54,000,000

DatAvengers Cybersecurity

Amaan Ansari
[Am-aan]
Lead Analyst

MS in Business Analytics and Information Management

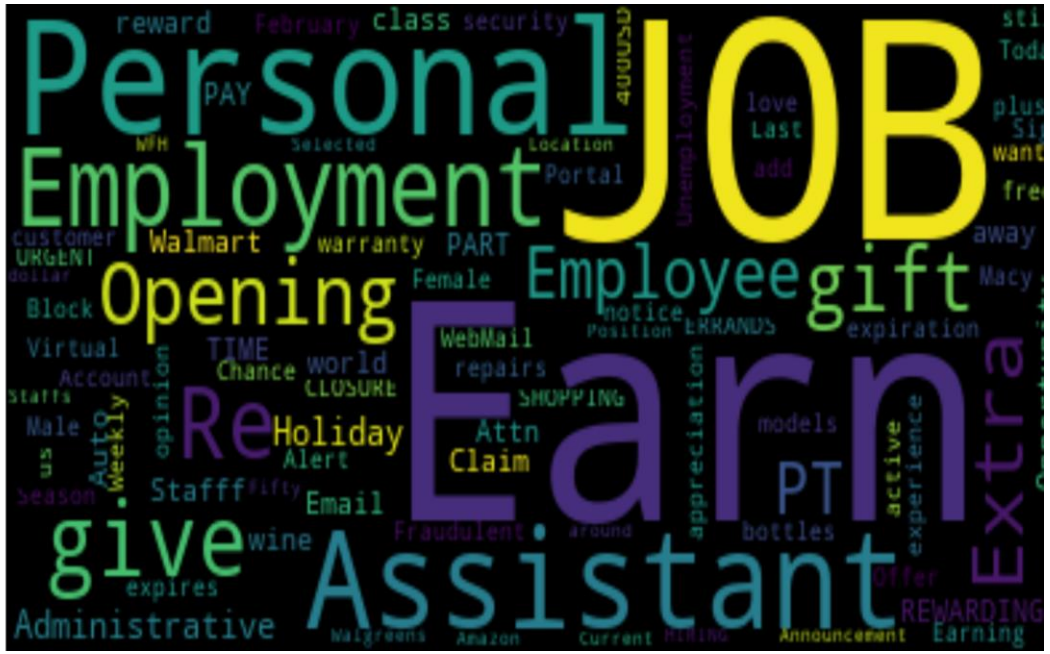
Jai Woo Lee
[Jay]
Data Analyst

MS in Business Analytics and Information Management

Paul Chen
[Paul]
Data Scientist

MS in Business Analytics and Information Management

WHAT IS PHISHING?



