Université de Rennes

WIDE

Inria

UMR IRISA

GOUVERNEMENT
Liberté
Égalité
Fraternité

PEReN
Pôle d'Expertise de la Régulation Numérique

Manipulation-proof auditing

Under manipulations, are there models harder to audit?

PrivSec seminar · April. 7th 2024


Augustin Godinot

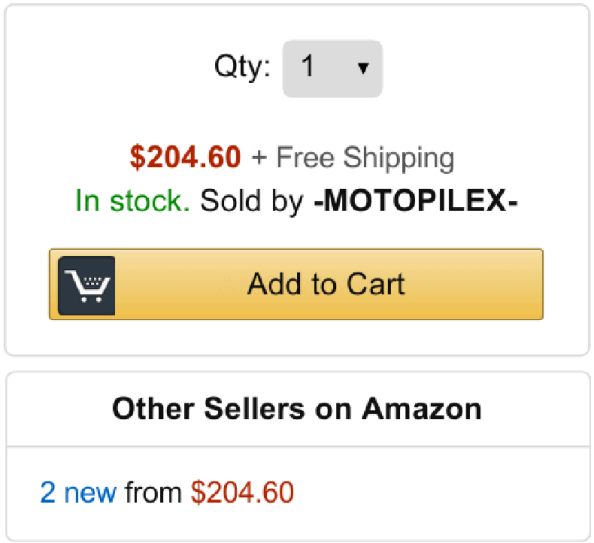

Erwan Le Merrer


Gilles Tredan


Camilla Penzo

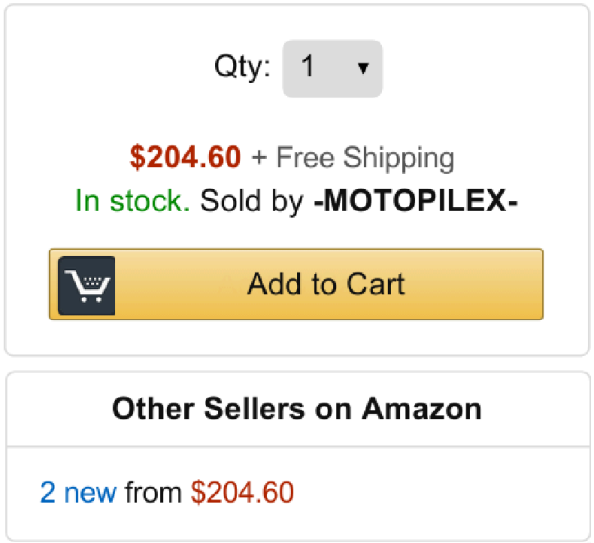

François Taïani

A first example



Metric Demographic parity between amazon and the other sellers

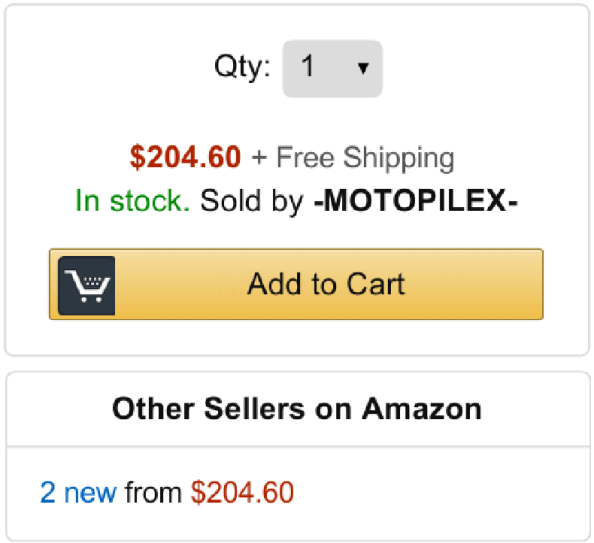
A first example



Metric Demographic parity between amazon and the other sellers

Audit queries Top-k best selling products

A first example



Metric Demographic parity between amazon and the other sellers

Audit queries Top-k best selling products

Data collection shameless scraping

In this talk

Context

How are audits currently conducted?

🔨 Framework

What do we mean by robust auditing: manipulation-proofness.

In this talk

Context

How are audits currently conducted?

🔨 Framework

What do we mean by robust auditing: manipulation-proofness.

Our contributions

🔎 A theoretical peek

Large models cannot be audited more efficiently that by random sampling.

📊 Empirical study

In practice, the cost to evade a black-box audit is mild.

In this talk

Context

How are audits currently conducted?

🔨 Framework

What do we mean by robust auditing: manipulation-proofness.

Our contributions

🔎 A theoretical peek

Large models cannot be audited more efficiently that by random sampling.

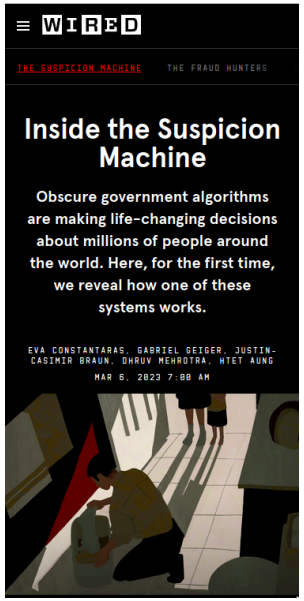
📊 Empirical study

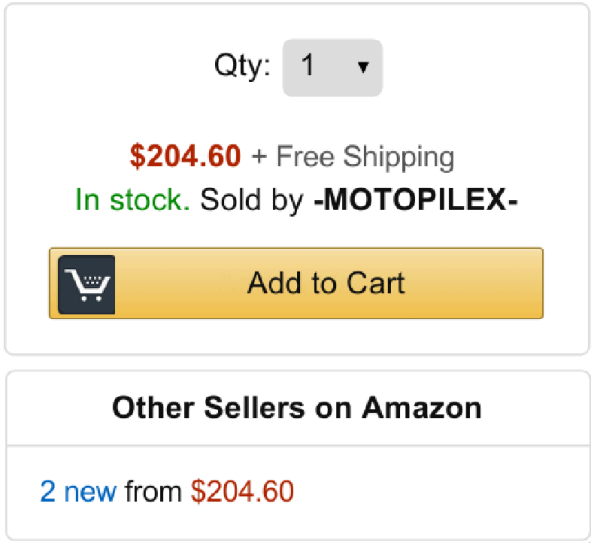
In practice, the cost to evade a black-box audit is mild.

Concluding remarks

The implications for AI regulation.

Algorithm audits





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

Context

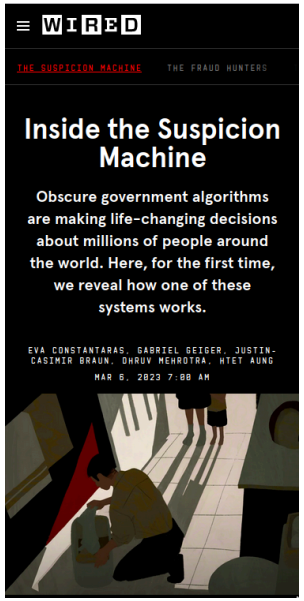
🔨 Framework

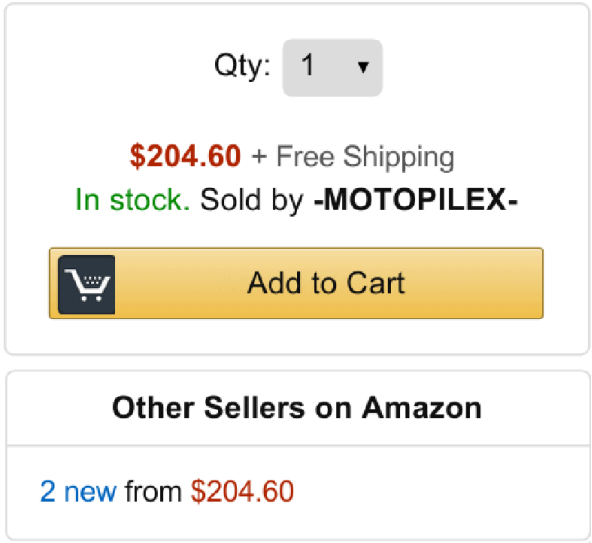
🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Algorithm audits





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

J. Dastin, L. Chen, A. Mislove, and C. Wilson, , J. Larson, S. Mattu, L. Kirchner, and J. Angwin, Rédaction

Context

🔨 Framework

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Audit steps

Choosing the metric

- FairML book [6]
- Political implications of the metric [7]
- Data minimization [8]
- Privacy auditing [9]

Choosing the queries

- Classical random sampling [10]
- Crafted datasets
- Active learning [11]
- Fairness by betting [12]

Data collection

- Do we get explanations? [13], [14]
- Do we have access to private API? [15]
- What if the platform lies? [11]

Audit steps

Context

🔨 Framework

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Choosing the metric

- FairML book [6]
- Political implications of the metric [7]
- Data minimization [8]
- Privacy auditing [9]

Choosing the queries

- Classical random sampling [10]
- Crafted datasets
- Active learning [11]
- Fairness by betting [12]

Data collection

- Do we get explanations? [13], [14]
- Do we have access to private API? [15]
- What if the platform lies? [11] **this talk**

my PhD topic: how to design audit methods that are not easily gamed by platforms?

Context

🔨 Framework

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

my PhD topic: how to design audit methods that are not easily gamed by platforms?

- detecting lies, robust auditing, robust estimation $\Leftrightarrow$ comparing the observations to a prior.

my PhD topic: how to design audit methods that are not easily gamed by platforms?

- detecting lies, robust auditing, robust estimation $\Leftrightarrow$ comparing the observations to a prior.
- which prior to choose and how does it impact the audit process in practice?
  $\Rightarrow$ This talk: one choice of prior

my PhD topic: how to design audit methods that are not easily gamed by platforms?

