Augustin Godinot

—

PhD Defense

Feb. 10, 2026

Thinking Out of the (Black)-Box:

Tools for machine learning audits in the presence of *deceptive* model providers.

Université de Rennes

IRISA

GOUVERNEMENT
Liberté
Égalité
Fraternité

PEReN
Pôle d'Expertise de la Régulation Numérique

Inria

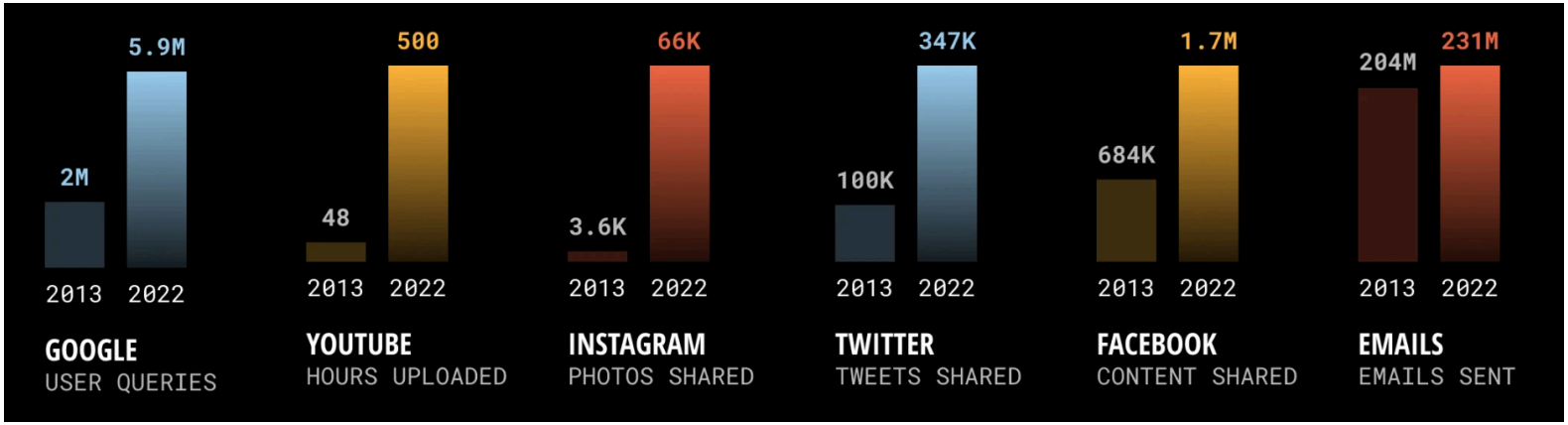
Jury

Présidente	Mme Aline **Roumy**	*INRIA*	Directrice de Recherche
Rapporteur	M. Damien **Garreau**	*University of Würzburg*	Professor
Rapporteur	M. Aurélien **Bellet**	*INRIA*	Directeur de Recherche
Examinateur	M. Nicolas **Papernot**	*University of Toronto*	Assistant Professor
Directeur de thèse	M. Gilles **Trédan**	*LAAS-CNRS*	Directeur de Recherche
Directeur de thèse	M. Erwan **Le Merrer**	*INRIA*	Chercheur

Invités

Encadrant	M. François **Taïani**	*Université de Rennes, IRISA*	Professeur
Encadrante	Mme. Gohar **Dashyan**	*PEReN*	Lead Recherche et Éthique

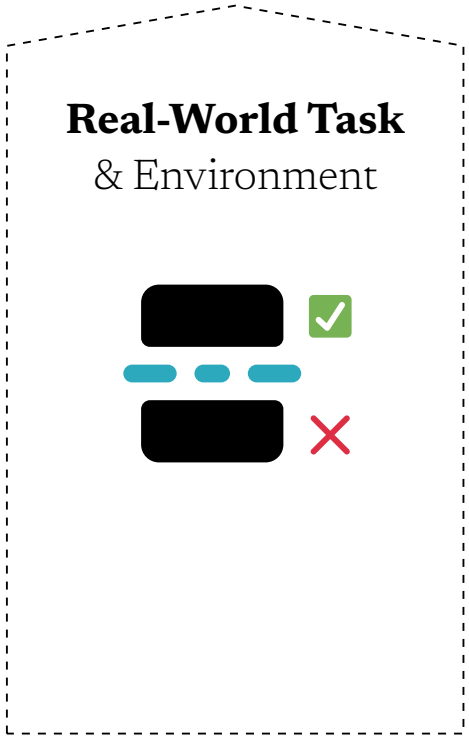
Data never sleeps



- explosion of data produced, collected and stored
- good algorithm to exploit this data for commercial applications: **machine learning**

