JPAAWG 5TH MEETING - 2022

Using AI in Threat Detection

Examples and 2022 Email Threat Review

Severin Walker – Vade Director of Products (ISP)

Maxime Meyer
Lead Research Scientist

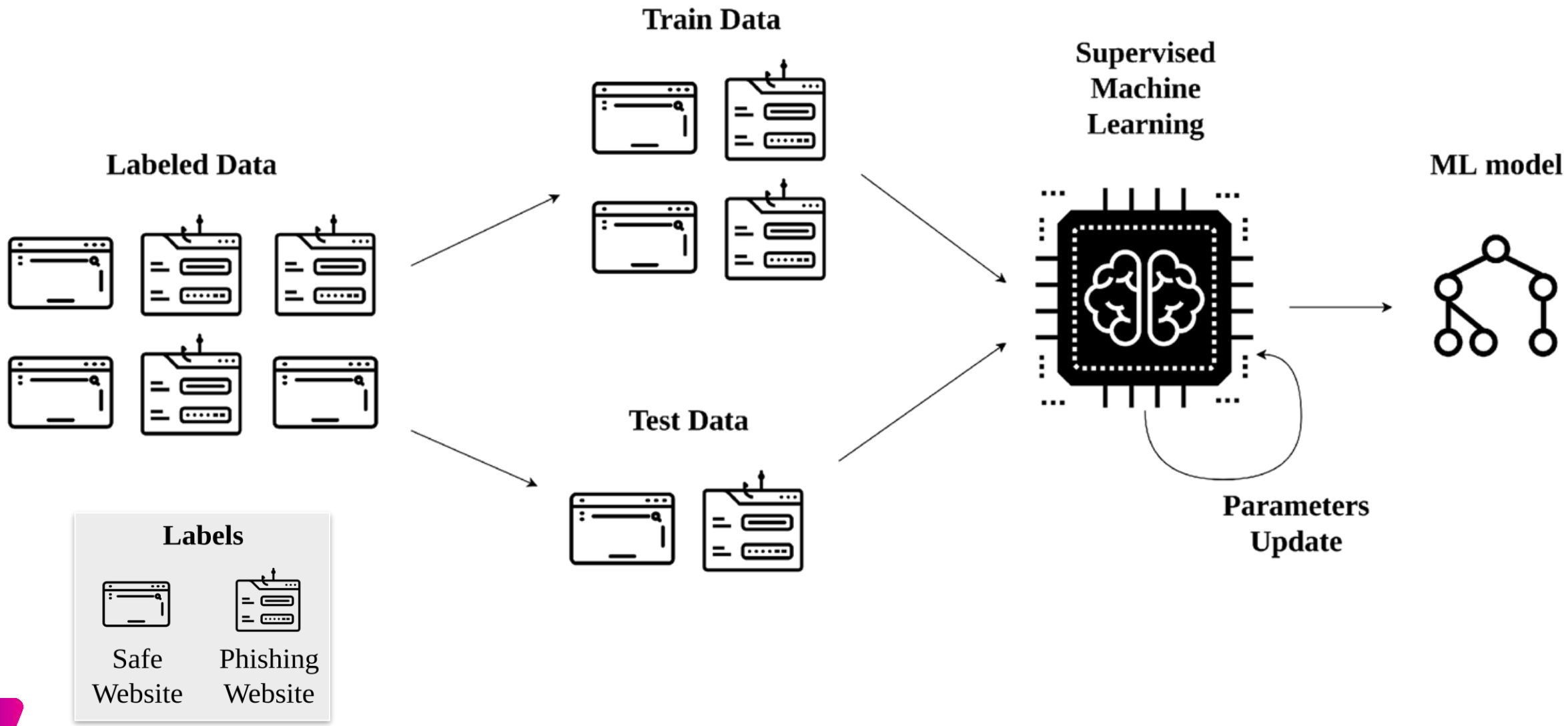
Gabriel Loiseau
Research Scientist

Outline

- Detecting phishing using Random Forest
- Using Deep Learning to improve phishing detection
- Phishing detection in production
- Q&A

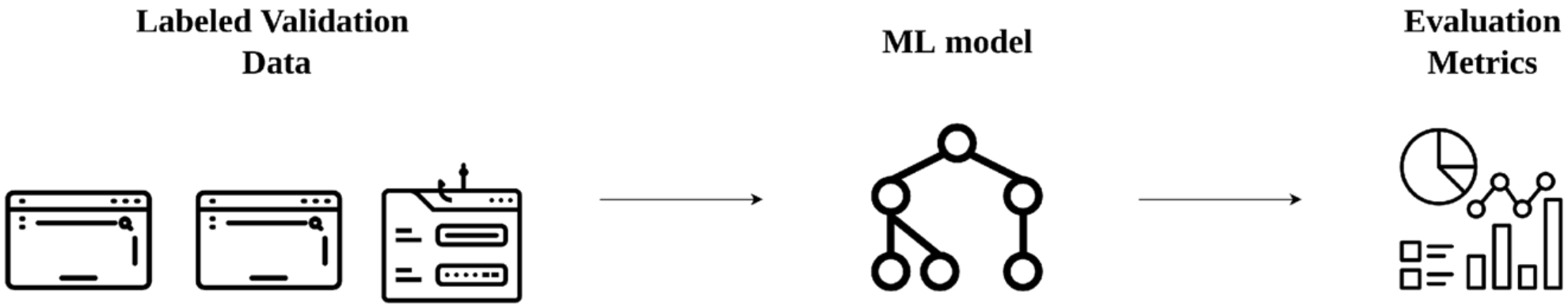
Detecting Phishing Using Random Forest

Supervised Machine Learning - Training



Supervised Machine Learning - Evaluation

- Evaluate on a dataset different from the training data
- Measure metrics (FPR, precision, …)
- Ensure non regression

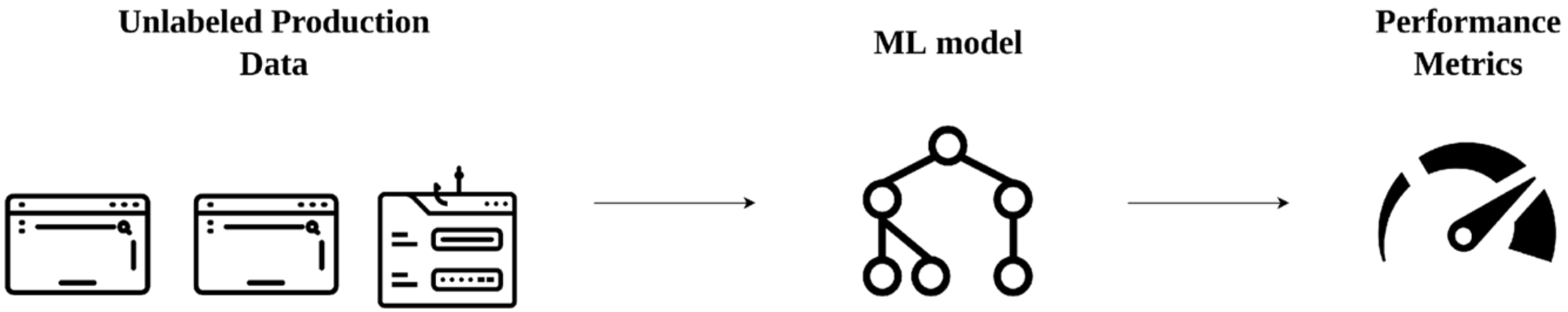


Supervised Machine Learning - Metrics

Truth / Prediction	Phishing	Safe
Phishing	True Positive (TP)	False Positive (FP)
Safe	False Negative (FN)	True Negative (TN)

- FPR = FP / (FP+TN)
- Recall = TP / (TP+FN)
- Precision = TP / (TP+FP)

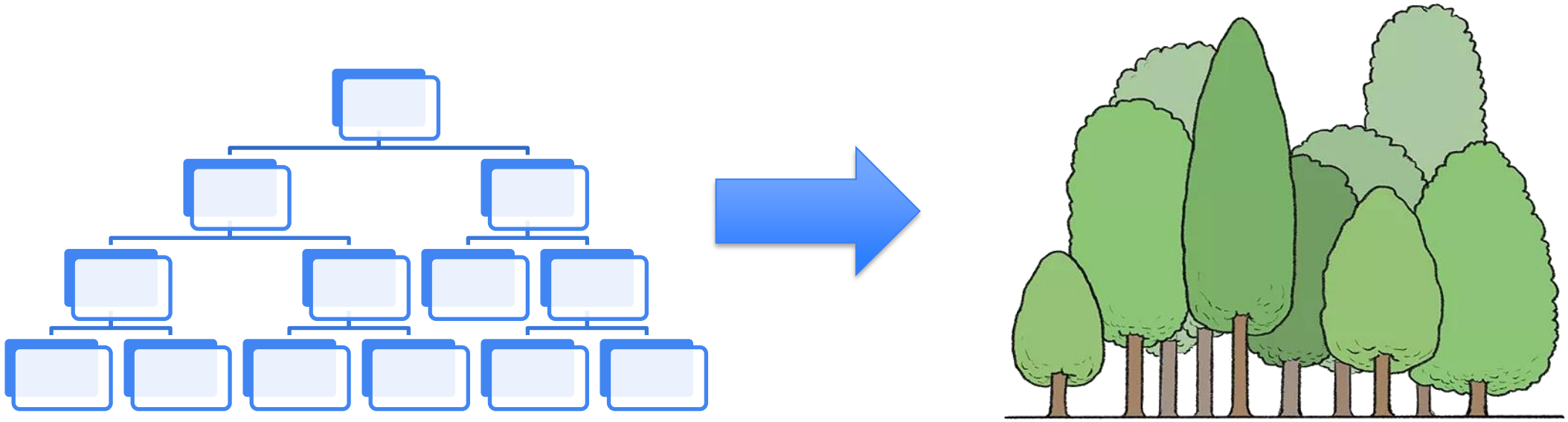
Supervised Machine Learning - Production



- Make prediction on production data
- Measure metrics (speed, memory usage, …)
- Evaluation based on emails from FBL

