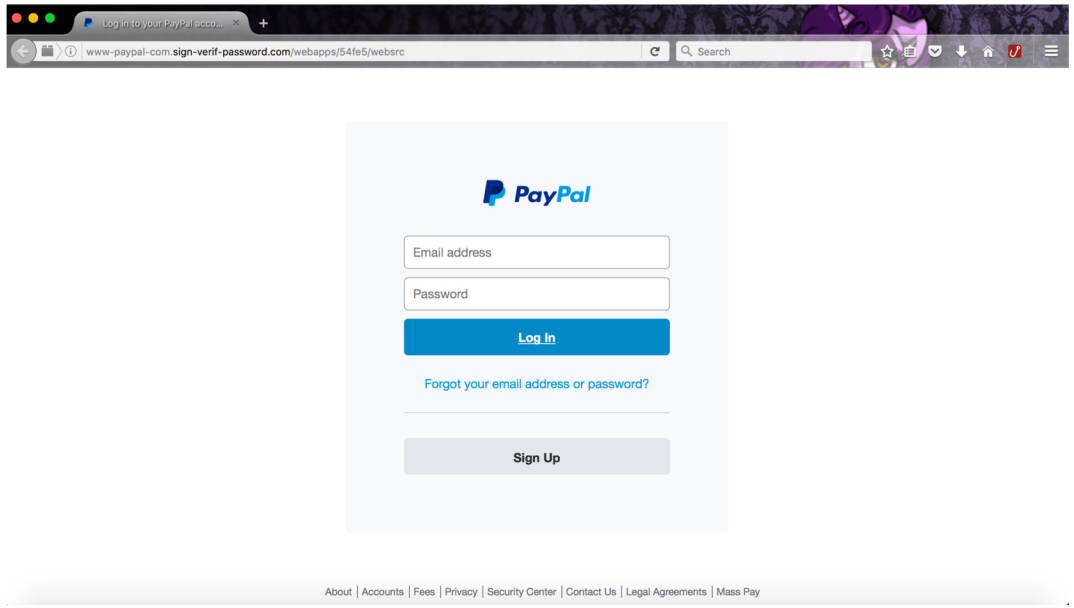
Aalto University

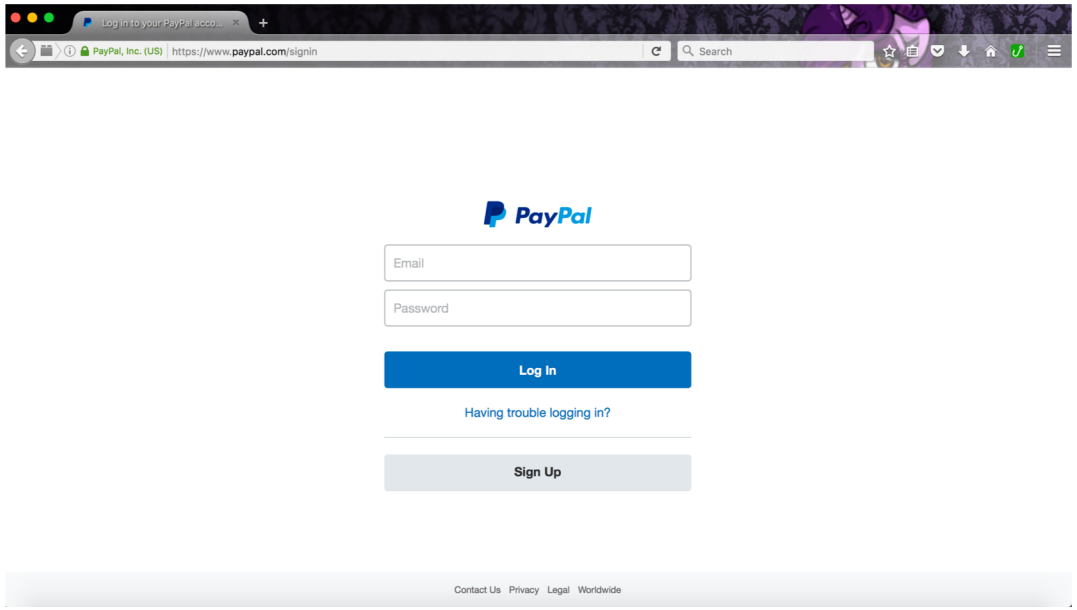
On Designing and Evaluating Phishing Webpage Detection Techniques for the Real World