per minute in 2021, https://explodingtopics.com/blog/data-generated-per-day

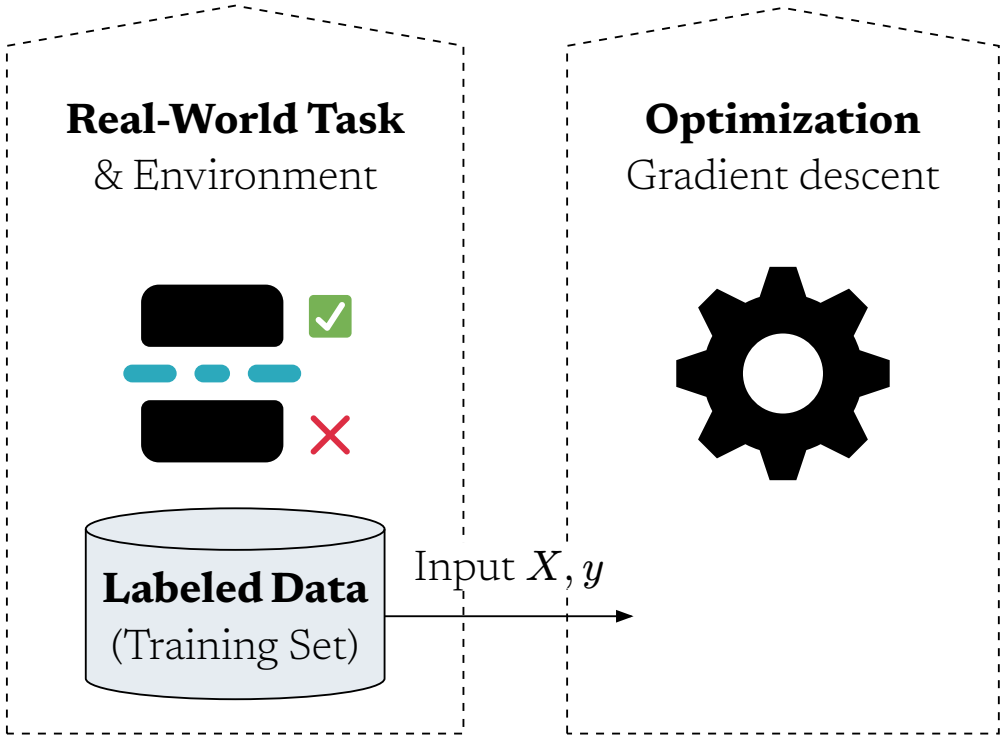
ML systems Data in, Prediction out



Example: Moderation on 𝕏

Data: Tweets with hate annotation

ML systems Data in, Prediction out

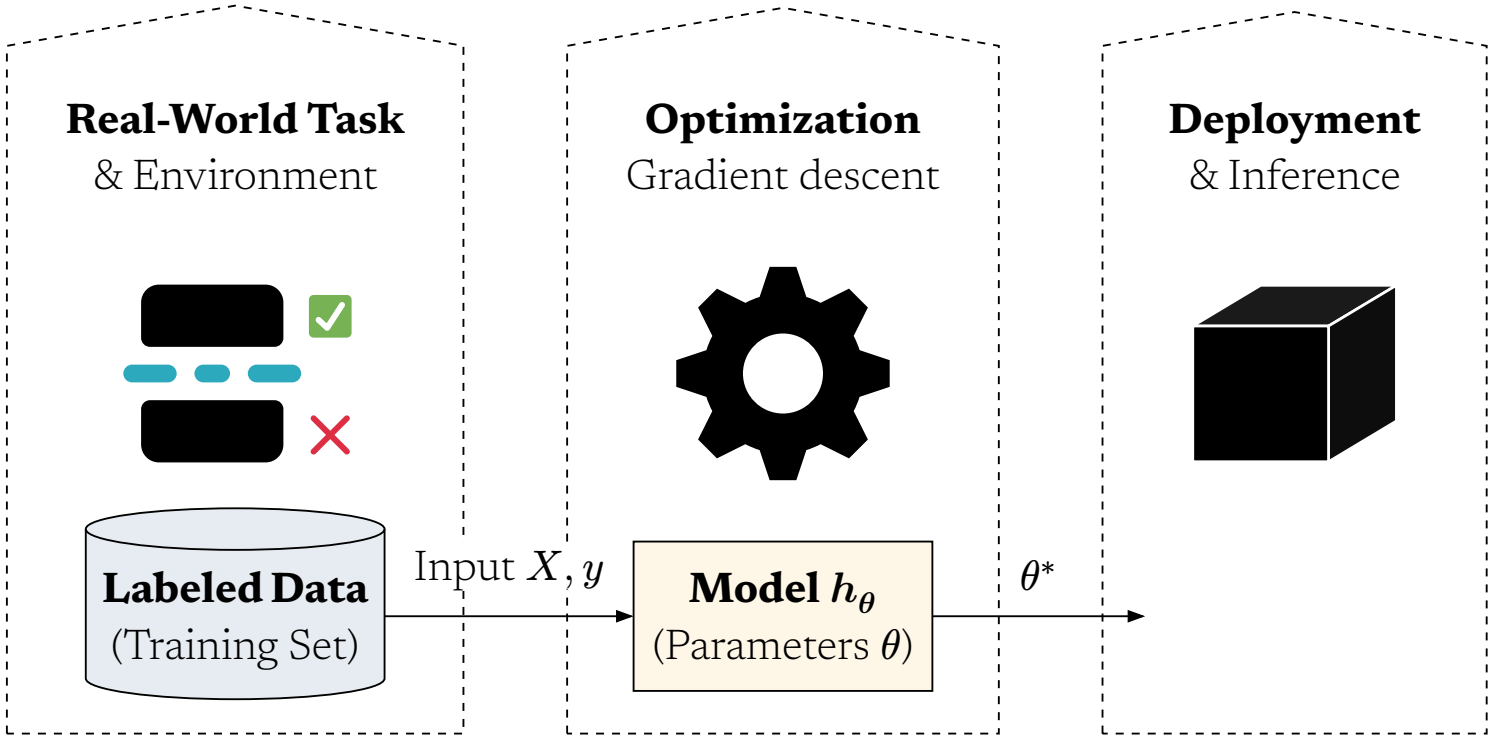


Example: Moderation on 𝕏

Data: Tweets with hate annotation

Model: Transformer & gradient descent

ML systems Data in, Prediction out



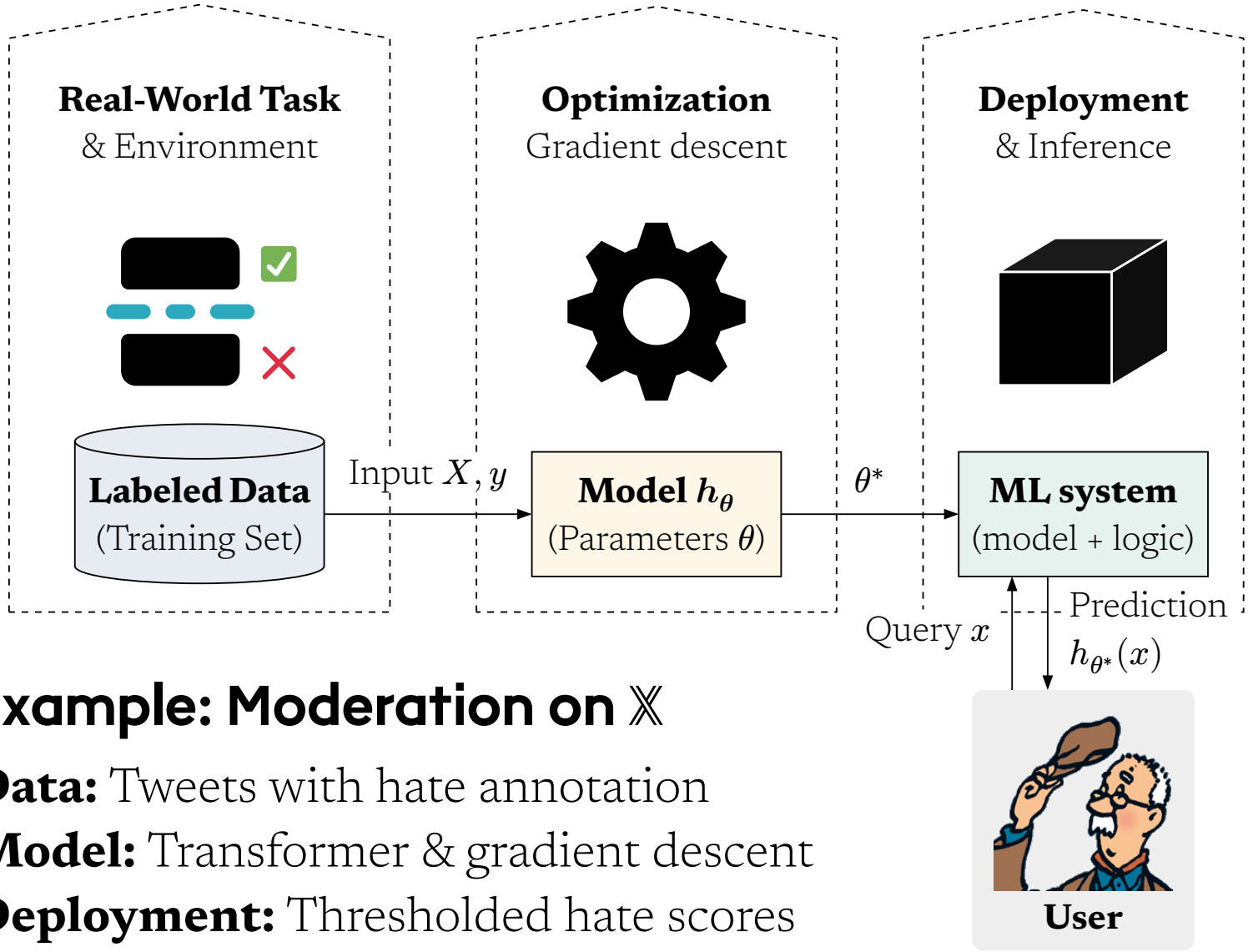
Example: Moderation on 𝕏

Data: Tweets with hate annotation

Model: Transformer & gradient descent

Deployment: Thresholded hate scores

ML systems Data in, Prediction out



Example: Moderation on 𝕏

Data: Tweets with hate annotation

Model: Transformer & gradient descent

Deployment: Thresholded hate scores

Inside the Suspicion Machine

Obscure government algorithms are making life-changing decisions about millions of people around the world. Here, for the first time, we reveal how one of these systems works.

