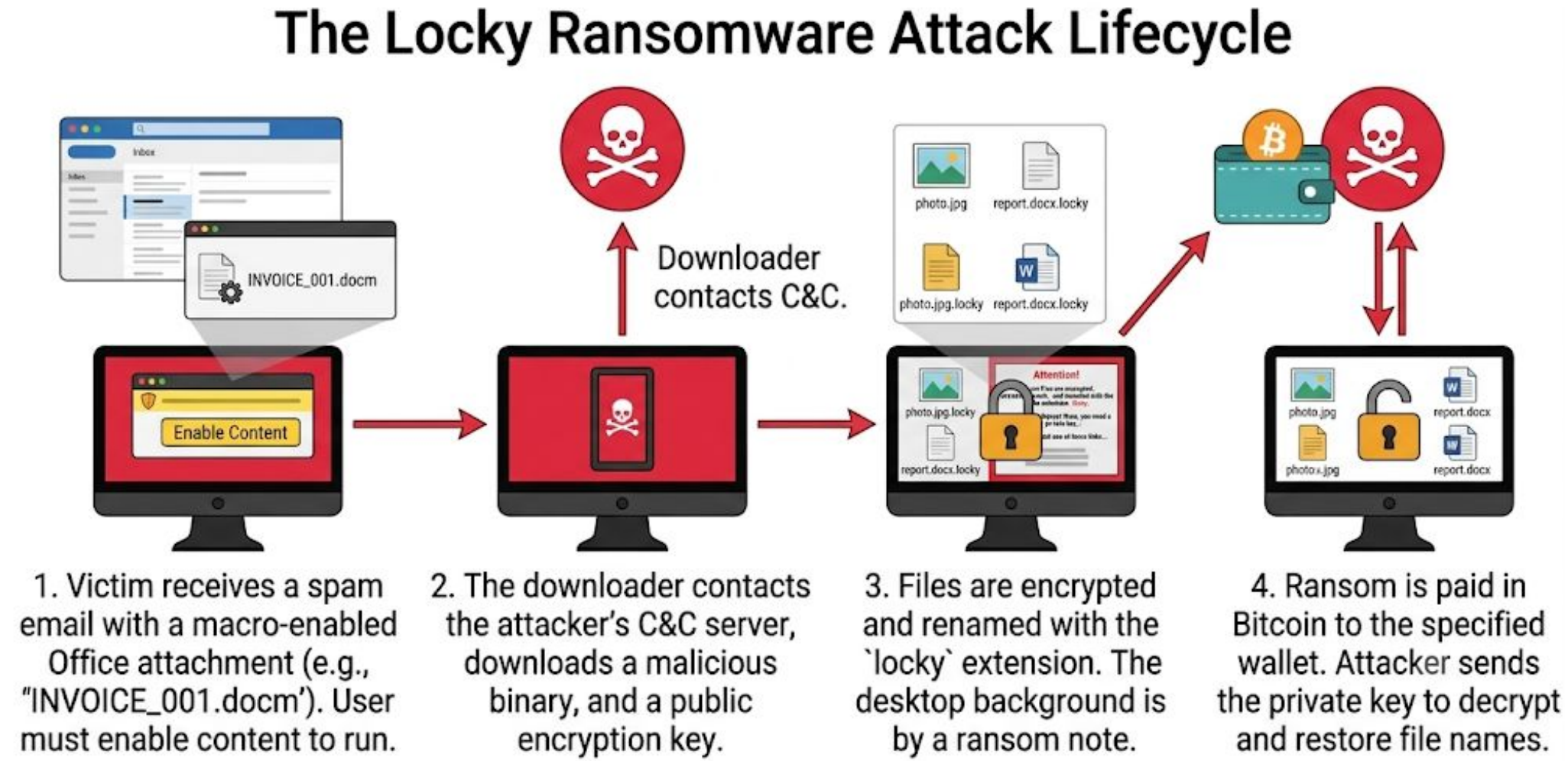
Prevent, Detect, Respond

@NJCybersecurity

cyber.nj.gov

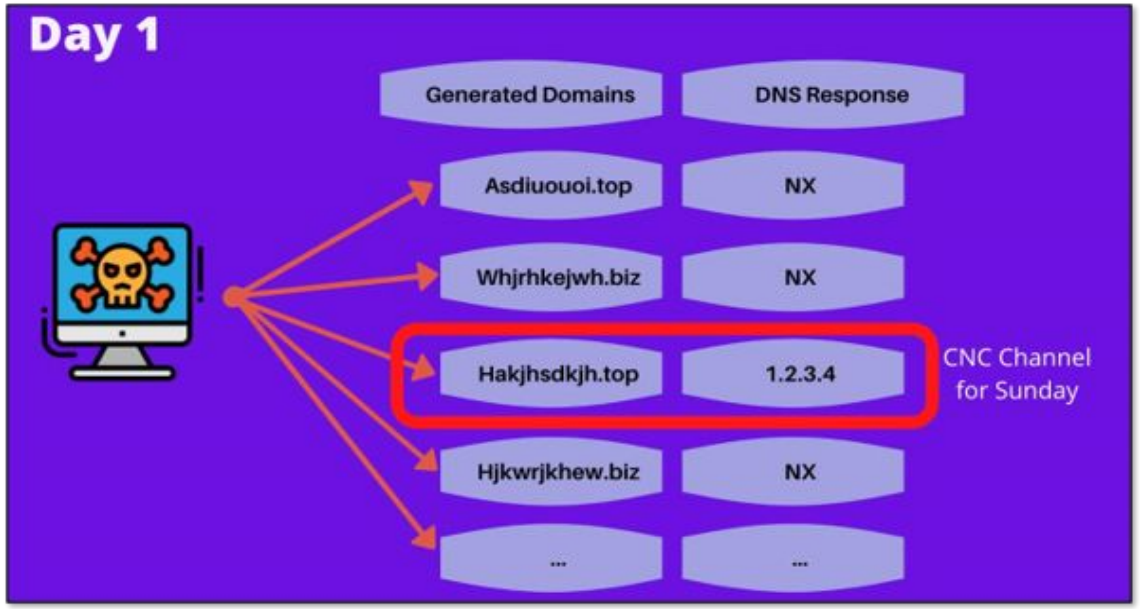
NJCCIC@cyber.nj.gov

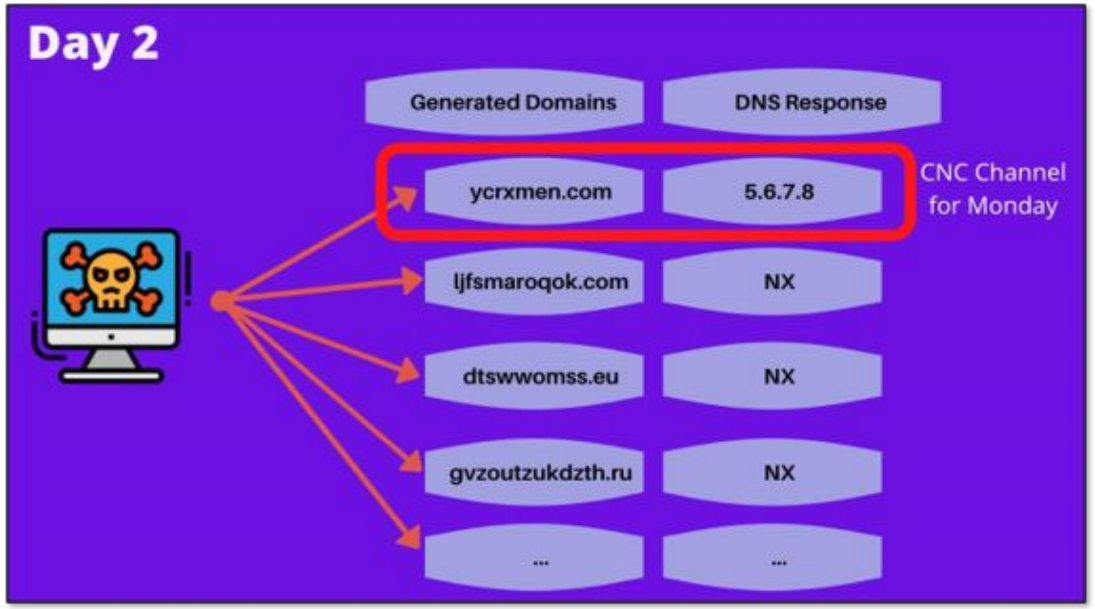
Locky Ransomware



The Problem

- Millions of connections every day
- Many domains are indicators of malicious activity
- Reliance on third parties to detect domains (Palo, Zscaler, Virus Total)
- Many potential threats go unnoticed
  - DGAs are designed to evade detection
  - Can be created quickly
- Can we improve DGA detection to help analysts find threats faster?

