Université de Rennes

WIDE

Inria

UMR IRISA

GOUVERNEMENT
Liberté
Égalité
Fraternité

PEReN
Pôle d'Expertise de la Régulation Numérique

Are there models harder to audit ?

Change-relaxed active fairness auditing

PFIA – RJCIA 2023 · 7 juillet 2023


Augustin Godinot

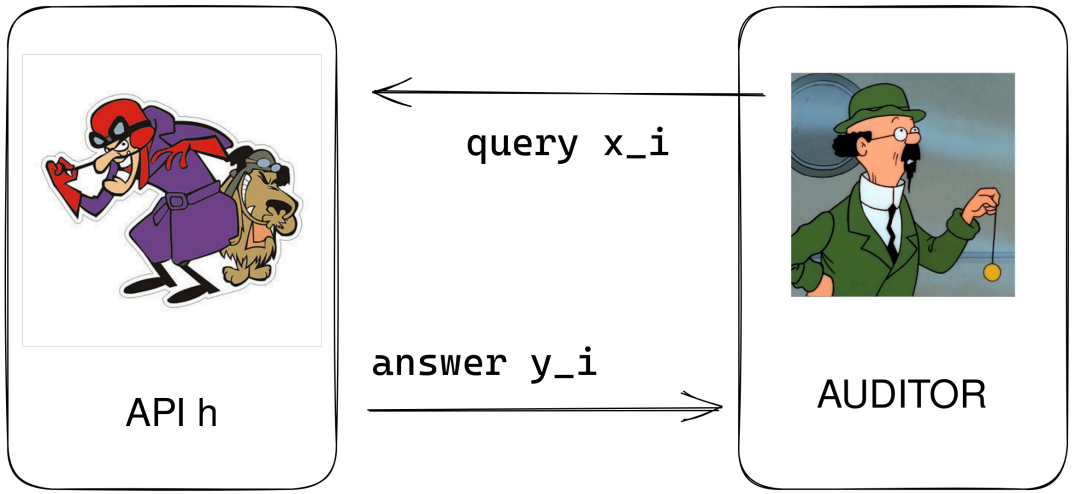

Erwan Le Merrer


Gilles Tredan

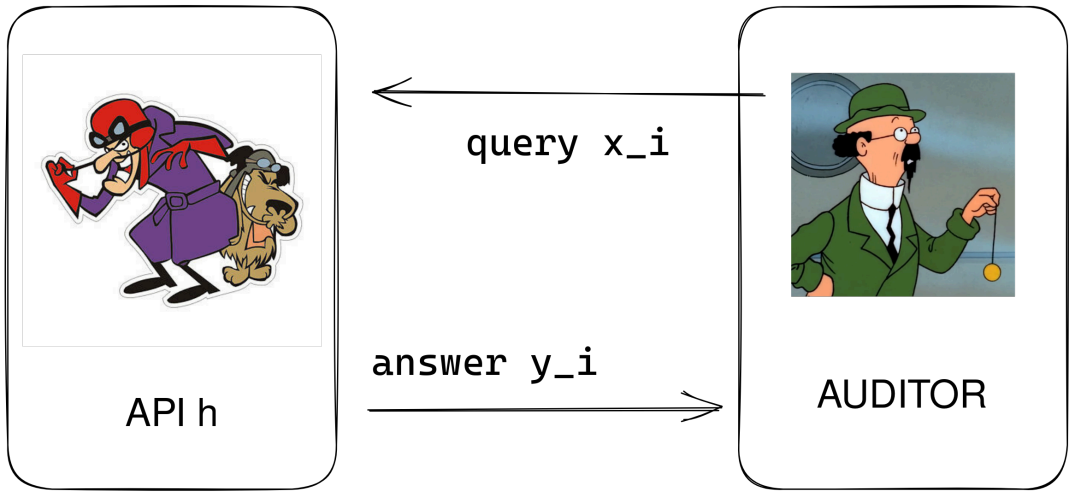

Camilla Penzo


François Taïani

The Auditing Game



The Auditing Game