Most frequent words in phishing email

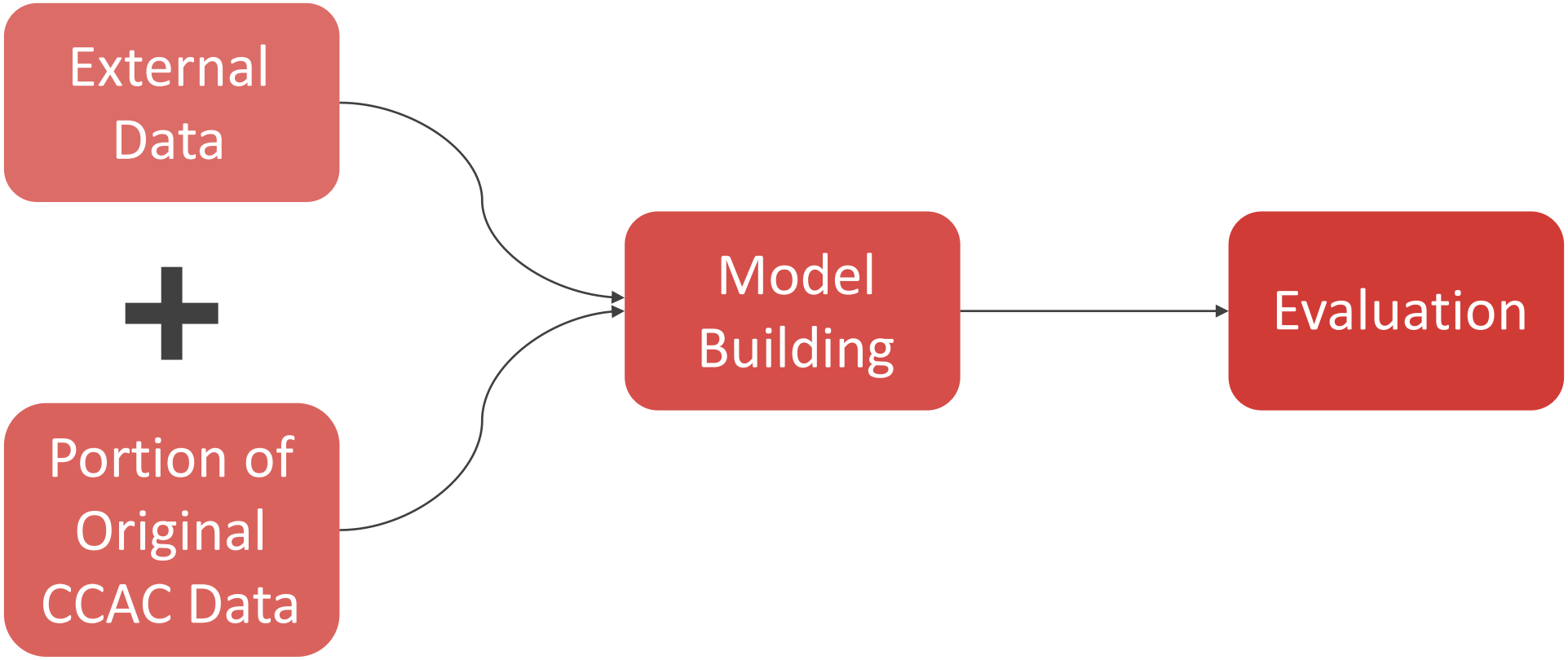
Job offer is the most frequent phishing email