- detecting lies, robust auditing, robust estimation $\Leftrightarrow$ comparing the observations to a prior.
- which prior to choose and how does it impact the audit process in practice?
  $\Rightarrow$ This talk: one choice of prior
- how to verify that the prior holds ?
  $\Rightarrow$ Model change detection

Manipulation-proof auditing

Threat model

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$

Hypothesis space

$\mathcal{H} \subset \{0,1\}^{\mathcal{X}}, h \in \mathcal{H}$

[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Manipulation-proof auditing

Threat model

Context

🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$

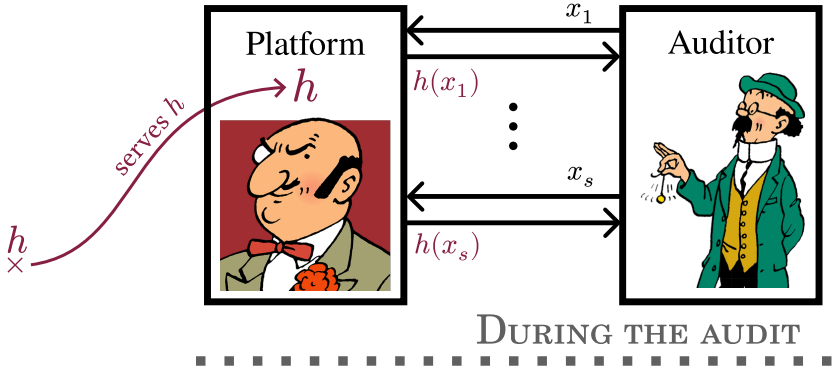
Hypothesis space

$\mathcal{H} \subset \{0,1\}^{\mathcal{X}}, h \in \mathcal{H}$

Audit set

$S = (x_1, ..., x_s) \subset \mathcal{X}$

$Y = (h(x_1), ..., h(x_s))$



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Manipulation-proof auditing

Threat model

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$

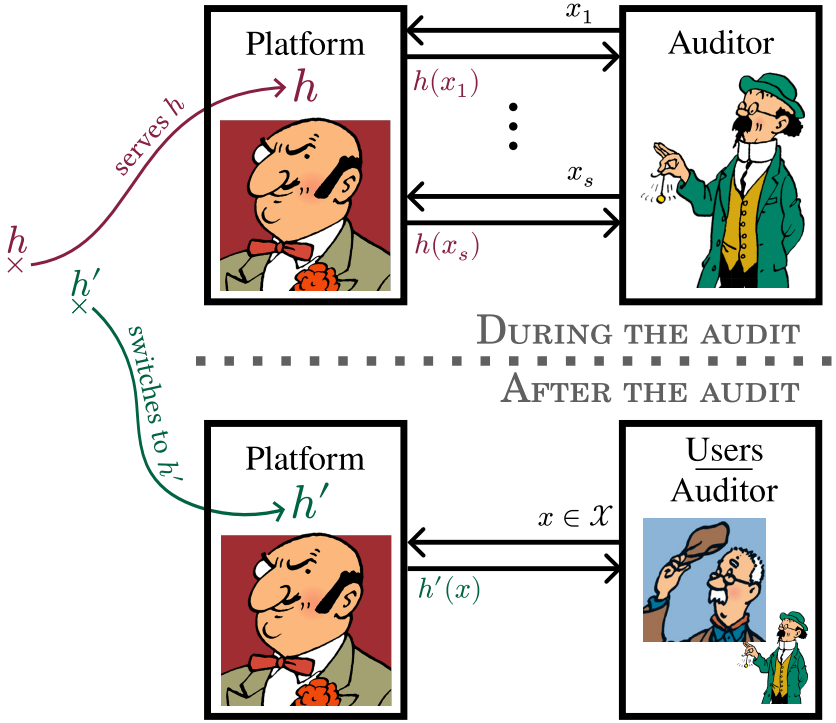
Hypothesis space

$\mathcal{H} \subset \{0,1\}^{\mathcal{X}}, h \in \mathcal{H}$

Audit set

$S = (x_1, ..., x_s) \subset \mathcal{X}$

$Y = (h(x_1), ..., h(x_s))$



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Manipulation-proof auditing

Threat model

Audit metric $\mu(h, S) = \mathbb{P}(\text{💵} \mid X \in S, \text{a}) - \mathbb{P}(\text{💵} \mid X \in S, \text{🇫🇷})$

Hypothesis space

$\mathcal{H} \subset \{0,1\}^{\mathcal{X}}, h \in \mathcal{H}$

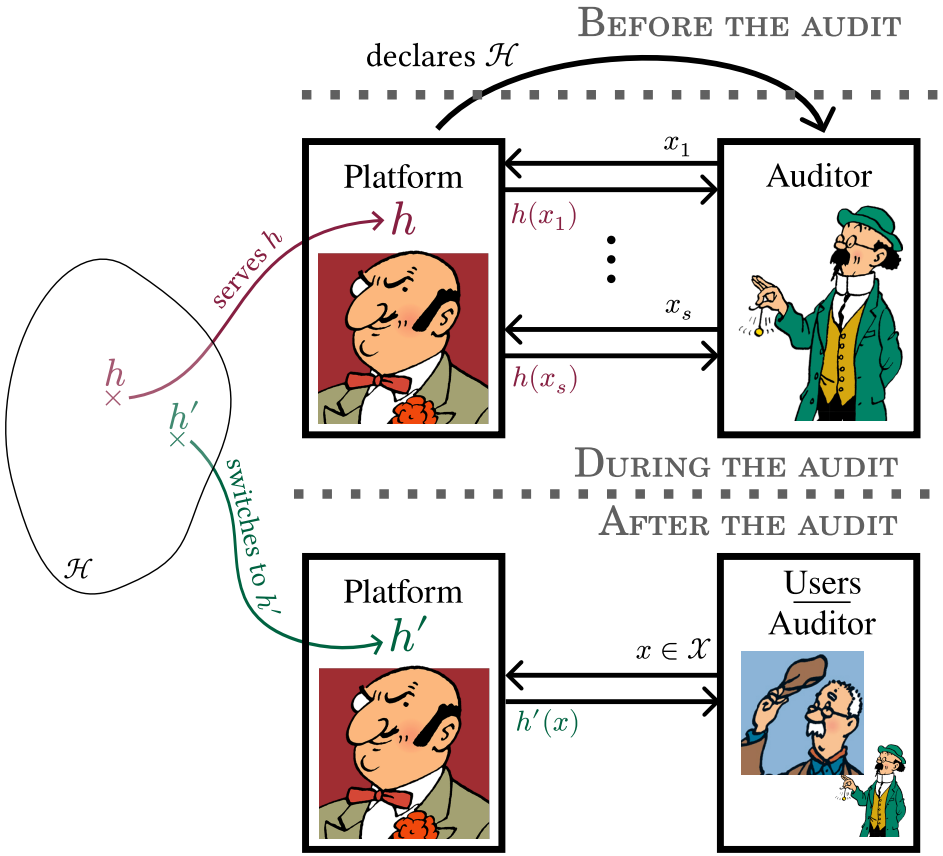
Audit set

$S = (x_1, ..., x_s) \subset \mathcal{X}$

$Y = (h(x_1), ..., h(x_s))$

Assumptions

1. *Auditor prior:* $\mathcal{H}$ is known



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Manipulation-proof auditing

Threat model

Audit metric $\mu(h, S) = \mathbb{P}(\text{💵} \mid X \in S, \text{a}) - \mathbb{P}(\text{💵} \mid X \in S, \text{🇫🇷})$

Hypothesis space

$\mathcal{H} \subset \{0,1\}^{\mathcal{X}}, h \in \mathcal{H}$

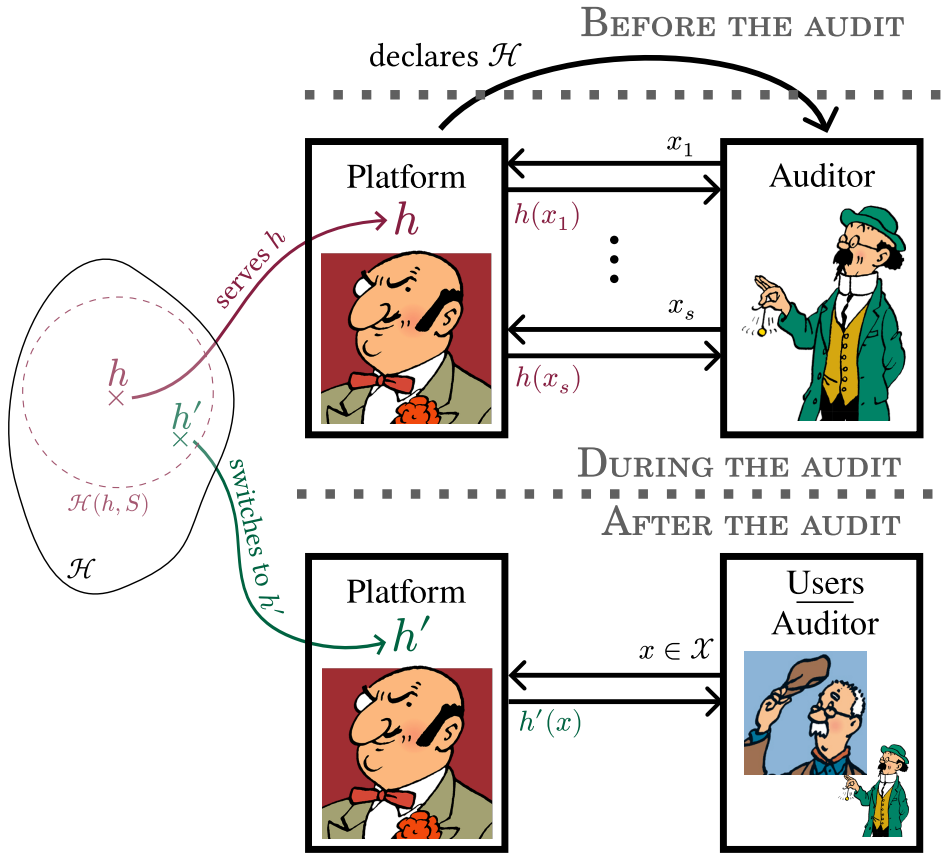
Audit set

$S = (x_1, ..., x_s) \subset \mathcal{X}$

$Y = (h(x_1), ..., h(x_s))$

Assumptions

1. *Auditor prior:* $\mathcal{H}$ is known
2. *Self-consistency:* once platform reveals its labeling of x, cannot change it.



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Manipulation-proof auditing

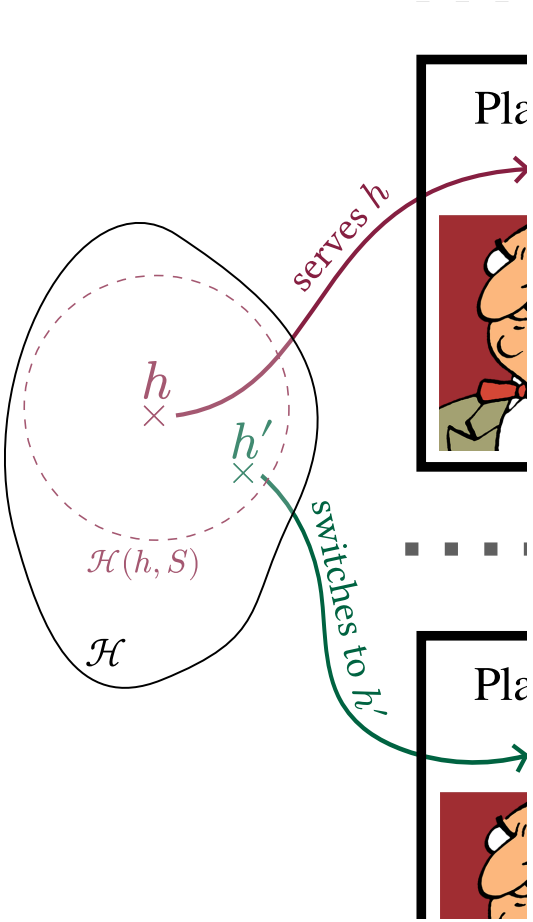
Audit manipulability

How much can μ change after the audit?

