LABSIN

An Introduction to Machine Learning and Deep Learning with R

HARPO (AKA Carlos A. Catania Ph.D)
LABSIN - Ingeniería - UNCuyo

@harpolabs
harpo@ingenieria.uncuyo.edu.ar

AGENDA Day 1: Machine Learning

- Basic concepts
- The Machine Learning Workflow
- Supervised vs Unsupervised
- R language
- Tidyverse
- The Caret Package
- LAB 1:
  - Wine Quality prediction

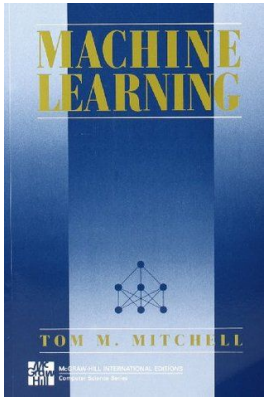
Materials Day 1: Machine Learning

- **Rstudio** desktop or server version 1.4
- **R** version >= 4.0
- **Rstudio** Notebook operational
- `install.packages("caret")`
- `install.packages("tidyverse")`
- Data for day 1 at http://bit.ly/mllab2019

Machine Learning

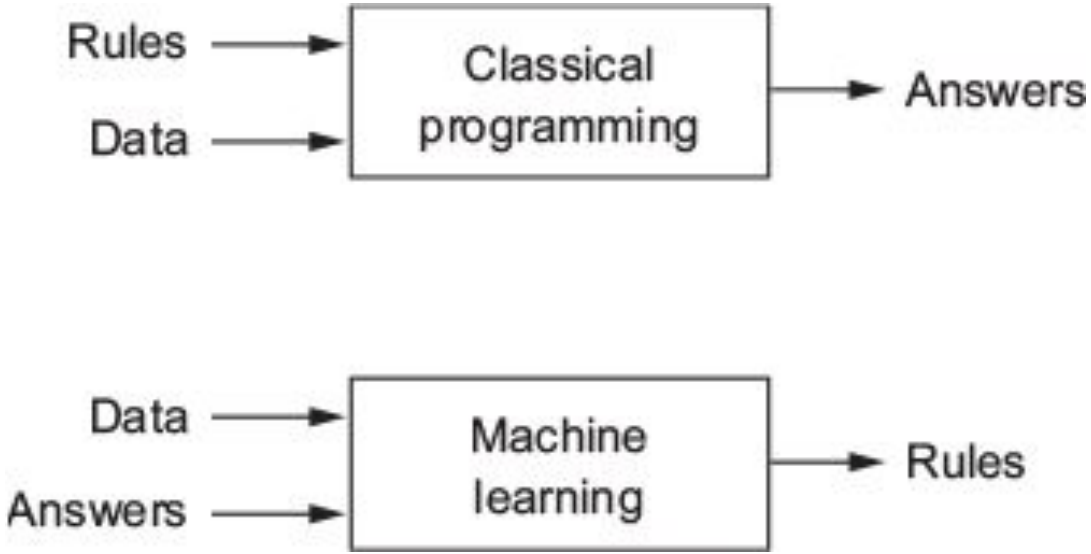
Machine Learning: A definition

"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."