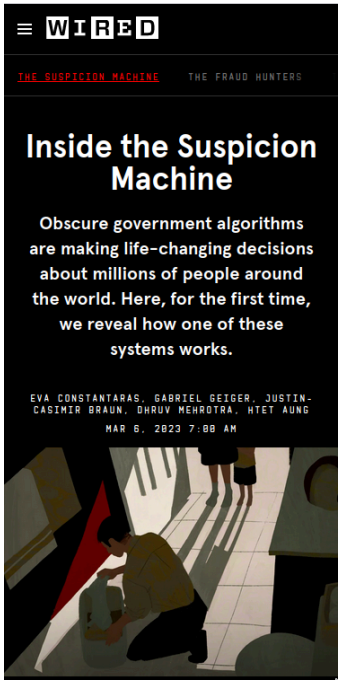


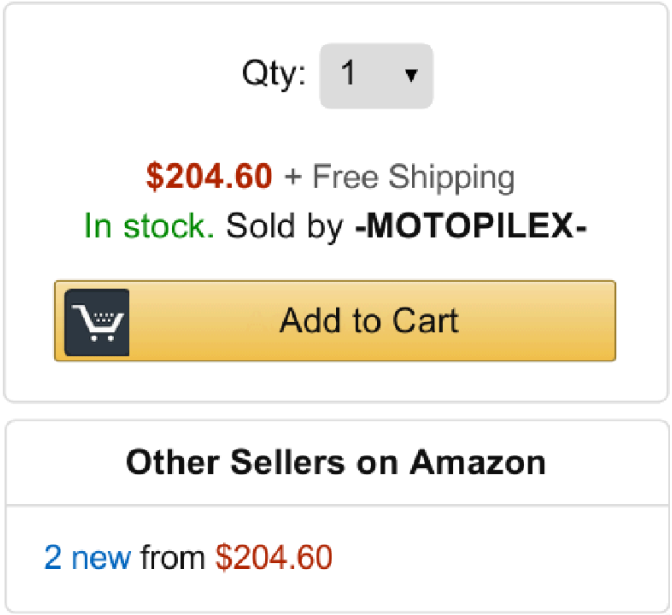
$$S = (x_1, ..., x_n), h(S) = (h(x_1), ..., h(x_n))$$

$$\hookrightarrow \mu(S, h(S)) = 0.035$$

Context

Automated Decision Systems



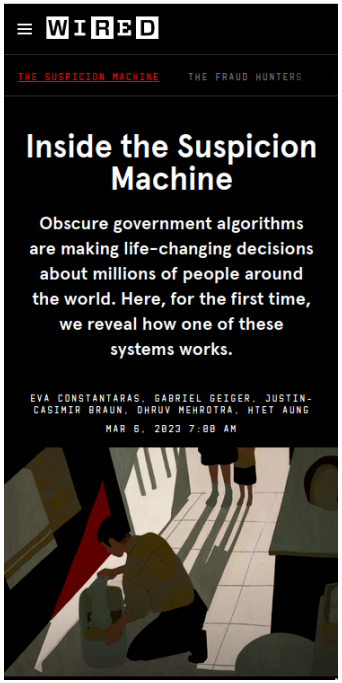


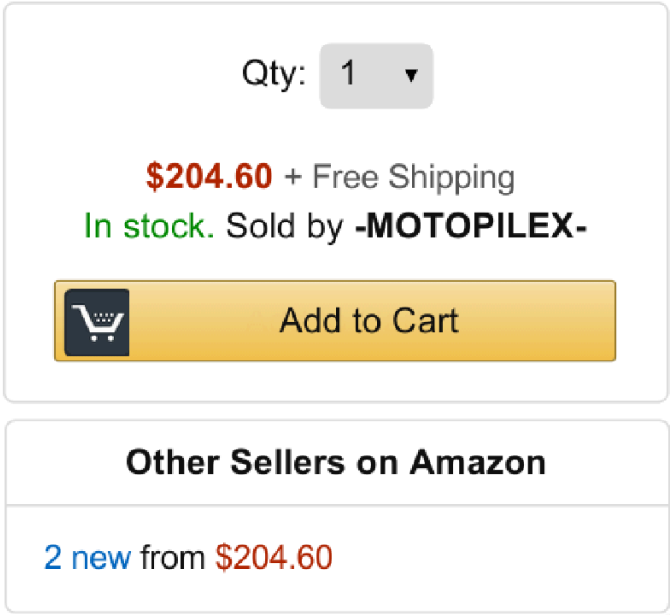
Hirevue claims it is "Fast. Fair. Flexible."

[1], [2], [3], [4], [5]

Context

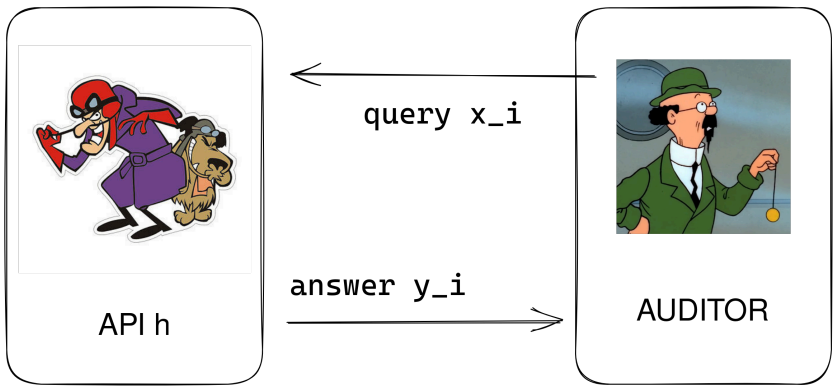
Automated Decision Systems





Hirevue claims it is "Fast. Fair. Flexible."