Promotion and Rewards are not only spam but also phishing

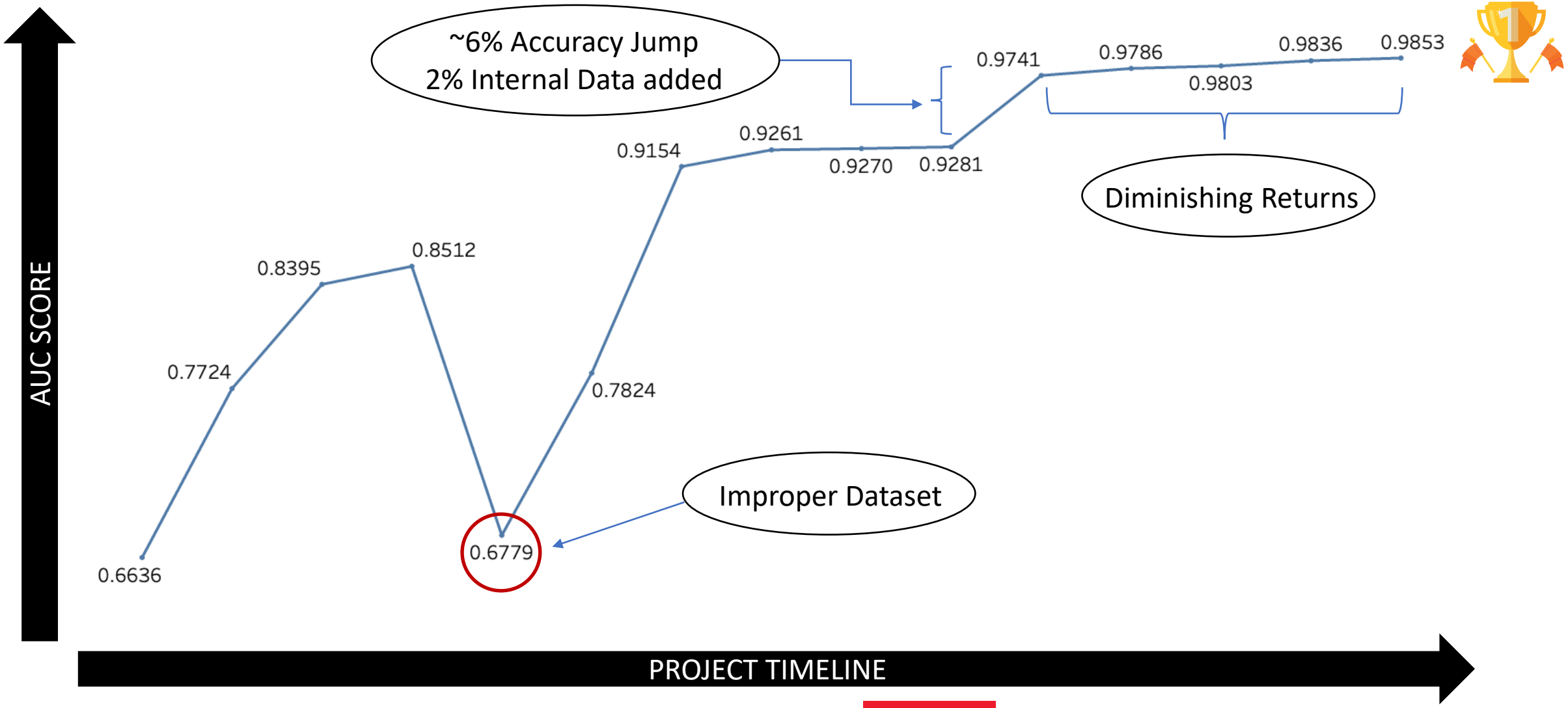
CAPITALIZATION!!

Upper letter and ! are widely use to show the urgency and get users' attention

APPROACH

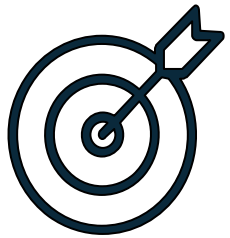

External Data
+
Portion of Original CCAC Data
Model Building
Evaluation

ROADMAP


SVM+XGBoost
~6% Accuracy Jump
2% Internal Data added
0.6636
0.7724
0.8395
0.8512
0.6779
0.7824
0.9154
0.9261
0.9270
0.9281
0.9741
0.9786
0.9803
0.9836
0.9853
Diminishing Returns
Improper Dataset
AUC SCORE
PROJECT TIMELINE

WHY INTERNAL DATA?

Internal Data Patterns



Higher Accuracy

Creates custom fit

URL Link Discounting

F.I.S.H PROTOTYPE

Filter

FISH's high accuracy model can filter most of the phishing email

Identify

The hackers are improving everyday so is FISH. FISH identifies the recent phishing key words and update our model everyday