Phishing detection with Random Forest

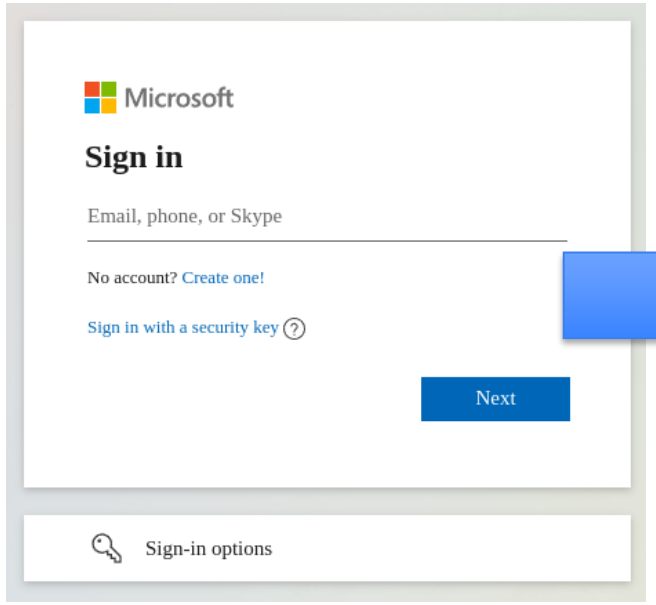
Random Forest[1, 2] use an ensemble of Decision Trees



[1]Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. doi: 10.1023/A:1010933404324
[2]Doyen Sahoo, Chenghao Liu, and Steven C. H. Hoi. Malicious URL Detection using Machine Learning: A Survey. 2017. doi: 10.48550/ARXIV.1701.07179.

Phishing detection with Random Forest

http://fx09092022fax81docs-1312962597.cos.ap-guangzhou.myq...



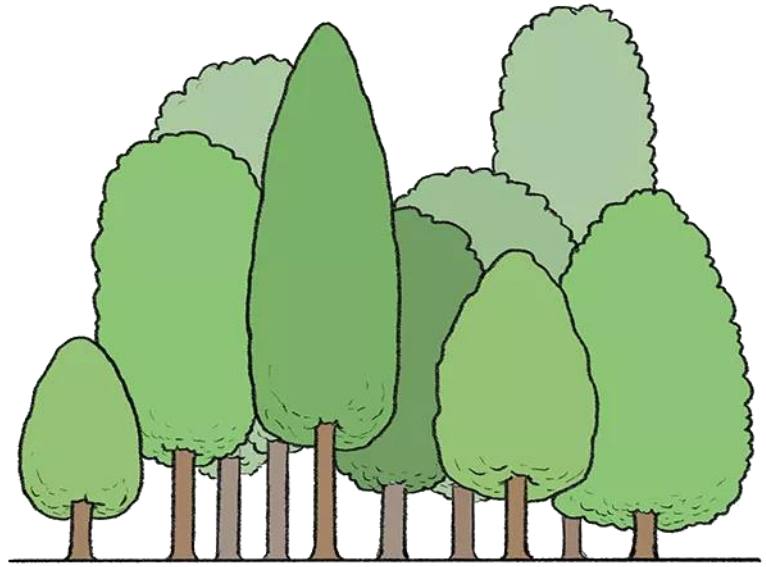
Feature	Value
Document Size	145
Number of tags	27
"login" in subdomain	False
TLD	".com"
...	

Phishing detection with Random Forest

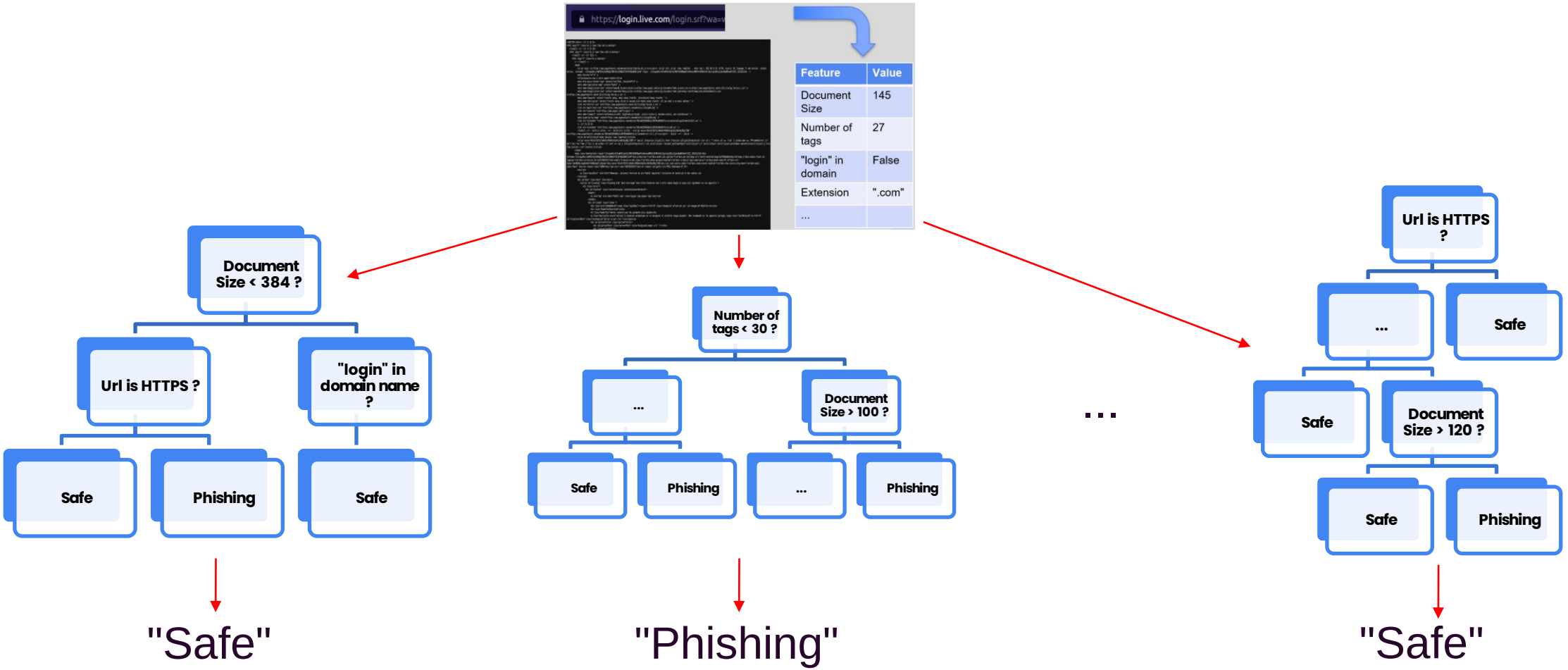
Feature / Document	Document Size	Number of tags	...	"login" in subdomain	Label
Document 1	145	27	...	False	Phishing
Document 2	801	50	...	True	Safe
...	...	...	...	...	...
Document n	59	10	...	False	Safe

Phishing detection with Random Forest

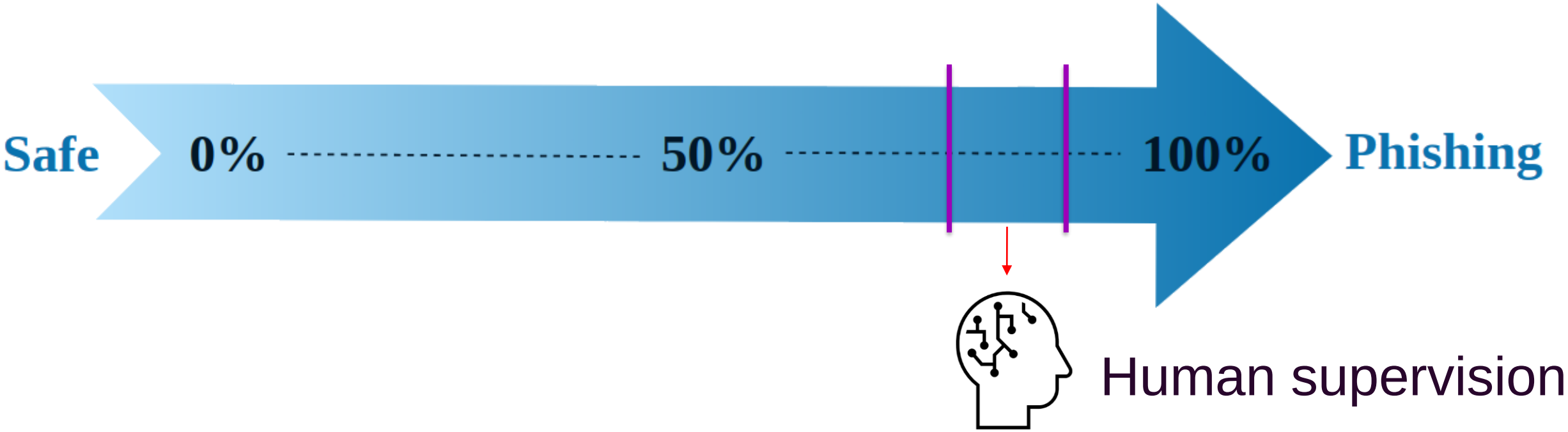
Feature / Document	Document Size	Number of tags	...	"login" in subdomain	Label
Document 1	145	27	...	False	Phishing
Document 2	801	50	...	True	Safe
...	...	...	...	...	...
Document n	59	10	...	False	Safe



Phishing detection with Random Forest



Decision boundary



Model outputs a "Phishing Risk" between 0% and 100%

50% ? 75% ? 90% ?