[1], [2], [3], [4], [5]

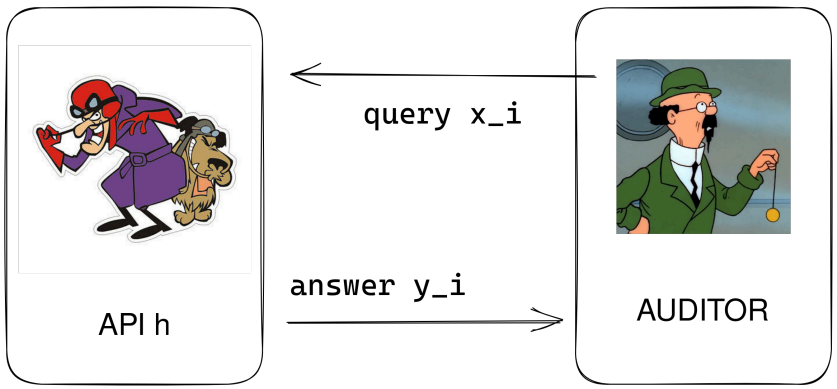


Input space $\mathcal{X}$

Output space $\mathcal{Y} = \{0, 1\}$

Model $h \in \mathcal{H}$

Metric $\mu : \mathcal{H} \times 2^{\mathcal{X}} \to \mathbb{R}$



Auditor prior

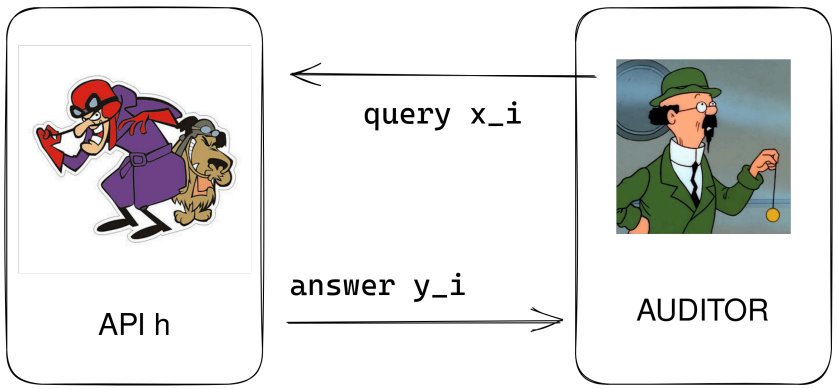
$\mathcal{H}$ known by the auditor

Input space $\mathcal{X}$

Output space $\mathcal{Y} = \{0, 1\}$

Model $h \in \mathcal{H}$

Metric $\mu : \mathcal{H} \times 2^{\mathcal{X}} \to \mathbb{R}$



Input space $\mathcal{X}$