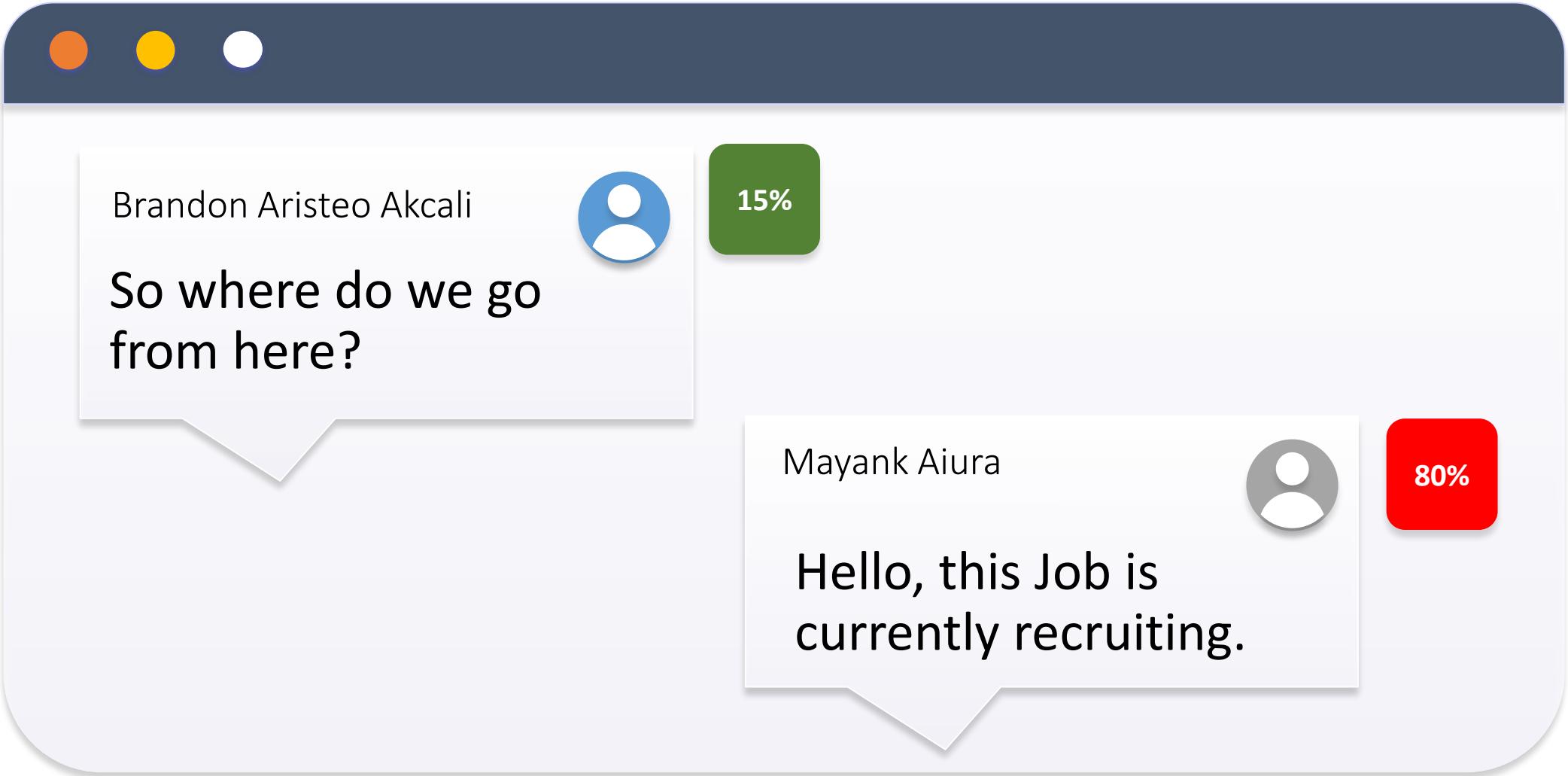
Summary

FISH provides monthly summary of the latest phishing news and word cloud

Hook-free

FISH provides user friendly interface letting user decide what is phishing and the adjustable threshold design