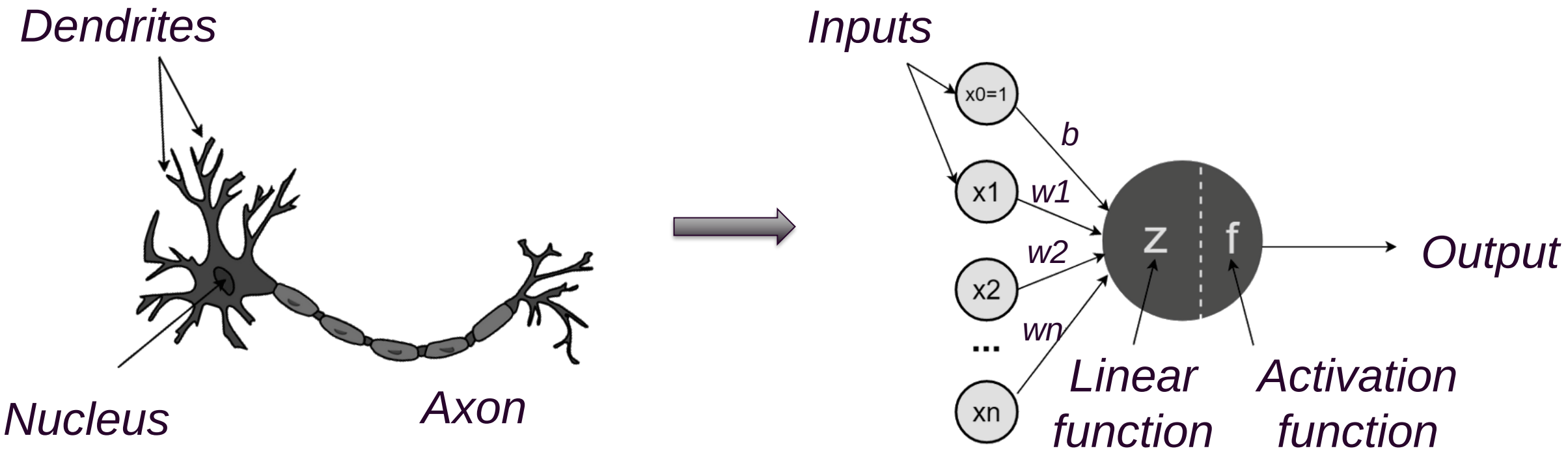
Limitations

Random Forest in production generalizes well and gives good results but …

- Still some FPs, quite a lot of FNs
- Too many samples sent yearly to analysts (> 250 000)
- New supervised learning technologies seem promising

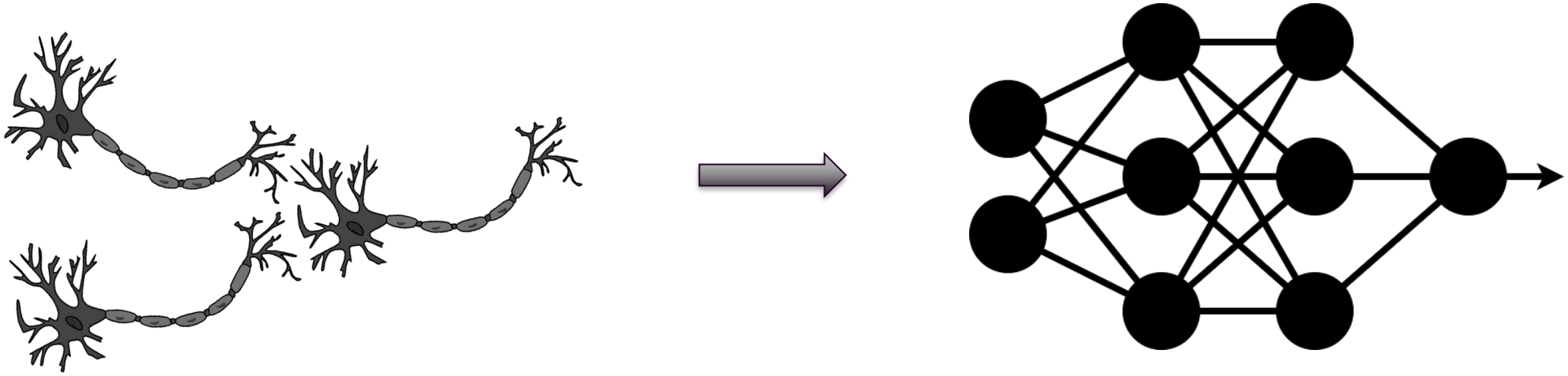
Using Deep Learning to Improve Phishing Detection

Deep Learning and Neural Networks

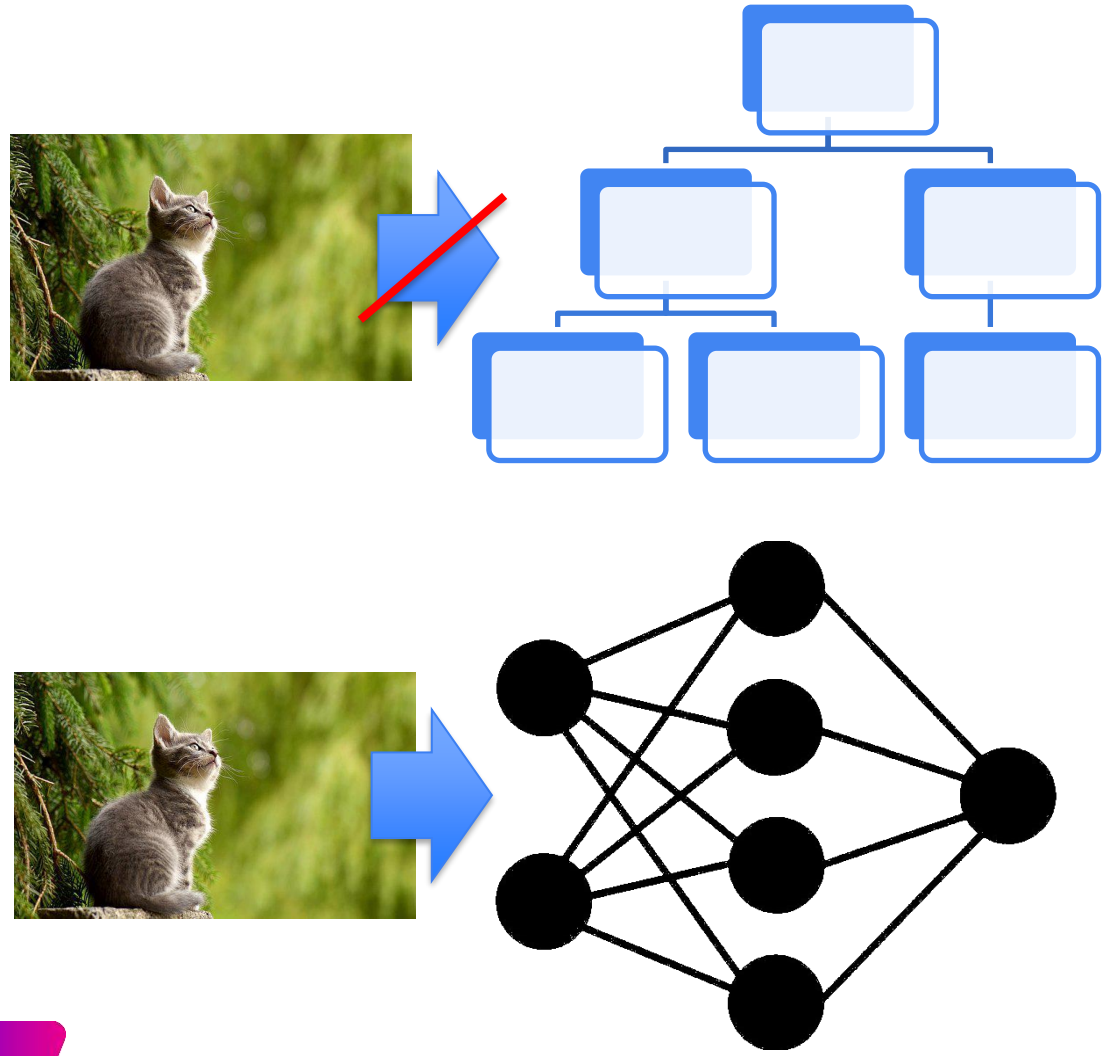


Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Deep Learning and Neural Networks



Deep Learning and Neural Networks

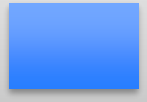


- Deep Learning can make sense of raw data by generating its own "abstract" features
- Automatic feature extraction
- Cannot be done with classical machine learning

Pros and Cons of Deep Learning

Pros

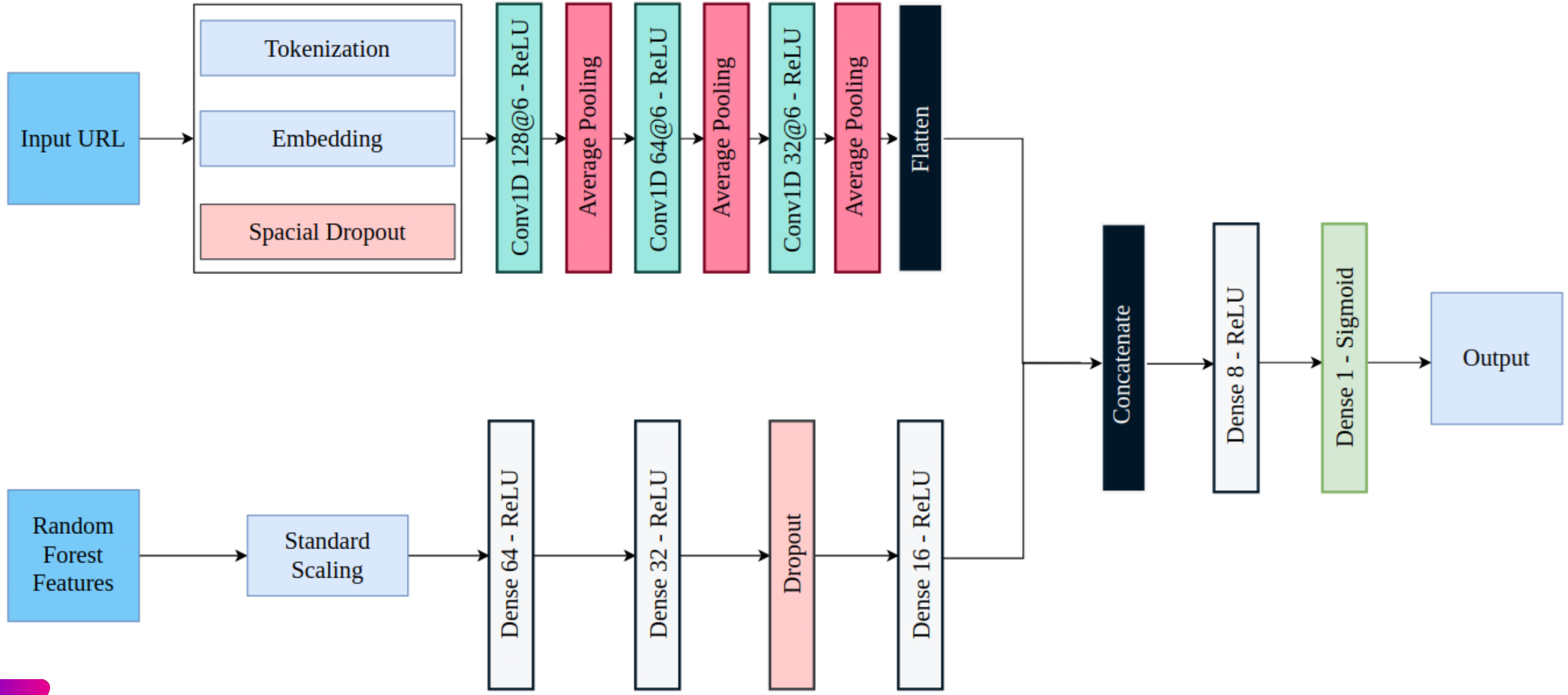
- Fewer preprocessing
- Less expert knowledge
- Can be used with all numerical data points