Samuel Marchal, *N. Asokan*
Aalto University, Finland

samuel.marchal@aalto.fi

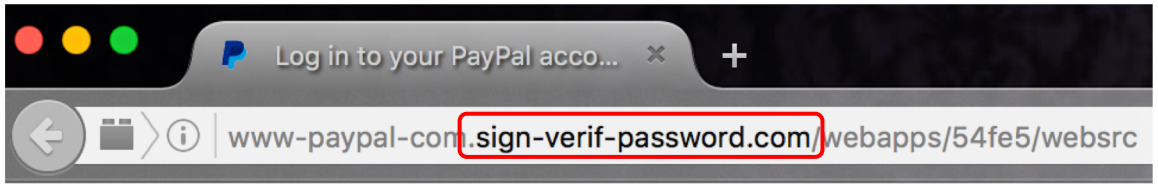
Phishing webpage

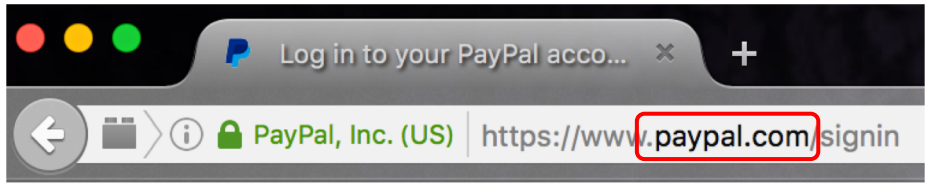




Phishing webpage (phish)

Legitimate webpage





State of research on phishing detection

- **Threat known since late 1990s**
- **First protection technique[1] early 2000s**

- **> 4,000 articles on “phishing”**
  - Half as popular as "malware”
- **Many solutions report high accuracy**
  - Cantina[2] (2007): 97%
  - Whittaker et al.[3] (2010): 99.9%
  - Off-the-Hook[4] (2017): 99.9%

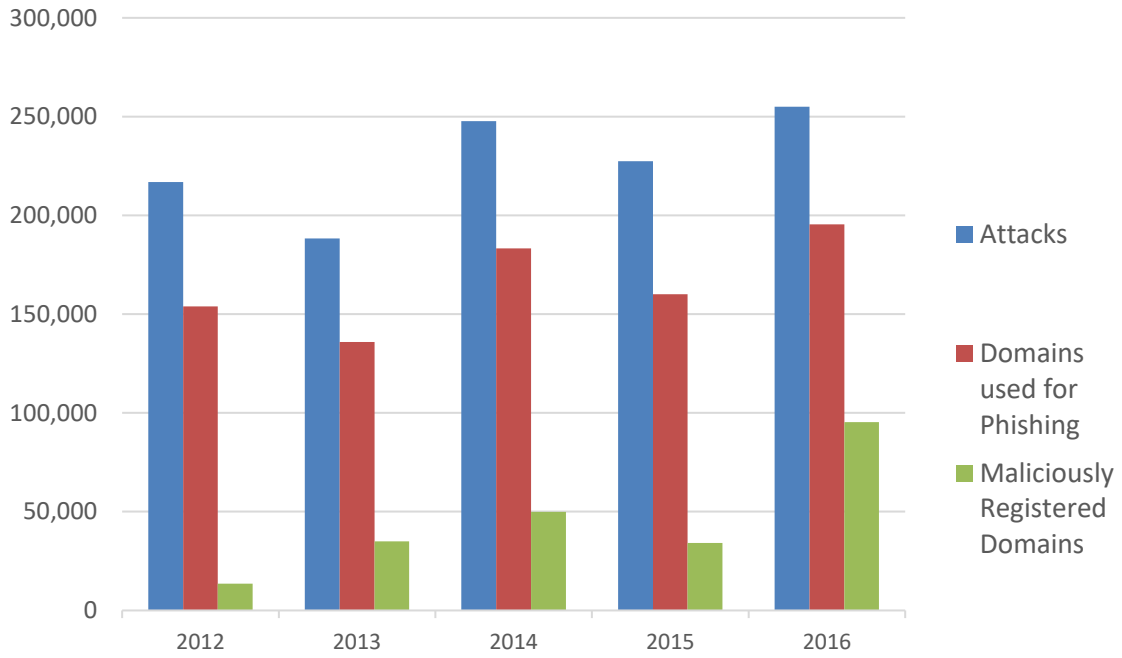
[1] Herzberg and Gbara, ”*Trustbar: Protecting (even naive) web users from spoofing and phishing attacks”* in Cryptology ePrint Archive, 2004.
[2] Zhang et al., *”CANTINA: A content-based approach to detecting phishing web sites”* in WWW, 2007.
[3] Whittaker et al., *”Large-scale automatic classification of phishing pages”* in NDSS Symposium, 2010.
[4] Marchal et al., *”Off-the-hook: An efficient and usable client-side phishing prevention application”* in IEEE Transactions on Computers 66, 10, 2017.