Measured during the audit	**True value** after the audit
$\mu(h)$	$\mu(h')$

Audit manipulability

$$|\mu(h) - \mu(h')| \leq ?$$

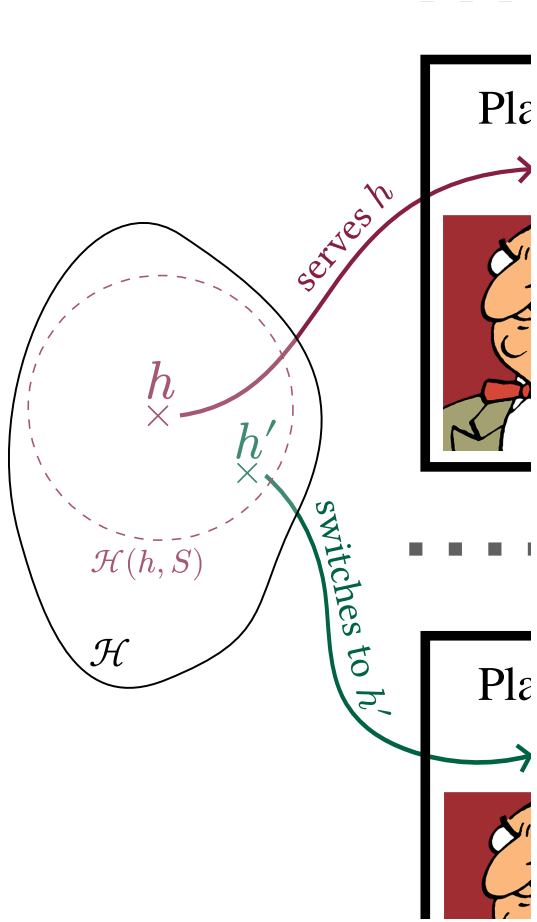


Manipulation-proof auditing

Audit manipulability

How much can μ change after the audit?

Measured during the audit	**True value** after the audit
$\mu(h)$	$\mu(h')$



Audit manipulability

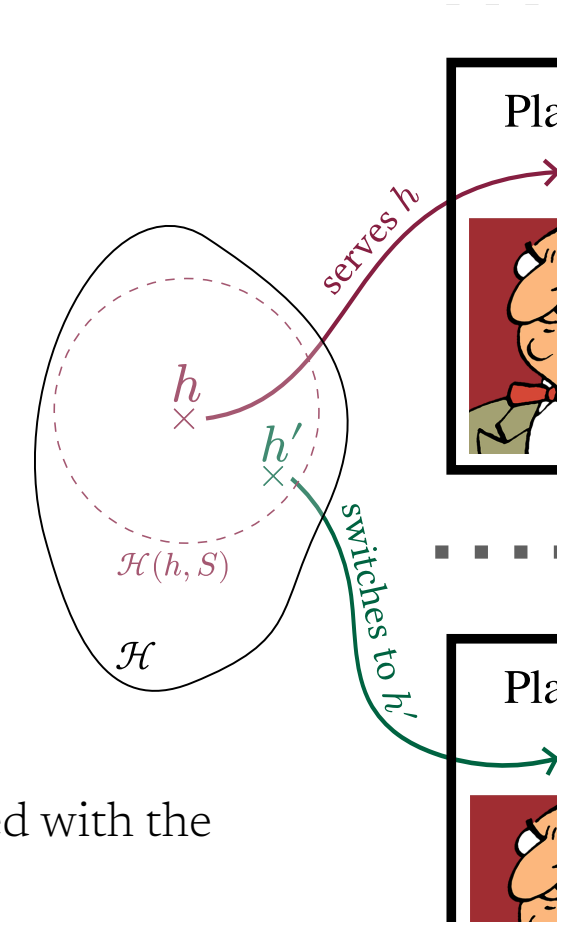
$$|\mu(h) - \mu(h')| \leq \operatorname{diam}_{\mu} \mathcal{H}(S, Y)$$

Manipulation-proof auditing

Audit manipulability

How much can μ change after the audit?

Measured during the audit	**True value** after the audit
$\mu(h)$	$\mu(h')$



Audit manipulability

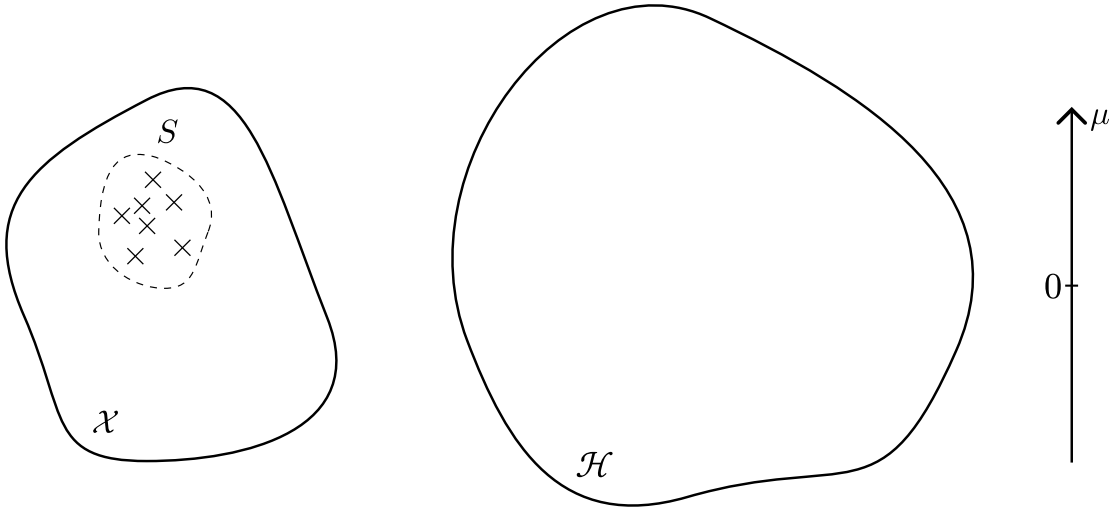
$$|\mu(h) - \mu(h')| \leq \operatorname{diam}_{\mu} \mathcal{H}(S, Y)$$

diameter of the version space (equipped with the pseudo metric $h, h' \mapsto |\mu(h) - \mu(h')|$)

Manipulation-proof auditing

Version Space 🌠

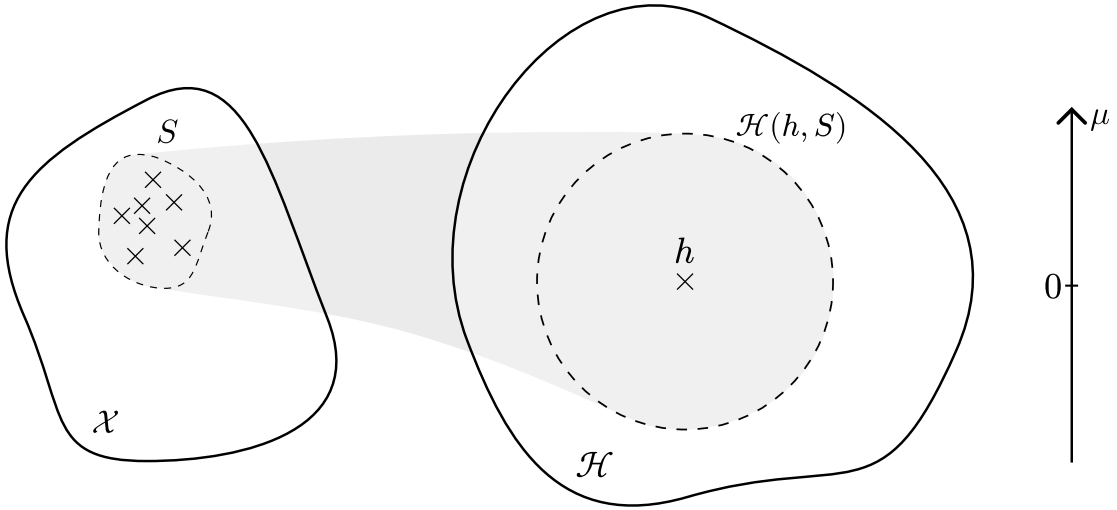
Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$



Manipulation-proof auditing

Version Space 🌠

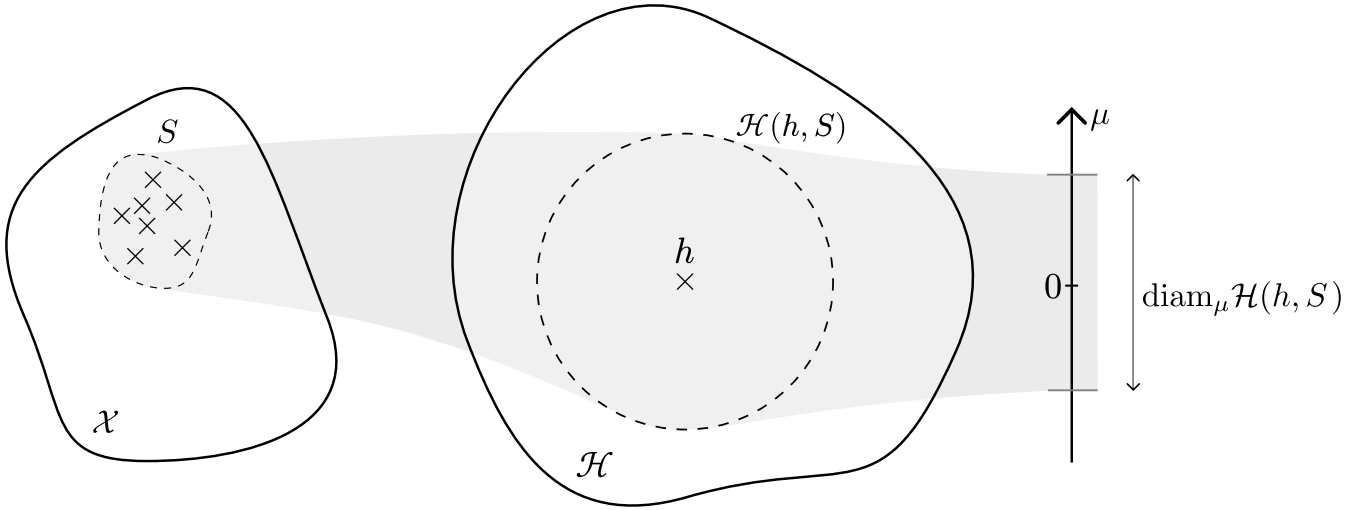
Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$



