Large models are impossible to regulate. CHANGE MY MIND

Augustin Godinot (WIDE)
Université de Rennes, IRISA, INRIA, PEReN

🏛 What regulators ask for…

- Digital Services Act (**DSA**): Large platforms induce risks for society, they have to implement risk mitigation meechanisms.
- Digital Markets Act (**DMA**): Large platforms have a lot of power, we must avoid power imbalance.
- Artificial Intelligence Act (**AI Act**): limit the use of some algorithms.

🕵 ML audit you said ?

- **Input space** $\mathcal{X}$. *Example: The space of all possible* 1000×1000 *images.*
- **Hypothesis** $h : \mathcal{X} \to \{0,1\}$. *Example: a deep neural network.*
- **Hypothesis class** $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$. *Example: all the ResNet models with* 50 *blocks.*

Audit a parity metric

$$\mu(h,S) = \mathbb{P}(h(X)=1 \mid X \in S, E) - \mathbb{P}(h(X)=1 \mid X \in S, \overline{E})$$

Example: make sure that in average, men are not advantaged compared to women by a resume screening algorithm.

🤖 Large Machine Learning models

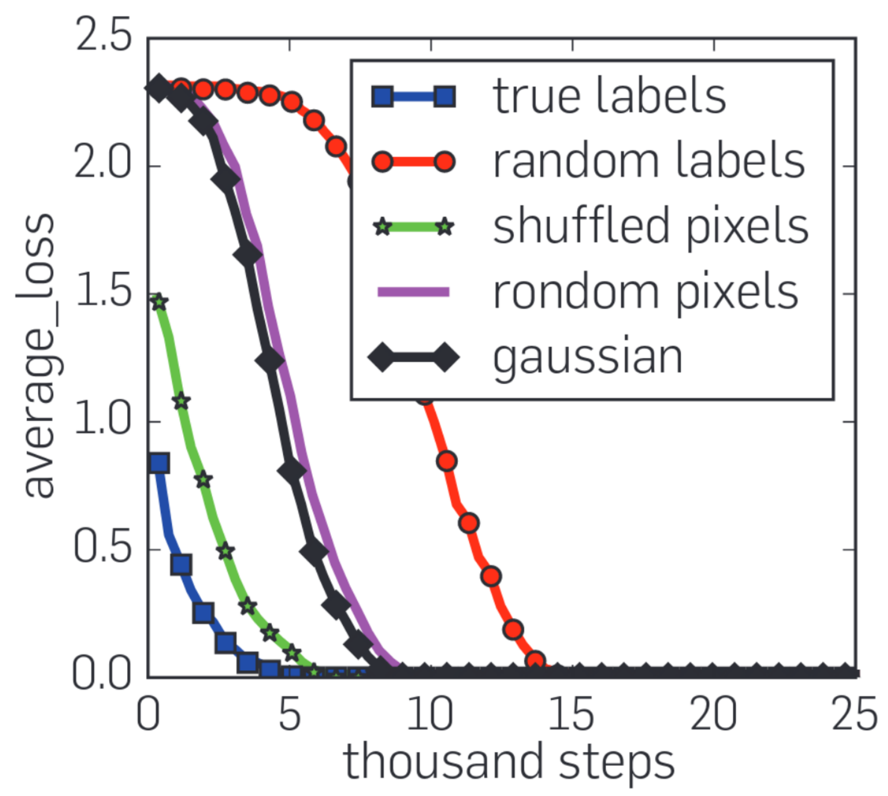
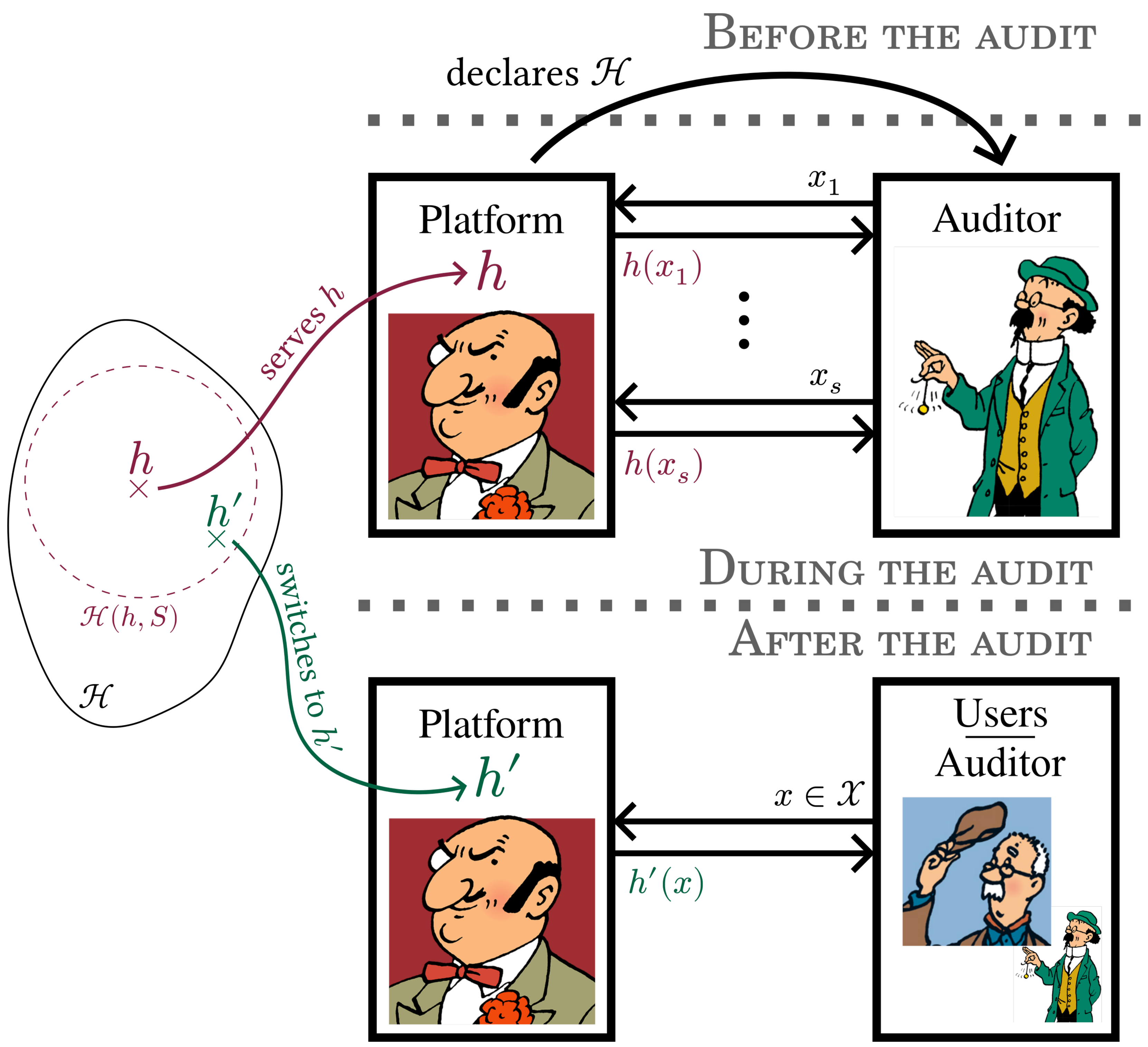


Figure 1: The training loss of an Inception model trained on CIFAR10. After enough steps, the loss reaches 0 *even when trained on random labels.*