Tom Mitchell circa 1997

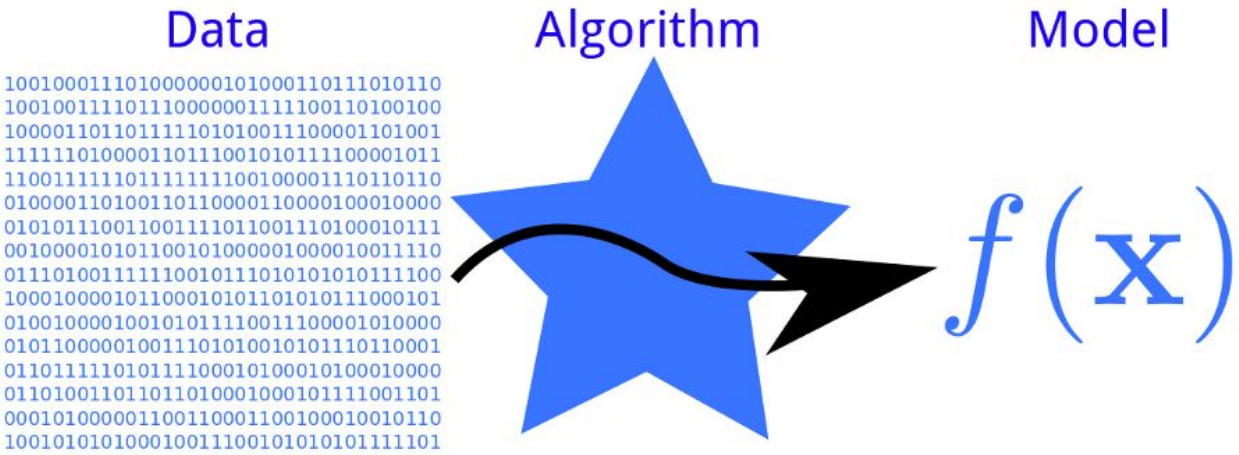
A Computer Science Perspective on Machine Learning



More Formally...

*A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

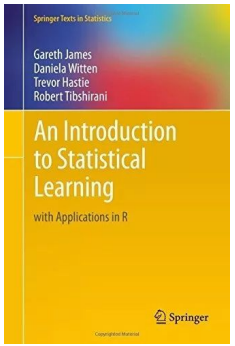
The general idea behind ML



More Formally... (again)

suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon.$$



Why Estimate f ?

Two main reasons:

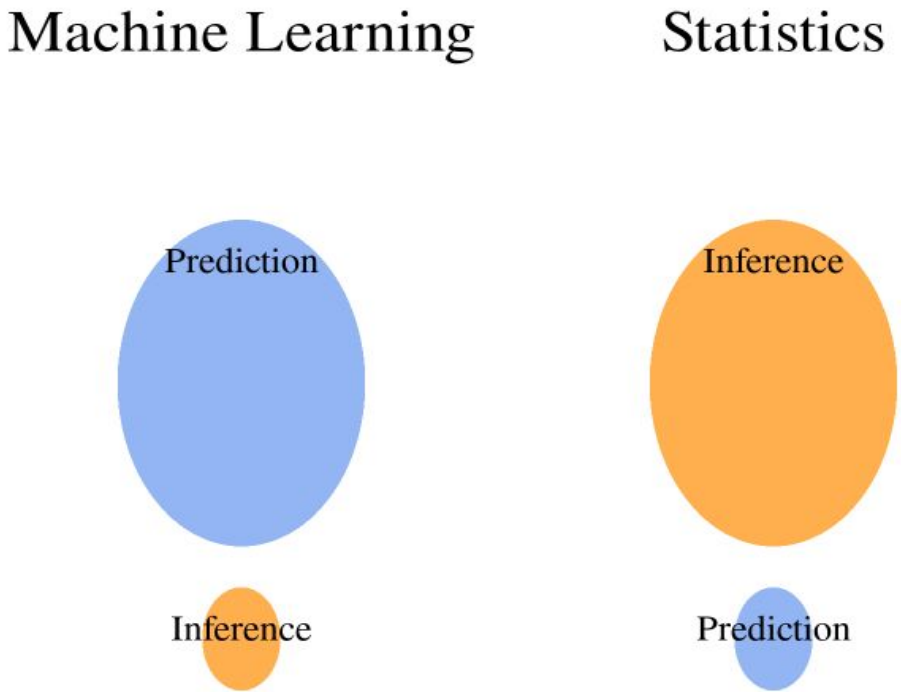
Prediction: Use the model to predict the outcomes for new data points. One **is not typically concerned with the exact form of f,** provided that it yields accurate predictions for Y.

$$\hat{Y} = \hat{f}(X),$$

Inference: We are often interested in understanding the way that Y is affected as X_1,...,X_p change.