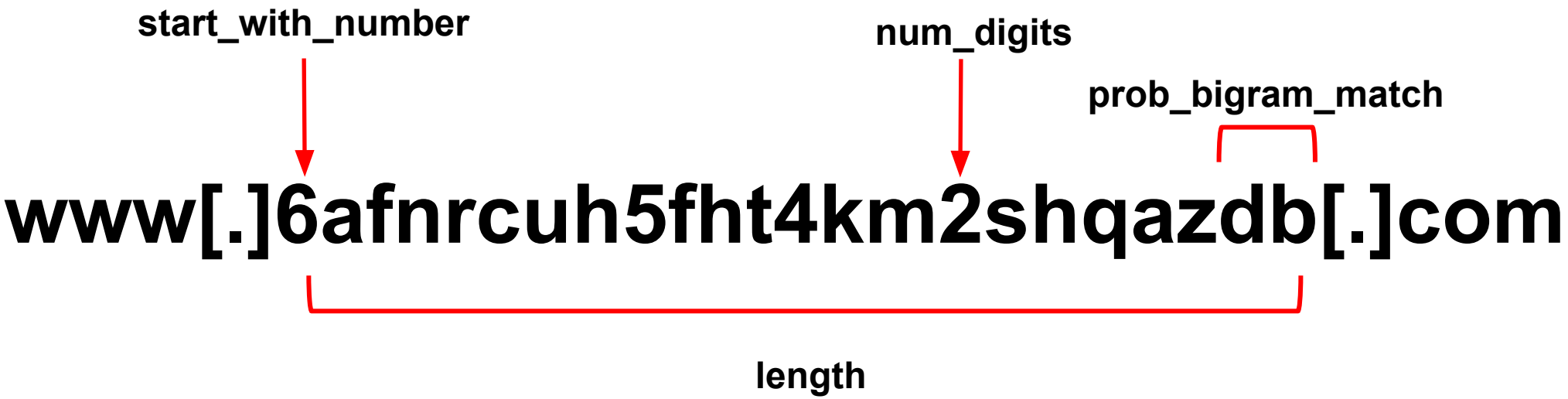




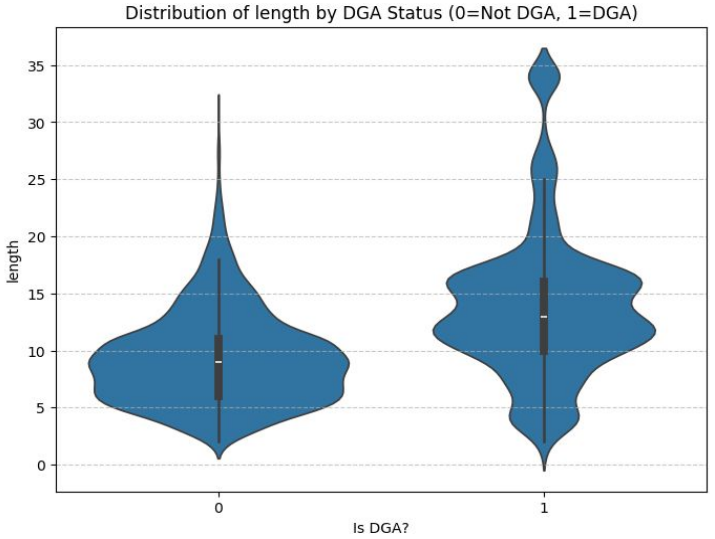
Step 1: Get your data

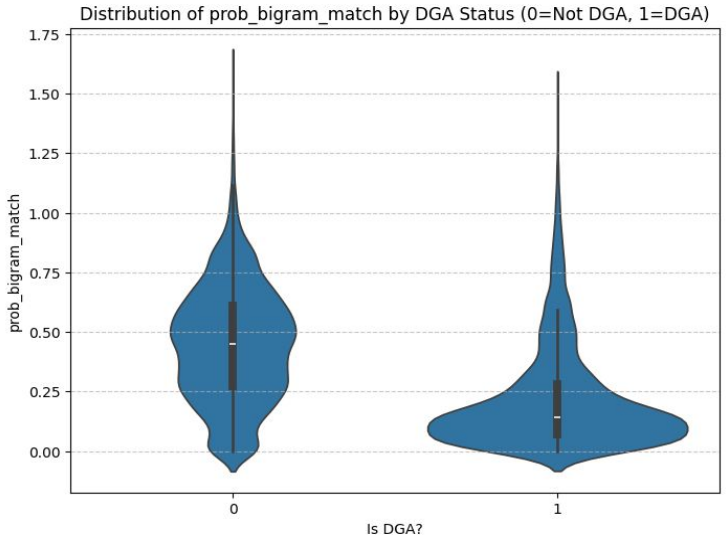
- Supervised Learning = Labeled Dataset
- 3 datasets combined
  - Alexa Top 1 Million
  - Cisco Top 1 Million
  - University of Murcia Domain Generation Algorithm Dataset (UMUDGA)
    - Contains 30 million DGA examples
- Even distribution of DGA and non-DGA examples in training set