Output space $\mathcal{Y} = \{0, 1\}$

Model $h \in \mathcal{H}$

Metric $\mu : \mathcal{H} \times 2^{\mathcal{X}} \to \mathbb{R}$

Auditor prior

$\mathcal{H}$ known by the auditor

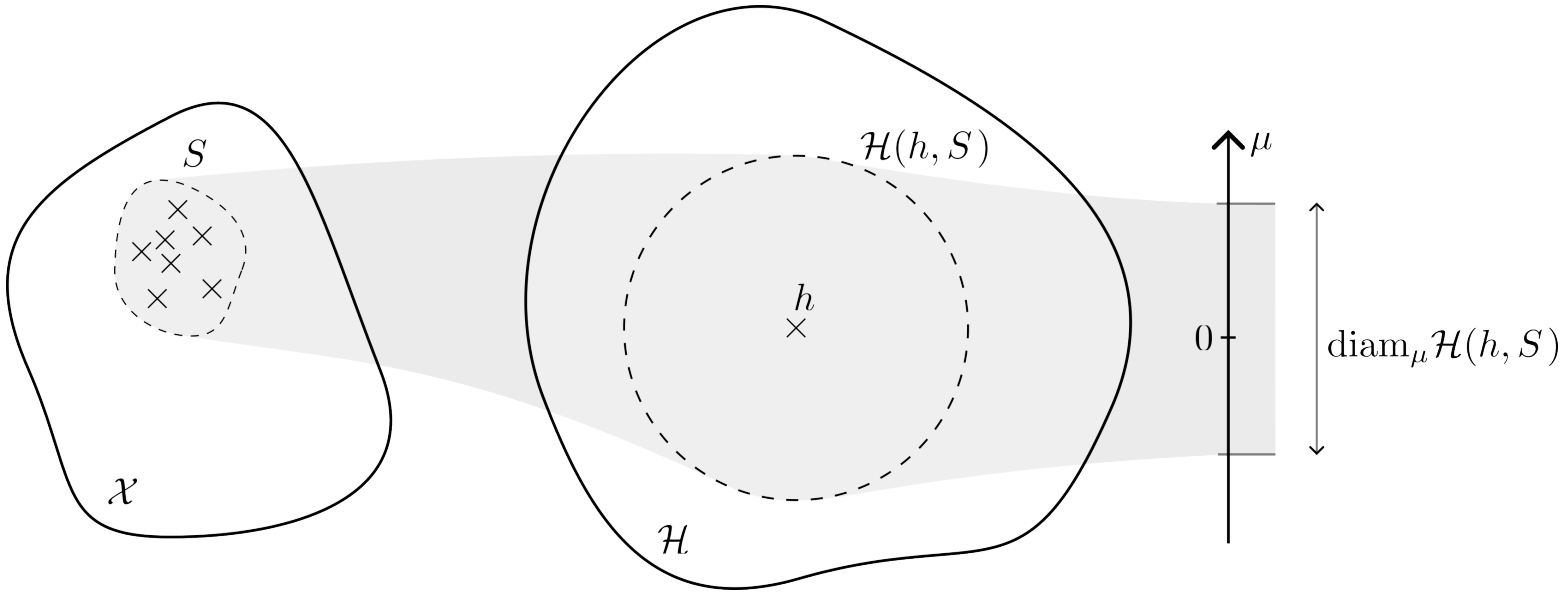
Consistency

$$h^{\text{API}}_{t_{\text{audit}}}(x) = y$$

$$\Rightarrow \forall t \geq t_{\text{audit}}, h^{\text{API}}_{t(x)} = y$$

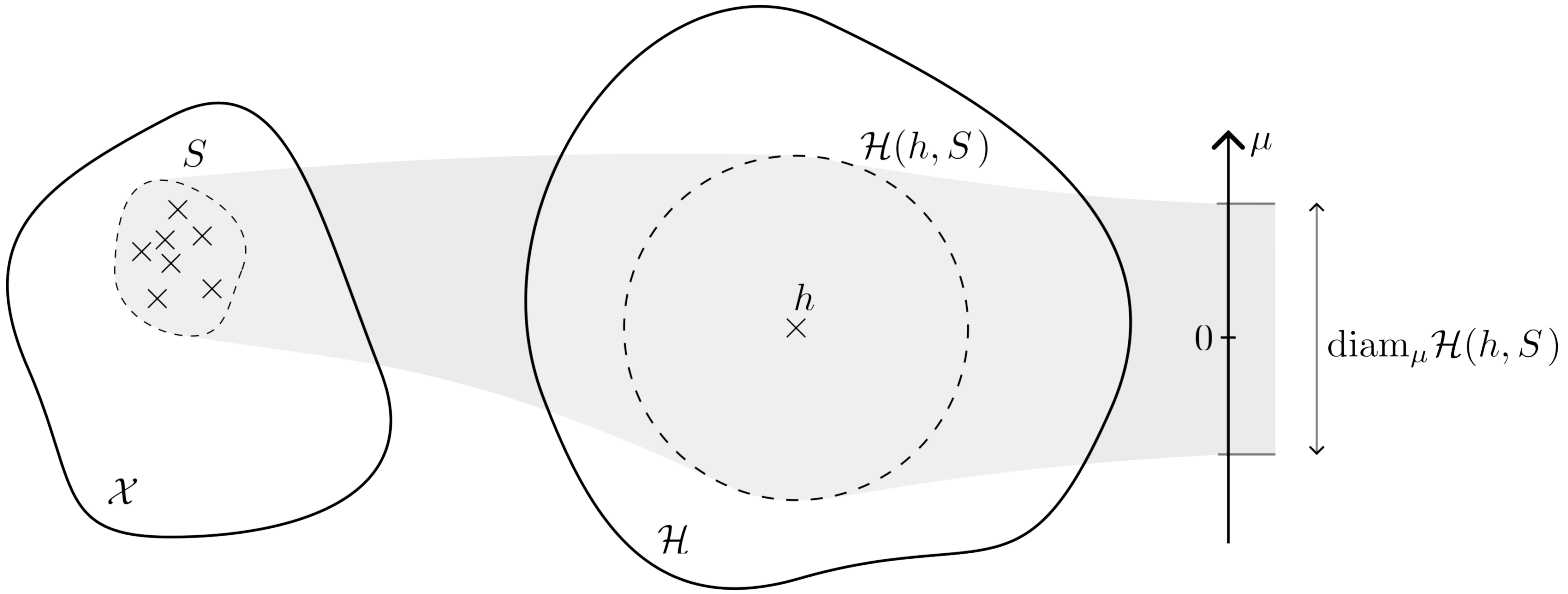
Are there models harder to audit ?

Measuring the audit robustness



Are there models harder to audit ?

