From Counterfactuals to Trees: Competitive Analysis of Model Extraction Attacks

Awa Khouna[1,2] Julien Ferry[1,2] Thibaut Vidal[1,2]

[1]Department of Mathematical and Industrial Engineering,
Polytechnique Montréal, Montréal, Canada

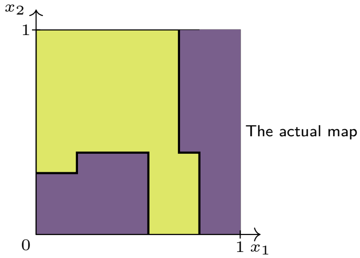
[2]CIRRELT & SCALE-AI Chair in Data-Driven Supply Chains,

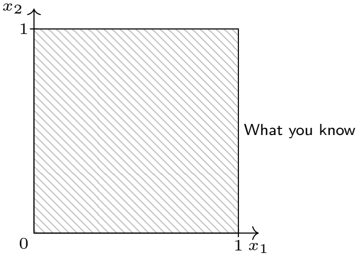
October 31, 2025

Introduction

The Map Riddle

Setup. Imagine a hidden map (a colored map).
You can't see the full map.



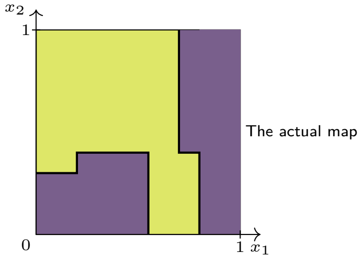


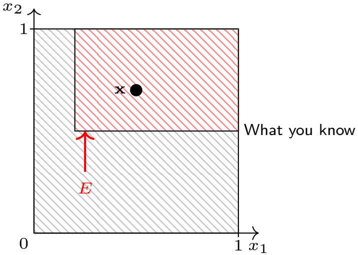
The Map Riddle

Setup. Imagine a hidden map (a colored map). You can't see the full map.

Oracle. For any point $\mathbf{x}$ and rectangle E, the oracle returns the *nearest* point $\mathbf{x}'$ of a different color within E if it exists.

Question. Can repeated queries ***exactly*** *reconstruct the entire map*?



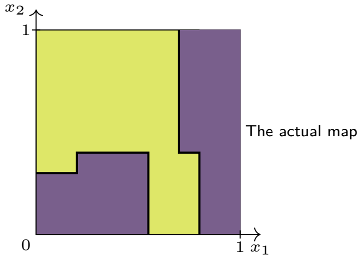


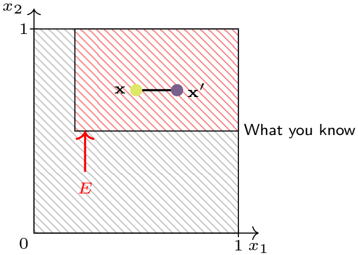
The Map Riddle

Setup. Imagine a hidden map (a colored map). You can't see the full map.

Oracle. For any point $\mathbf{x}$ and rectangle E, the oracle returns the *nearest* point $\mathbf{x}'$ of a different color within E if it exists.

Question. Can repeated queries ***exactly*** *reconstruct the entire map*?





Model Extraction Attack

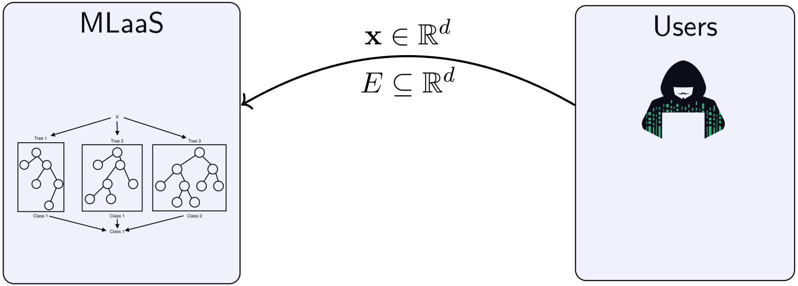


Figure: Model extraction attacks framework.

Model Extraction Attack

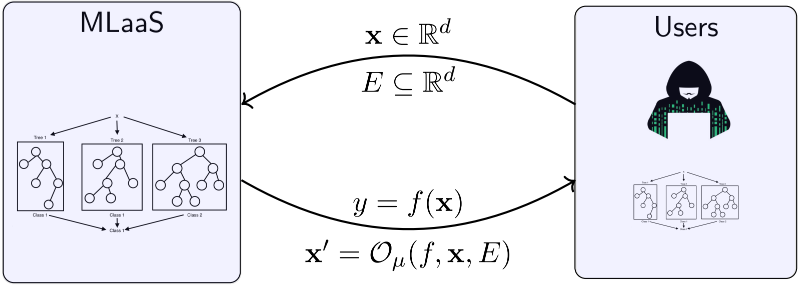


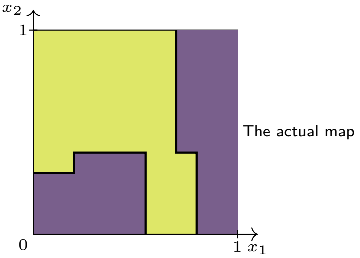
Figure: Model extraction attacks framework.