EVA CONSTANTARAS, GABRIEL GEIGER, JUSTIN-CASIMIR BRAUN, DHRUV MEHROTRA, HTET AUNG

MAR 6, 2023 7:00 AM

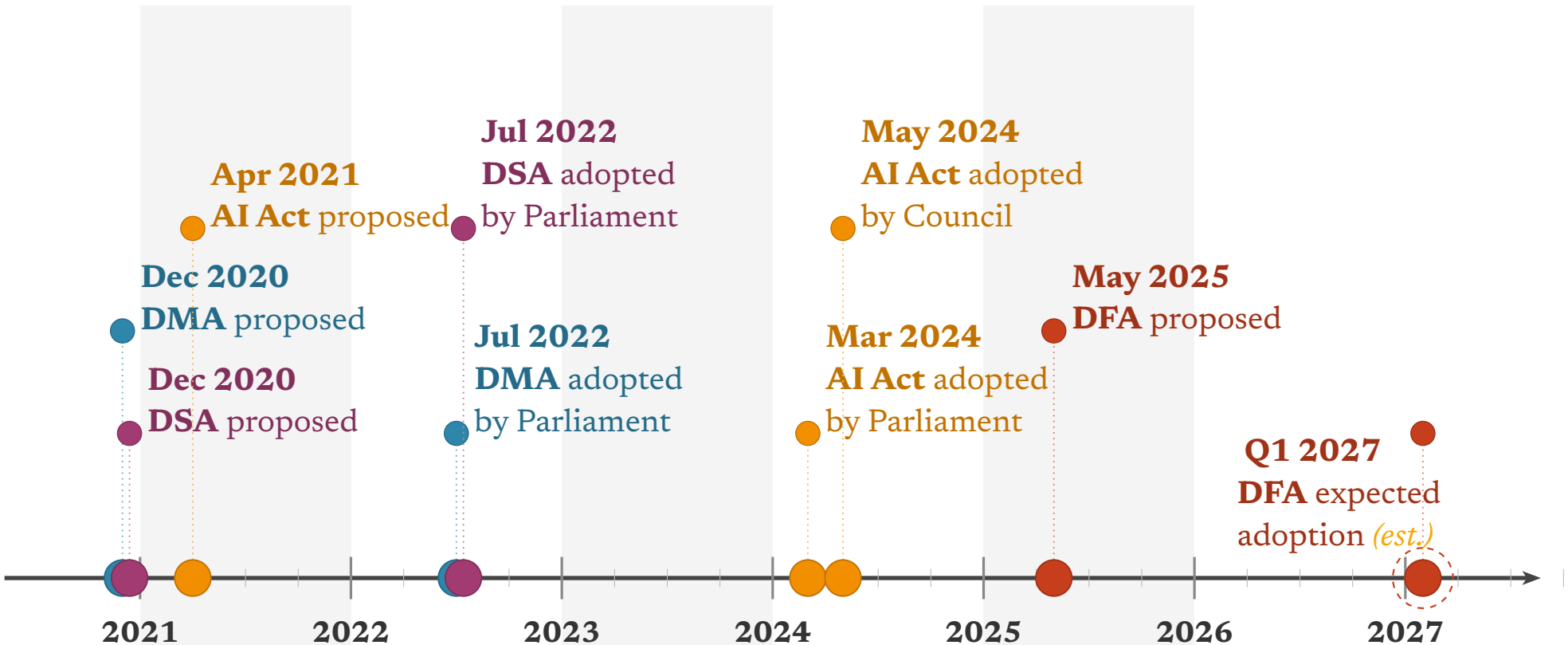
Main EU Regulation on ML systems

DMA
Digital Markets Act

DSA
Digital Services Act

AI Act
Artificial Intelligence Act

DFA
Digital Fairness Act



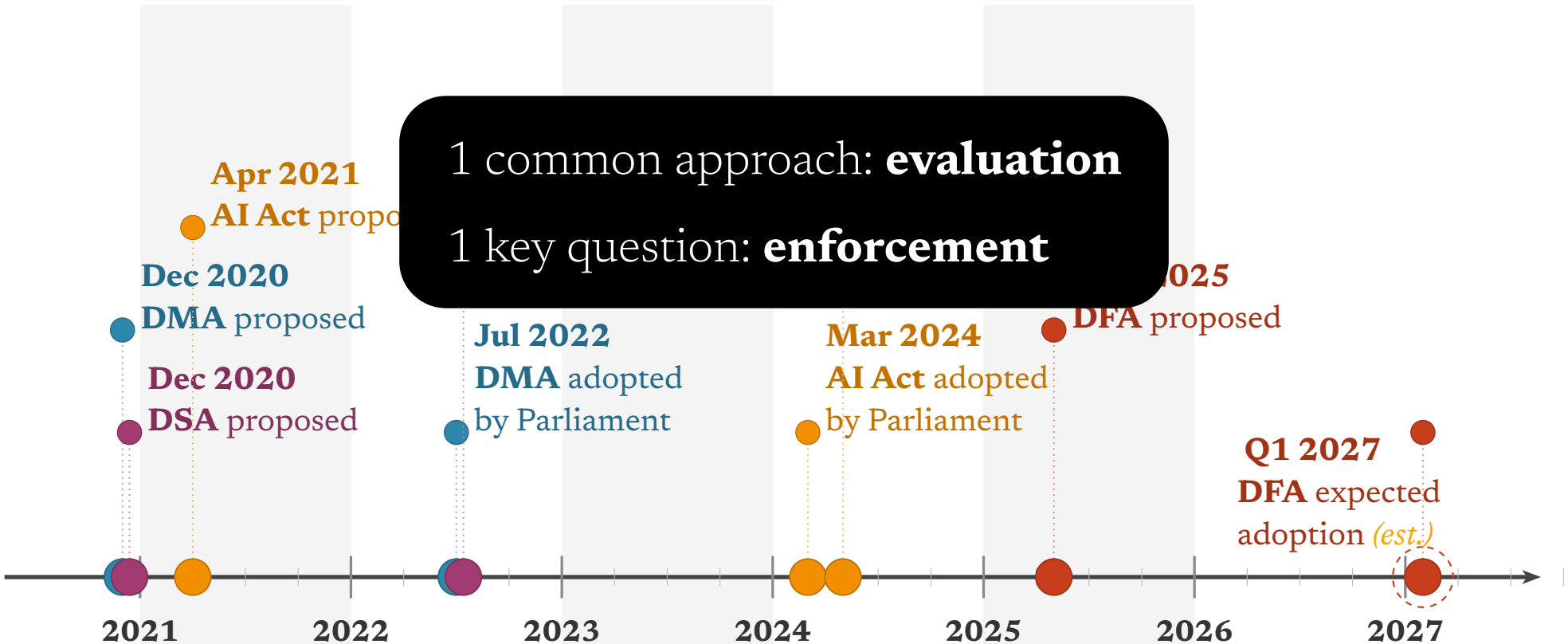
Main EU Regulation on ML systems

DMA
Digital Markets Act

DSA
Digital Services Act

AI Act
Artificial Intelligence Act

DFA
Digital Fairness Act



Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

ML audits

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

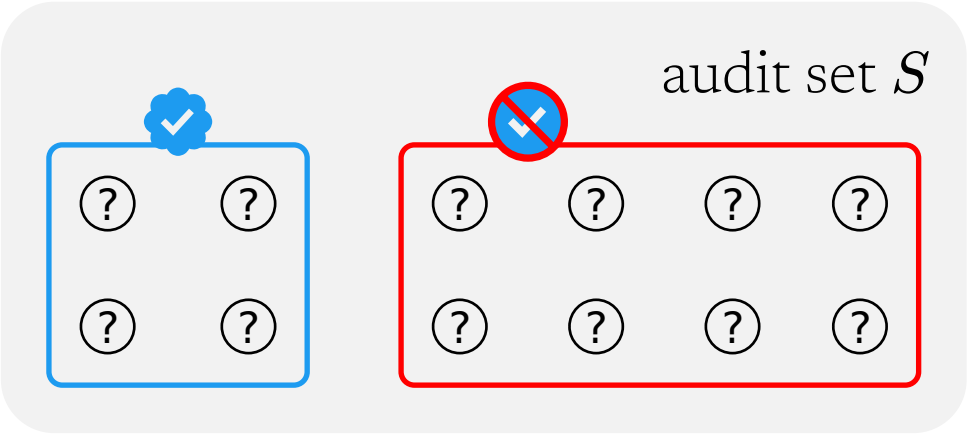
ML audits

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

Demographic parity

$$\mu(h, S) = \mathbb{P}\left(\text{#\%!\&} \;\middle|\; X \in S, \text{✓} \right) - \mathbb{P}\left(\text{#\%!\&} \;\middle|\; X \in S, \text{⊘} \right)$$



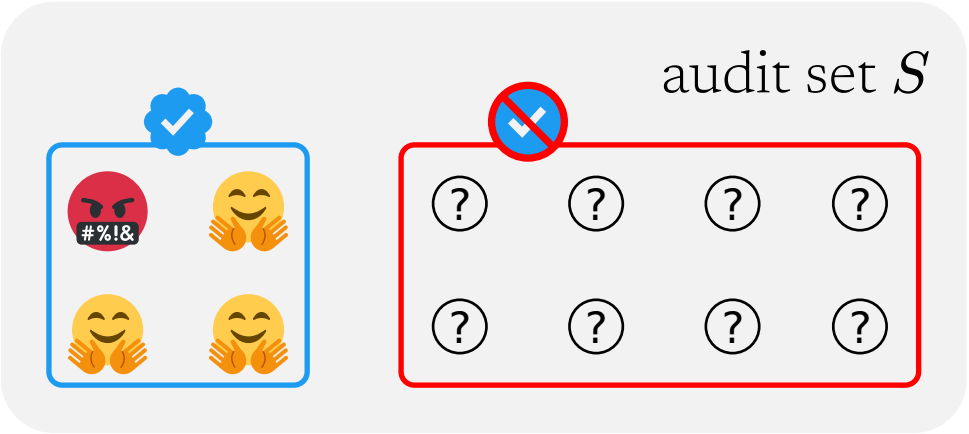
Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

Demographic parity

$$\mu(h, S) = \mathbb{P}\left(\text{😡} \mid X \in S, \text{✅}\right) - \mathbb{P}\left(\text{😡} \mid X \in S, \text{🚫}\right)$$

$$= \frac{1}{4}$$



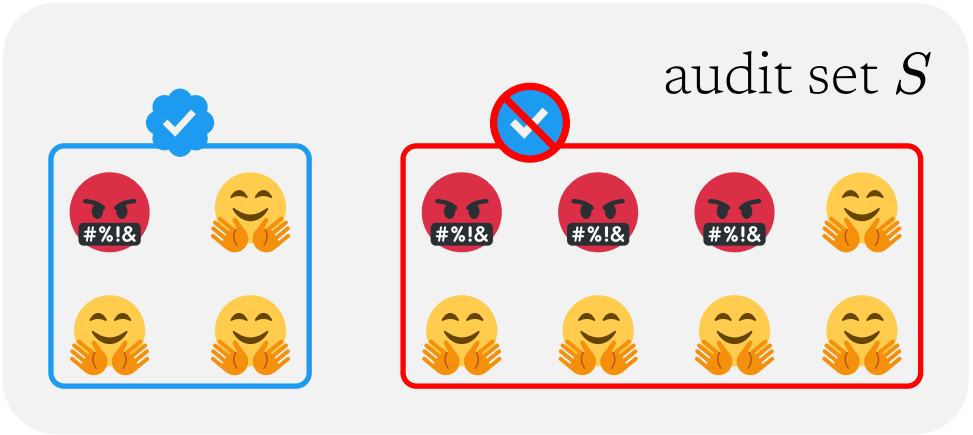
Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

Demographic parity

$$\mu(h, S) = \mathbb{P}\left(\text{#\%!\&} \,\middle|\, X \in S, \text{✓}\right) - \mathbb{P}\left(\text{#\%!\&} \,\middle|\, X \in S, \text{⊘}\right)$$

$$= \frac{1}{4} - \frac{3}{8} =$$



[2] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI Auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024. 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

ML audits

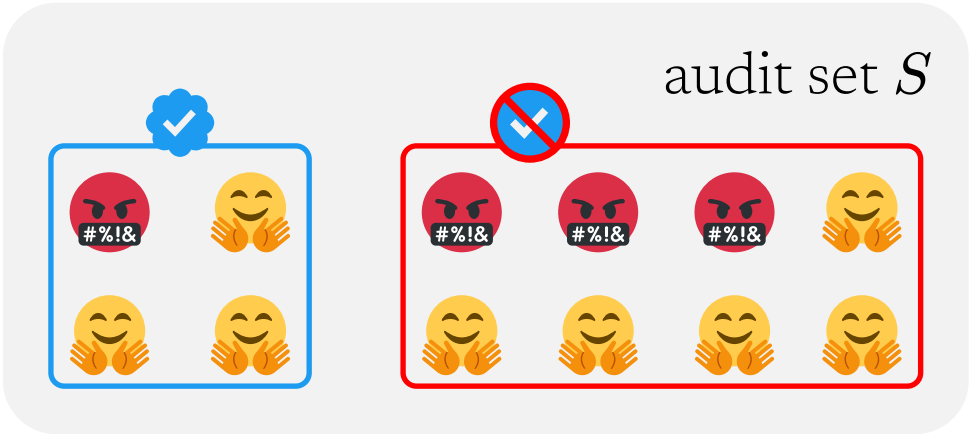
Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Audit target performance, disparity, robustness, privacy, ...