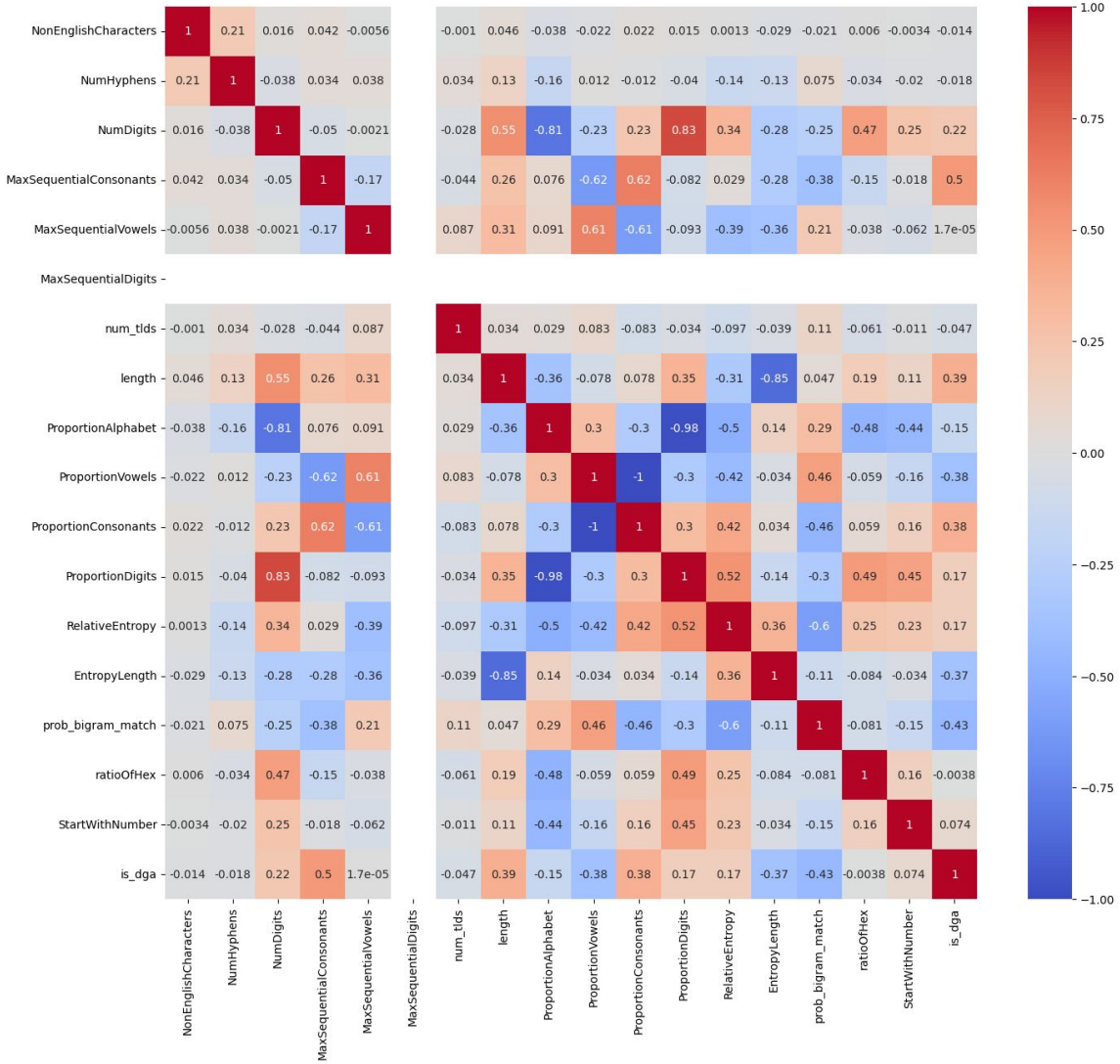
Step 2: Feature Engineering



Step 3: Feature Analysis



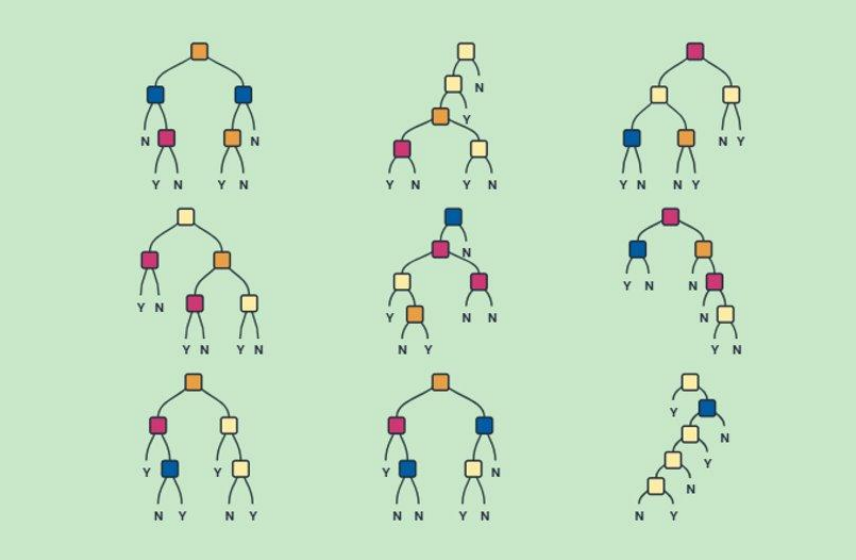




Step 4a: Algorithms

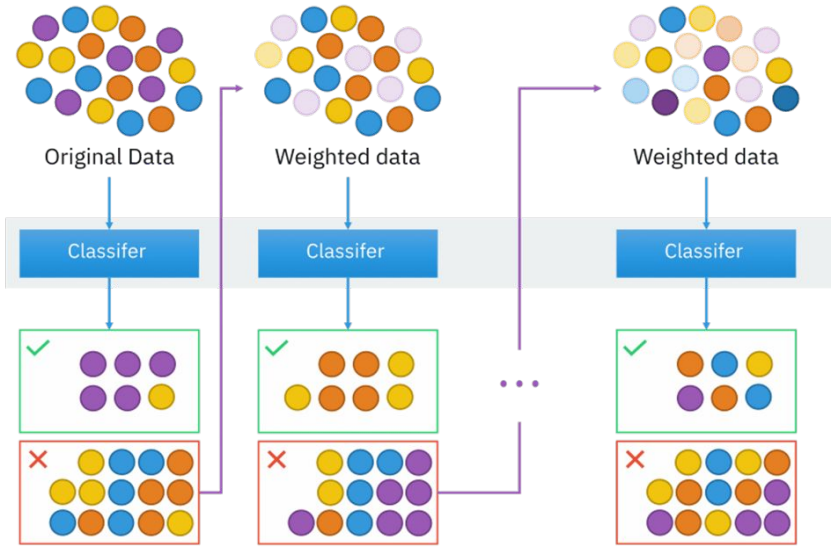
Random Forest

- Uses random feature subsets at each split to explicitly decorrelate the trees.
- **Primary Impact:** Reduces model variance without increasing bias; highly robust to overfitting and handles high-dimensional data well.