Manipulation-proof auditing

Version Space 🌠

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$



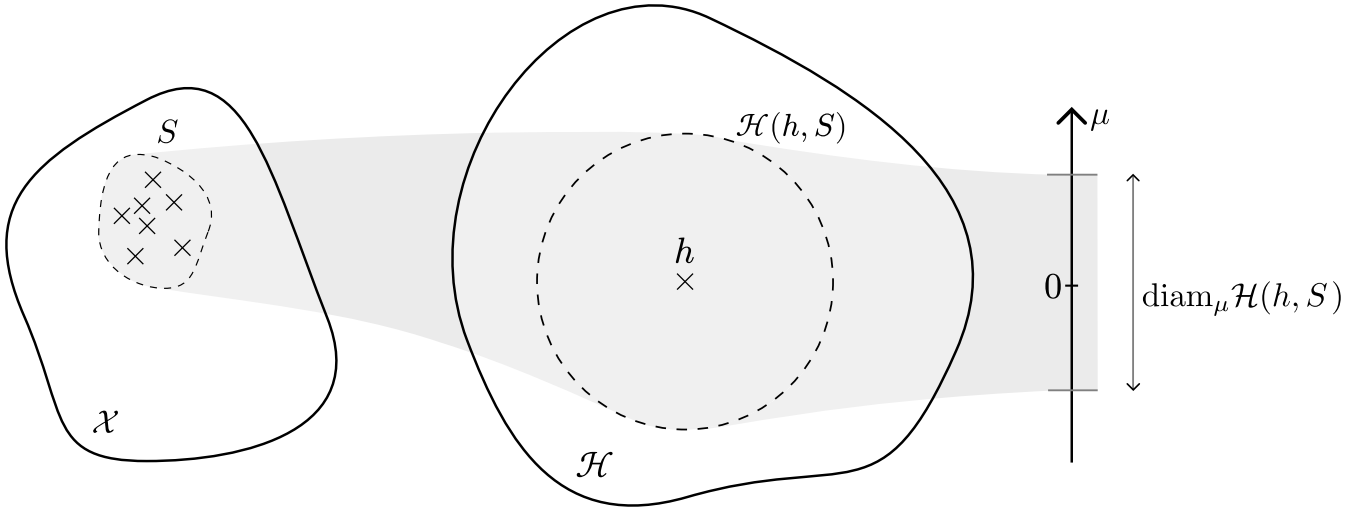
$$\mathcal{H}(S, h) = \{h' \in \mathcal{H} : \forall x \in S, h'(x) = h(x)\}$$

$$\text{diam}_\mu \ \mathcal{H}(S, h) \quad = \max_{h' \in \ \mathcal{H}(S,h)} |\mu(h') - \mu(h)|$$

Manipulation-proof auditing

Version Space 🌠

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$



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

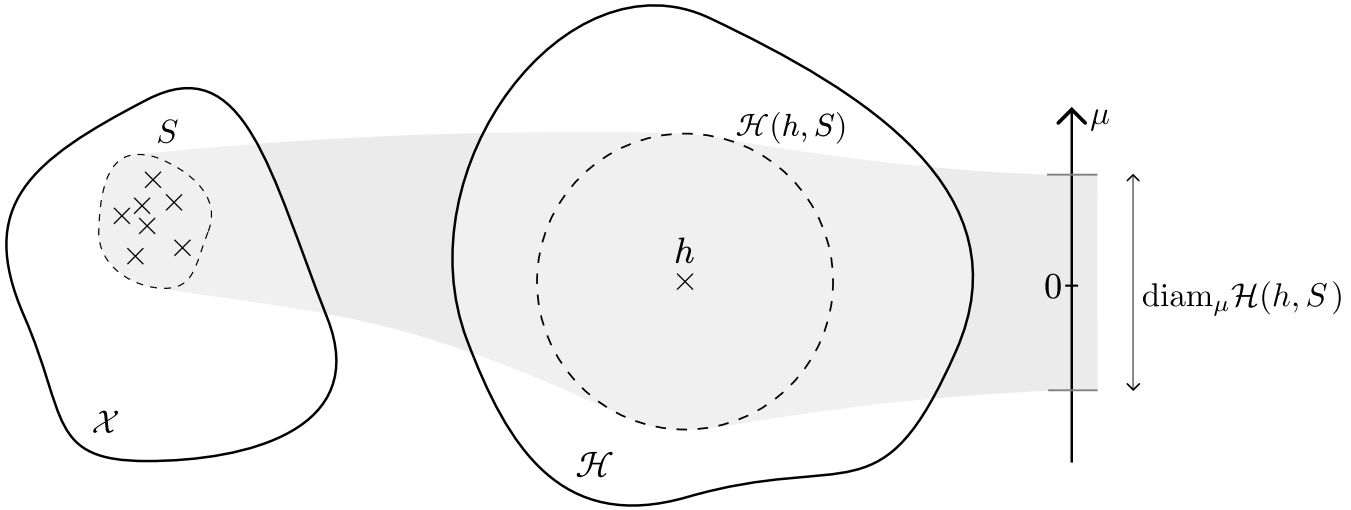
$$\mathrm{diam}_\mu \, \mathcal{H}(S, h) = \max_{h' \in \mathcal{H}(S,h)} |\mu(h') - \mu(h)|$$

Version space

Manipulation-proof auditing

Version Space 🌠

Audit metric $\mu(h, S) = \mathbb{P}(💵 \mid X \in S, \text{a}) - \mathbb{P}(💵 \mid X \in S, 🇫🇷)$