Demographic parity

$$\mu(h, S) = \mathbb{P}\left(\text{#\%!\&} \,\middle|\, X \in S, \text{✓}\right) - \mathbb{P}\left(\text{#\%!\&} \,\middle|\, X \in S, \text{⊘}\right)$$

$$= \frac{1}{4} - \frac{3}{8} = -12.5\%$$



[2] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI Auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024. 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

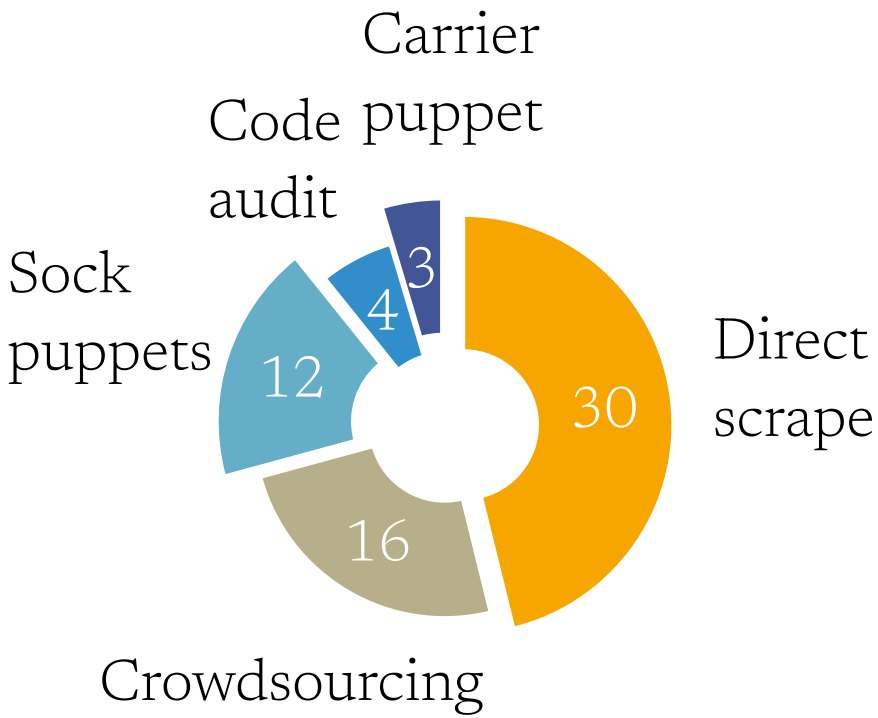
Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Evaluation limited access to model, data, documentation

ML audits

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Evaluation limited access to model, data, documentation

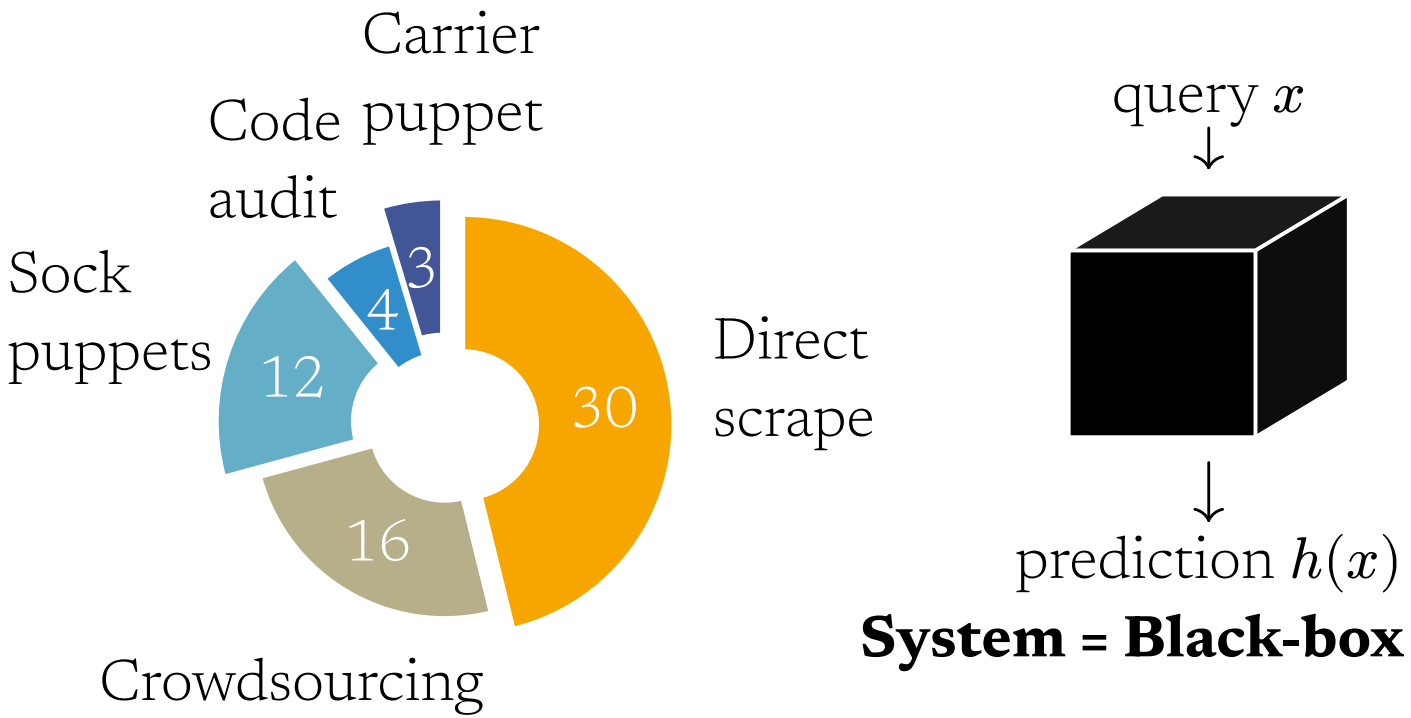


[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

ML audits

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

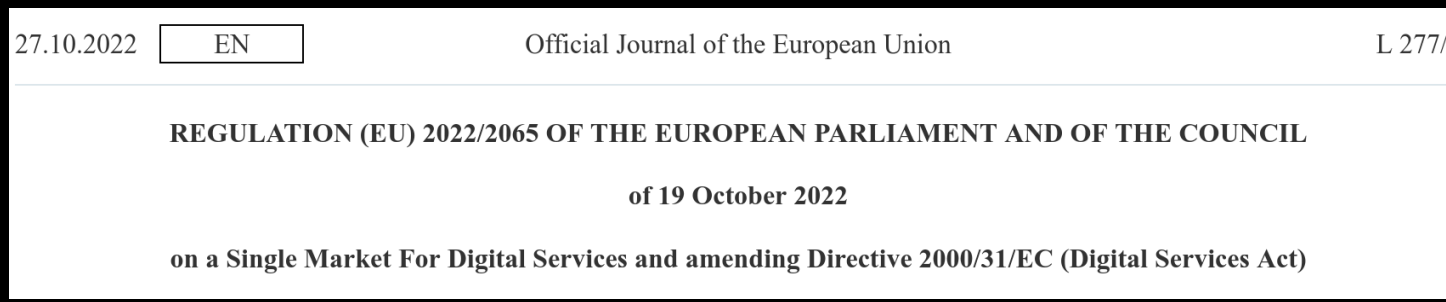
Evaluation limited access to model, data, documentation



[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Power/**data asymmetry** + **Regulation**



27.10.2022 EN Official Journal of the European Union L 277/1

REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 19 October 2022

on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

prediction $h(x)$

System = Black-box

Crowdsourcing

[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

ML audits

Definition 1.2 (ML Audit) *Any* **independent** *assessment of an identified* **audit target** *via an* **evaluation** *of articulated expectations with the implicit or explicit* **objective of accountability**.

Power/**data asymmetry** + **Regulation**

⇒ ***rationalization risk***

27.10.2022 EN Official Journal of the European Union L 277/1

REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 19 October 2022

on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

prediction $h(x)$

System = Black-box

Crowdsourcing

[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

Evading black-box audits

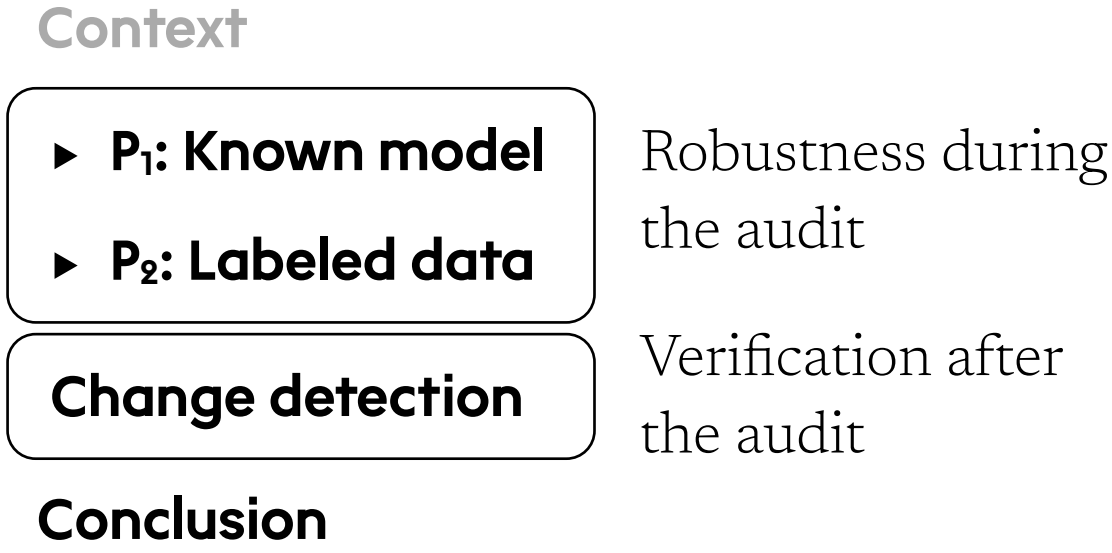
Auditor

Platform

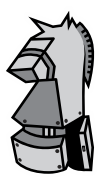
User

The M$ question

To go beyond the black-box model, what minimal additional information should the auditor request to achieve meaningful audit validity guarantees?