From the Riddle to the Attack

Setup. Imagine a hidden map (a colored map). You can't see the full map.

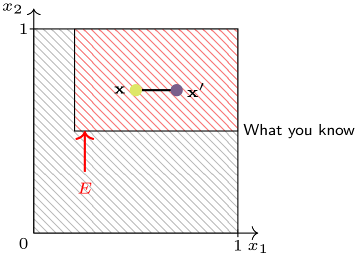
Oracle. For any point $\mathbf{x}$ and rectangle E, the oracle returns the *nearest* point $\mathbf{x}'$ of a different color within E if it exists.

Question. Can repeated queries ***exactly*** *reconstruct the entire map*?



Our analogy: the colored map $\leftrightarrow$ classifier decision regions; oracle $\leftrightarrow$ counterfactual explanation.

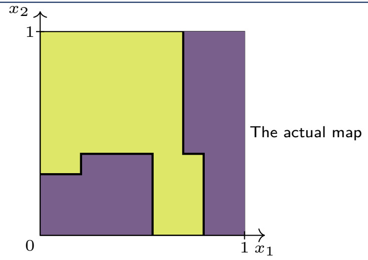
Takeaway: Counterfactuals reveal *where* the nearest boundary is. With a smart querying strategy, this can be enough to reconstruct the model.

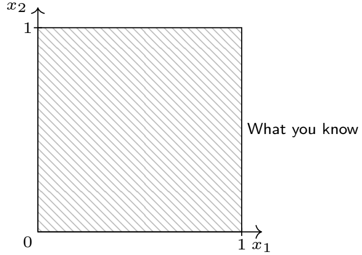


Contributions

TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

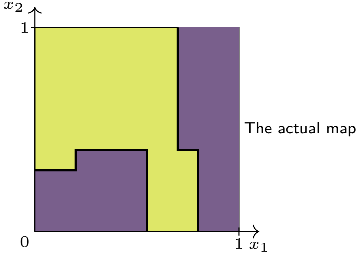


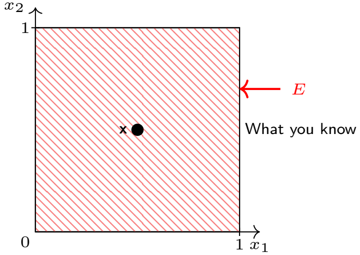


TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.

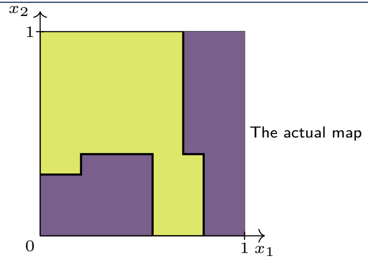


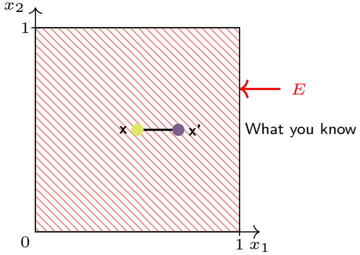


TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.

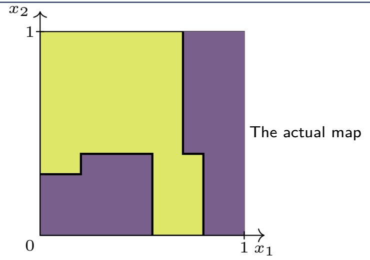


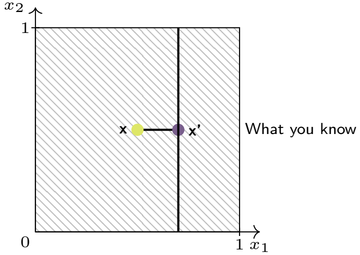


TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.
2. **Split E via CF:** cut on features where $\mathbf{x} \neq \mathbf{x}'$.

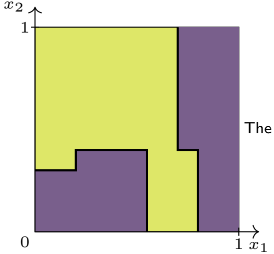




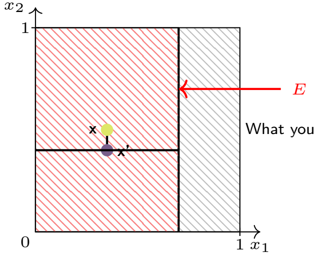
TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.
2. **Split E via CF:** cut on features where $\mathbf{x} \neq \mathbf{x}'$.



The actual map

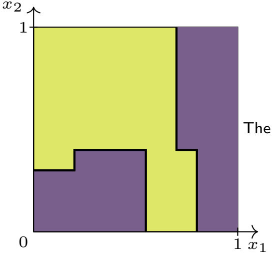


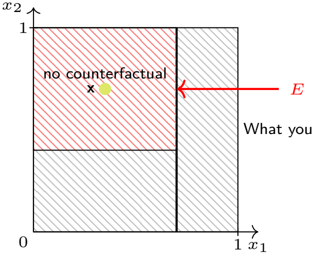
What you know

TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.
2. **Split E via CF:** cut on features where $\mathbf{x} \neq \mathbf{x}'$.
3. **No CF $\Rightarrow$ label E:** assign $f(\mathbf{x})$; continue with remaining rectangles (BFS).

