/threading-and-caret-buring-your-cpu-to-improve-model-training-speed
- Deploying a model RDA files for use in production side predictions: https://draperanddash.com/machinelearning/deploying-a-trained-supervised-ml-model/

INTERESTED IN DEEP LEARNING WITH H2O.ai AND K WITH R?

SIGNING OFF