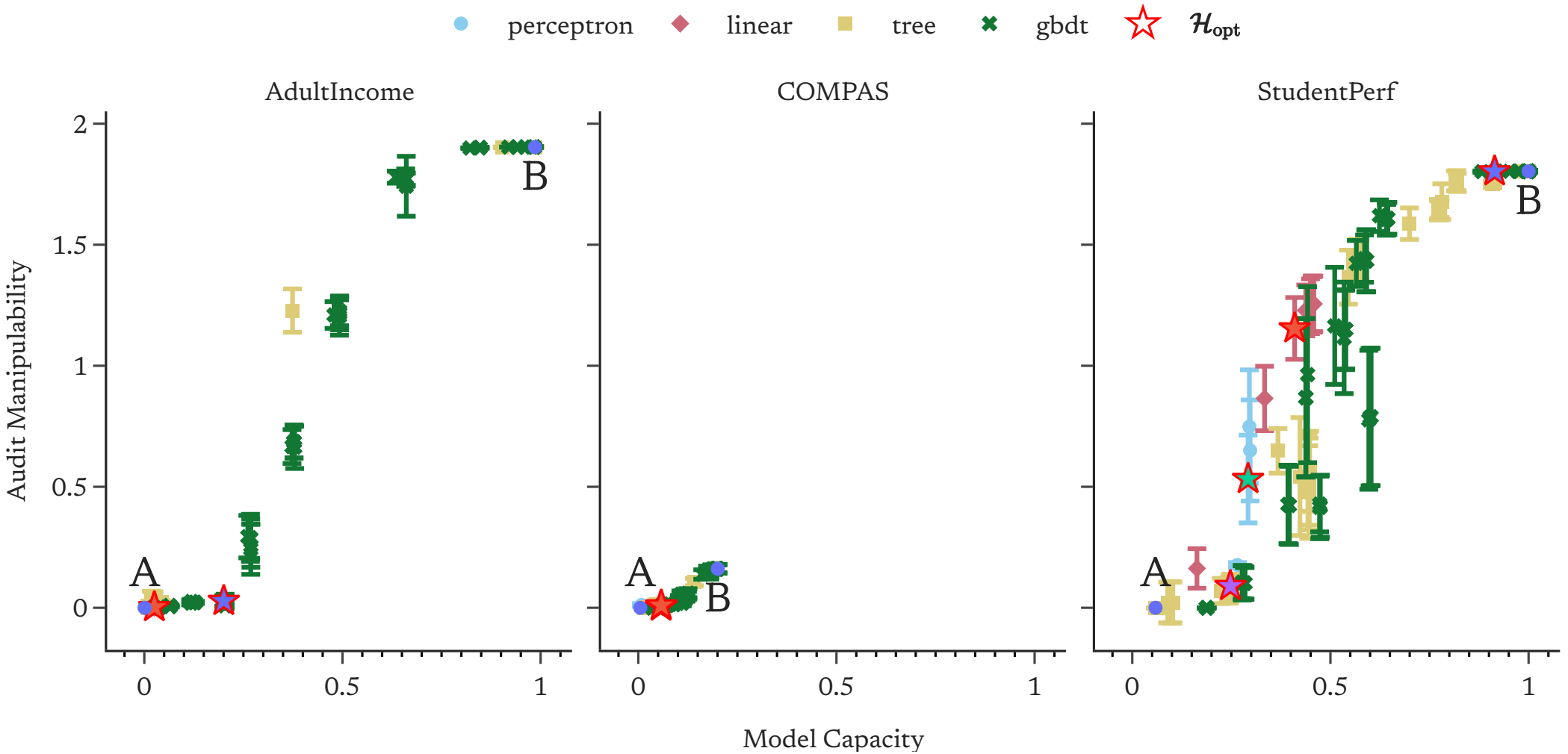
P_1: Known model

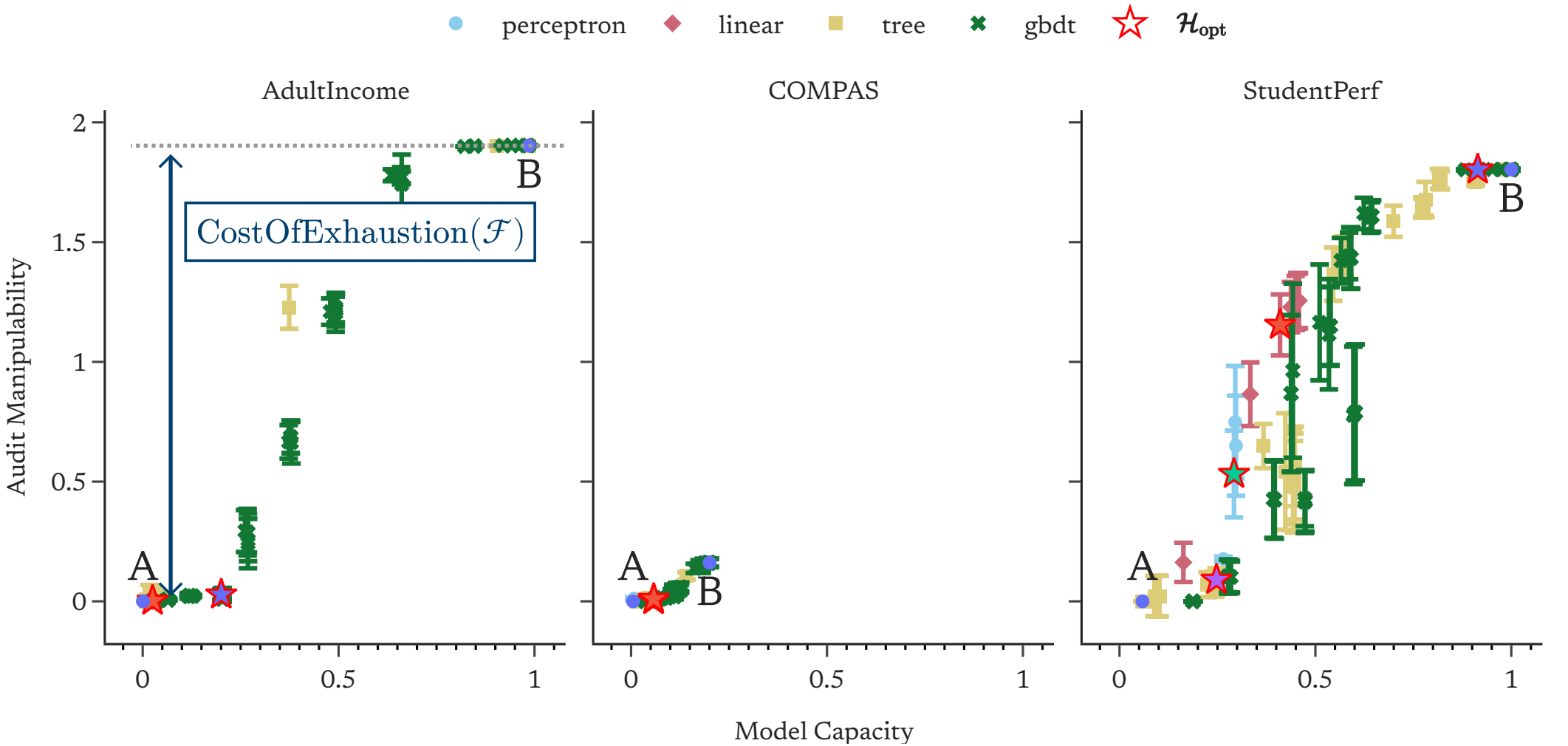


SaTML 2024

Under manipulations, are some AI models harder to audit?

Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani


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



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

Model capacity is key

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

Conclusion

- Easy to overfit the audit set with low utility cost
- Active auditing is not robust despite computational cost
- Regulators need to have more information than just the hypothesis class

P_2: Labeled data

ICML 2025

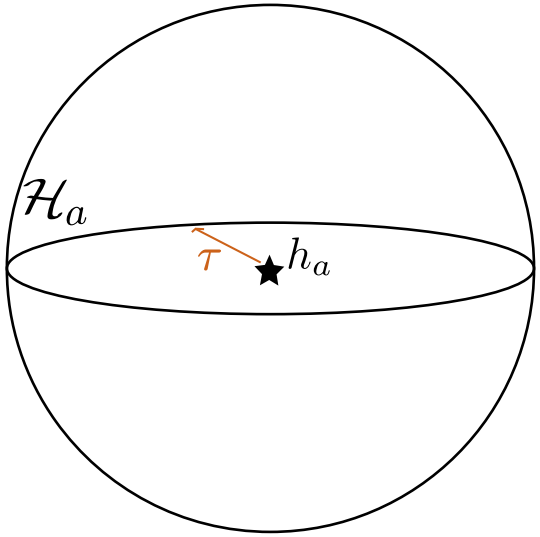
Robust ML Auditing using Prior Knowledge
Jade Garcia Bourrée, Augustin Godinot*, Sayan Biswas, Anne-Marie Kermarrec, Erwan Le Merrer, Gilles Tredan, Martijn de Vos, Milos Vujasinovic* (Spotlight poster)

Data prior

Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$



Data prior

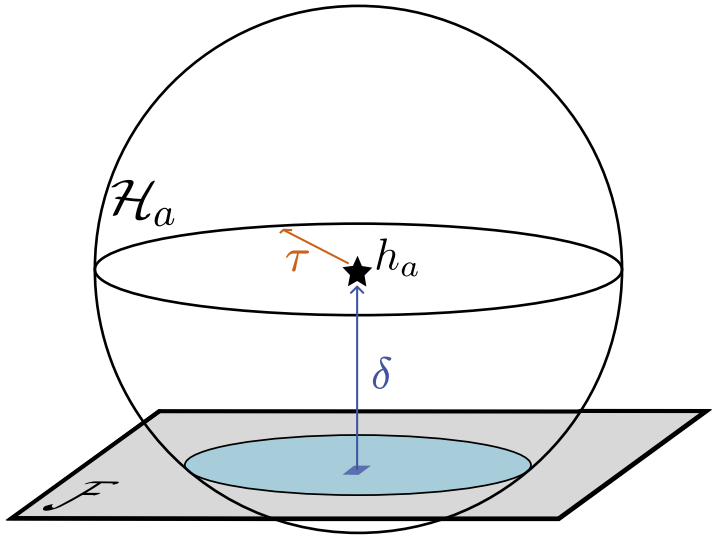
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{ h \in y^{|D_a|} \mid \mu(h, D_a) = 0 \right\}$$