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

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

μ-diameter

Version space

State of the Art

Active auditing algorithms

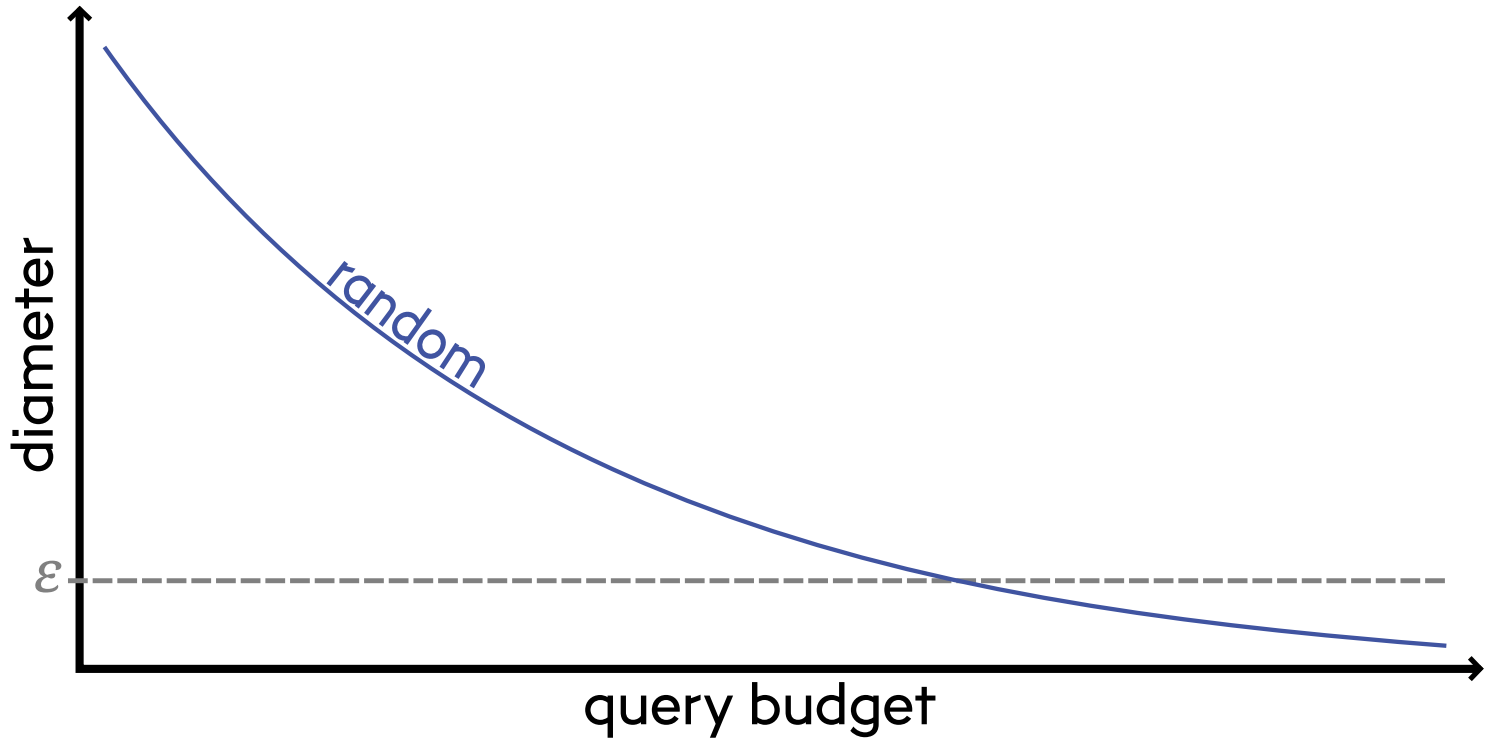
Context

🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

State of the Art

Active auditing algorithms

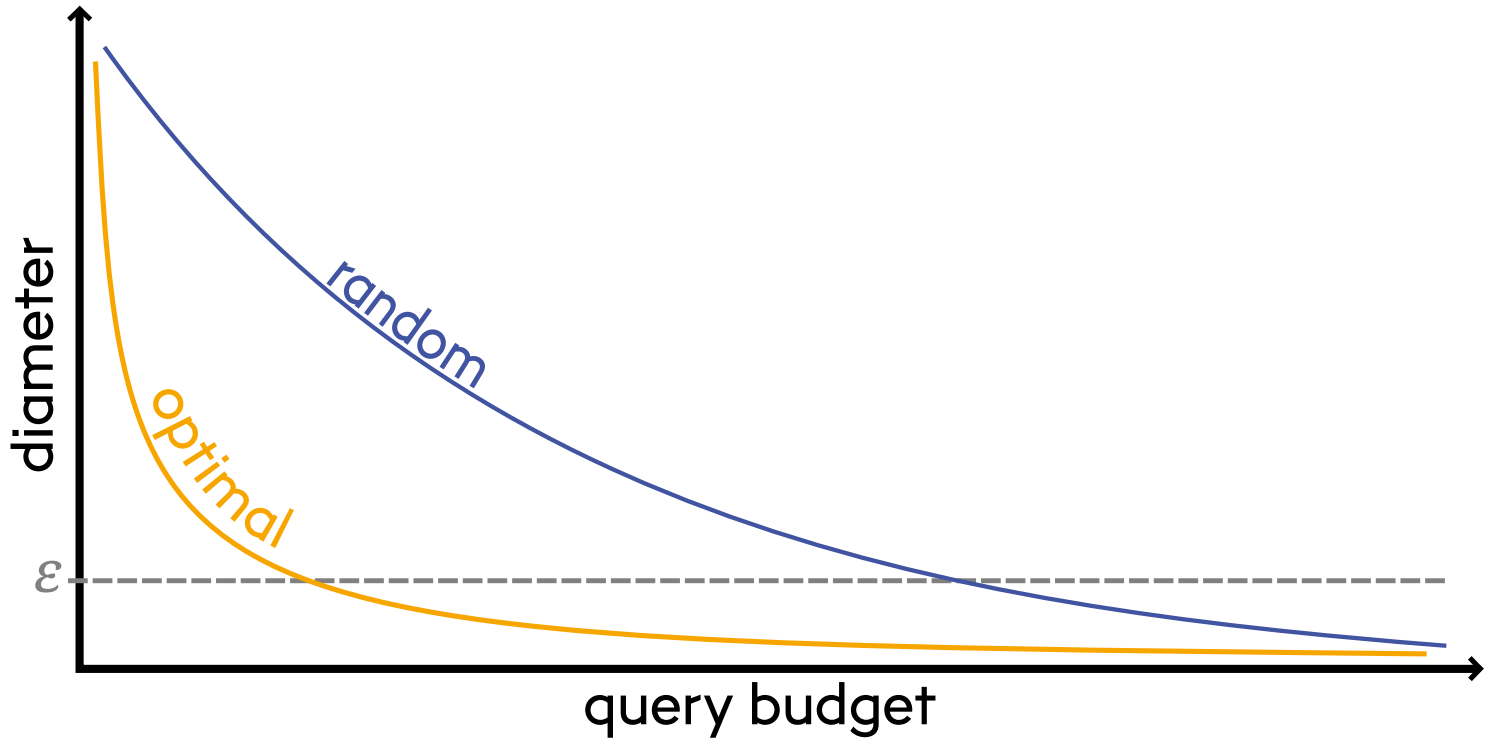
Context

🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

State of the Art

Active auditing algorithms

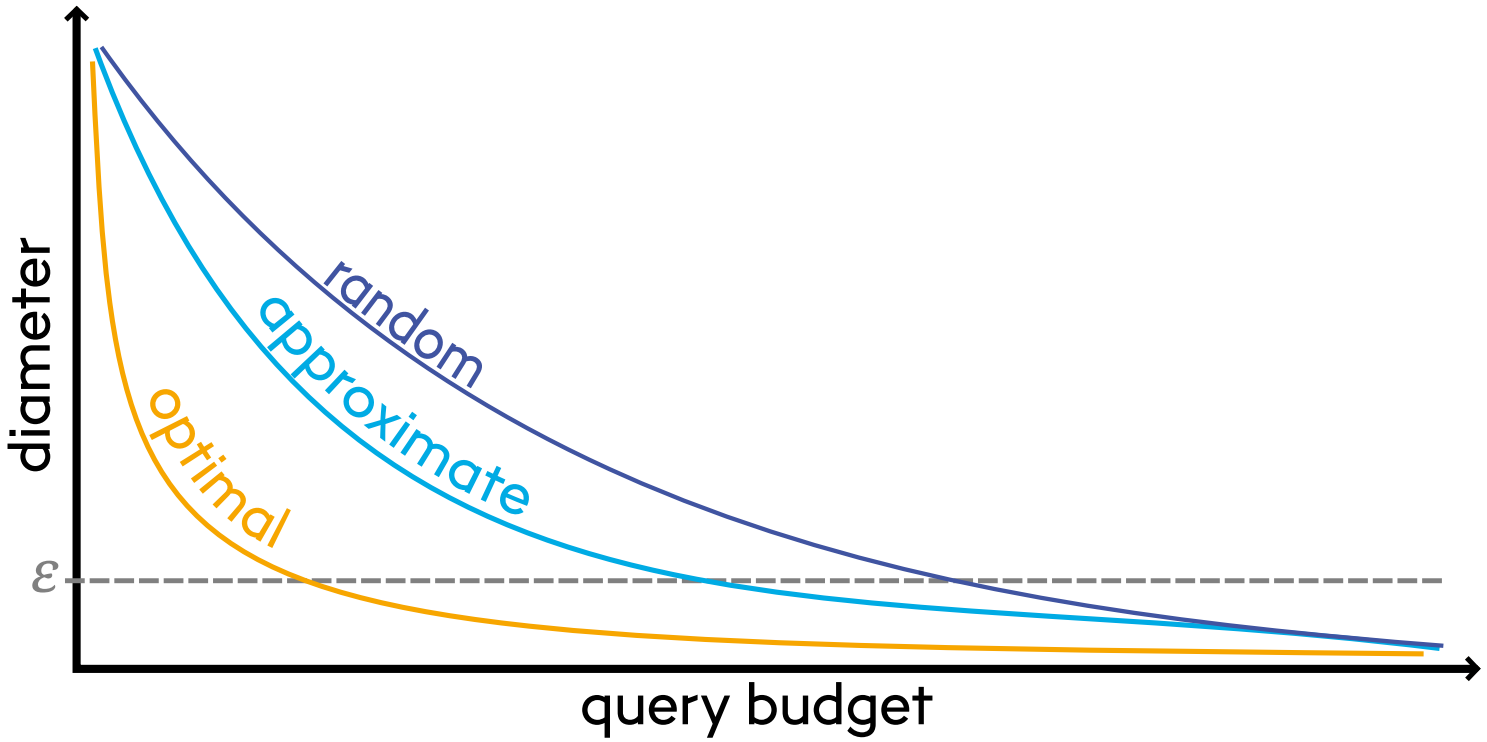
Context

🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

State of the Art

Active auditing algorithms

Context

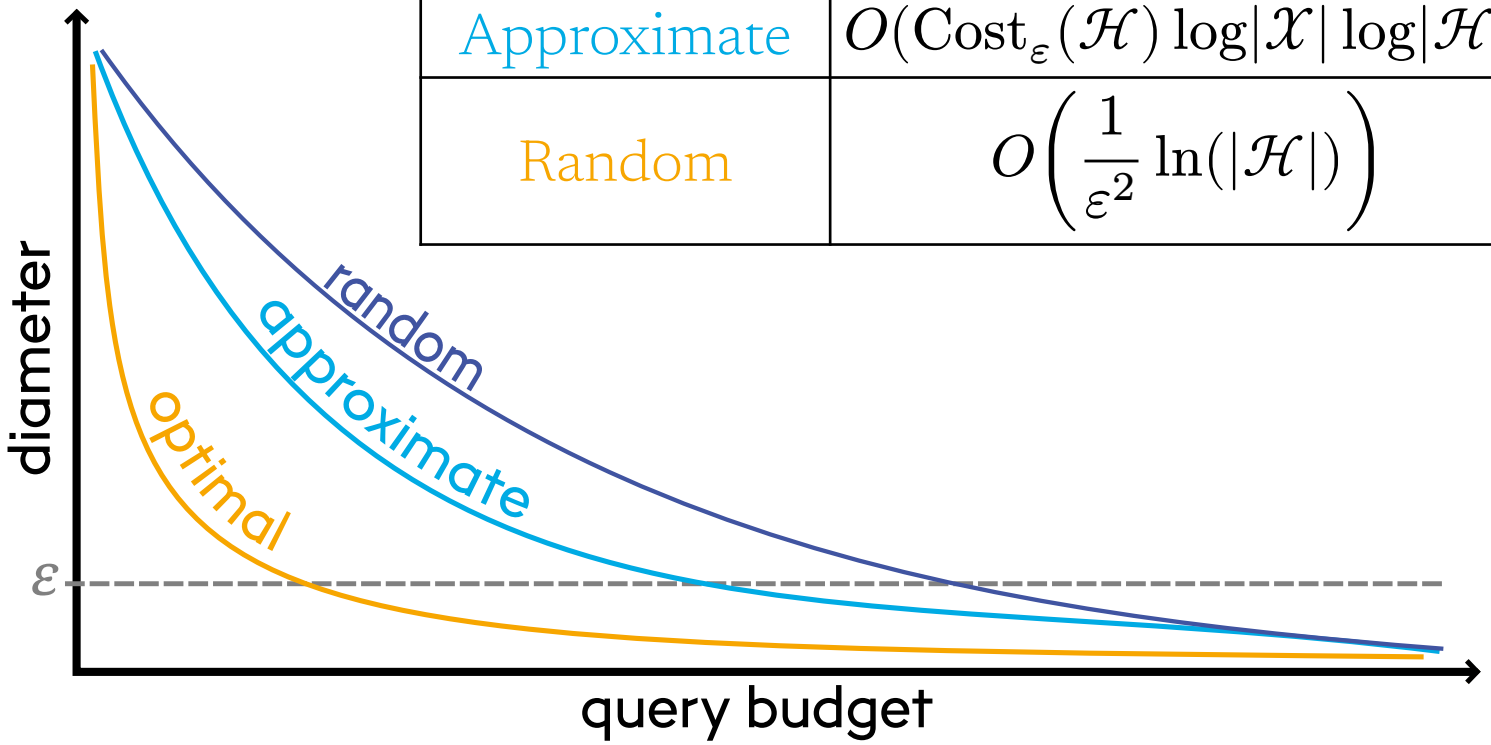
🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Audit method	Query complexity
Optimal	$\mathrm{Cost}_\varepsilon(\mathcal{H})$
Approximate	$O(\mathrm{Cost}_\varepsilon(\mathcal{H}) \log\lvert\mathcal{X}\rvert \log\lvert\mathcal{H}\rvert)$
Random	$O\left(\frac{1}{\varepsilon^2}\ln(\lvert\mathcal{H}\rvert)\right)$