Cons

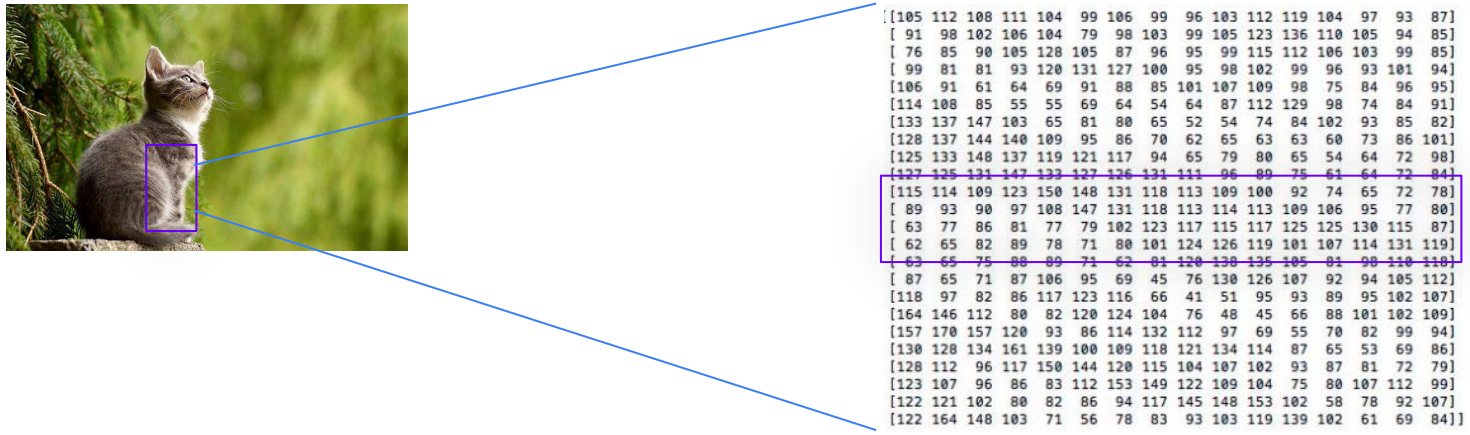
- More computational power (GPU)
- Hardly explainable
- Requires a lot of data to work properly

Architecture



Data inputs for Deep Learning

- Transform real life objects into numerical values
- Images→Pixel values

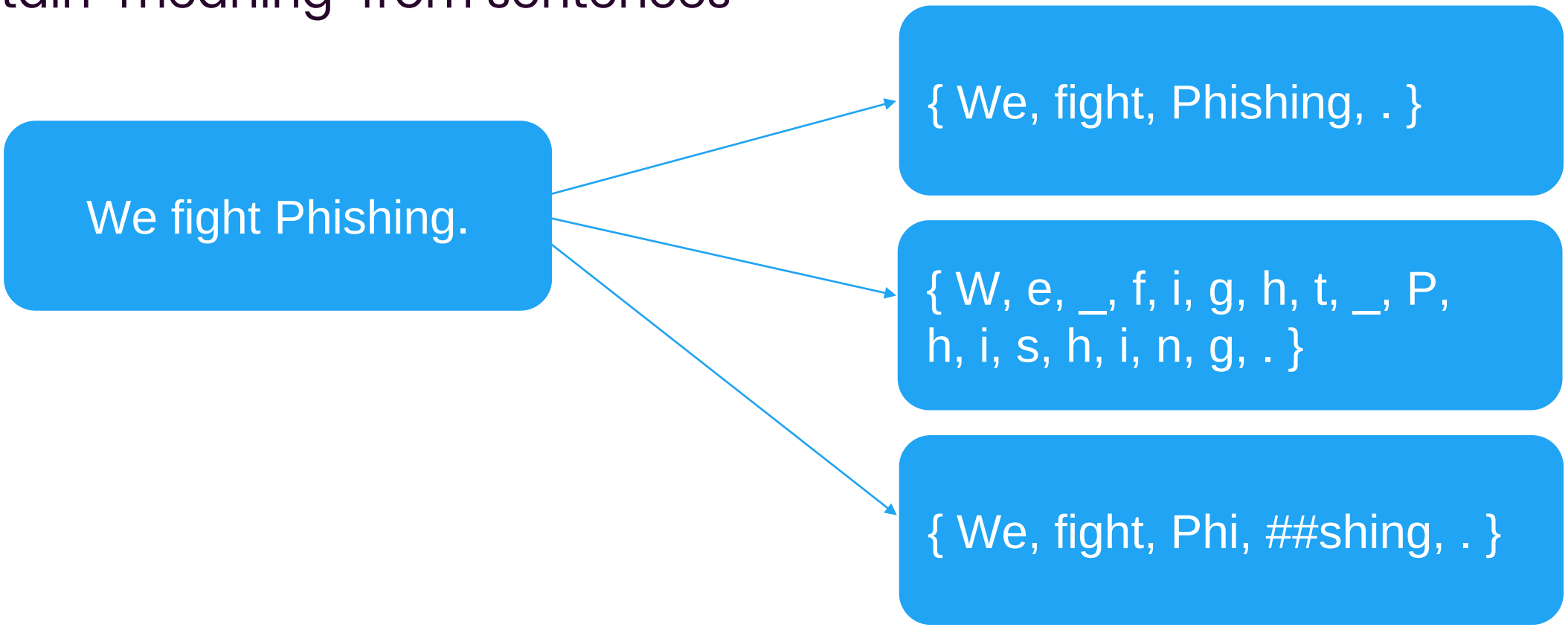


```
[[105 112 108 111 104  99 106  99  96 103 112 119 104  97  93  87]
 [ 91  98 102 106 104  79  98 103  99 105 123 136 110 105  94  85]
 [ 76  85  90 105 128 105  87  96  95  99 115 112 106 103  99  85]
 [ 99  81  81  93 120 131 127 100  95  98 102  99  96  93 101  94]
 [106  91  61  64  69  91  88  85 101 107 109  98  75  84  96  95]
 [114 108  85  55  55  69  64  54  64  87 112 129  98  74  84  91]
 [133 137 147 103  65  81  80  65  52  54  74  84 102  93  85  82]
 [128 137 144 140 109  95  86  70  62  65  63  63  60  73  86 101]
 [125 133 148 137 119 121 117  94  65  79  80  65  54  64  72  98]
 [127 125 131 147 133 127 126 131 111  96  89  75  61  64  72  84]
 [115 114 109 123 150 148 131 118 113 109 100  92  74  65  72  78]
 [ 89  93  90  97 108 147 131 118 113 114 113 109 106  95  77  80]
 [ 63  77  86  81  77  79 102 123 117 115 117 125 125 130 115  87]
 [ 62  65  82  89  78  71  80 101 124 126 119 101 107 114 131 119]
 [ 63  65  75  80  89  71  62  81 120 138 135 105  81  98 110 110]
 [ 87  65  71  87 106  95  69  45  76 130 126 107  92  94 105 112]
 [118  97  82  86 117 123 116  66  41  51  95  93  89  95 102 107]
 [164 146 112  80  82 120 124 104  76  48  45  66  88 101 102 109]
 [157 170 157 120  93  86 114 132 112  97  69  55  70  82  99  94]
 [130 128 134 161 139 100 109 118 121 134 114  87  65  53  69  86]
 [128 112  96 117 150 144 120 115 104 107 102  93  87  81  72  79]
 [123 107  96  86  83 112 153 149 122 109 104  75  80 107 112  99]
 [122 121 102  80  82  86  94 117 145 148 153 102  58  78  92 107]
 [122 164 148 103  71  56  78  83  93 103 119 139 102  61  69  84]]
```

- Text→???
- Tokenization step

What is tokenization ?