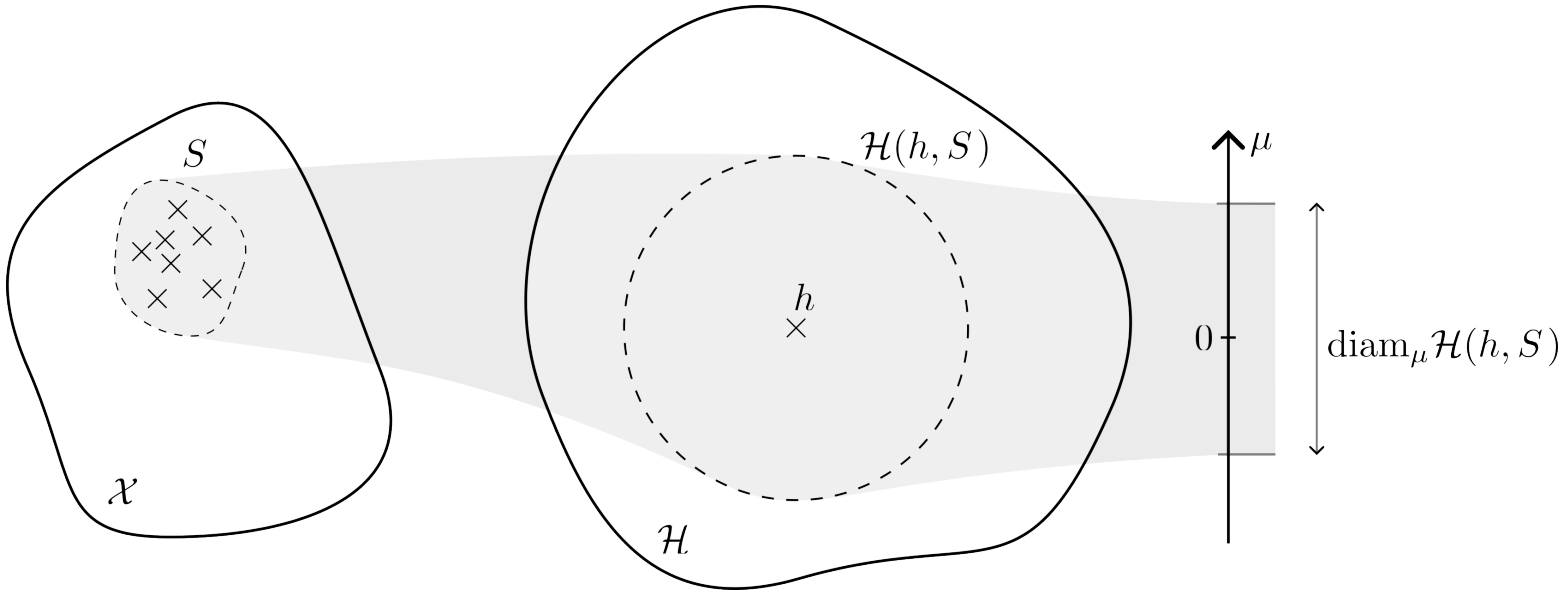
Measuring the audit robustness



$$\mathcal{H}(S, h^*) = \{h \in \mathcal{H} : \forall x \in S, h(x) = h^*(x)\}$$

Are there models harder to audit ?

Measuring the audit robustness



$$\mathcal{H}(S, h^*) = \{h \in \mathcal{H} : \forall x \in S, h(x) = h^*(x)\}$$

$$\mathrm{diam}_\mu \mathcal{H}(S, h^*) = \max_{h \in \mathcal{H}(S, h^*)} |\mu(h) - \mu(h^*)|$$

Prior art

Audit as a set covering problem

Algorithm 1 Minimax optimal deterministic auditing

Require: Finite hypothesis class $\mathcal{H}$, target error ϵ, fairness measure μ