Data prior

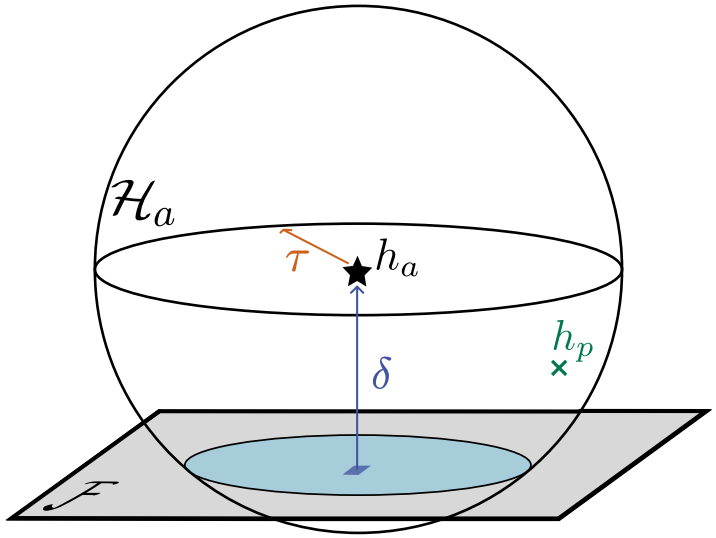
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{ h \in y^{|D_a|} \mid \mu(h, D_a) = 0 \right\}$$



Data prior

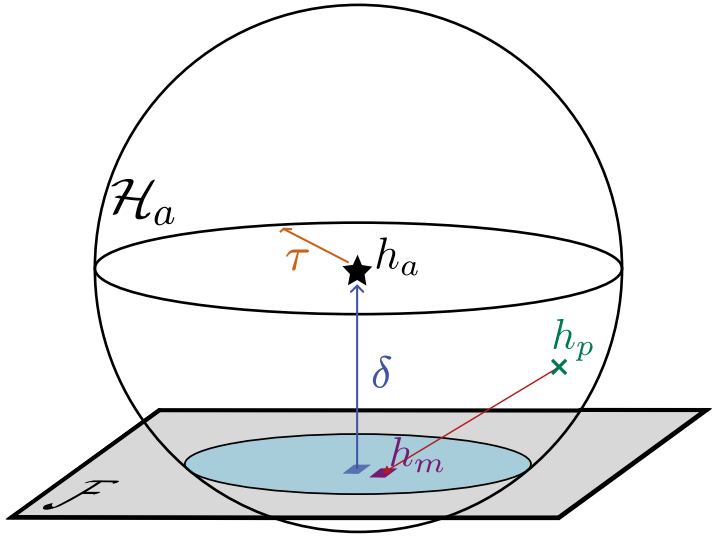
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{ h \in y^{|D_a|} \mid \mu(h, D_a) = 0 \right\}$$



Data prior

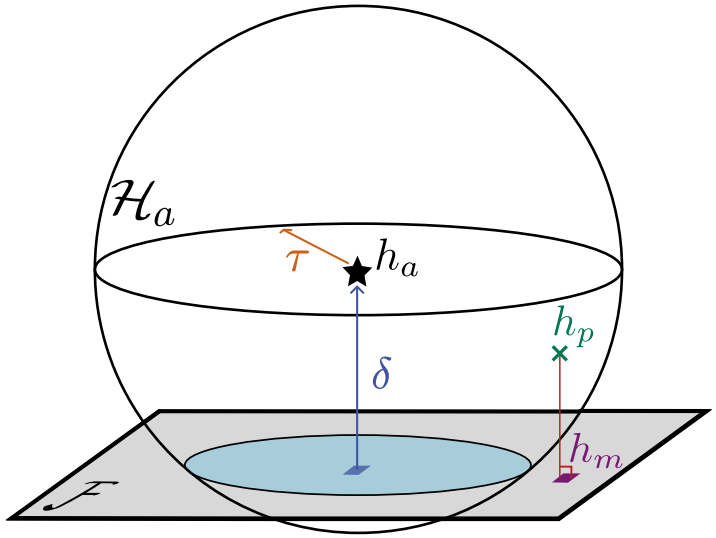
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{ h \in y^{|D_a|} \mid \mu(h, D_a) = 0 \right\}$$



Data prior

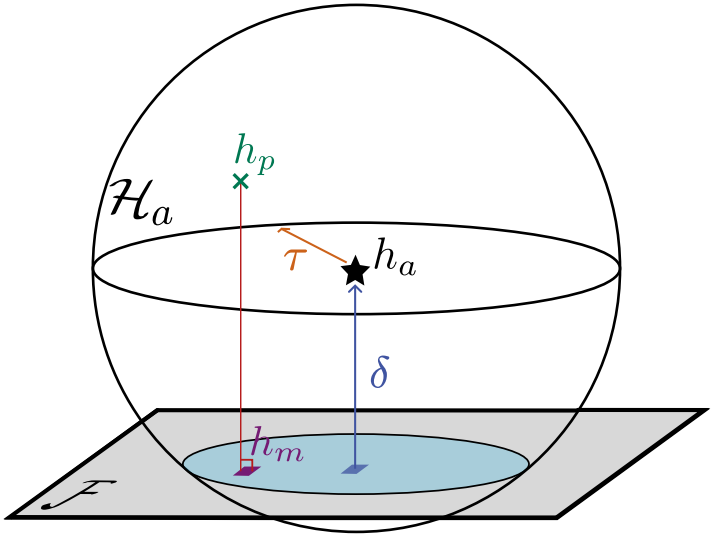
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{ h \in y^{|D_a|} \mid L(h, D_a) < \tau \right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{ h \in y^{|D_a|} \mid \mu(h, D_a) = 0 \right\}$$



Data prior

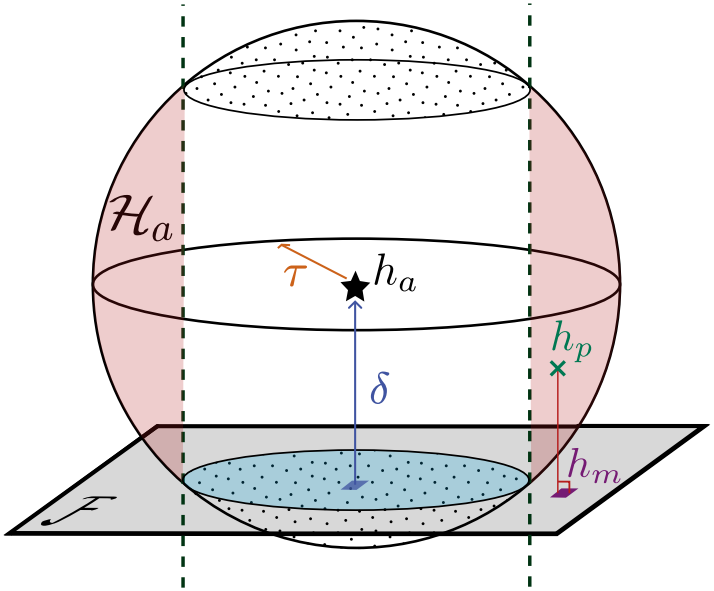
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{h \in y^{|D_a|} \mid L(h, D_a) < \tau\right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{h \in y^{|D_a|} \mid \mu(h, D_a) = 0\right\}$$



Data prior

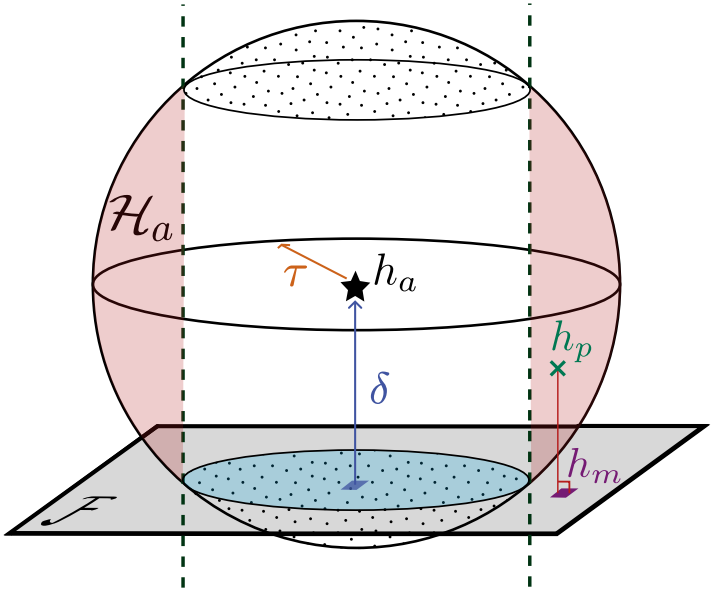
Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \left\{h \in y^{|D_a|} \mid L(h, D_a) < \tau\right\}$$

Set of fair labelings

$$\mathcal{F} = \left\{h \in y^{|D_a|} \mid \mu(h, D_a) = 0\right\}$$



Interesting cases

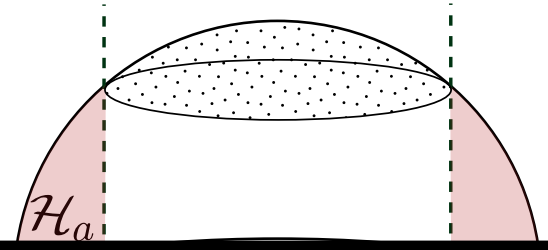
- Fair audit set ($\delta = \mu(h_a, D_a) = 0$) $\Rightarrow$ *impossible detection*
- Perfect threshold ($\tau = \delta$) $\Rightarrow$ *always detected*
- In between: exact formulation (Thm. 2.2), bounds (Cor 2.3)

Data prior

Auditors have **labeled data**, how to take it into account?