State of phishing threat

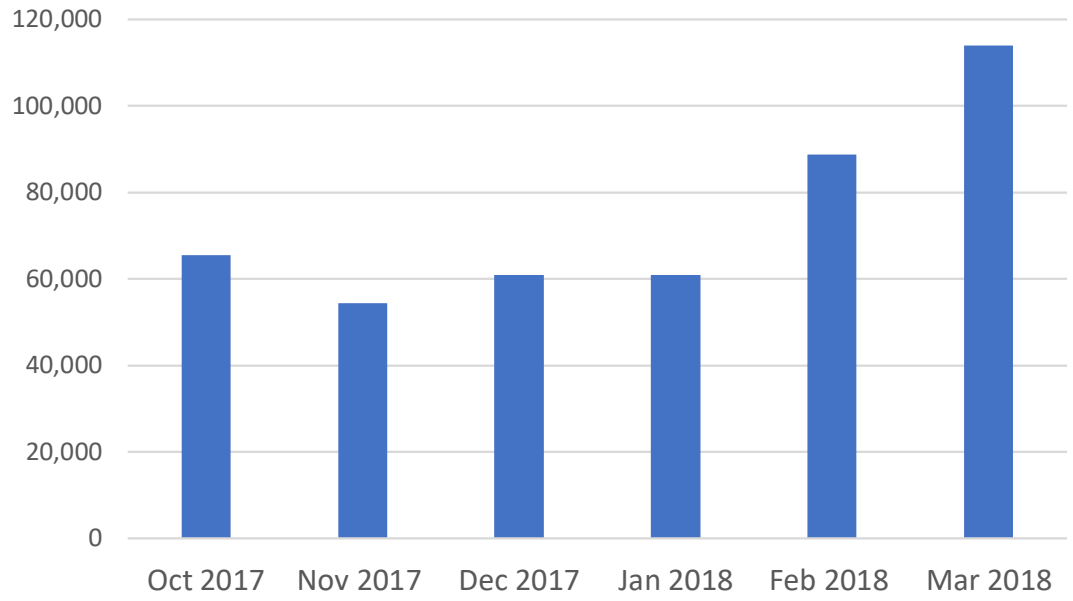
Monetary damage:

- 2013-2016: $1.6 billion loss for businesses (US only)
- Most expensive attack (2015): $100 million cost (US defense department)

Phishing attacks 2012-2016



Phishing websites 2017-2018



Source: Anti Phishing Worgin Group (APWG).