In Machine Learning we mostly care about Prediction!

Differences between ML and Statistics



Differences between ML and Statistics

- ML tends to deal with **large, complex datasets** (such as a dataset of millions of images, each consisting of tens of thousands of pixels)
- **Little mathematical theory**—maybe too little—
- ML **is engineering oriented**. ideas are proven empirically more often than theoretically.

Differences between ML and Statistics

ML algorithms are often treated as black boxes.

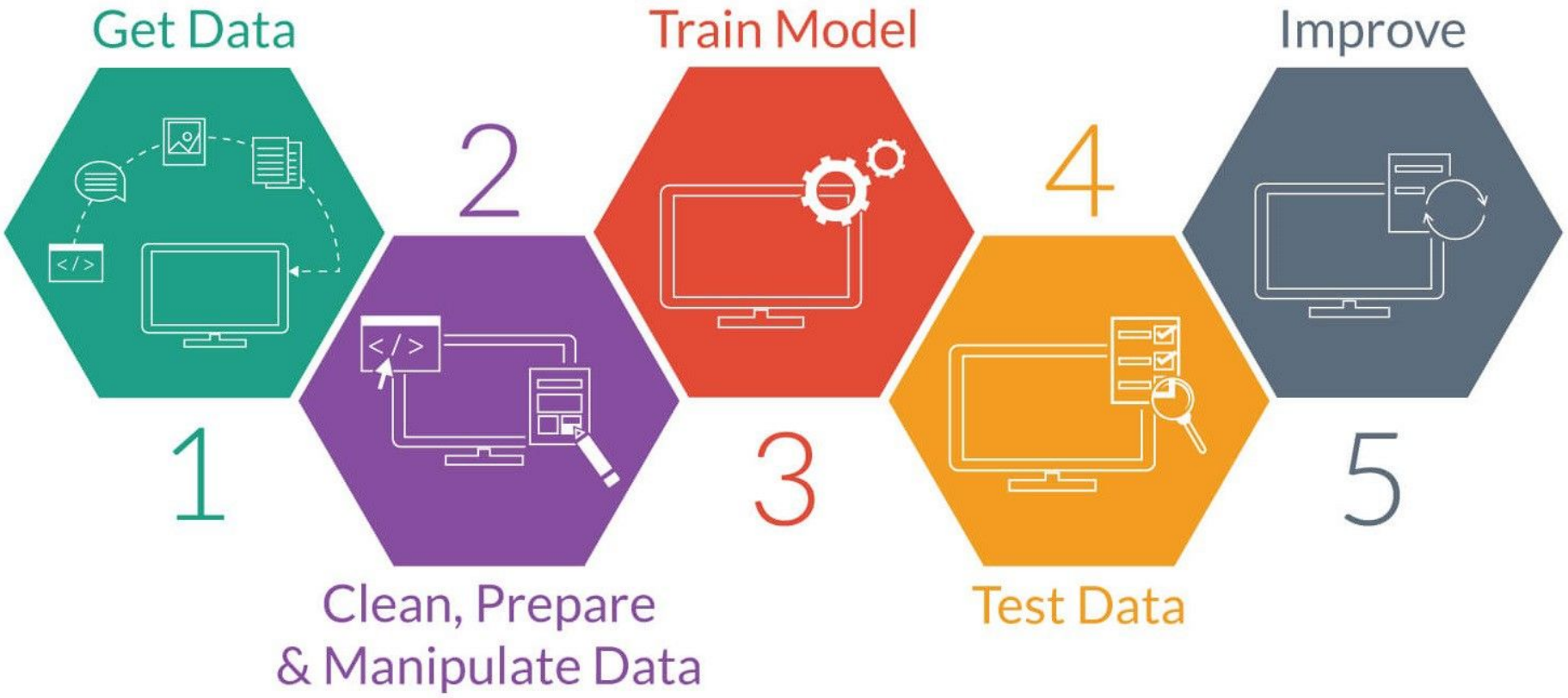
Machine Learning Algorithms

The algorithms and method come from areas such as:

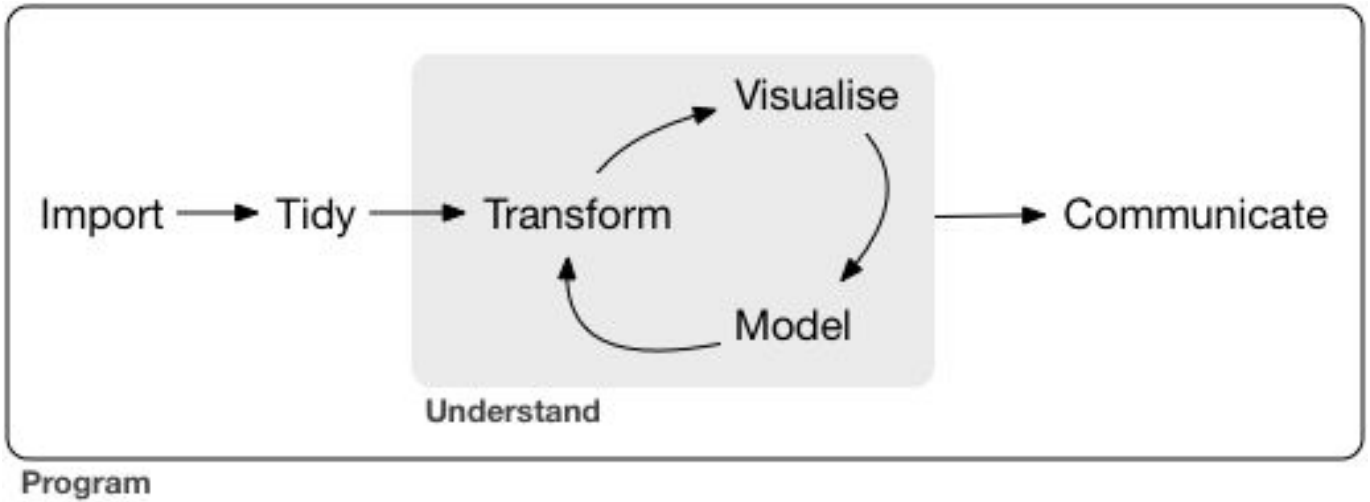
- Pattern Recognition
- Applied Statistics
- Artificial Intelligence

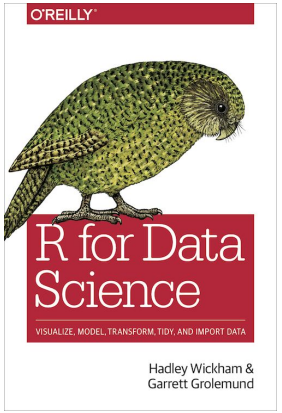
The borderline between disciplines has become diffuse.