Ensure: $\hat{\mu}$, an estimate of $\mu(h^*)$

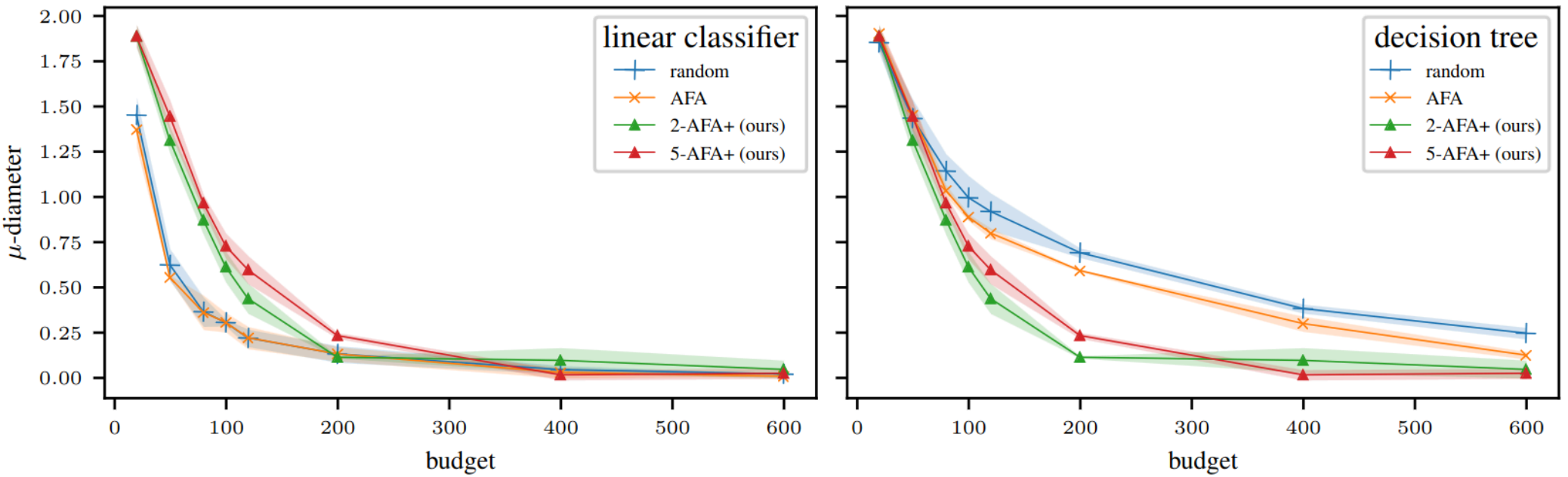
1: Let $V \leftarrow \mathcal{H}$
2: **while** $\text{diam}_\mu(V) > 2\epsilon$ **do**
3: Query $x \in \text{argmin}_x \max_y \text{Cost}\,(V_x^y)$, obtain label $h^*(x)$
4: $V \leftarrow V(h^*, \{x\})$
5: **return** $\frac{1}{2}\left(\max_{h \in V} \mu(h) + \min_{h \in V} \mu(h)\right)$

Active Fairness Auditing, Yan Le et al. [6]

Audit with explanations

Auditor	Query Complexity
Baseline	$O(\frac{1}{\epsilon}\log\frac{1}{\delta})$
$\texttt{AlgLC}_c$	1
$\texttt{AlgLC}_a$	$O(d\log(\frac{2c}{\epsilon}))$
AlgDT	$O(V)$

A learning-theoretic framework for certified auditing of machine learning models, Chhavi Yadav et al. [7]



$$\mathcal{H}(S, h^*, r) = \{h \in \mathcal{H} : \|h(S) - h^*(S)\| \leq r\}$$

$$\operatorname{diam}_\mu \mathcal{H}(S, h^*, r) = \max_{h \in \mathcal{H}(S,h^*,r)} |\mu(h) - \mu(h^*)|$$

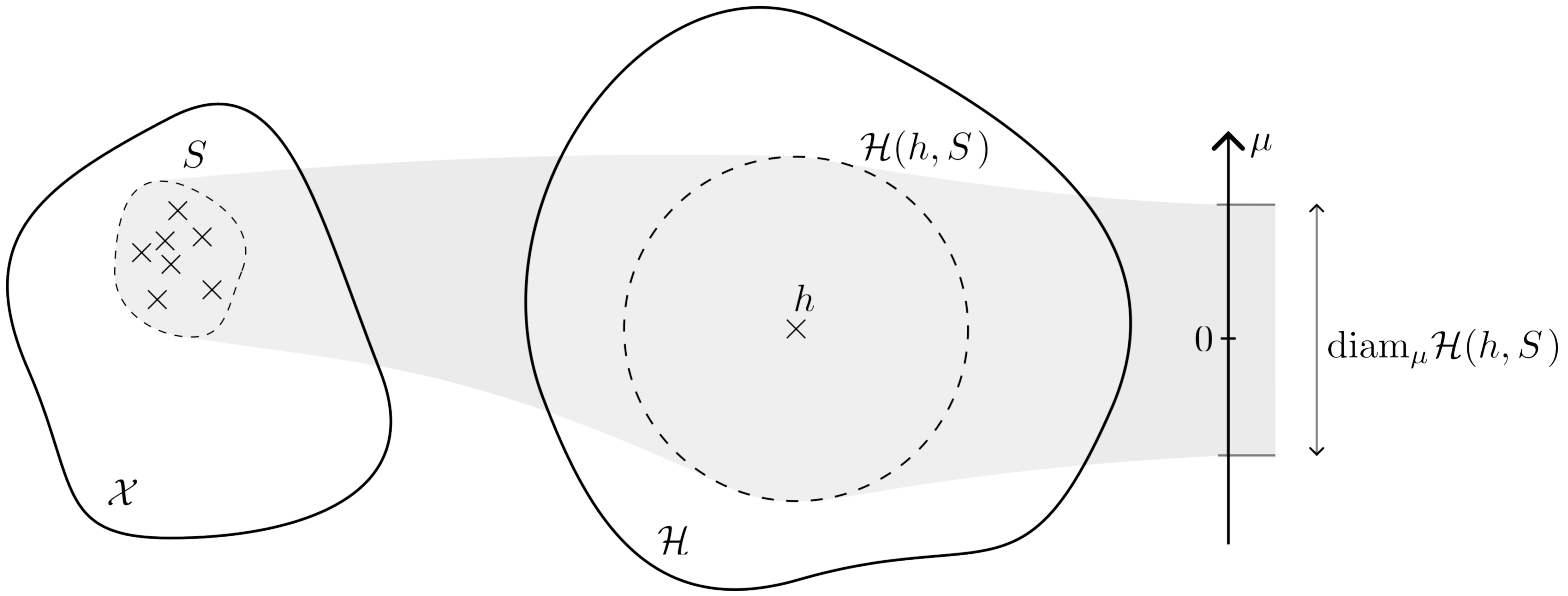
Are there models harder to audit ?