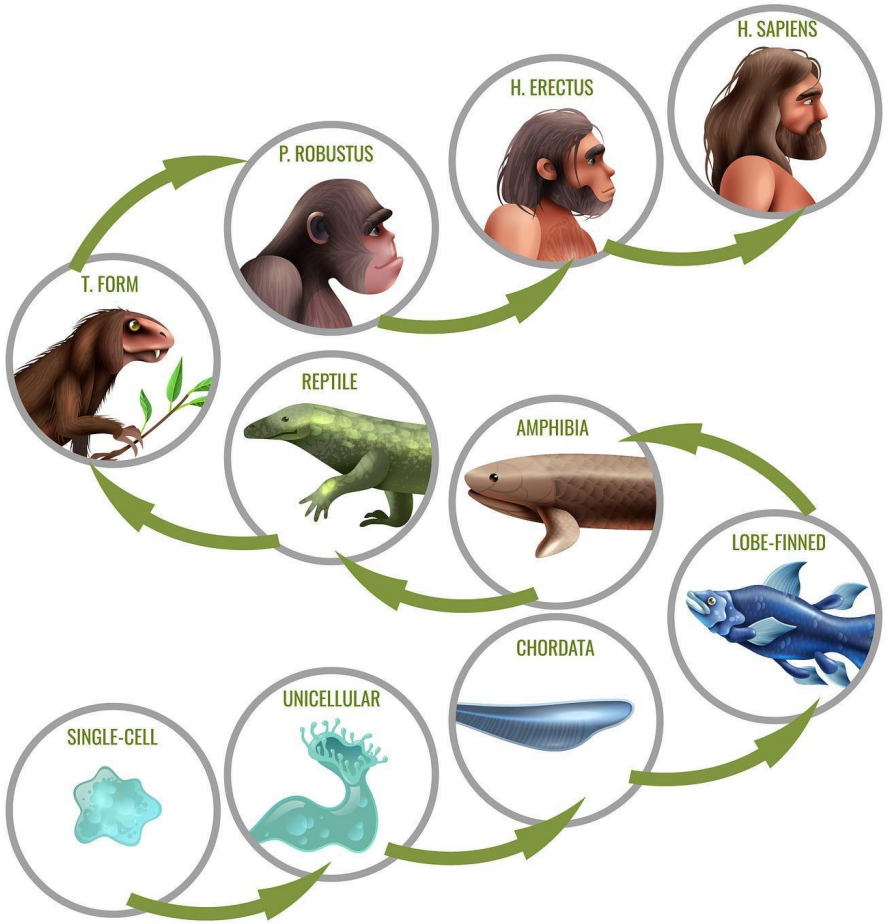


AdaBoost (Adaptive Boosting)

- Iteratively updates sample weights, penalizing misclassifications so each new tree focuses strictly on the previous model's errors.
- **Primary Impact:** Aggressively reduces model bias to fit complex boundaries, though it is much more sensitive to noisy data and outliers than bagging.