Detection of phishing webpages

Gap between

- High accuracy reported in literature
- Low effectiveness when applied to the real-world

What goes wrong during design & evaluation?

- Design choices only driven by high detection accuracy level
- Evaluation not representative of the real-world

Effective phishing detection

Requirements for effectiveness

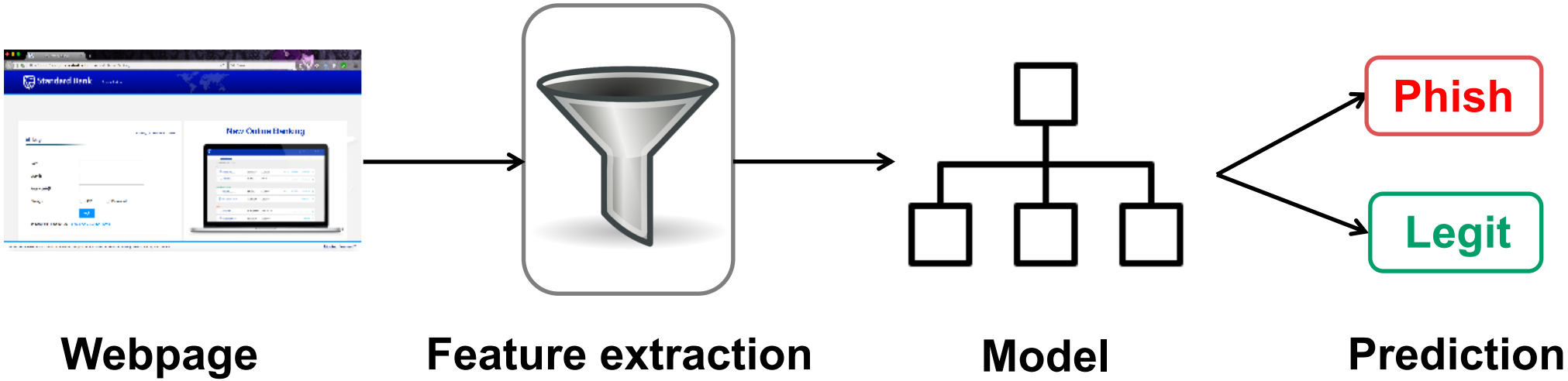
- Detection performance
- Temporal resilience
- Deployability
- Usability

Recommendations

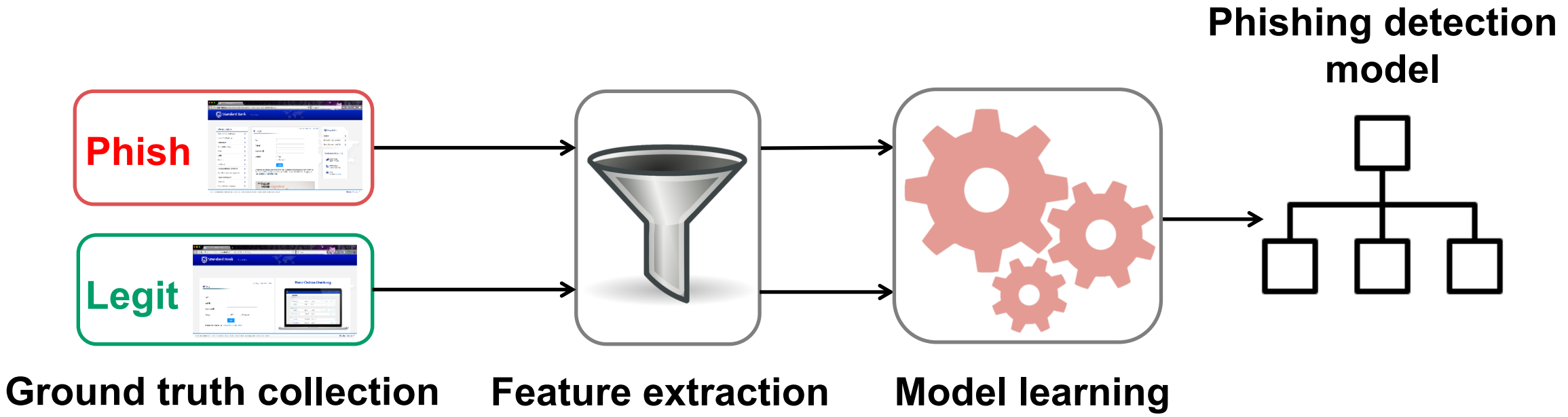
- Design of detection method
- Evaluation
  - Ground truth selection
  - Assessment methodology