- Split text into smaller units (words)
- Obtain "meaning" from sentences



From text to values

- BERT Wordpiece Tokenizer[1]
- Used in other URL-based approaches[2]

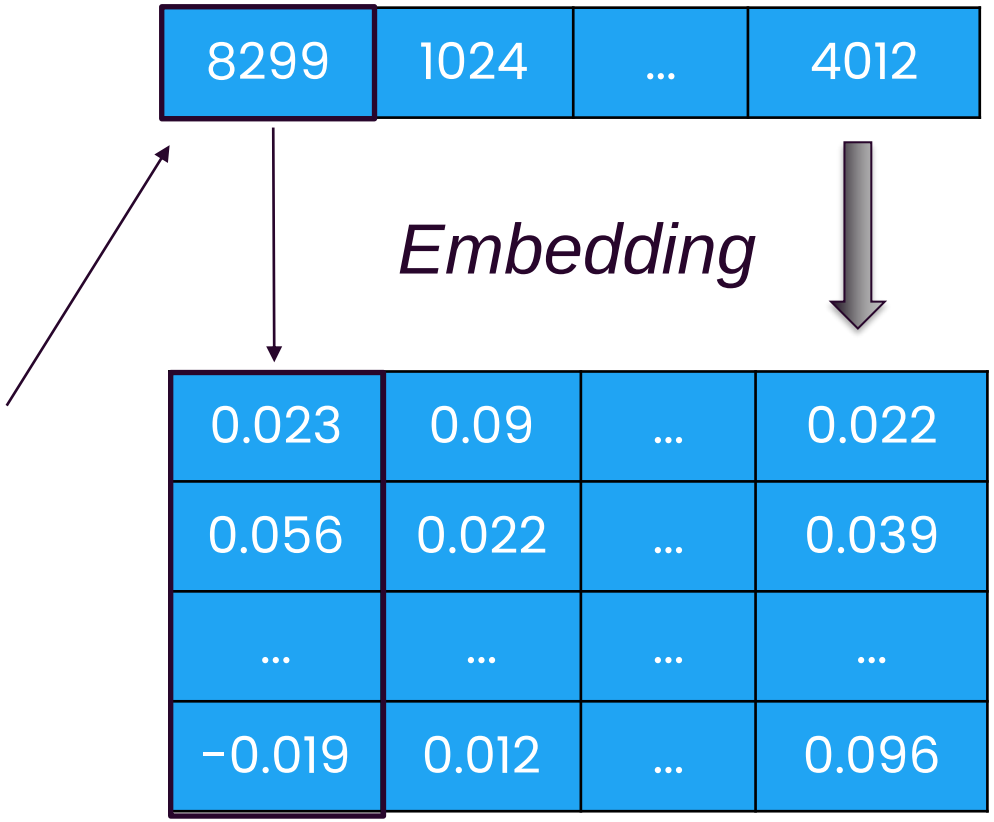
Original URL	`http://thisisaphishingurl.com`
Tokenized URL	`[http, :, /, /, this, ##isa, ##phi, ##shing, ##ur, ##l, ., com]`
Token indexes from vocabulary	`[8299,1024,1013,1013,2023,14268,21850,12227,3126,2140,1012, 4012]`

[1]Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2

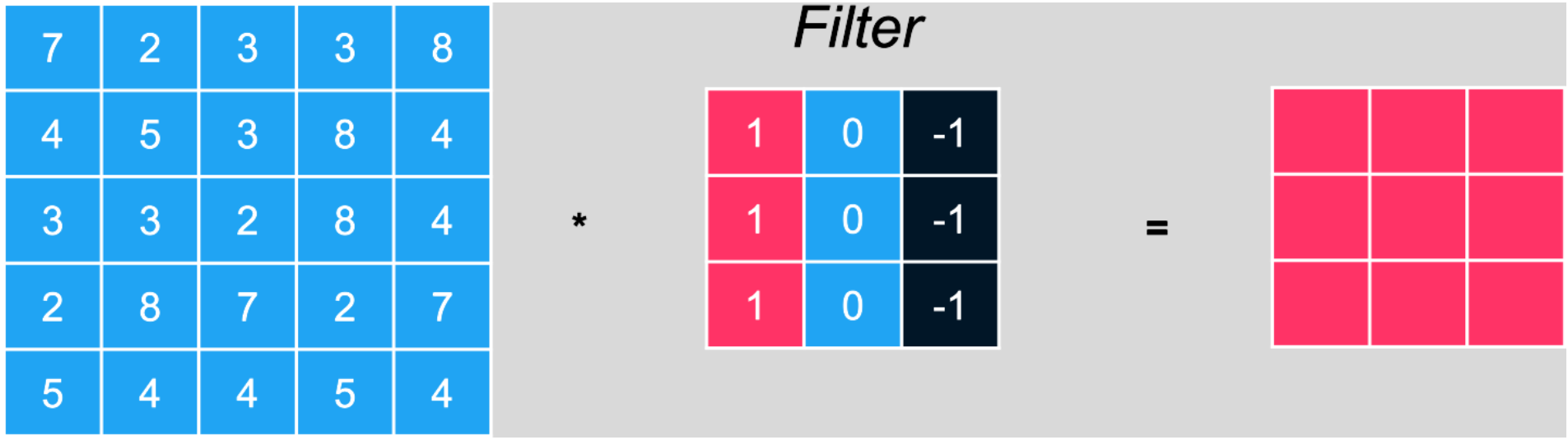
[2]Maneriker, Stokes, Lazo, Carutasu, Tajaddodianfar, Gururajan (*27 August 2021) "*URLTran: Improving Phishing URL Detection Using Transformers". arXiv:2106.05256

Embedding Layer

Tokenized URL	[http, :, /, /, this, ##isa, ##phi, ##shing, ##ur, ##l, ., com]
Token indexes from vocabulary	[8299,1024,1013,1013,2023,14268,21850,12227,3126,2140,1012,4012]



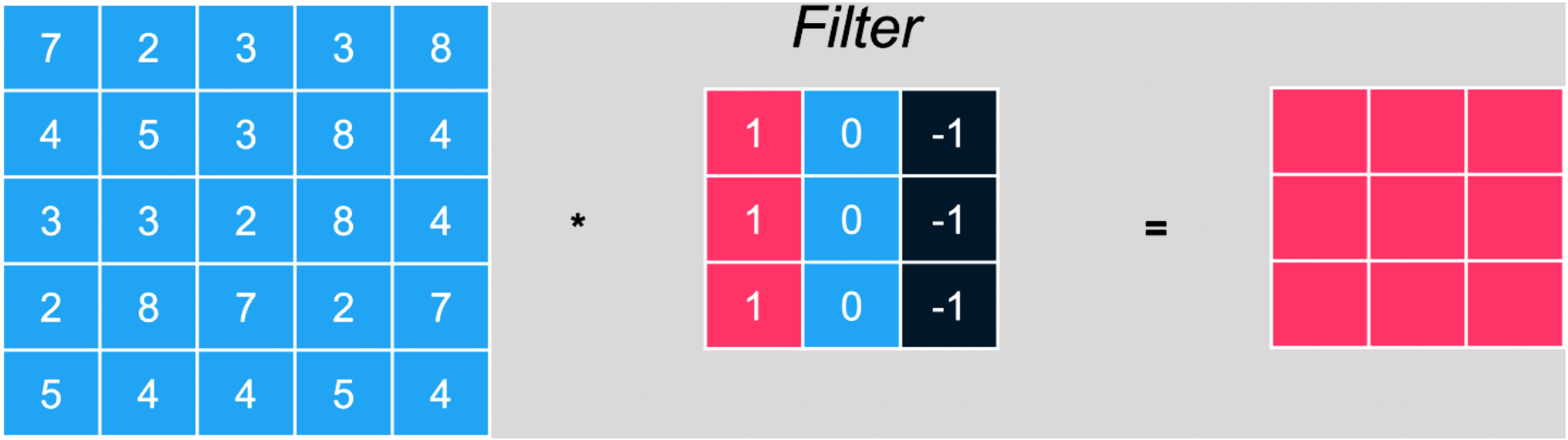
Convolution Layer[1]



_ x 1 + _ x 0 + _ x (-1) +
_ x 1 + _ x 0 + _ x (-1) +
_ x 1 + _ x 0 + _ x (-1) = ?

[1]LeCun, Yann; Bengio, Yoshua (1995). "Convolutional networks for images, speech, and time series"

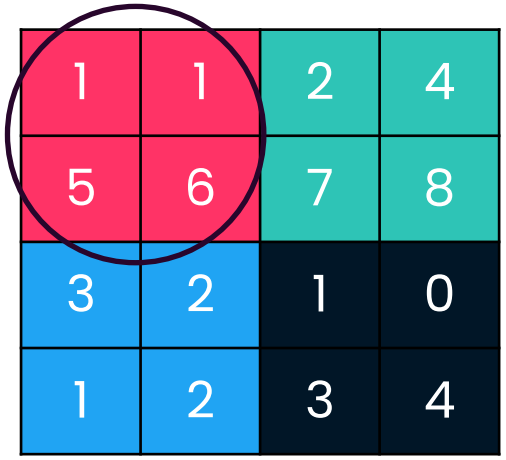
Convolution Layer[1]



_ x 1 + _ x 0 + _ x (-1) +
_ x 1 + _ x 0 + _ x (-1) +
_ x 1 + _ x 0 + _ x (-1) = ?

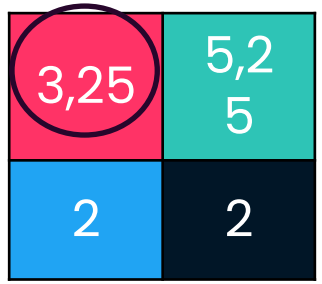
[1] LeCun, Yann; Bengio, Yoshua (1995). "Convolutional networks for images, speech, and time series"

Average Pooling Layer[1]