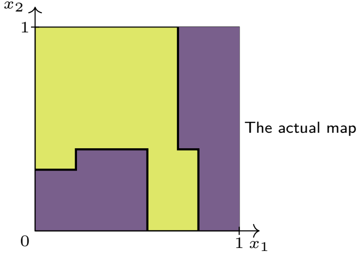


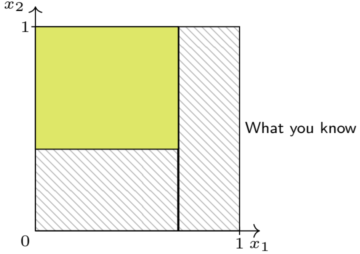


TRA : Tree Reconstruction Attack (in 3 steps)

The algorithm in 3 steps:

1. **Probe center:** query $(\mathbf{x}, E)$.
2. **Split E via CF:** cut on features where $\mathbf{x} \neq \mathbf{x}'$.
3. **No CF $\Rightarrow$ label E:** assign $f(\mathbf{x})$; continue with remaining rectangles (BFS).





Theory: Query Complexity & Competitiveness

Let n be the number of split levels and s_i splits on feature i, $\sum_i s_i = n$.

- **Worst-case queries:** $O\Big(\prod_{i=1}^{m}(s_i+1)\Big) \leq O\Big((1+\frac{n}{m})^m\Big)$.
- **Competitive ratio:**

$$C_{\mathsf{TRA}}^{(n,m)} = \frac{2\prod_{i=1}^{m}(s_i+1)-1}{n+1} \leq \frac{2(1+\frac{n}{m})^m-1}{n+1}.$$

- **Tight for D&C:** no pure divide-and-conquer method can beat $C_{\mathsf{TRA}}^{(n,m)}$.

Key Insight

CFs provide *precise local boundary* information; TRA converts local probes into a *global reconstruction* with provable efficiency.

Results

Anytime Performance (Fidelity vs Queries)

- TRA reaches 100% fidelity *faster* (orders of magnitude fewer queries).
- Outperforms CF / DualCF (no functional equivalence) and PathFinding (equivalence but many queries).
- Also works with **non-optimal** CFs (practical APIs).

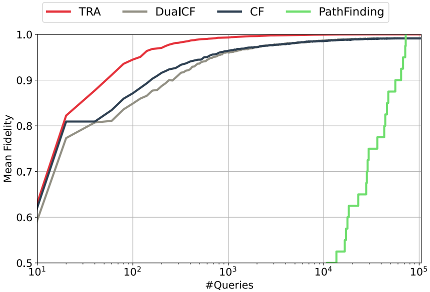


Figure: Mean Fidelity vs Number of queries over 40 trained classifier on Adult dataset

Functional Equivalence: Trees & Forests

- **Decision Trees:** TRA extracts exact decision boundaries with far fewer queries than PathFinding.
- **Random Forests:** TRA reconstructs an equivalent *tree* with perfect fidelity; *sub-linear* query growth vs nodes (empirically).

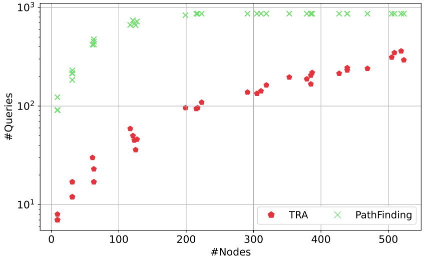


Figure: # Queries vs # Nodes for 40 classifiers.

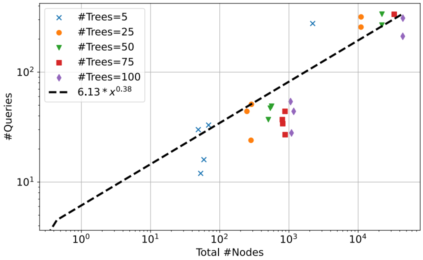


Figure: # Queries vs # Nodes for 25 trained Random forests.

Implications

Implications

- Counterfactual explanations *can fully expose* tree/ensemble decision boundaries.
- **Design challenge:** preserve *recourse and explainability* while *limiting leakage*.

Towards Safer Explanations

IP-preserving CFs by rate-limiting the oracle, region restrictions or query auditing.

Takeaway

Once you can ask *"where is the nearest boundary?"*,

you can reconstruct the **entire map**.

Explainability and Security must be co-designed.

Thank you!

Meet us at our poster at NeurIPS 2025!

Scan for the paper on arXiv

Poster: From Counterfactuals to Trees

Competitive Analysis of Model Extraction Attacks

Wednesday, December 3, 2025

4:30–7:30 p.m. PST

Exhibit Hall C,D,E – San Diego Convention Center

awa.khouna@polymtl.ca

`github.com/vidalt/Tree-Extractor`