The machine learning workflow



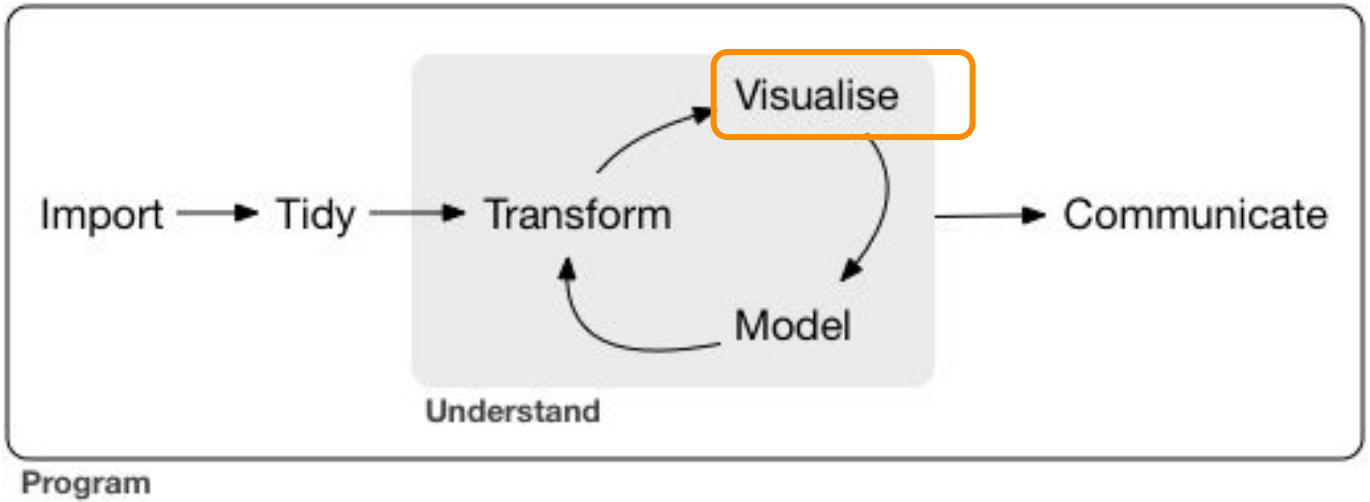
In R we have our own way to do it. (Actually, Hadley's way)



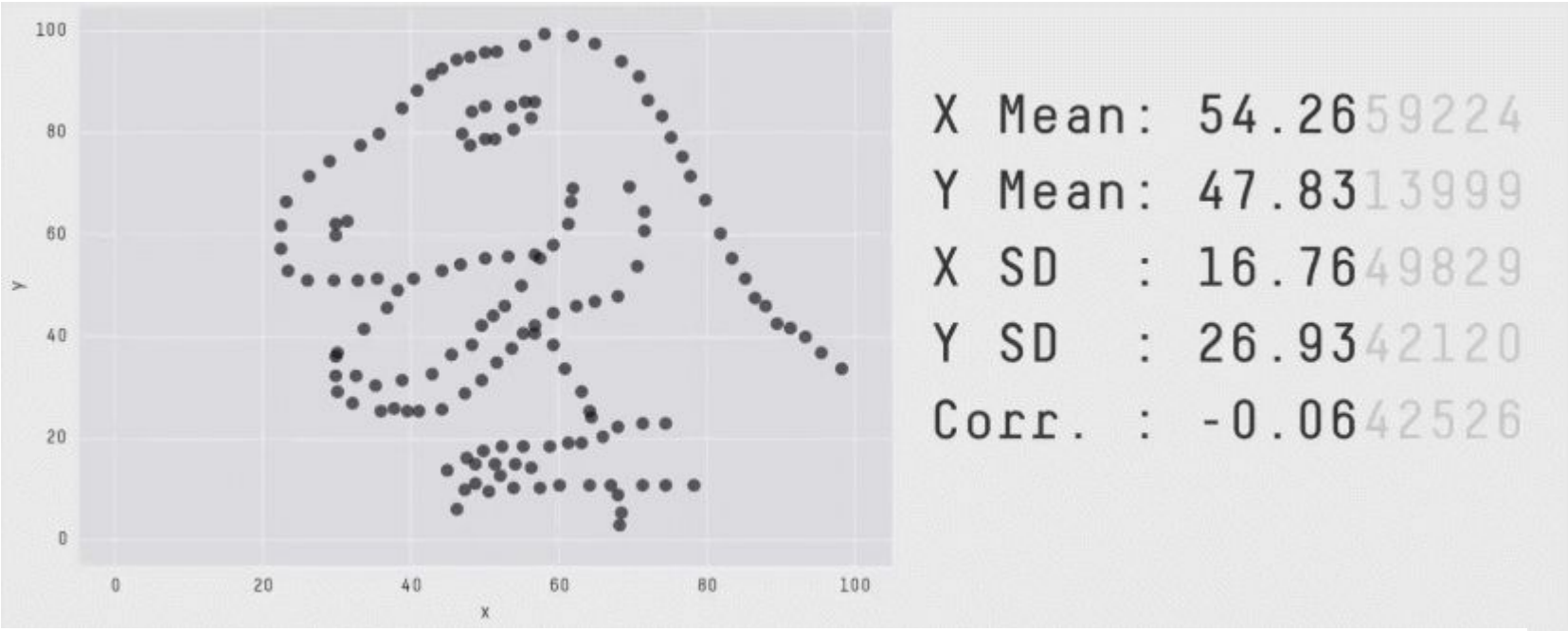


https://r4ds.had.co.nz

In R we have our own way to do it. (Actually, Hadley's way)