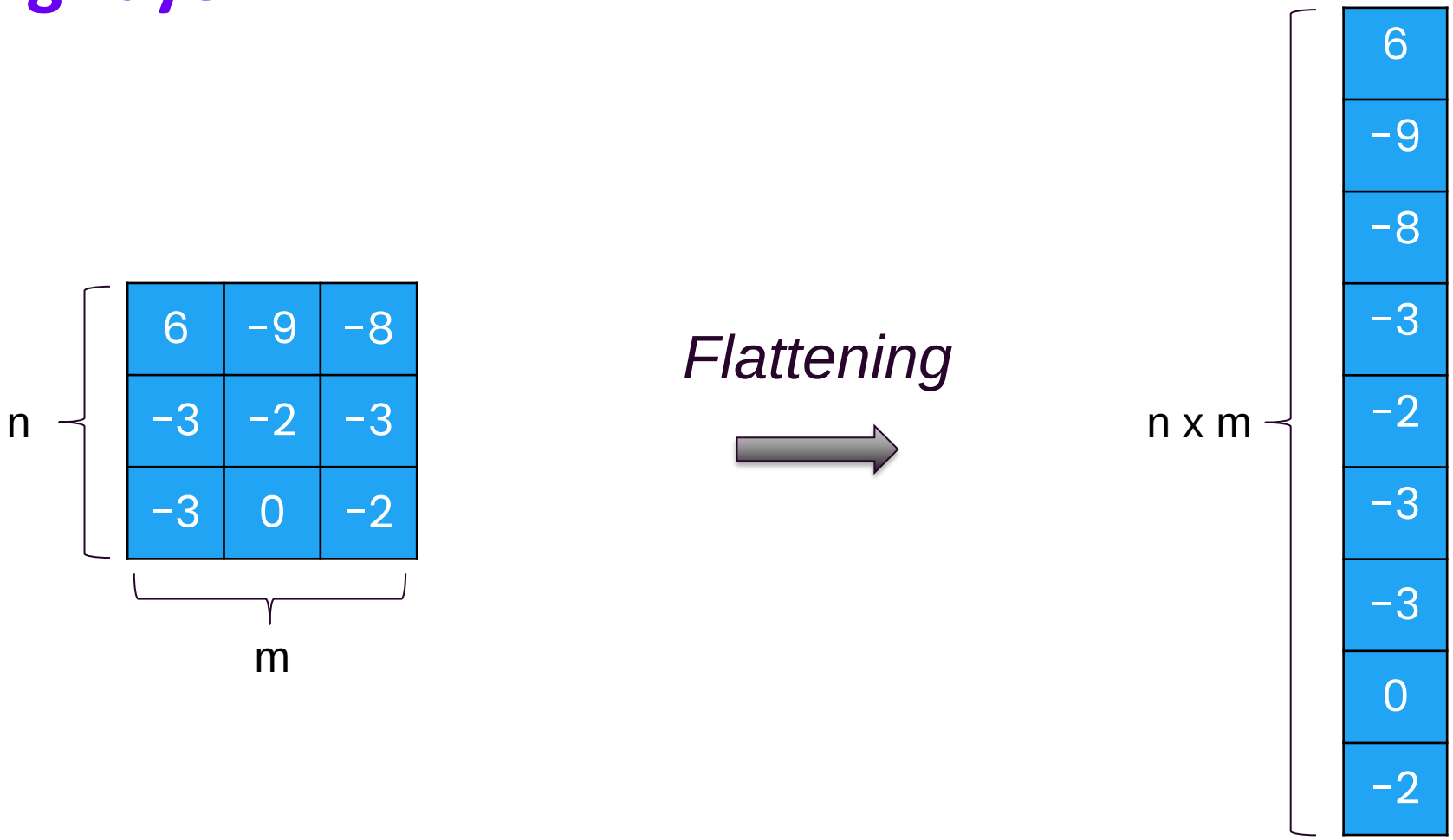
Average Pooling

(1+1+5+6)/4 = 3.25



[1] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

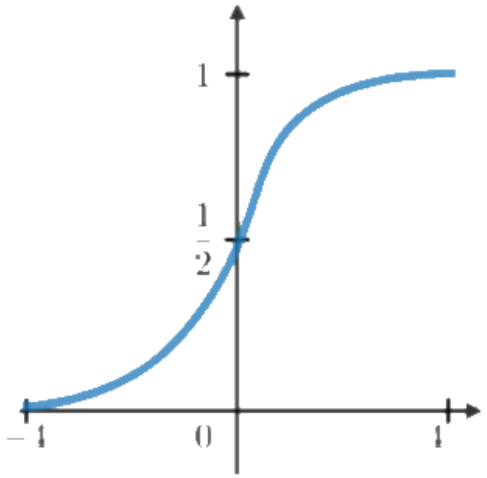
Flattening Layer



Activation Functions

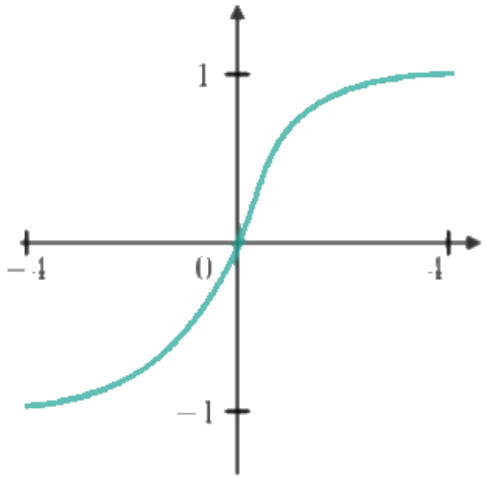
Sigmoid

$$g(z) = \frac{1}{1 + e^{-z}}$$



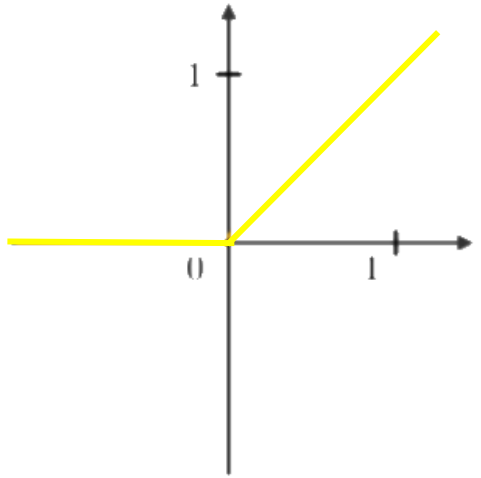
Tanh

$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$



ReLU

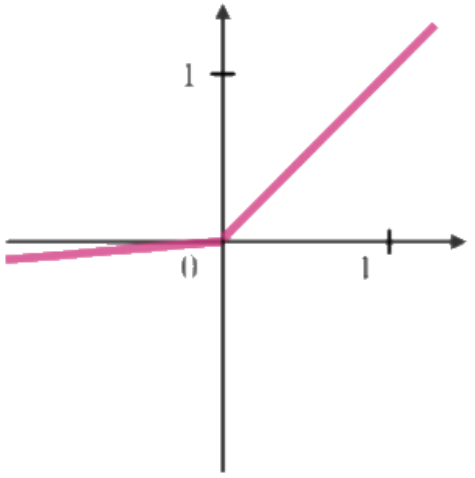
$$g(z) = \max(0, z)$$



Leaky ReLU

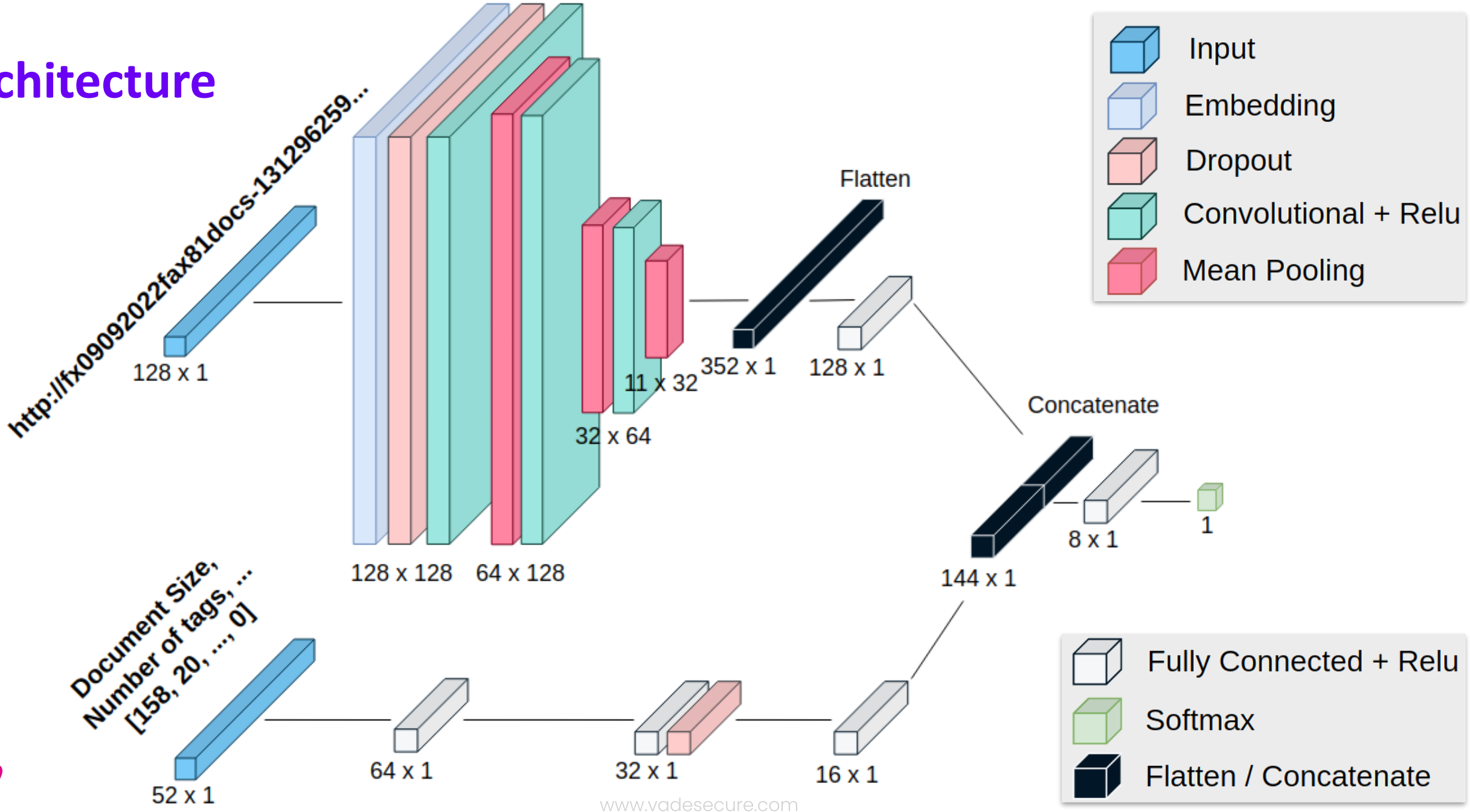
$$g(z) = \max(\epsilon z, z)$$

with $\epsilon \ll 1$



Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. (2018). URLNet: Learning a URL representation with deep learning for malicious URL detection. *arXiv preprint arXiv:1802.03162*.