Definition 2.3 (Dataset prior)

$$\mathcal{H}_a = \{h \in \mathcal{Y}^{|D_a|} \mid L(h, D_a) < \tau\}$$



The good regulator theorem [3]

"Every good regulator of a system must be a model of that system"

Interesting cases

- Fair audit set ($\delta = \mu(h_a, D_a) = 0$) $\Rightarrow$ *impossible detection*
- Perfect threshold ($\tau = \delta$) $\Rightarrow$ *always detected*
- In between: exact formulation (Thm. 2.2), bounds (Cor 2.3)

Change detection

Queries, Representation & Detection: the next 100 model fingerprinting schemes

Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani

Audit-then-verify

During the audit

- Code access
- Data access
- Model access
- Weights access

After the audit

- Query access

Existing methods

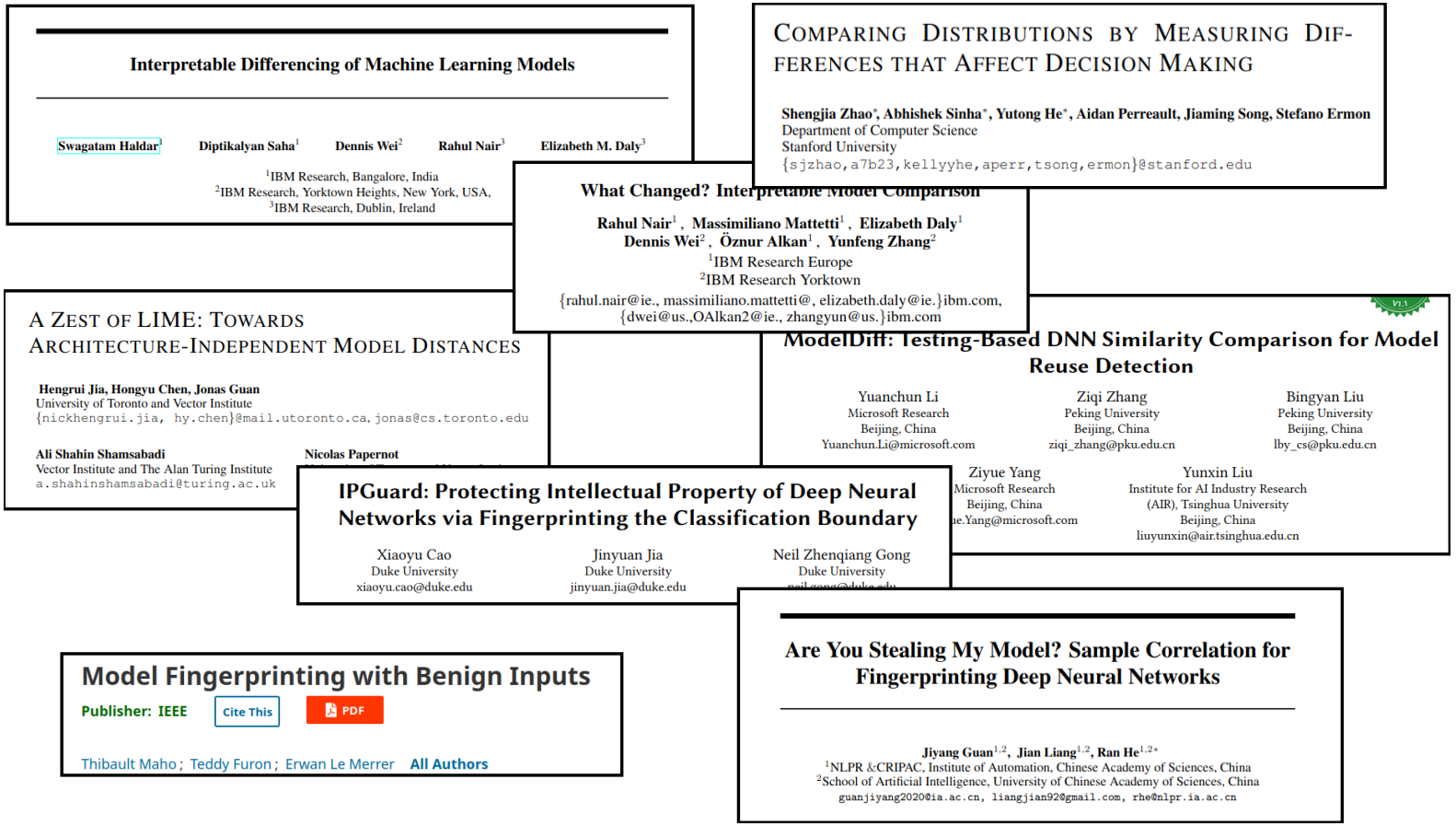
Context

P_1: Known model

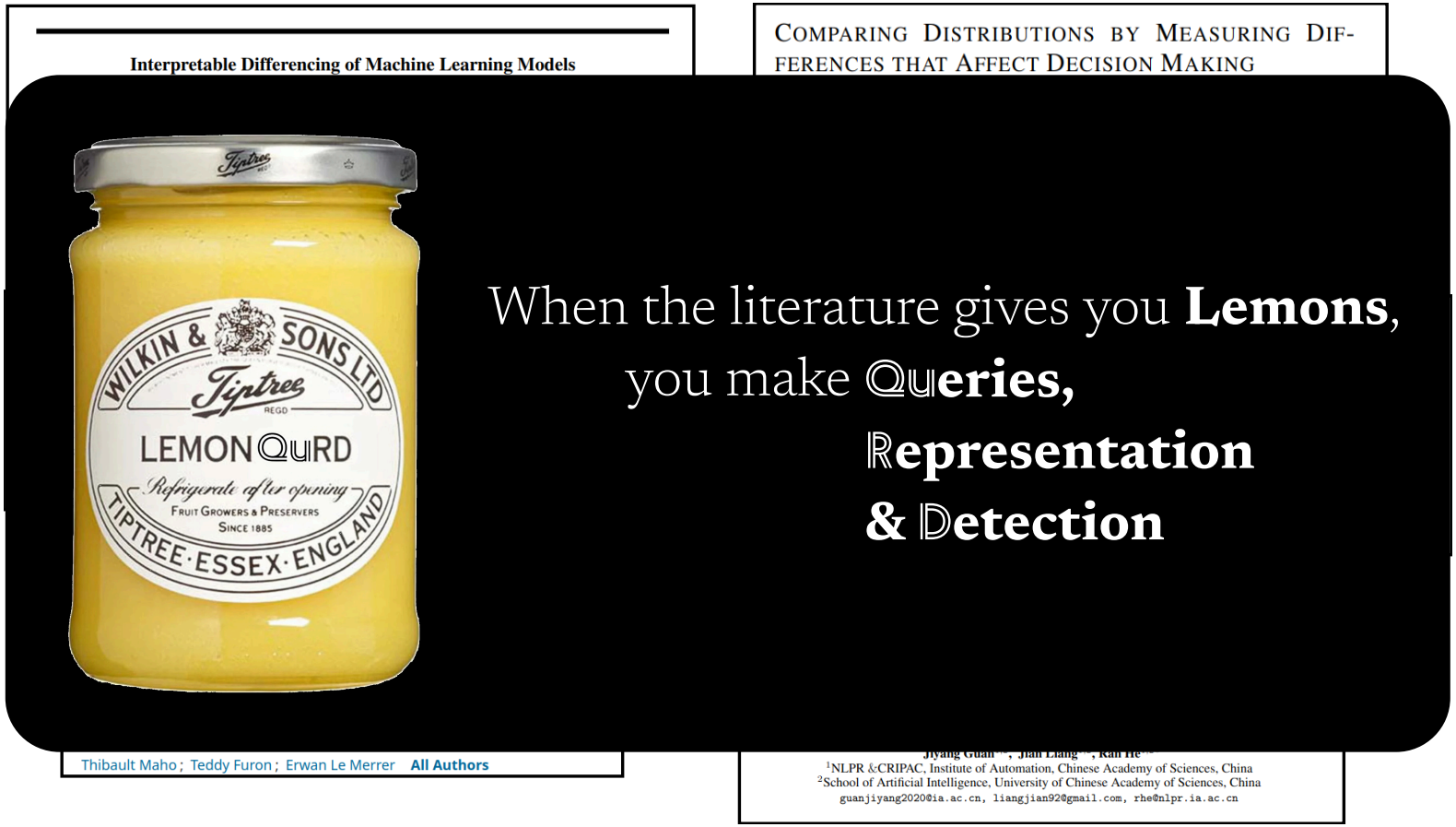
P_2: Labeled data

Change detection

Conclusion



Existing methods



A model fingerprint baseline

Requires Sampling access to $\mathcal{D}$, white-box h, black-box h'

Draw $x \sim \overline{\mathcal{D}}_h$ (distribution on $\mathcal{X}$ such that $h(x) \neq c(x)$)

If $h(x) = h'(x)$

  Return 1 (`Stolen`)

Else Return 0 (`Benign`)

Figure 4.2: The proposed baseline, AKH.

$$\mathcal{H}_0 : h \neq h' \qquad \mathcal{H}_1 : h = h'$$

ℚueries

How to select
$S = (x_1, ..., x_s) \subset \mathcal{X}$?