F.I.S.H PROTOTYPE



F.I.S.H PROTOTYPE

Low Risk

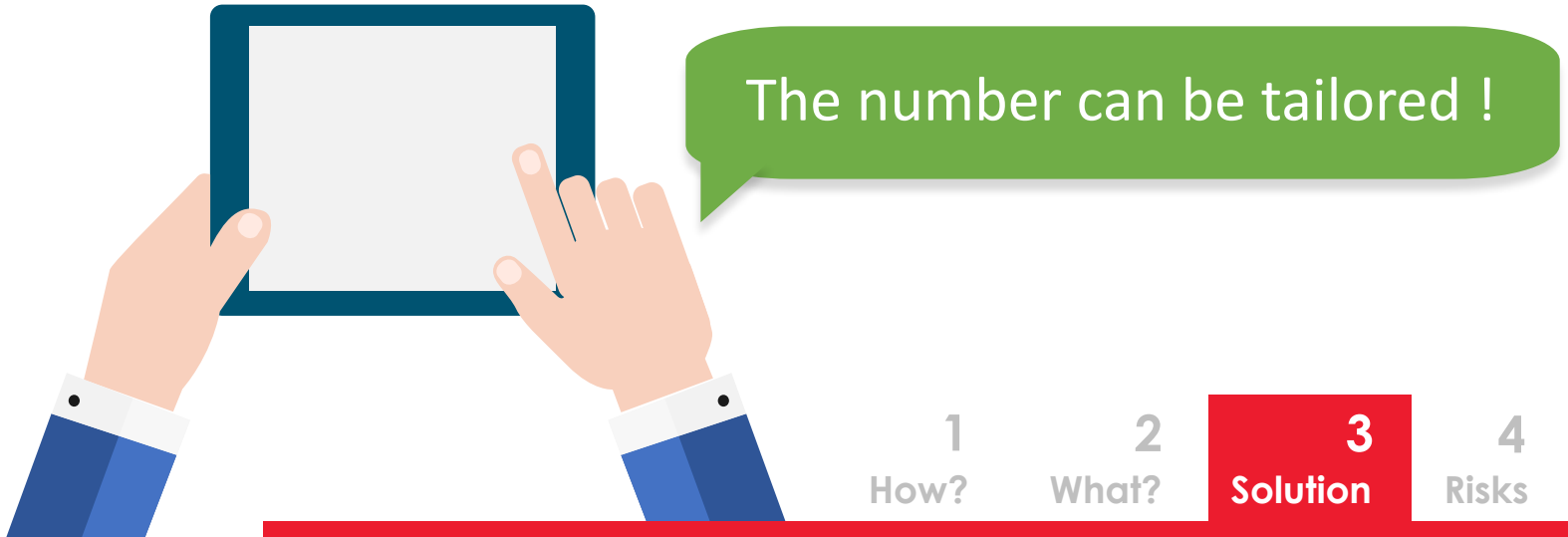
Less than 30%

Neutral

Between 30% to 60%

High Risk

Higher than 60%



RISK ANALYSIS

1 Confusion Matrix

Predicted \ Actual	Phishing	Non-Phishing
Phishing	9700(TP)	60(FP)
Non-Phishing	90(FN)	150(TN)

2 Cost Matrix

Predicted \ Actual	Phishing	Non-Phishing
Phishing	$0	$0
Non-Phishing	$224*	$0

Opportunity Cost

- Work-related emails
- Important personal emails

Exposed to Phishing Emails

- Malware/ransomware
- Data breach

3 Mitigation Strategy

- Adjust a threshold

Potential Cost for Misclassification

$20,160

* Average Loss in US (2021): $14.8M

*In 2020, Number of phishing complaints: 241,342 | Loss: $54 million in US (IC3 Report, 2020), $54M / 241,342 = 224

CONCLUSION

Identification of Phishing

Word Cloud and Pattern Analysis
Ex. Upper Letters, Job Offers, Promotions, !!!!!!!!!!!!

Strategic Data Selection

Incorporated internal data gradually and proved adding ~2% data generates the highest accuracy rate efficiently

98.53%

AUC Score based on Ensemble model

F.I.S.H

High accuracy Filter,
Identify new phishing key words,
Monthly Summary,
Adjustable threshold to provide Hook-free environment

Thank You