[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

State of the Art

Active auditing algorithms

Context

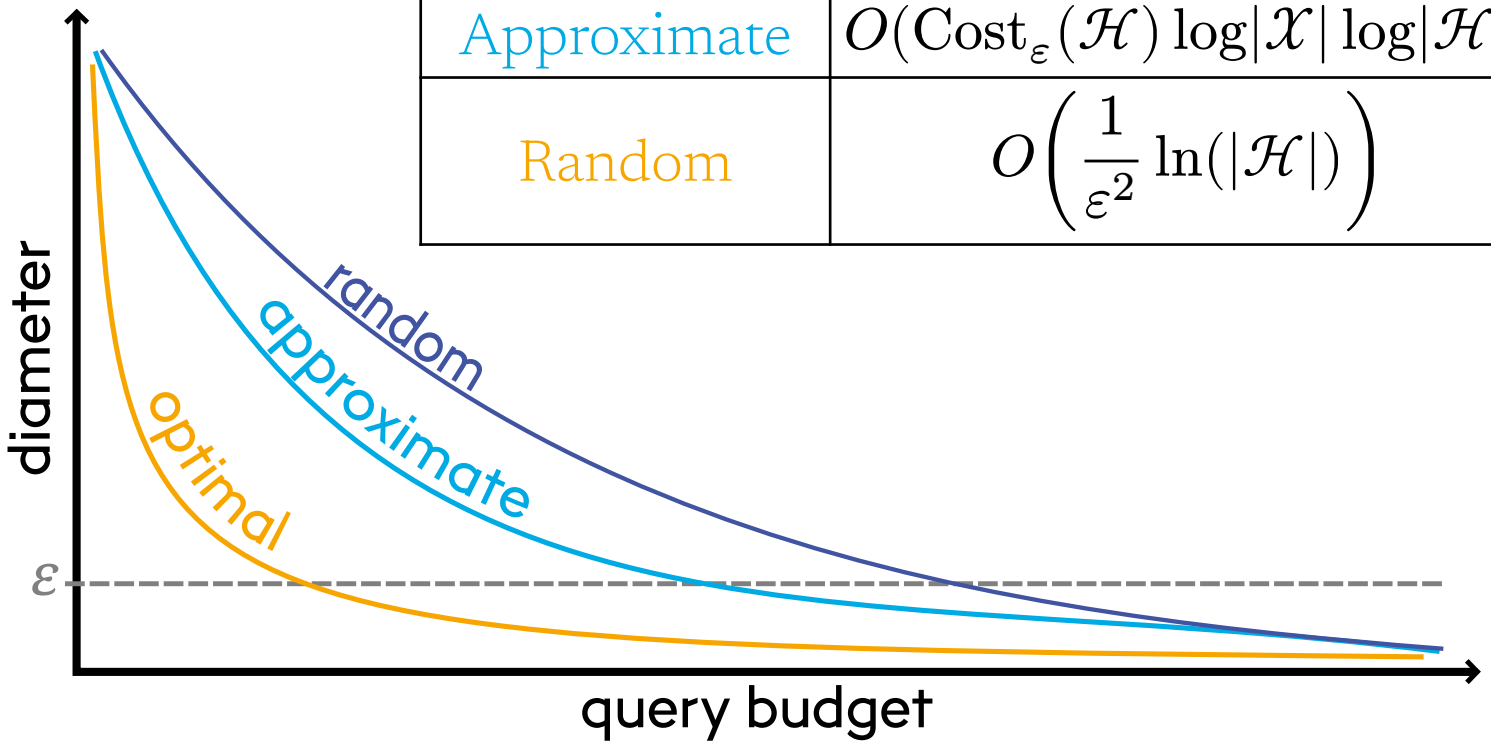
🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Audit method	Query complexity
Optimal	$\mathrm{Cost}_\varepsilon(\mathcal{H})$
Approximate	$O(\mathrm{Cost}_\varepsilon(\mathcal{H}) \log\lvert\mathcal{X}\rvert \log\lvert\mathcal{H}\rvert)$
Random	$O\left(\frac{1}{\varepsilon^2}\ln(\lvert\mathcal{H}\rvert)\right)$



Problem: Computational cost depends on $\mathcal{H}$!

[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

State of the Art

Active auditing algorithms

Context

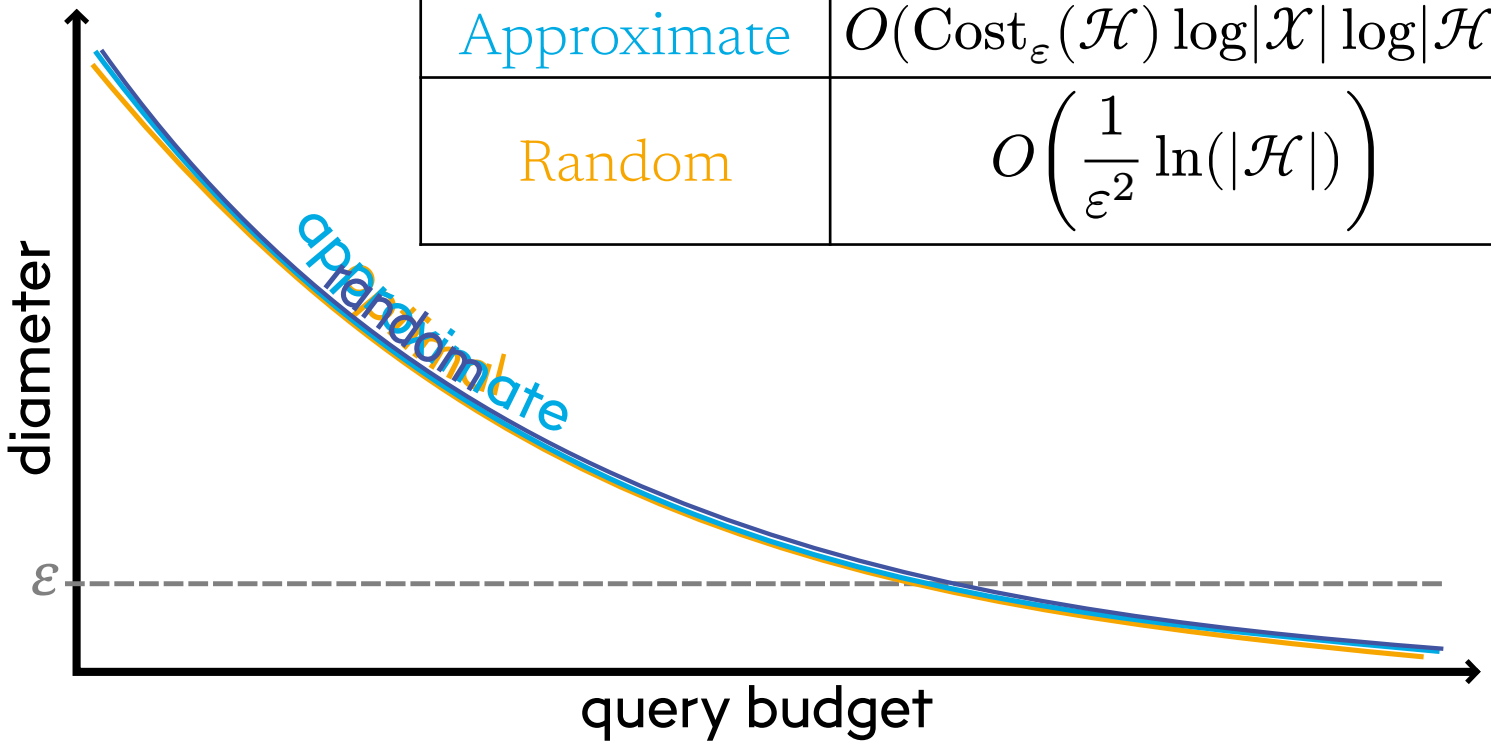
🔨 **Framework**

🔎 A theoretical peek

📊 Empirical study

Concluding remarks

Audit method	Query complexity
Optimal	$\mathrm{Cost}_\varepsilon(\mathcal{H})$
Approximate	$O(\mathrm{Cost}_\varepsilon(\mathcal{H}) \log\lvert\mathcal{X}\rvert \log\lvert\mathcal{H}\rvert)$
Random	$O\left(\frac{1}{\varepsilon^2}\ln(\lvert\mathcal{H}\rvert)\right)$



Problem: Computational cost depends on $\mathcal{H}$!

[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

Research questions

RQ1 $\exists\, \mathcal{H}$ such that $\begin{cases} \text{Complexity}(\mathcal{H}, \text{random audit}) \\ = \\ \text{Complexity}(\mathcal{H}, \text{optimal audit}) \end{cases}$?

RQ2 Do these $\mathcal{H}$ exist in practice ?

A simple case

Shattering hypothesis class

Theorem 1: No need to aim

If $\mathcal{H} = \{0,1\}^{\mathcal{X}}$, then

$$\operatorname{diam}_{\mu}\mathcal{H}(h^*, S) = 2 - (\mathbb{P}(X \in S \mid X_A = 1) + \mathbb{P}(X \in S \mid X_A = 0))$$

A simple case

Shattering hypothesis class

Theorem 1: No need to aim

If $\mathcal{H} = \{0,1\}^{\mathcal{X}}$, then

$$\operatorname{diam}_{\mu}\mathcal{H}(h^*, S) = 2 - (\mathbb{P}(X \in S \mid X_A = 1) + \mathbb{P}(X \in S \mid X_A = 0))$$

Proof intuition:

1. Split the value of the μ-diameter on S and $\overline{S}$
2. Constuct the "optimal" hypotheses $h^{\uparrow}$ and $h^{\downarrow}$
3. Express the result as a function of $\mathbb{P}(X \in S \mid X_A = 0 \text{ or } 1)$

A more refined case

Dictionnary models

Theorem 2: Little Robert (informal)

Let $d \in \{0,1\}^x$ be a dictionnary of memory m.