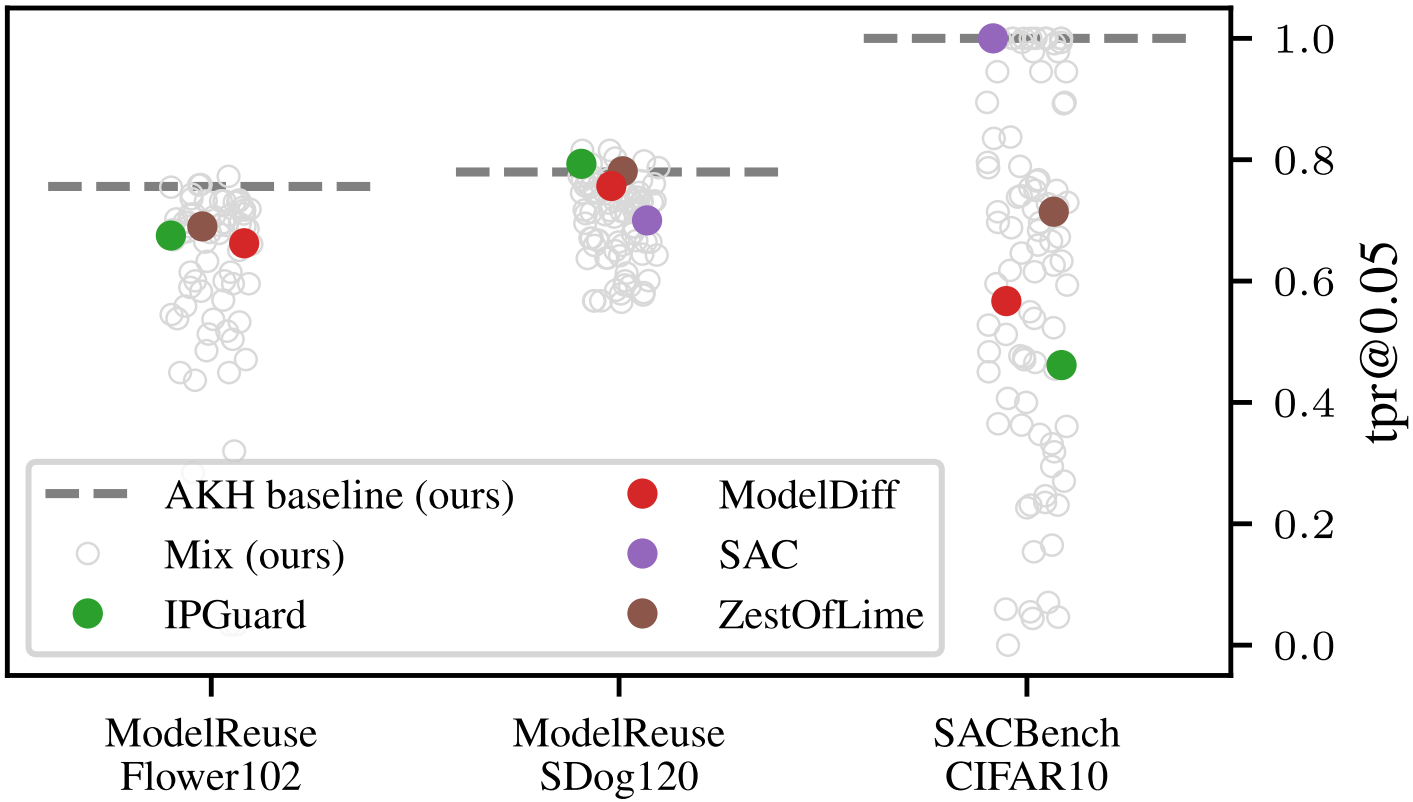
ℝepresentation

How to "compress" $Y = (h(x_1), ..., h(x_s))$ into representation Z?

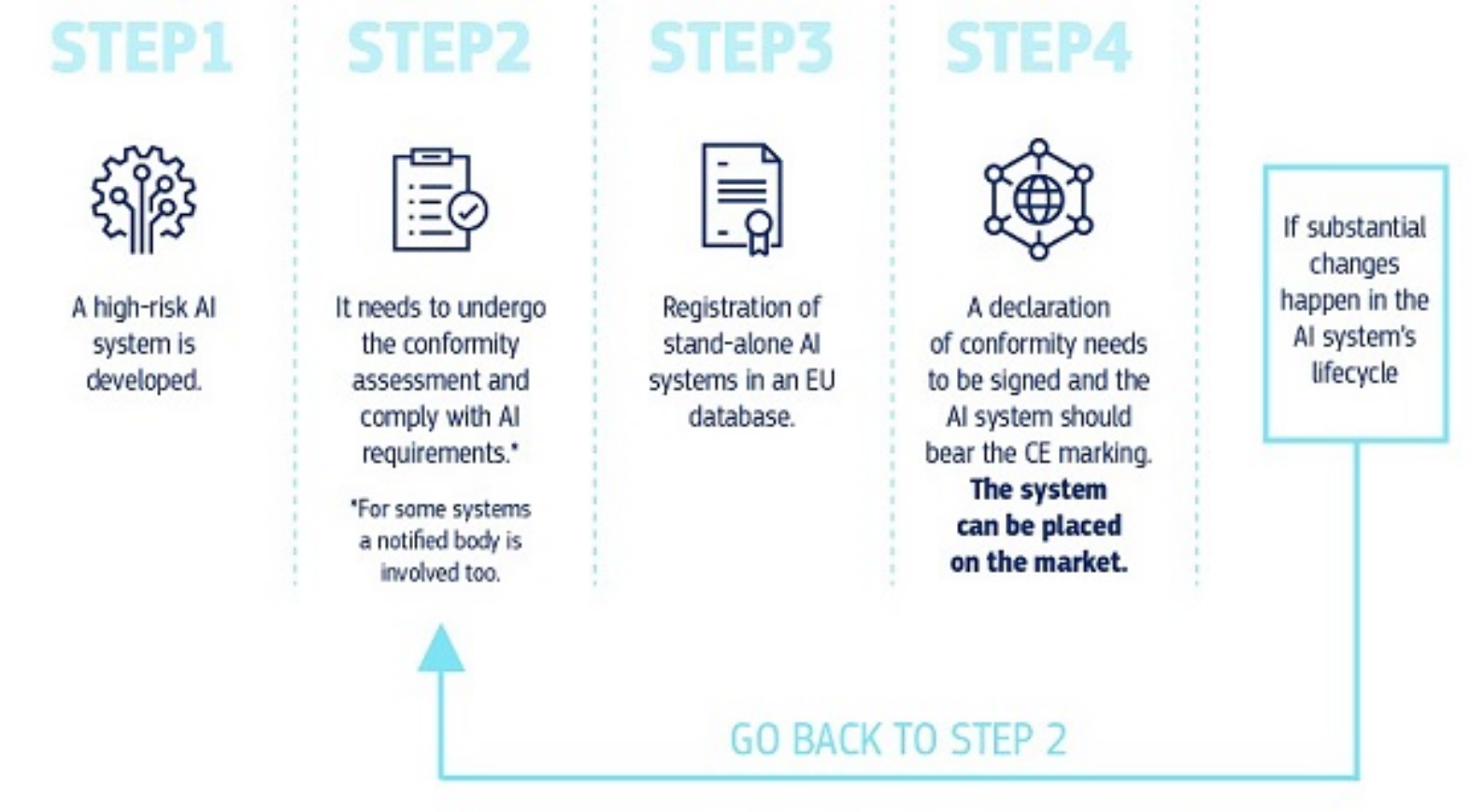
𝔻etection

How to make the decision $h = h'$ vs. $h \neq h'$ based on $d(Z, Z')$?

LemonQuRD
A model fingerprint benchmark



Implications for re-audits



https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Conclusion

Audits are a matter of

1. Comparison to priors
2. Information gain

Contributions

Conferences


SaTML 2024

Under manipulations, are some AI models harder to audit?
Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani


AAAI 2025

Queries, Representation & Detection: the next 100 model fingerprinting schemes
Augustin Godinot, Gilles Tredan, Erwan Le Merrer, Camilla Penzo, Francois Taiani


ICML 2025

Robust ML Auditing using Prior Knowledge
Jade Garcia Bourrée, Augustin Godinot*, Sayan Biswas, Anne-Marie Kermarrec, Erwan Le Merrer, Gilles Tredan, Martijn de Vos, Milos Vujasinovic* (Spotlight poster)

Preprints


Soon

Manipulation-Proof Oblivious Audits against Deceptive Model Providers
Augustin Godinot, Sofiane Azogagh, Julien Ferry, Sébastien Gambs

Δ-`audits`: adaptive and manipulation-proof performance monitoring
Augustin Godinot, Mohammad Yaghini Nicolas Papernot

And also

Open Source

- https://github.com/grodino/QuRD
- https://github.com/huggingface/pytorch-image-models

Reading group

- https://gitlab.inria.fr/WIDE/auditia/reading-group

Teaching

- Convex & proximal optimization labs

Perspectives
Post audit verification

- Sensitivity of audit metric / platform utility
- Links with performance prediction

Perspectives Post audit verification

Tools beyond evaluation

- Pre-audits: incident databases
- During the audit: distributed user auditing
- Links with system specification

Auditing trade-offs

- Beyond estimation of the impact, reveal the choices
- Scaling laws against data/compute asymetry
- Prediction uncertainty and fallback mechanisms

Appendix

Bibliography

[1] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 74:1–74:34. https://doi.org/10.1145/3449148

[2] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI Auditing: The Broken Bus on the Road to AI Accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024. 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

[3] Roger C. Conant and W. Ross Ashby. 1970. Every good regulator of a system must be a model of that system. *International journal of systems science* 1, 2 (1970), 89–97.