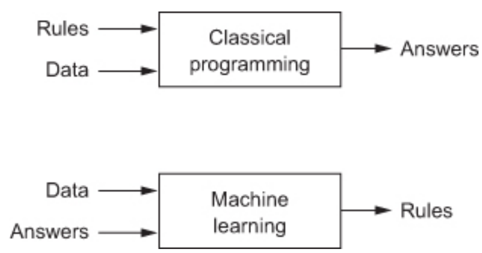
Intro to Classification

Questions that Data Science Methods Can Answer

- How Many or How Much of something (Regression)
- **Is this observation A or B, or C or D or E... (Classification)**
- What groupings exist in the data already (Clustering)
- What should we expect to happen next? (Time Series Analysis)
- Is this weird? (Anomaly Detection)

What are Classification algorithms?

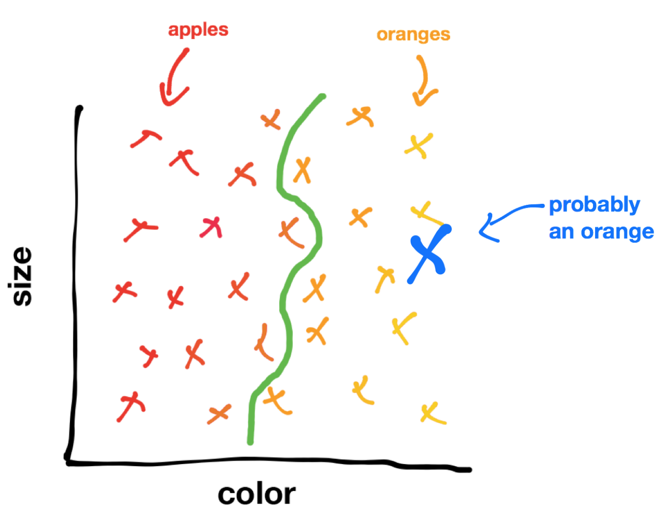
- Classification is a supervised learning task. That means we train on data w/ answers/labels



- We train with ansers/labels to produce a `decision rule` we'll use to classify future data.

Main Ideas

- With classification, we use labeled data to train algorithms to classify future data points.
- The decision boundary becomes becomes the decision rule that discerns A from B



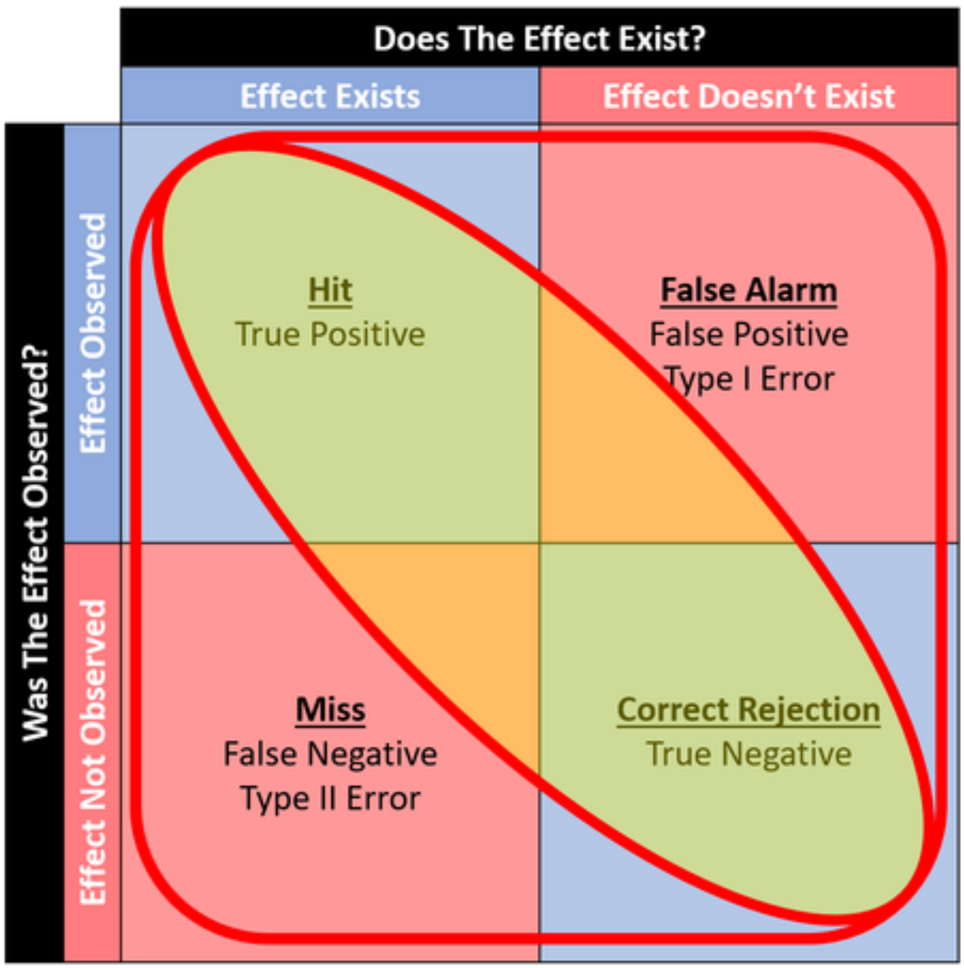
Vocabulary

- Classifier - an algorithm that maps the input data to a specific category
- Classification model, a trained model predicts the class of future datapoints.
- Feature - an input, independent varaible, or column of data to predict the target
- Binary classifier determines if an observation belongs to one class or another
- Multiclass classifiers determine if somethign is A or B or C (or something else)

Correct and Incorrect Classification

For each observation we classify with a classification model, there are four possible outcomes:

1. True Positive - *Correct prediction* = check engine light goes on, engine has trouble.
2. True Negative - *Correct rejection* = check engine light remains off, engine is OK
3. False Positive - *False alarm* = check engine light goes on, but engine is OK.
4. False Negative - *Miss* = check engine light remains off, but engine has trouble.



Different Measures of correctness

When evaluating classification errors, the measure of correctness we choose depends on the cost/benefit of true predictions or Type I (false alarm) or Type II (miss) errors.

What's the cost/benefit of misses or false positives in autocorrect for predictive text? Maybe some embarassing moment, but no big deal.

What's the cost/benefit of misses or false positives for a cancer screen?

Is a false positive better than a false negative?

It depends on the domain, the circumstance, and the costs/benefits involved.

Accuracy: #(TP+TN)/#Total Observations. Describes overall, how often the classifier correct

Sensitivity or Recall = #TP/#(TP + FN). When it's actually yes, how often does it predict yes

Specificity = #TN/#(TN + FP) .When it's actually no, how often does it predict no?

Precision = #TP/# of predicted yes = # True Positives / (#True Positives + #False Positives). the higher this number is, the more you were able to pinpoint all positives correctly. If this is a low score, you predicted a lot of positives where there were none.