No prior, no gain

Theorem:

IF $\mathcal{H} = y^{\mathcal{X}}$ (no prior)

THEN $\operatorname{diam}_{\mu}\mathcal{H}(S, h^*) \underset{|S| \ll |A|}{\approx} 2\left(1 - \frac{|S|}{|A|}\right)$



Are there models harder to audit ?

Empirical study

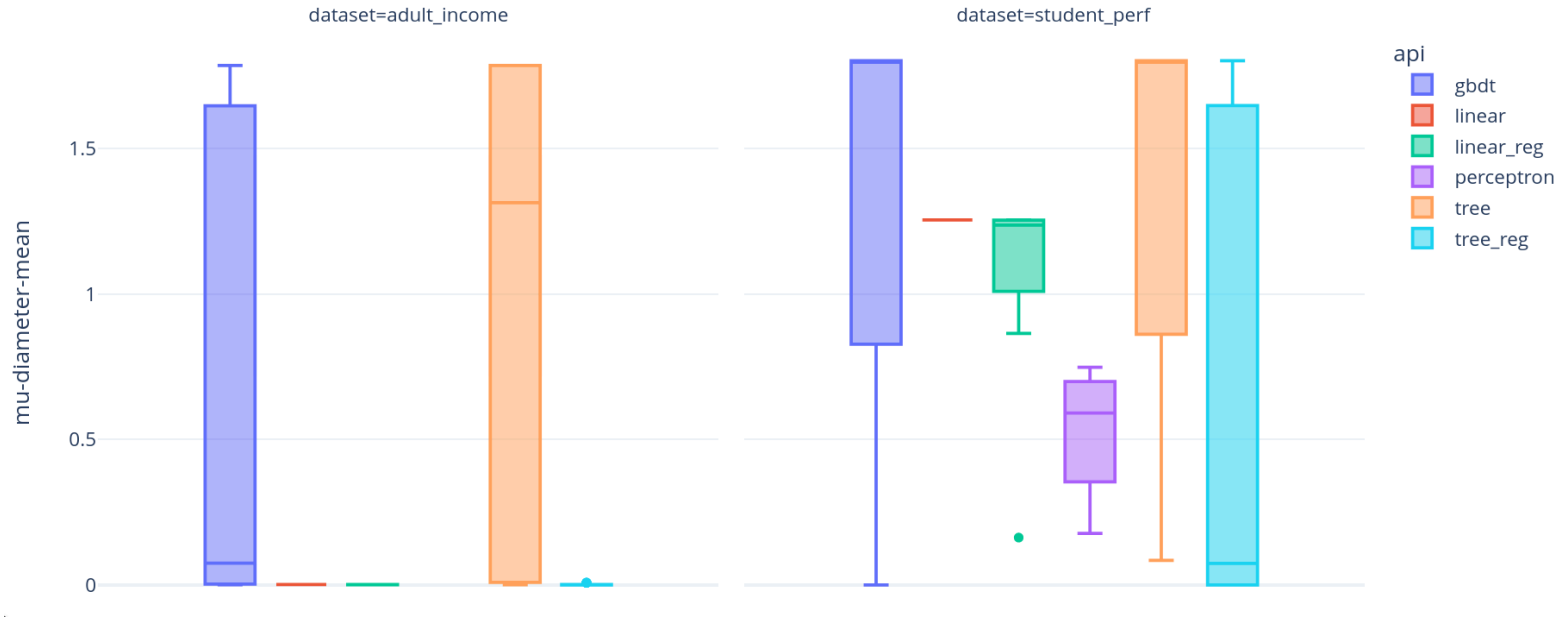
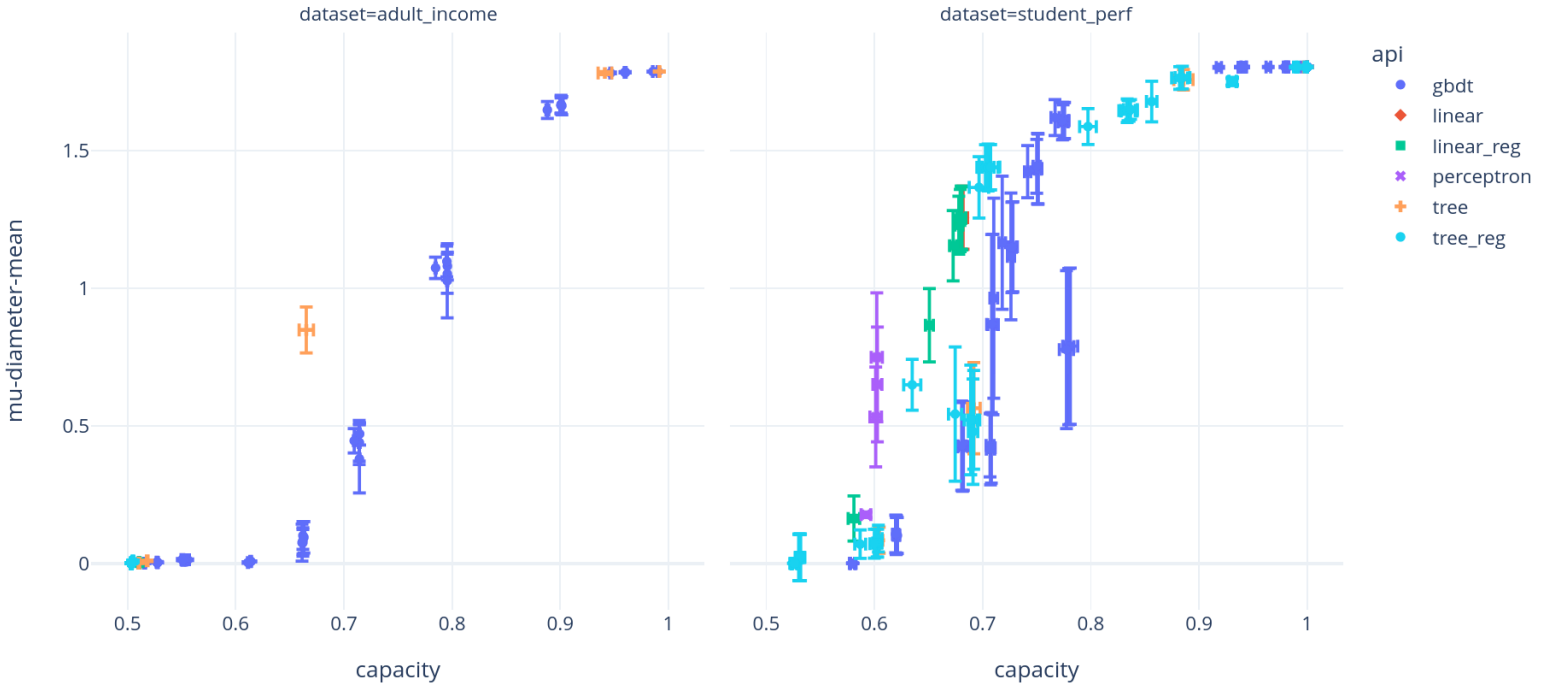