Architecture



Motivations to not use Deep Learning alone

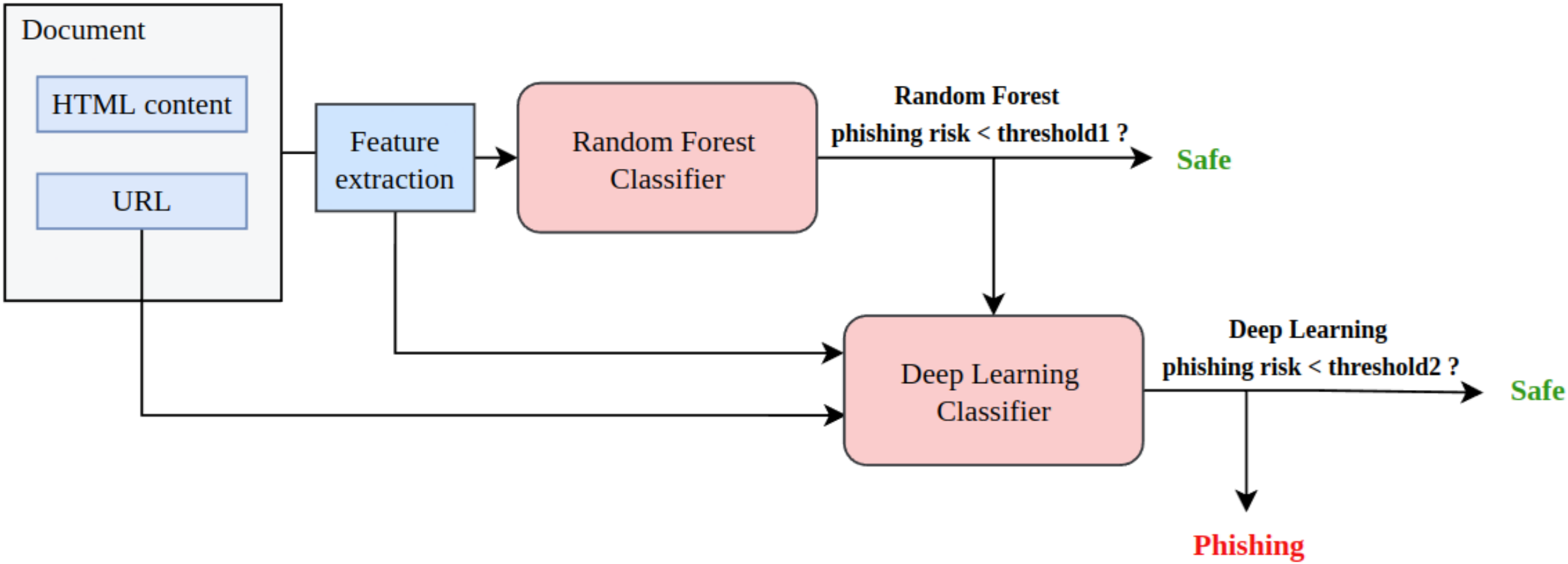
- Deep Learning achieves better recall for Phishing detection …
- … but with a slight decrease of precision
- Helps to remove some Random Forest FPs
- Combine Random Forest and Deep Learning
- Hybrid methods gives better results for phishing detection[1]

[1]Sara Afzal · Muhammad Asim · Abdul Rehman Javed · Mirza Omer Beg · Thar Baker (4 March 2021). "URLdeepDetect: A Deep Learning Approach for Detecting Malicious URLs Using Semantic Vector Models". Journal of Network and Systems Management (2021) https://doi.org/10.1007/s10922-021-09587-8

Best choice to combine decision

- Voting
- Weighted Averaging
- Distinguish cases
- Use one after the other (verification step)

Resulting Pipeline



Phishing Detection in Production

Experimentation Protocol

- Can this model be used to reduce supervision workload
- ~ 5000 phishing documents over 6 months are analyzed
- Apply our pipeline for all documents in this period
- Compare the agreements / disagreements

Agreements / Disagreements

Agreement	# of documents
DL predicts Phishing & Supervision says Phishing	4073 (80,21%)
DL predicts Safe & Supervision says Safe	167 (3,29%)

Disagreement	# of documents
DL predicts Phishing & Supervision says Safe	137 (2.7%)
DL predicts Safe & Supervision says Phishing	664 (13,08%)

Total	5078

Agreements / Disagreements

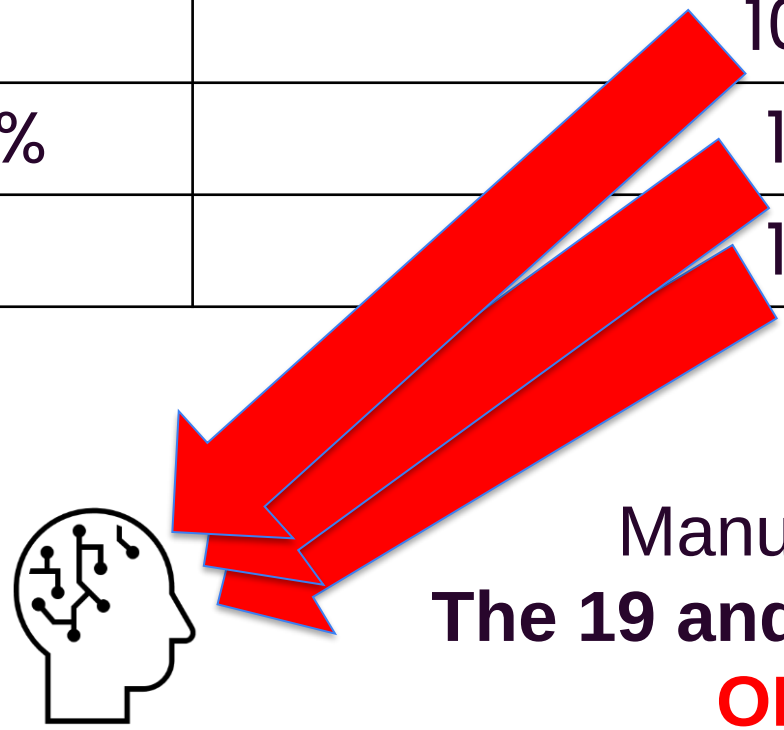
Agreement	# of documents
DL predicts Phishing & Supervision says Phishing	4073 (80,21%)
DL predicts Safe & Supervision says Safe	167 (3,29%)

Disagreement	# of documents
DL predicts Phishing & Supervision says Safe	137 (2.7%)
DL predicts Safe & Supervision says Phishing	664 (13,08%)

Total	5078

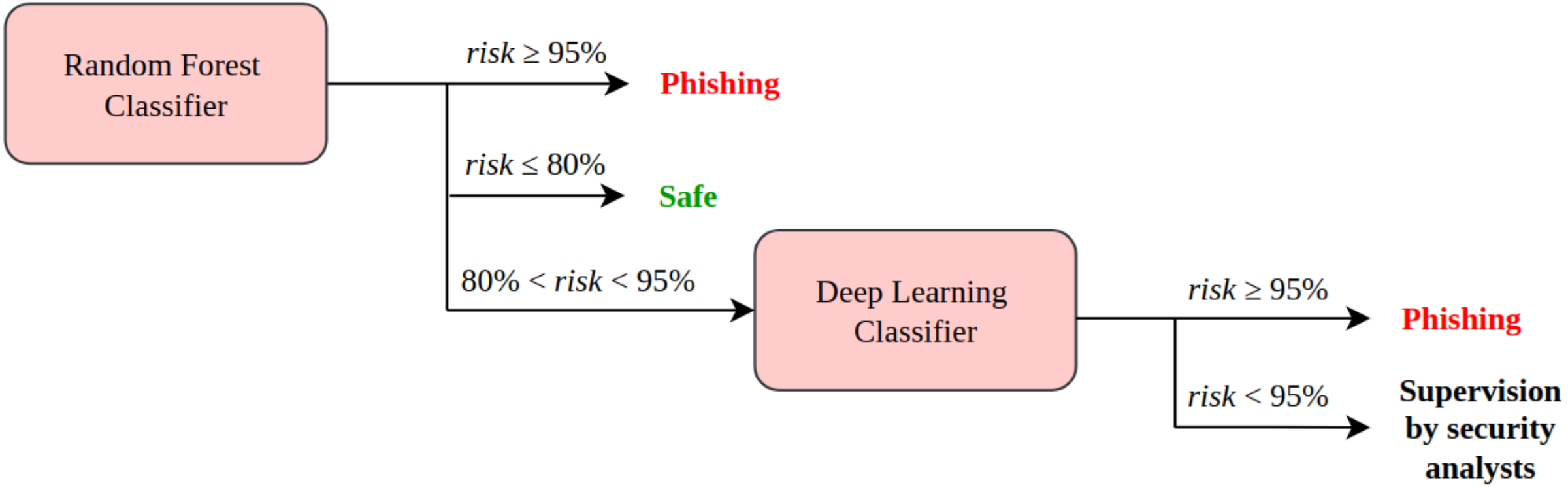
Focus on FPs

Phishing risk percentage interval	Number of documents where supervision answered Safe
Under 90%	100
Between 90% and 95%	19
Over 95%	18



Manual analysis:
The 19 and 18 samples are ONLY FN.