A more refined case

Dictionnary models

Theorem 2: Little Robert (informal)

Let $d \in \{0,1\}^{\mathcal{X}}$ be a dictionnary of memory m.
Then, for m large enough, $\mathbf{diam}_{\mu}\mathcal{H}(S,Y)$ is a function of

- sensitive groups $\mathbb{P}(X \in S \mid X_A = 1,2)$ ($\searrow$)
- the dictionary's size m ($\nearrow$)

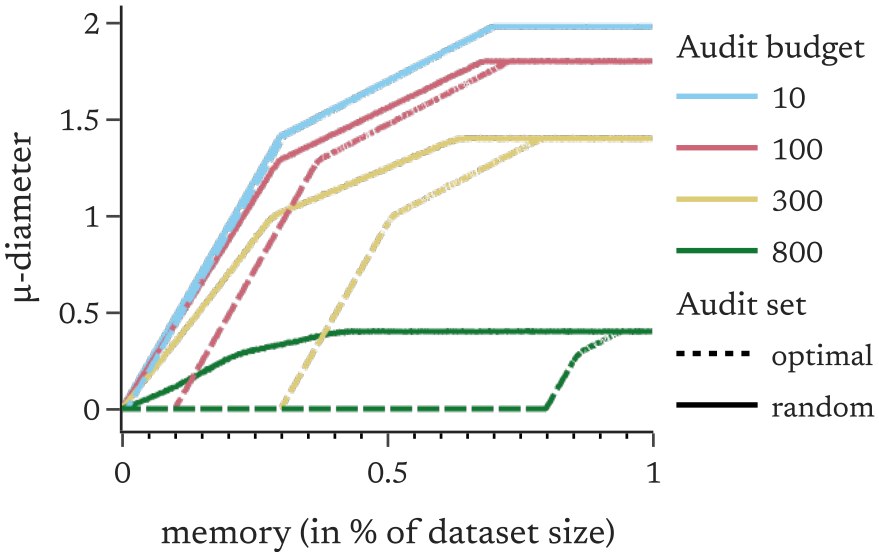
A more refined case

Dictionnary models

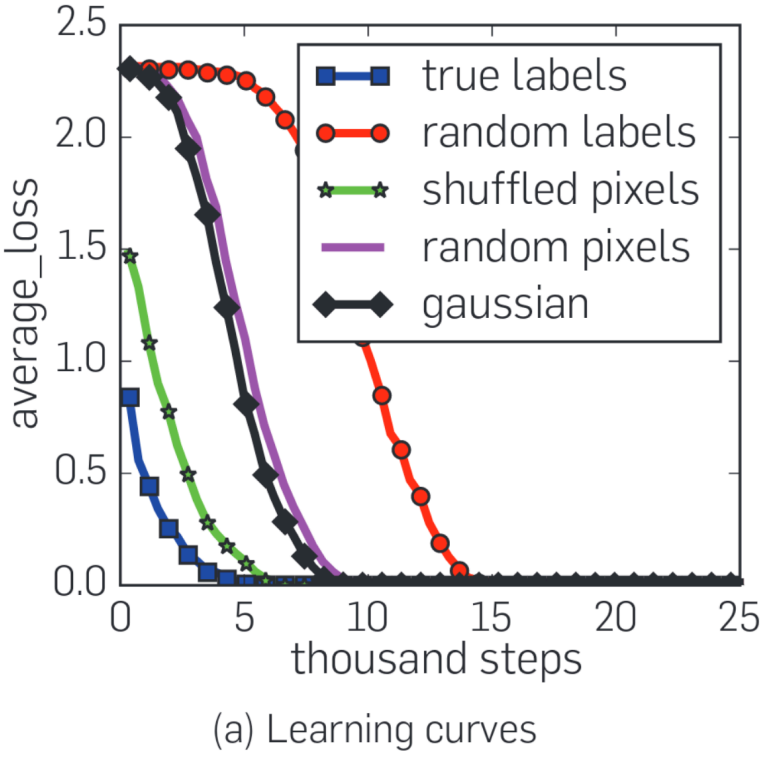
Theorem 2: Little Robert (informal)

Let $d \in \{0,1\}^{\mathcal{X}}$ be a dictionnary of memory m.
Then, for m large enough, $\mathbf{diam}_\mu \mathcal{H}(S, Y)$ is a function of

- sensitive groups $\mathbb{P}(X \in S \mid X_A = 1, 2)$ ($\searrow$)
- the dictionary's size m ($\nearrow$)



Benign overfitting



(a) Learning curves

Train loss of an Inception net on CIFAR10

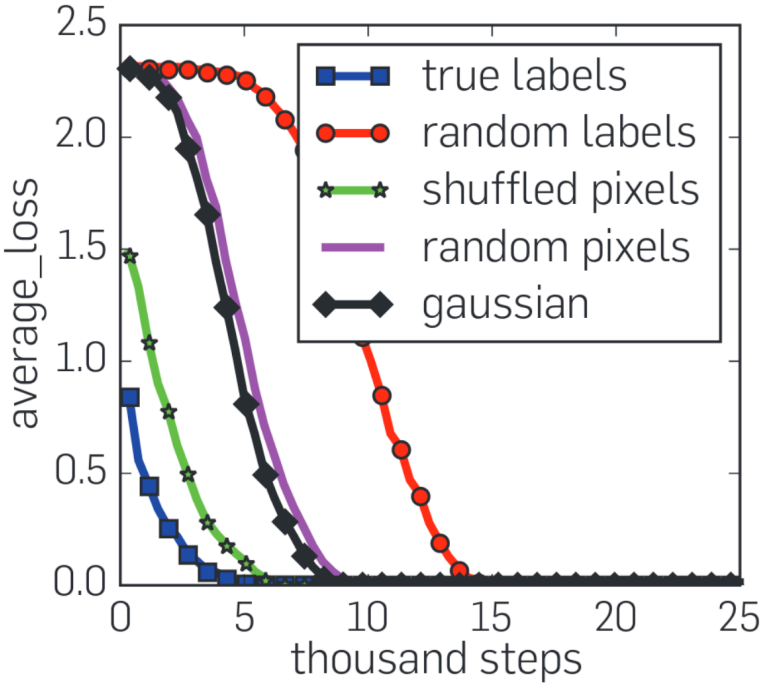
Taken from [16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding Deep Learning (Still) Requires Rethinking Generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, Feb. 2021, doi: 10.1145/3446776.

Benign overfitting

Definition 2: Benign overfitting on c

$\mathcal{H}$ exhibits benign overfitting iif $\exists h^* \in \mathcal{H}$ such that

1. $\forall D \subset \mathcal{X}, |D| \leq d_0, \text{error}(h, D) = 0$
2. $\text{error}(h^*, \mathcal{X}) \leq \varepsilon$



Train loss of an Inception net on CIFAR10

Taken from [16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding Deep Learning (Still) Requires Rethinking Generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, Feb. 2021, doi: 10.1145/3446776.

Benign overfitting

and audit difficulty

Corollary 1: Large models are difficult to audit

If $\mathcal{H}$ exhibits benign overfitting with respect to the sensitive attribute, then (for $\varepsilon < \frac{1}{2}$), $\mathbf{diam}_{\mu}(h, S)$ is a function of

- sensitive groups $\mathbb{P}(X \in S \mid X_A = 1, 2)$ ($\searrow$)
- the model's error-rate ε ($\searrow$)

Research questions

RQ1 $\exists\, \mathcal{H}$ such that $\begin{cases} \text{Complexity}(\mathcal{H}, \text{random audit}) \\ = \\ \text{Complexity}(\mathcal{H}, \text{optimal audit}) \end{cases}$?

$\Rightarrow$ ***Yes:*** *models with high capacity*

RQ2 Do these $\mathcal{H}$ exist in practice ?

Metrics

Simulated models

Model family $\mathcal{F}$ e.g. decision trees

Hypothesis class $\mathcal{H} \in \mathcal{F}$ e.g. trees with `max_depth = 2`

Metrics

- AuditManipulability($\mathcal{H}$)
- ModelCapacity($\mathcal{H}$)

Metrics

Simulated models

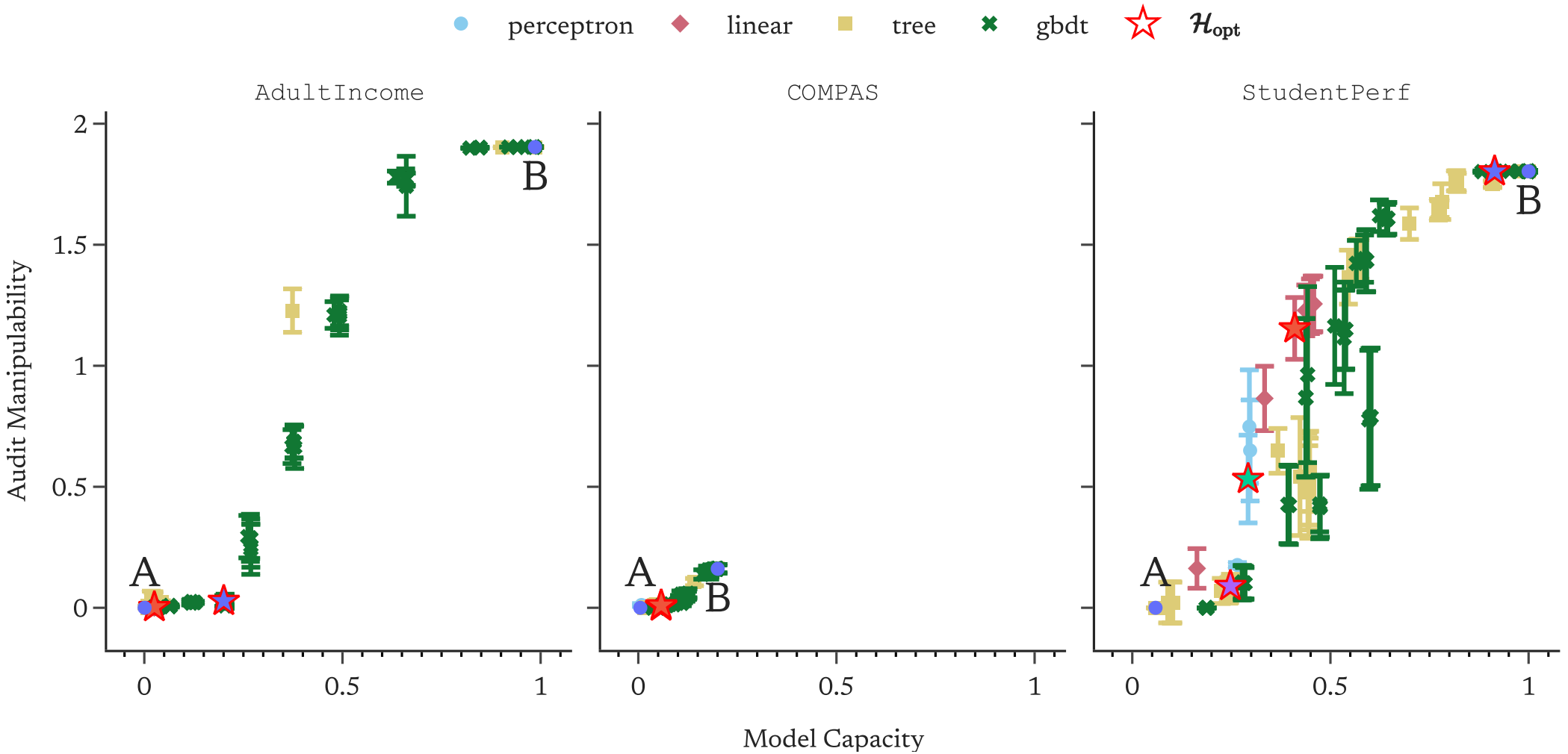
Model family $\mathcal{F}$ e.g. decision trees