Figure 1: $\mathrm{diam}_\mu \mathcal{H}(S, h)$ for different type of models with varying hyperparameters, on AdultIncome and student perf datasets. Bootstrapped with 15 realizations. $|S| = .1\ |\mathcal{X}|$

Are there models harder to audit ?

Empirical study



Capacity (empirical Rademacher):

$$\mathcal{R}(\mathcal{H}, m) = \mathbb{E}_{\substack{x_i \\ y_i}} \left[\frac{1}{m} \max_{h \in \mathcal{H}} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) = y_i\} \right]$$

Merci ! Questions ?

Bibliography

[1] J. Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women”, Reuters, Oct. 2018, Accessed: Mar. 06, 2023. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] L. Chen, A. Mislove, and C. Wilson, “An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace”, in WWW '16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2016, pp. 1339–1349. doi: 10.1145/2872427.2883089.

[3] “EU AI Act: First Regulation on Artificial Intelligence | News | European Parliament”. Accessed: Jun. 21, 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[4] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, “How We Analyzed the COMPAS Recidivism Algorithm”, ProPublica, May 2016, Accessed: Mar. 06, 2023. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[5] Rédaction, “Numérique : que sont le DMA et le DSA, les règlements européens qui visent à réguler internet ?”. Accessed: Jun. 21, 2023. [Online]. Available: https://www.touteleurope.eu/societe/numerique-que-sont-le-dma-et-le-dsa-les-reglements-europeens-qui-veulent-reguler-internet/

[6] T. Yan and C. Zhang, “Active Fairness Auditing”, presented at the International Conference on Machine Learning, PMLR, Jun. 2022, pp. 24929–24962. Accessed: Dec. 01, 2022. Available: https://proceedings.mlr.press/v162/yan22c.html

[7] C. Yadav, M. Moshkovitz, and K. Chaudhuri, “A Learning-Theoretic Framework for Certified Auditing with Explanations”. Accessed: Dec. 20, 2022. [Online]. Available: http://arxiv.org/abs/2206.04740