Decision Pipeline



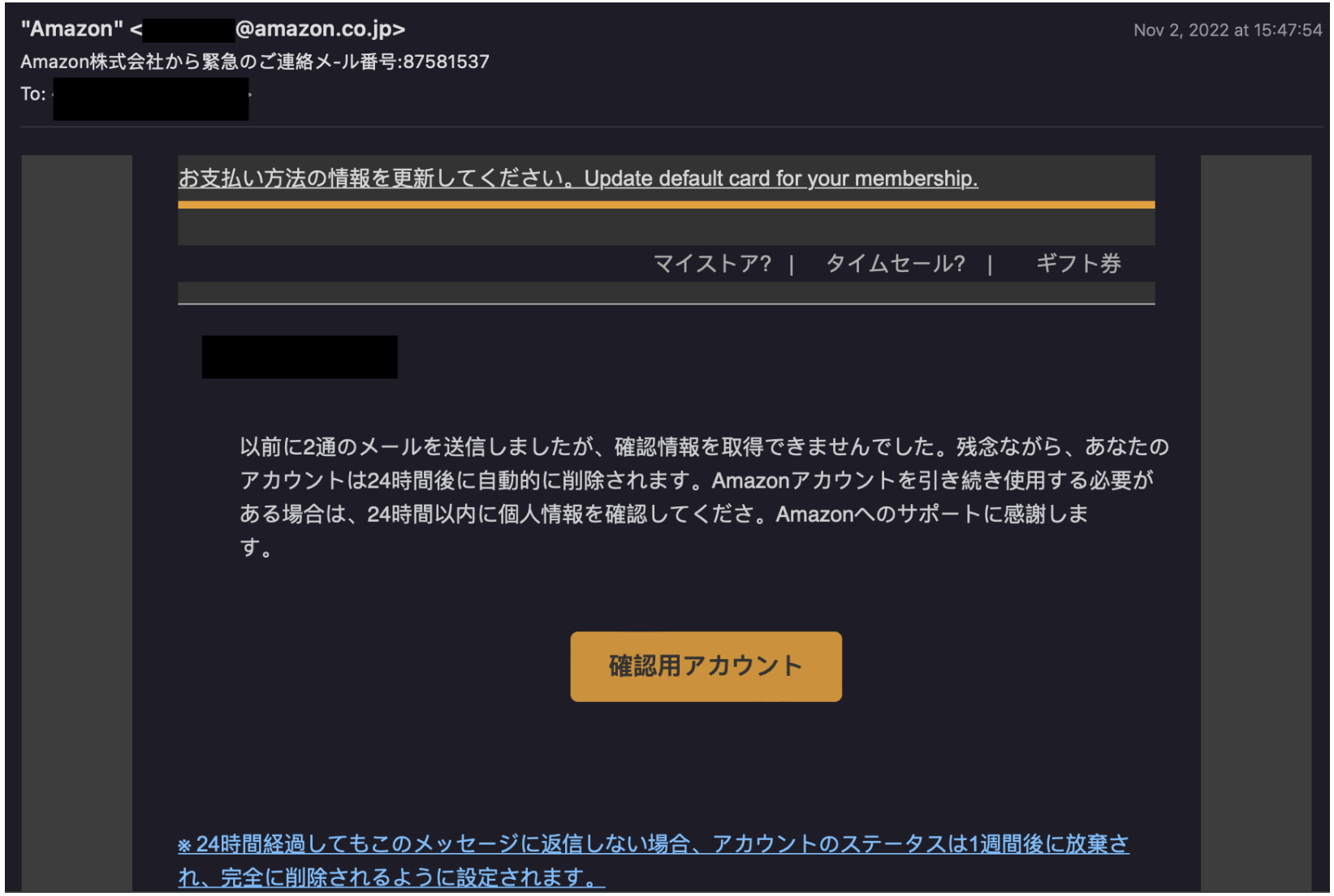
Metrics in production

- Daily process of 13M documents
- 5ms to extract features
- 33ms to obtain a Random Forest prediction
- 15ms to obtain Deep Learning prediction
- Between 1ms and 5ms of latency

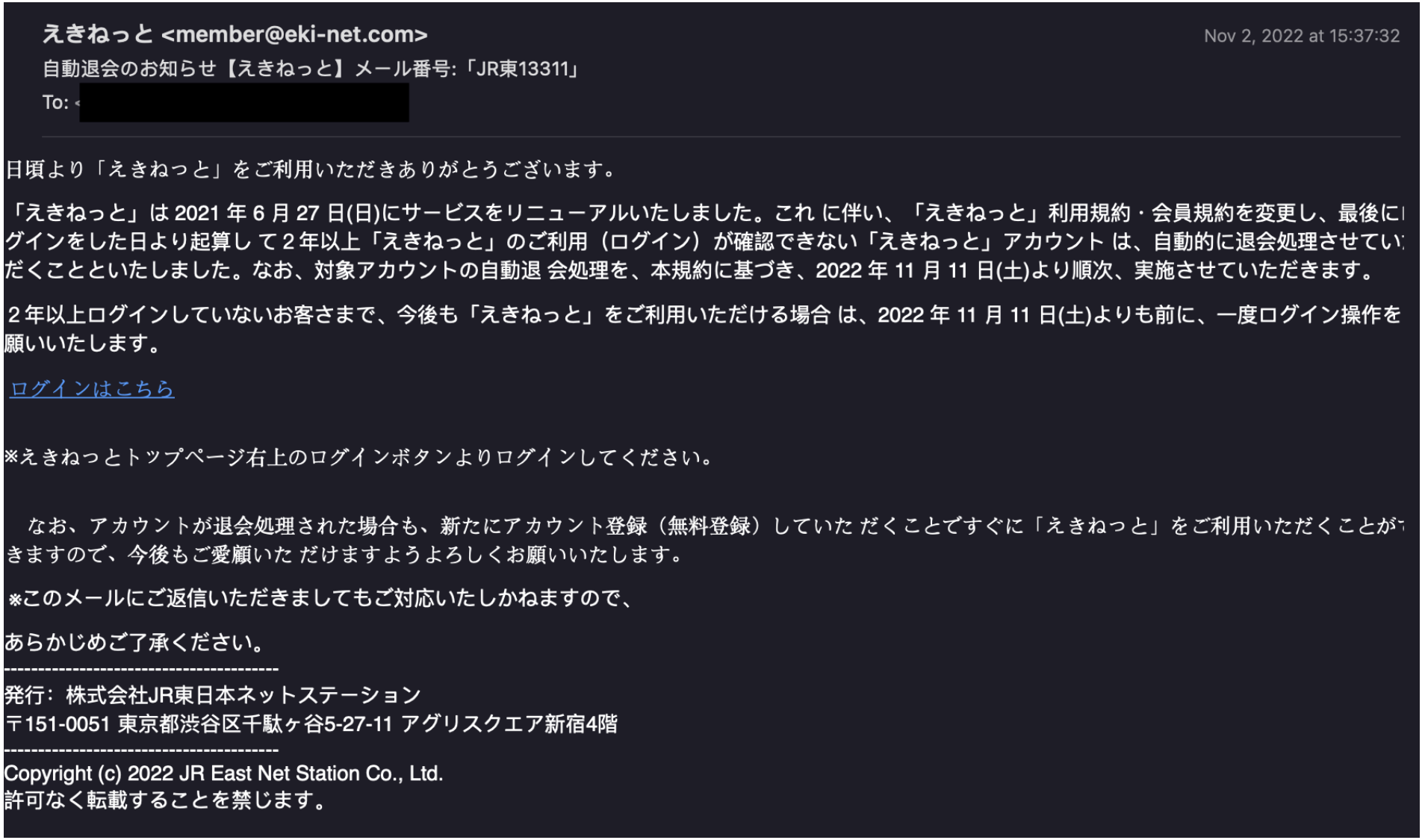
Conclusion

- Deep Learning can capture new phishing
- Combination of multiple tools helps to reduce supervision workload
- 30% of removed documents from supervision
- Instant supervision from Deep Learning
- Best strategy: classify documents as phishing automatically only when the two models agree

Examples in Recent Trends



Examples in Recent Trends



えきねっと <member@eki-net.com> Nov 2, 2022 at 15:37:32

自動退会のお知らせ【えきねっと】メール番号:「JR東13311」

To: [redacted]

日頃より「えきねっと」をご利用いただきありがとうございます。

「えきねっと」は 2021 年 6 月 27 日(日)にサービスをリニューアルいたしました。これ に伴い、「えきねっと」利用規約・会員規約を変更し、最後にグインをした日より起算し て 2 年以上「えきねっと」のご利用（ログイン）が確認できない「えきねっと」アカウント は、自動的に退会処理させていただくことといたしました。なお、対象アカウントの自動退 会処理を、本規約に基づき、2022 年 11 月 11 日(土)より順次、実施させていただきます。

2 年以上ログインしていないお客さまで、今後も「えきねっと」をご利用いただける場合 は、2022 年 11 月 11 日(土)よりも前に、一度ログイン操作を願いいたします。

ログインはこちら

※えきねっとトップページ右上のログインボタンよりログインしてください。

なお、アカウントが退会処理された場合も、新たにアカウント登録（無料登録）していた だくことですぐに「えきねっと」をご利用いただくことができますので、今後もご愛顧いた だけますようよろしくお願いいたします。

※このメールにご返信いただきましてもご対応いたしかねますので、

あらかじめご了承ください。

発行：株式会社JR東日本ネットステーション

〒151-0051 東京都渋谷区千駄ヶ谷5-27-11 アグリスクエア新宿4階

Examples in Recent Trends

De Support <notifications@dypaekngfuj0lv7454q38y.onozuka.co.jp>
Sujet 通知：Appleアカウント(参照ID: APP-79457658)に関するご対応のお願い
Pour xxxxx@redacted.vadesecure.com

Répondre | Répondre à tous | Transférer | Autres
14/08/2022, 07:43

保護のため、 Apple IDは自動的に無効になります。

親愛な,

セキュリティ上の理由により、お客様の Apple ID がロックされています。システムがいくつかの不成功の試みを検出しました。

アカウントのロックを解除する前に、身元を確認する必要があります。

に行く (Apple ID) すぐにあなたの情報を確認してください

本人確認

※ 私たちは24時間以内にあなたからの応答を受信しない場合、アカウントがロックされます。

Apple ID .| サポート | プライバシーポリシー

Copyright © 2022 iTunes K.K. 〒106-6140 東京都港区六本木6丁目10番1号 六本木ヒルズ All rights reserved.

width="

Examples in Recent Trends

保護のため、 Apple IDは自動的に無効になります。

親愛な,

セキュリティ上の理由により、お客様の Apple ID がロックされています。システムがいくつかの不成功の試みを検出しました。

アカウントのロックを解除する前に、身元を確認する必要があります。

に行く (Apple ID) すぐにあなたの情報を確認してください

本人確認

※ 私たちは24時間以内にあなたからの応答を受信しない場合、アカウントがロックされます。

Apple ID .| サポート | プライバシーポリシー
Copyright © 2022 iTunes K.K. 〒106-6140 東京都港区六本木6丁目10番1号 六本木ヒルズ All rights reserved.

Questions?

For additional questions, please email:

maxime.meyer@vadesecure.com

gabriel.loiseau@vadesecure.com