ML-based phishing webpage detection

Machine learning based phishing detection



Phishing detector training



Design of detection method

System design

	Centralized	Client-side
Pros	• High computational power • Easy model updates • Confidentiality of detection model	• User privacy • Fast decision • Website data availability
Cons	• Delay in decision • Impacts user privacy (browsing history)	• Degrades client device performance • Lack of model confidentiality

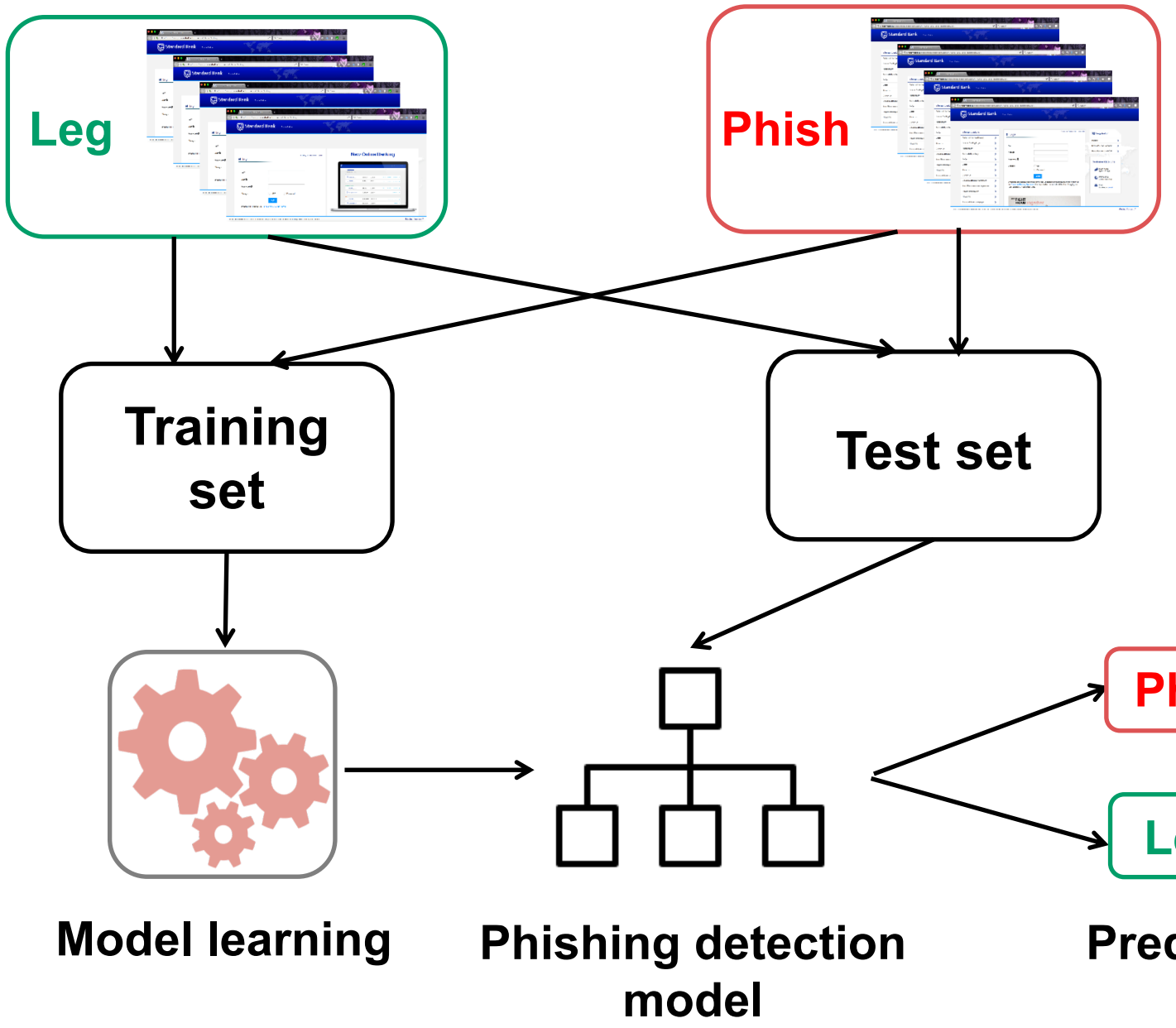
➡ **Centralized solution currently favored by industry….**