Step 4b: Evolutionary Algorithm



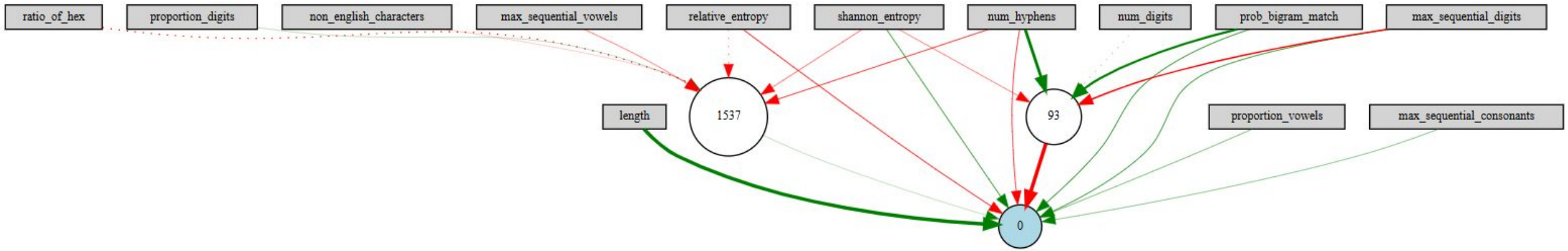
* Image by Freepik

4 Key Drivers of Evolution

- Genetic Encoding
- Mutation
- Natural Selection
- Reproduction and Heredity

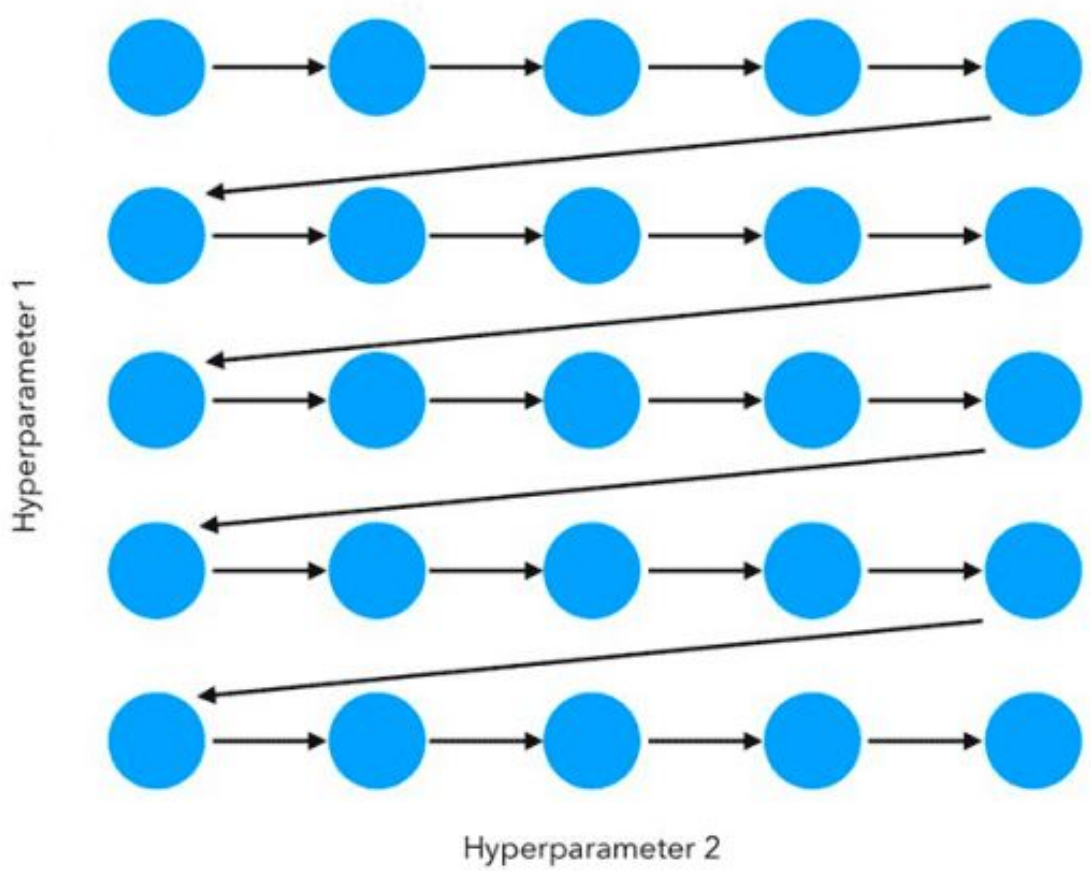
Step 4c: NEAT Algorithm

- **N**euro**E**volution of **A**ugmenting **T**opologies
- 4 Key Drivers of Evolution
  - **Genetic Encoding**: Neural network structure. Input/Output nodes and connections.
  - **Mutation**: Random chance to add/remove connections and hidden nodes.
  - **Natural Selection**: Fitness/Evaluation function.
  - **Reproduction/Heredity**: NEAT pairs surviving members to pass down structural traits.



Step 5: Hyperparameter Tuning

Technique:
GridSearchCV



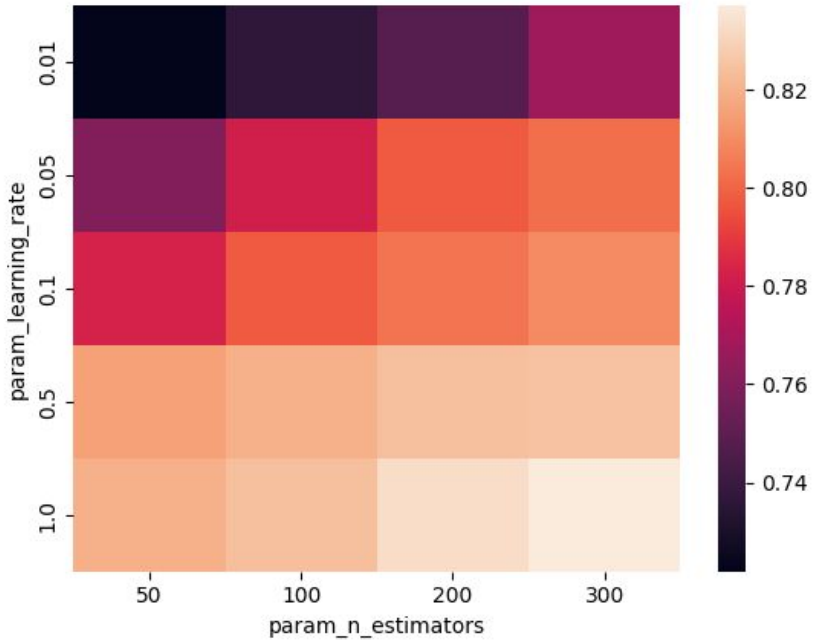
Adaboost Parameters:

- Max Depth, Learning Rate, # Estimators

RandomForest Parameters:

- Max Depth, # Estimators

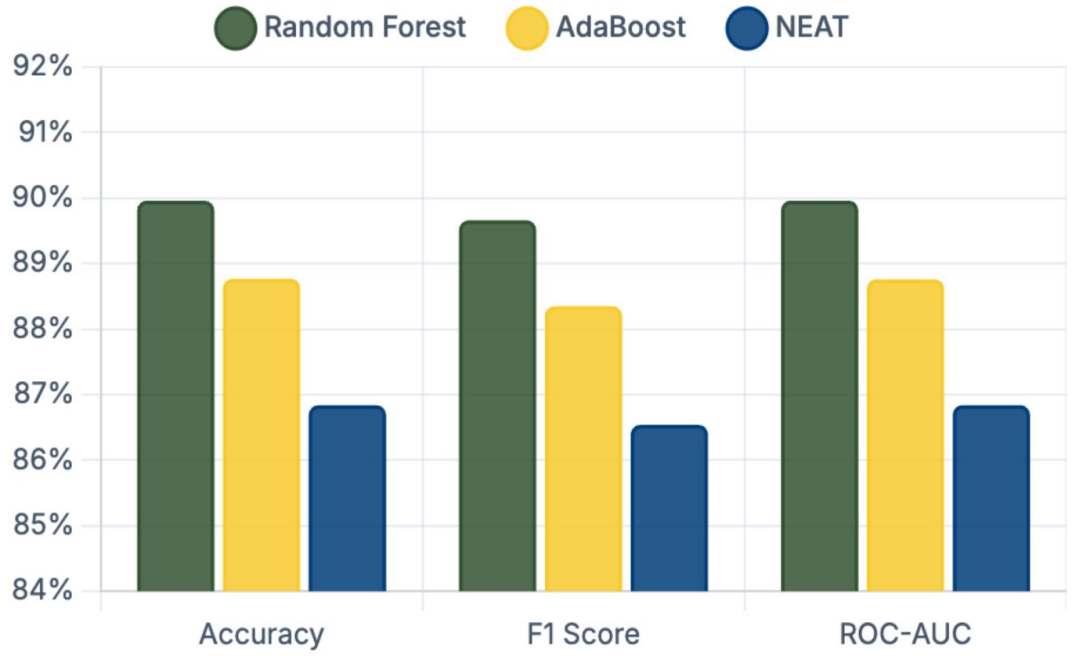
Example GridSearchCV Output



Model Metrics

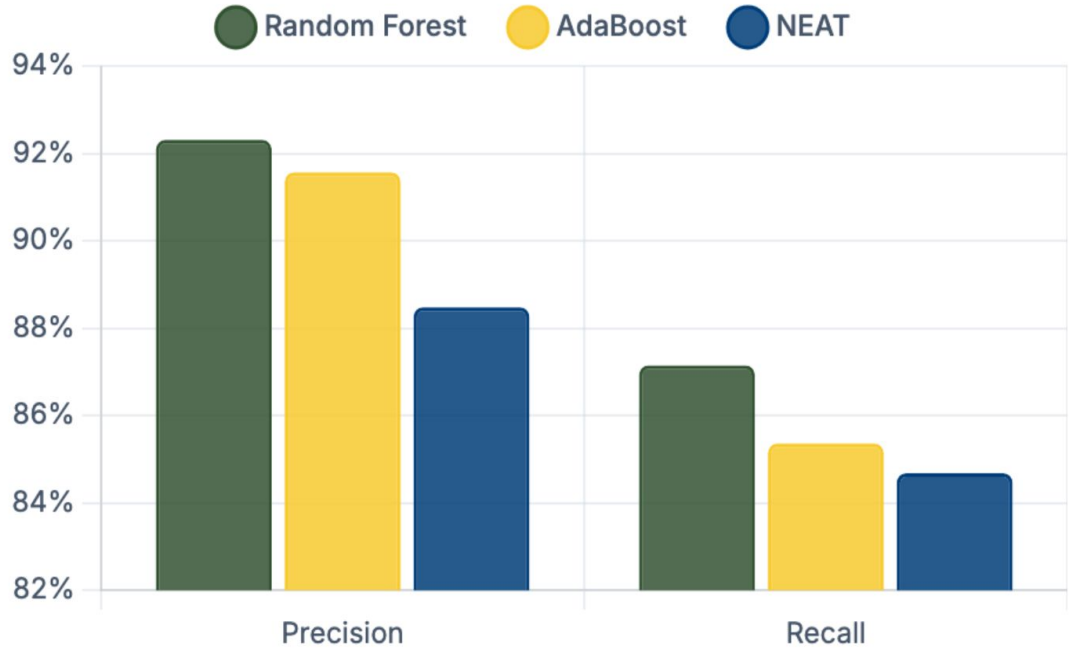
Overall Performance

Accuracy, F1 Score, and ROC-AUC (Higher is better)



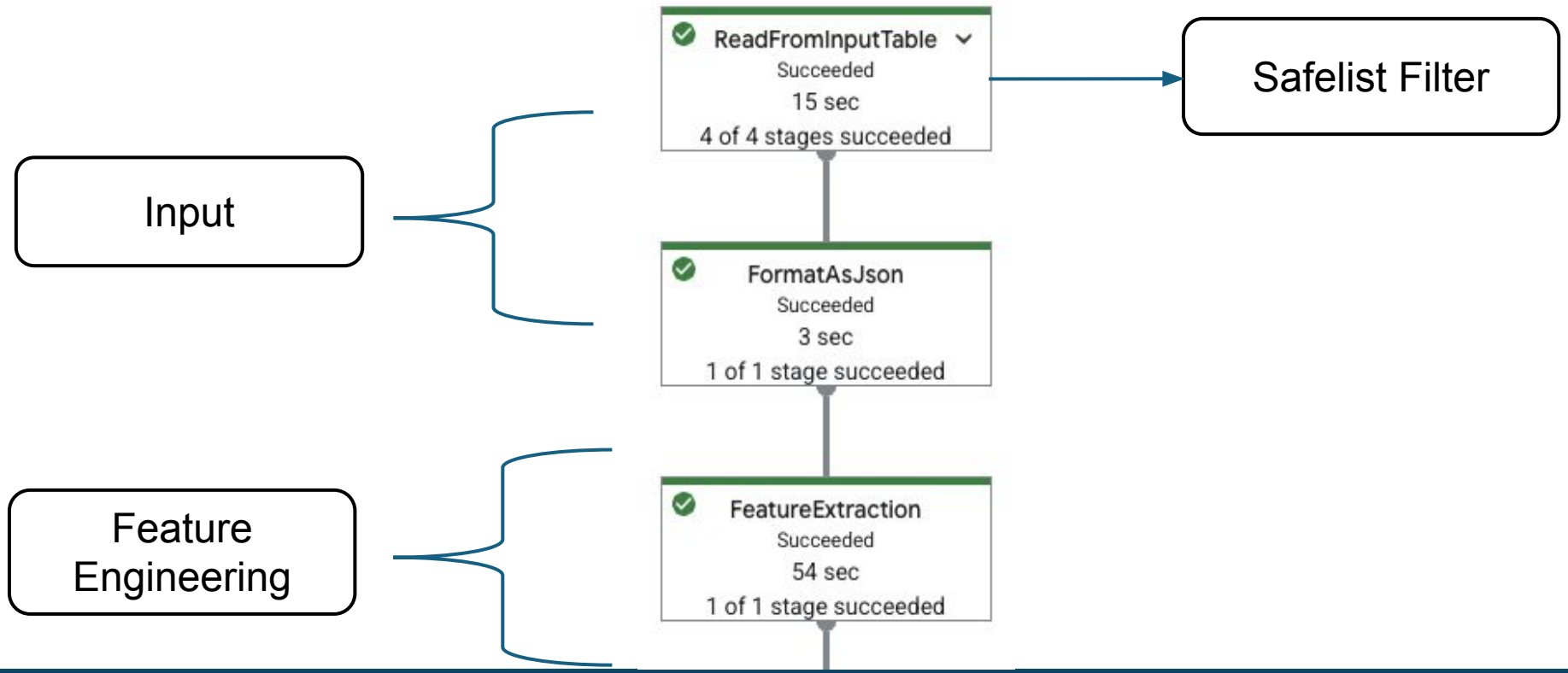
Precision vs. Recall Trade-off

Detailed classification metrics (Higher is better)