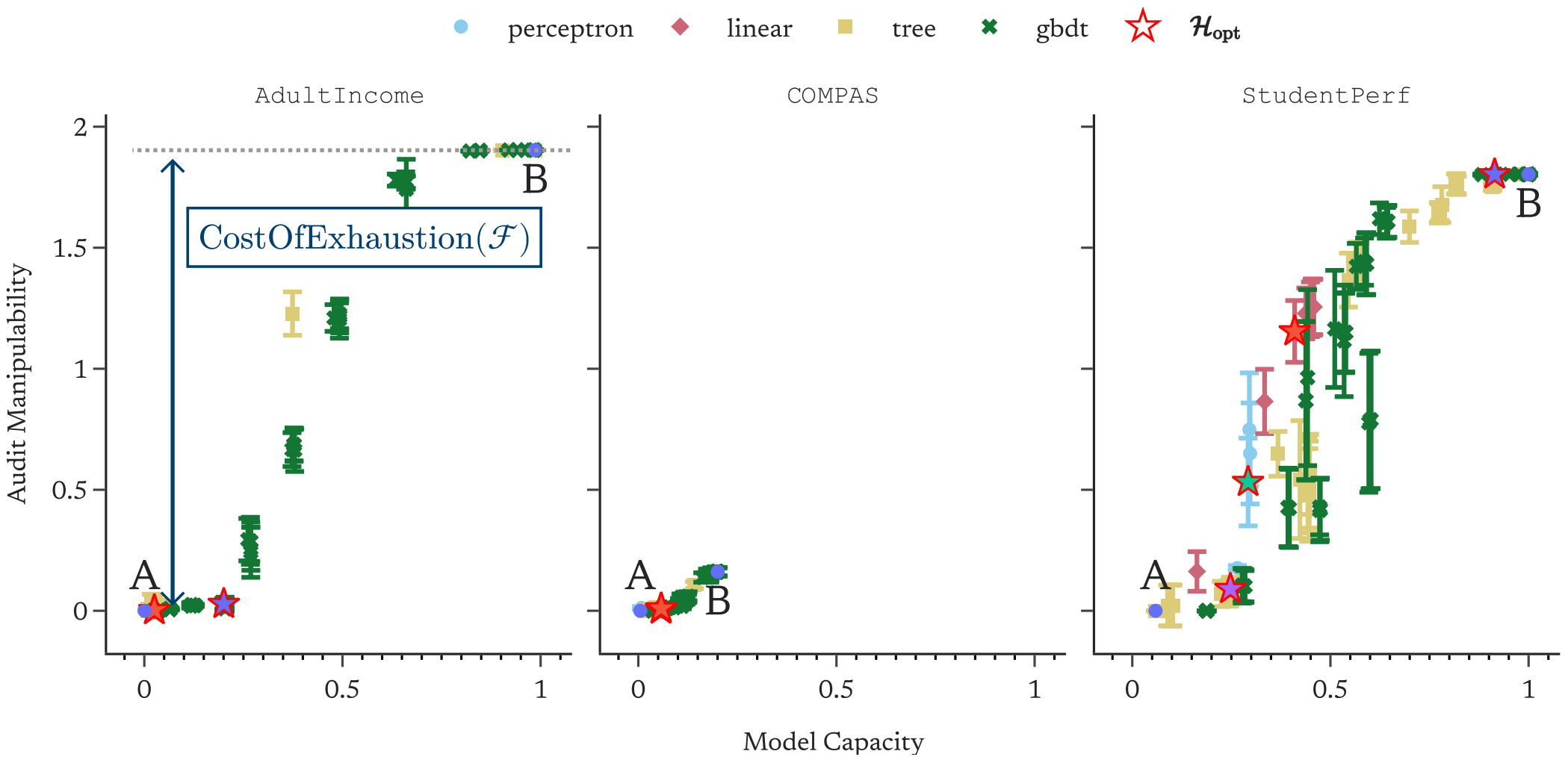
● perceptron ◆ linear ■ tree ✖ gbdt ☆ $\mathcal{H}_{\text{opt}}$

Hypothesis class $\mathcal{H} \in \mathcal{F}$ e.g. trees with `max_depth = 2`

Metrics

- $\text{AuditManipulability}(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{X}^m} \left[\text{diam}_\mu(h^*, S) \right]$
- $\text{ModelCapacity}(\mathcal{H}) = \mathbb{E}_{D \sim \mathcal{X}^r} [\text{Rademacher}(\mathcal{H}, D)]$


perceptron
linear
tree
gbdt
$\mathcal{H}_{\text{opt}}$
AdultIncome
COMPAS
StudentPerf
Audit Manipulability
Model Capacity
A
B
0
0.5
1
1.5
2


perceptron
linear
tree
gbdt
$\mathcal{H}_{\text{opt}}$
AdultIncome
COMPAS
StudentPerf
Audit Manipulability
Model Capacity
CostOfExhaustion($\mathcal{F}$)
A
B
0
0.5
1
1.5
2

Cost of exhaustion

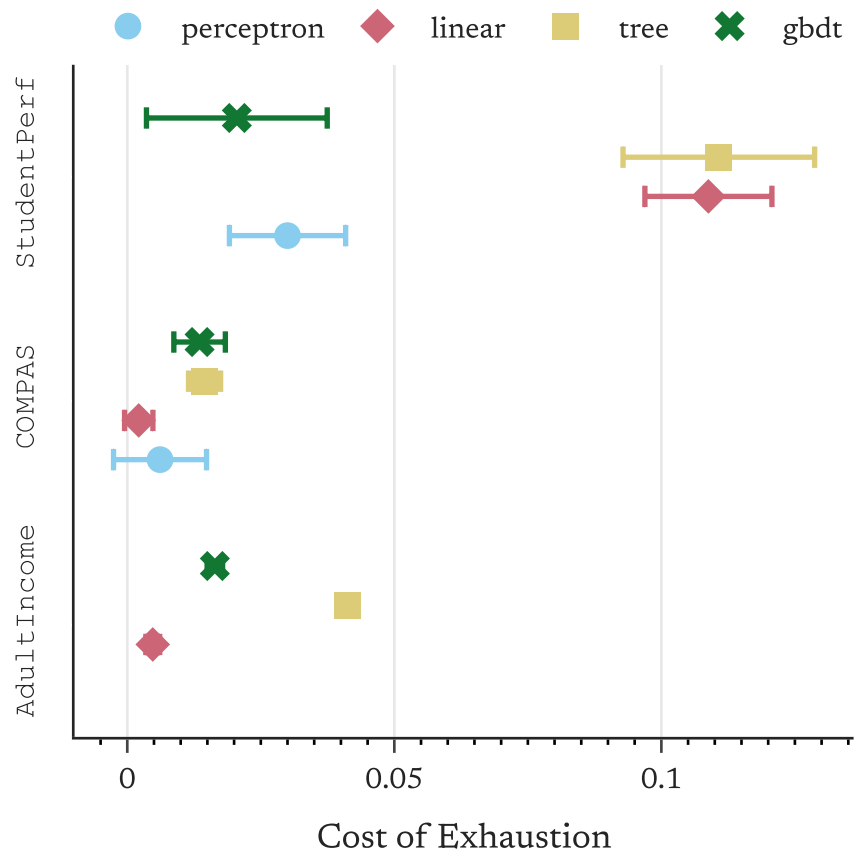
Context

🔨 Framework

🔎 A theoretical peek

📊 **Empirical study**

Concluding remarks



Conclusion

It seems [...] a platform could always *game the system* [...] *without sacrificing a lot of accuracy* of the model learnt.

– Anonymous reviewer

Paper link

Conclusions

- Robustness needs a prior
- Prior on the model
  ⇒ guarantees depend on the capacity of the model

Implications for AI regulation

- Need more access to the model
- And/Or anonymous auditor

Thanks for your attention !

Announcing a new team in SaTML town:

ARTISHAU

ARTificial Intelligence: Security, TrutHfulness and AUdit

Centre Inria de l'Université de Rennes, France

Bibliography

[1] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," *Reuters*, Oct. 2018, Accessed: Mar. 06, 2023. [Online]. 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 *Proceedings of the 25th International Conference on World Wide Web*, 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. [Online]. 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] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019.

[7] Arvind Narayanan, "Tutorial: 21 Fairness Definitions and Their Politics." Accessed: Oct. 12, 2023. [Online]. Available: https://www.youtube.com/watch?v=jIXIuYdnyyk

[8] B. Rastegarpanah, K. Gummadi, and M. Crovella, "Auditing Black-Box Prediction Models for Data Minimization Compliance," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 20621–20632. Accessed: Nov. 02, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/ac6b3cce8c74b2e23688c3e45532e2a7-Abstract.html

[9] F. Lu *et al.*, "A General Framework for Auditing Differentially Private Machine Learning," presented at the Advances in Neural Information Processing Systems, Dec. 2022, pp. 4165–4176. Accessed: Aug. 16, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/1add3bbdbc20c403a383482a665eb5a4-Abstract-Conference.html

[10] J. Bandy, "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–34, Apr. 2021, doi: 10.1145/3449148.

[11] T. Yan and C. Zhang, "Active Fairness Auditing," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Jun. 2022.

[12] B. Chugg, S. Cortes-Gomez, B. Wilder, and A. Ramdas, "Auditing Fairness by Betting." 2023.

[13] C. Yadav, M. Moshkovitz, and K. Chaudhuri, "XAudit : A Theoretical Look at Auditing with Explanations." Accessed: Sep. 18, 2023. [Online]. Available: http://arxiv.org/abs/2206.04740

[14] A. Shahin Shamsabadi *et al.*, "Washing The Unwashable : On The (Im)possibility of Fairwashing Detection," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 14170–14182. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/5b84864ff8474fd742c66f219b2eaac1-Paper-Conference.pdf

[15] J. G. Bourrée, E. L. Merrer, G. Tredan, and B. Rottembourg, "On the relevance of APIs facing fairwashed audits," *arXiv preprint arXiv:2305.13883*, 2023.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding Deep Learning (Still) Requires Rethinking Generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, Feb. 2021, doi: 10.1145/3446776.