Mom said, you should always visualize your dataset



Justin Matejka, George Fitzmaurice (2017) Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Supervised vs. Unsupervised Learning

SUPERVISED:

For each observation of the predictor measurement(s) x_i, $i = 1, \ldots, n$ there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors.

UNSUPERVISED:

For every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no associated response y_i. We seek to understand the relationships between the variables or between the observations.

SUPERVISED VS UNSUPERVISED

SUPERVISED LEARNING

All data has been labeled (supervised) by an expert. Thanks to this labeling process, we can help the network to realise the difference between classes (even though sometimes this does not happen).

Some techniques: NNs, SVM, etc.

UNSUPERVISED LEARNING

Our data are not labeled. Unsupervised algorithms consider confidence measures among samples in order to create homogeneous clusters.

Most famous technique: Clustering (k-means, hierarchical etc.)

For doing ML we need:

- *Input data points*
- *Examples of the expected output*
- *A way to measure whether the algorithm is doing a good job*

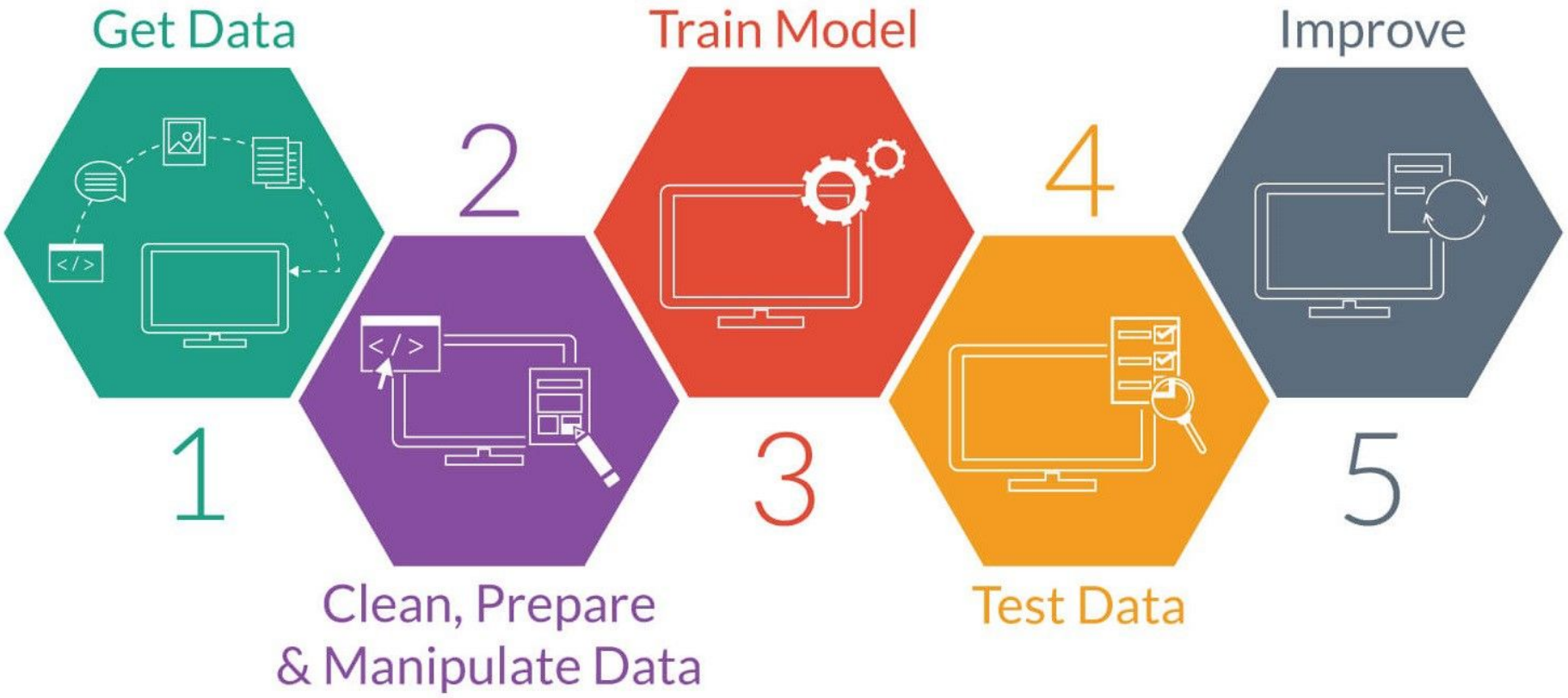
The Caret Package.

