**Research Problem**

- Machine learning (ML) promising when applied to the malware domain.
- In-lab experiments (CV, leave-one-out) show performance metrics, but:
  - Malware is an evolving target.
  - Unclear how ML would really perform once deployed.
- Real-world deployment hard to assess.
- New, previously unseen labels.
- Changes in the data distribution.

*We need statistical (quality) metrics*

**Conformal Evaluator: ML Evaluation with Confidence**

Build statistical metrics (RQ1) to highlight:

- **Statistical data distribution according to the algorithm** (RQ2.1)
  - Shows whether classes overlap.
  - Provides insights during ML design (RQ3).
- **Statistical confidence on the ML algorithm choices** (RQ2.2)
  - High confidence puts trust on the ML process (RQ3).
  - Provides thresholds to detect and contain ML decay during deployment (RQ4).

**RQ1: Statistical Metrics**

- **Algorithm credibility** – Captures the conformance of the algorithm’s prediction for a given data point.
- **Algorithm confidence** – Measure the distinguishability of a sample’s prediction among the other families.

**RQ2, RQ3: Statistical Metrics Insights (Design)**

**RQ4: Statistical Indicators of ML Decay (Deployment)**

- **Alpha assessment** – Shows the statistical distribution of data according to the algorithm under evaluation.
- **Decision assessment** – Shows the quality with statistical confidence of correct and incorrect choices of the algorithm under evaluation.
- **Testing in real-world scenarios** – Statistical metrics provide indications on ML decays when no labels are available.

This research has been supported by the UK research grants EPSRC EP/R033943/1, EP/R000265/1, and EPSRC EP/NS3/2701/1.