Step 6: Model Deployment

```
pipeline
| "ReadFromInputTable" >> beam.io.ReadFromBigQuery(table = input_table)
| "FormatAsJson" >> beam.Map(lambda row: json.dumps(row))
| "FeatureExtraction" >> beam.ParDo(dga_transforms.ExtractFeatures())
```

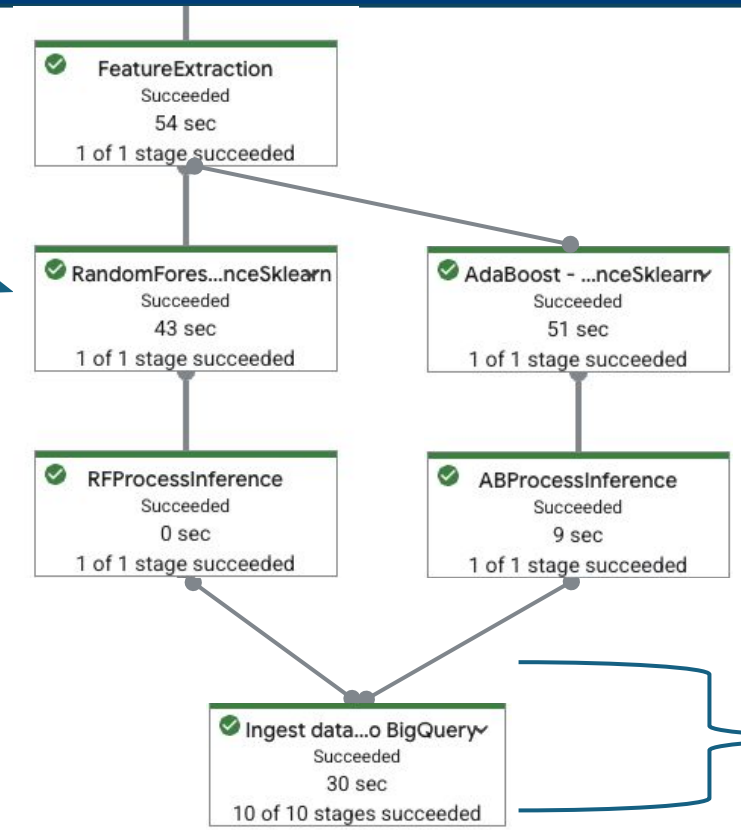


Step 6: Model Deployment

Run Inference Function

3 Requirements:

- Model Registry
- Infrastructure
- Function Tuning



Run our Model

Store results

```
| "RandomForest - RunInferenceSklearn" >> RunInference(model_handler=rf_keyed_sklearn_model_handler)
| "RFProcessInference" >> beam.ParDo(dga_transforms.RFPostProcessor())
| "AdaBoost - RunInferenceSklearn" >> RunInference(model_handler=ab_keyed_sklearn_model_handler)
| "ABProcessInference" >> beam.ParDo(dga_transforms.ABPostProcessor())
| 'Ingest data into BigQuery' >> beam.io.WriteToBigQuery(
    output_table,
    write_disposition=beam.io.BigQueryDisposition.WRITE_APPEND)
```

Production Results

Questionable DGAs

g10696554090.co
g11488891430.co
g10498469755.co
g10300385420.co
g10102301085.co
g1386590345.co
g1386590346.co

jbxgbdetzfgpn.space
us205.speedycdn.space
pl21.speedycdn.space
mpipnopvjuuee.space
api.umanoff-analytics.space

Legitimate DGAs

4dc5a13432ef8d11776524c928f32123.fp.measure.office.com
9bcfd446986a426363df26c280680da5.fp.measure.office.com
6b8af5bbc6a94335a068c6332f409954.fp.measure.office.com
6098ce6fe65a45c48c76a0c38f666873.fp.measure.office.com
deb68d6447d64d4989570305d1f596e1.fp.measure.office.com
e8afa997b74841cc96b79b8077930c8d.fp.measure.office.com
ab91b11198c88ac6625abd3e06e8d8b6.fp.measure.office.com
4a0b0e75ff776b5b462099e1c902adab.fp.measure.office.com
aa59ed645b514e29b3a4d8ab6a40065d.fp.measure.office.com
e63320674e9c49f2999385a854da3efc.fp.measure.office.com
9a05542c35f34fa5835a6dc057c93fa0.fp.measure.office.com
3b0c3acb2bd5f0f55678991ca93a0680.fp.measure.office.com
ca2d1af32b284e75a9aa0e17e7485746.fp.measure.office.com
3454c56860c94d5f93be4c63836a6a3d.fp.measure.office.com
82f76631b28e43e18d587cc0ed955494.fp.measure.office.com
f18c952c76d44e30bbdfd92569fac908.fp.measure.office.com
a7b8b99b5279b867bb5a373a65d74a96.fp.measure.office.com
a37ef3921eb9465b9b900117196506c1.fp.measure.office.com

Not DGAs

www.northjerseyhousehunt.com
www.nationalacademyleague.com
northcentralcardinals.com
thewashingtonstandard.com
cdn-assets.custompricecalculator.com
coloradooutdoorsmag.com
visitor.r20.constantcontact.com