The **caret** package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models.

The package contains tools for:

- Data splitting
- Pre-processing
- Feature selection
- Model tuning using resampling
- Variable importance estimation

The machine learning workflow



LAB1: Tenaris Consumption

1. Preprocessing
2. Visualize
3. A first step towards a methodology approach
4. A little an inaccurate model.

TRAIN THE MODEL: Split the dataset

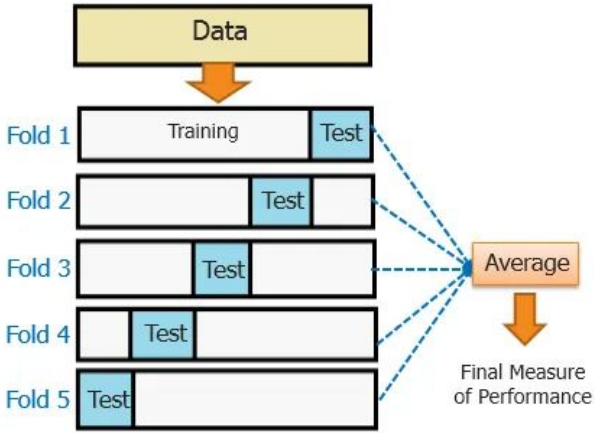
```
# TRAIN THE MODEL
## Split train and test
```{r}

trainIndex <- createDataPartition(as.factor(trainset$quality), p=0.80, list=FALSE)
data_train <- trainset[ trainIndex,]
data_test <-  trainset[-trainIndex,]
colnames(data_train) <- make.names(colnames(data_train))
colnames(data_test) <- make.names(colnames(data_test))
```


```

TRAIN THE MODEL: The train control object

```
ctrl_fast <- trainControl(method="cv",
                          repeats=1,
                          number=5,
                        #  summaryFunction=twoClassSummary,
                          verboseIter=T,
                          classProbs=F,
                          allowParallel = TRUE)
```



TRAIN THE MODEL: train()

```
ctrl_fast <- trainControl(method="cv",
                          repeats=1,
                          number=5,
                        #  summaryFunction=twoClassSummary,
                          verboseIter=T,
                          classProbs=F,
                          allowParallel = TRUE)
```

```
rfFitupsam<- train(train_formula,
                   data = data_train,
                   trControl = ctrl_fast,
                   method="rpart")
```

Bibliography

- An Introduction to Statistical Learning with Applications in R, 2022
- Data Science for R. 2022
- Resources for a Gentle Introduction to Machine Learning

Carlos A. Catania (PhD)
(AKA **Harpo**)
LABSIN - Ingeniería - UNCuyo
@harpolabs
harpo@ingenieria.uncuyo.edu.ar
https://harpomaxx.github.io

LABSIN

http://labsin.org

Feel free to ask questions anytime during the lecture.

1. Leave your question online via
   slides.app.goo.gl/jKpS7

2. 

   vs