….but increasing privacy concerns may change the game.

Evaluation

Evaluation setup

Ground truth collection



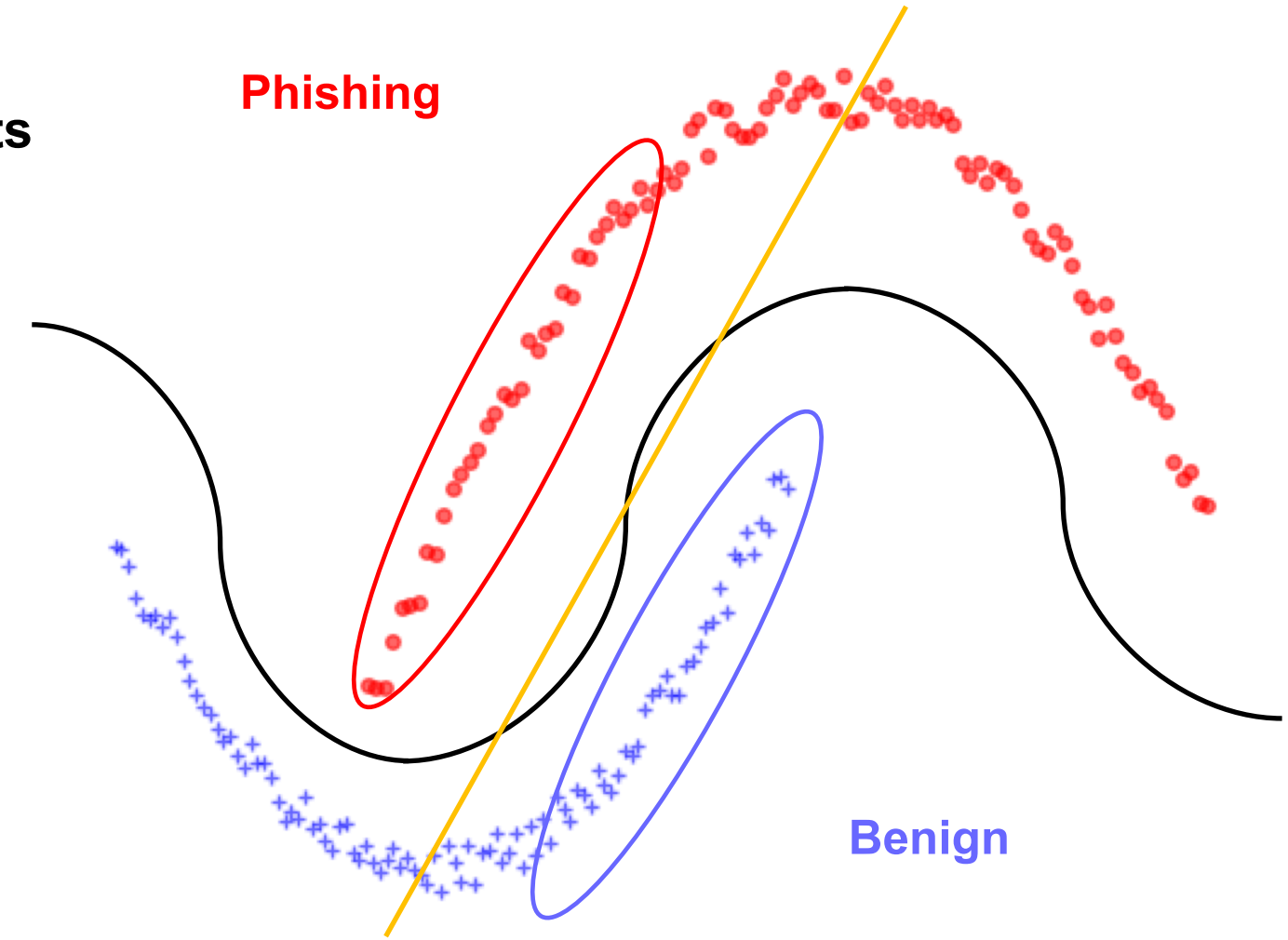
Ground truth selection

Improve relevance of accuracy results

- Validity
- Generalizability
- Reproducibility

Ground truth

- Validity of labels
- Representativeness
- Availability



Webpage selection

Generic guidelines

- Multi-lingual + different alphabet
- Publicly available sources (≠ static dataset)

Legitimate webpage

- Diverse popularity
- Real URLs: as browsed
  - www.amazon.com ≠ https://www.amazon.com/gp/cart/view.html?ref=nav_cart

Phishing webpage

- Targeting different brands
- Fresh and up-to-date
  - PhishTank (https://www.phishtank.com/)
  - OpenPhish (https://openphish.com)

Phishing webpage validity

Analysis of 23,118 phishing pages (source Phishtank)

- 59% valid (13,646)
- 41% invalid (9,472)
  - Content unavailable
  - Domain parking
  - Legitimate webpage

Phishing data requires sanitization

- Scrape and save webpages of fresh phishes
- Sanitization
  - Screenshot analysis
  - Google search with keywords
  - Later visit of URL

Dataset usage

Follow realistic use cases

- Train model with oldest data & test with newest data
  - No cross-validation to get accuracy metrics
- Larger testing set than training set → scalability
- Use real-world distribution: 1 phish / 100 legitimate pages → relevant accuracy metrics

Accuracy metrics

Positive (P) = identified as phish	**Negative (N) =** identified as benign
True positive (TP) = detected phish	**False positive (FP)** = benign detected as phish
False negative (FN) = missed phish	**True negative (TN)** = benign identified benign

- Phishing detection capability — True positive rate — $TPR = \frac{TP}{TP+FN}$
- Erroneous phishing warnings — False positive rate — $FPR = \frac{FP}{TN+FP}$
- Correctness of phishing warnings — Precision — $Precision = \frac{TP}{TP+FP}$

Temporal resilience

Ensure steadiness of effectiveness over time

Longitudinal study: readiness for deployment

- Data collection over extended period of time
- Recompute accuracy metrics
  - Steady accuracy without retraining → ready for deployment / low maintenance cost
  - Steady accuracy with retraining → ready for deployment / maintenance cost depends on retraining period
  - Decrease in accuracy with retraining → not ready for deployment

Resilience to adversaries

- Security assessment using adversarial machine learning attacks
- Evaluate manipulability of features

Conclusion

Recommendations

- Design of detection method
- Evaluation
  - Ground truth selection
  - Assessment methodology

Goals for research in phishing detection

- Relevant accuracy results + easy comparison
- More impactful research → technology transfer

Aalto University

On Designing and Evaluating Phishing Webpage Detection Techniques for the Real World

Samuel Marchal, *N. Asokan*
Aalto University, Finland

samuel.marchal@aalto.fi