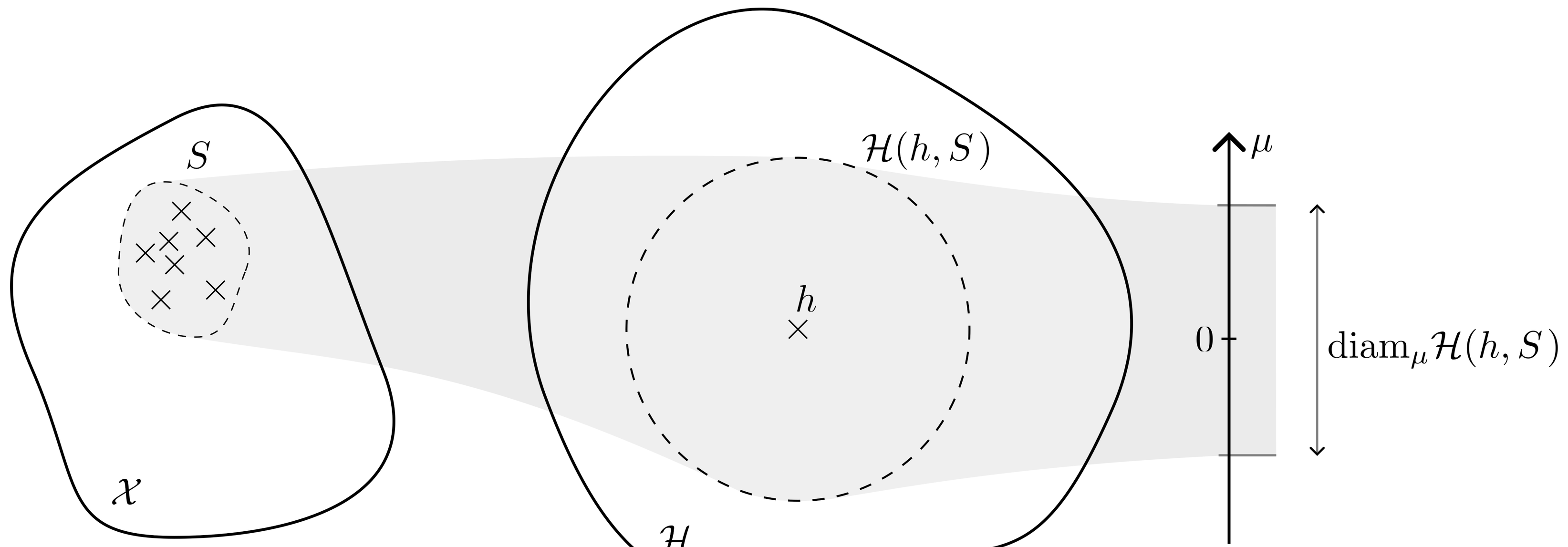
Taken from *Understanding deep learning requires rethinking generalization* (Zang et al, CACM 2021)

- Current ML models can reach **billions of parameters**.
- Current ML models can **overfit the train data** and have **good generalization** properties.
- Some explanation attempts: **benign overfitting** and **double descent**.

⚠ Threat model



📏 Mesuring the effect of potential manipulations



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

diam_μ [redacted] $= \max_{h' \in \text{[redacted]}} |\mu(h') - \mu(h)|$

μ-diameter — Version space

🔎 Impossibility theorem

Definition 2: Benign overfitting on c

$\mathcal{H}$ exhibits benign overfitting with respect to c iif fhere exists $d_0 \in \mathbb{N}_*$ and $\varepsilon \in [0,1)$ such that $\forall d \le d_0, S \in \mathcal{X}, \sigma \in \{0,1\}^d$,

$$\exists h \in \mathcal{H}, \begin{cases} \forall x_i \in S, h(x_i) = \sigma_i \text{ (fits any train set)} \\ \mathbb{P}\big(h(X) = c(X) \,\big|\, X \in \overline{S}\big) = 1 - \varepsilon \text{ (low error)} \end{cases}$$

Theorem: Better than random? No can do.

If $\mathcal{H}$ exhibits benign overfitting with respect to the sensitive attribute, then,

$$\forall S, |S| = |S_{\text{random}}|, \quad \mathrm{diam}_\mu(h,S) = \mathrm{diam}_\mu(h, S_{\text{random}})$$

📊 And in practice ?

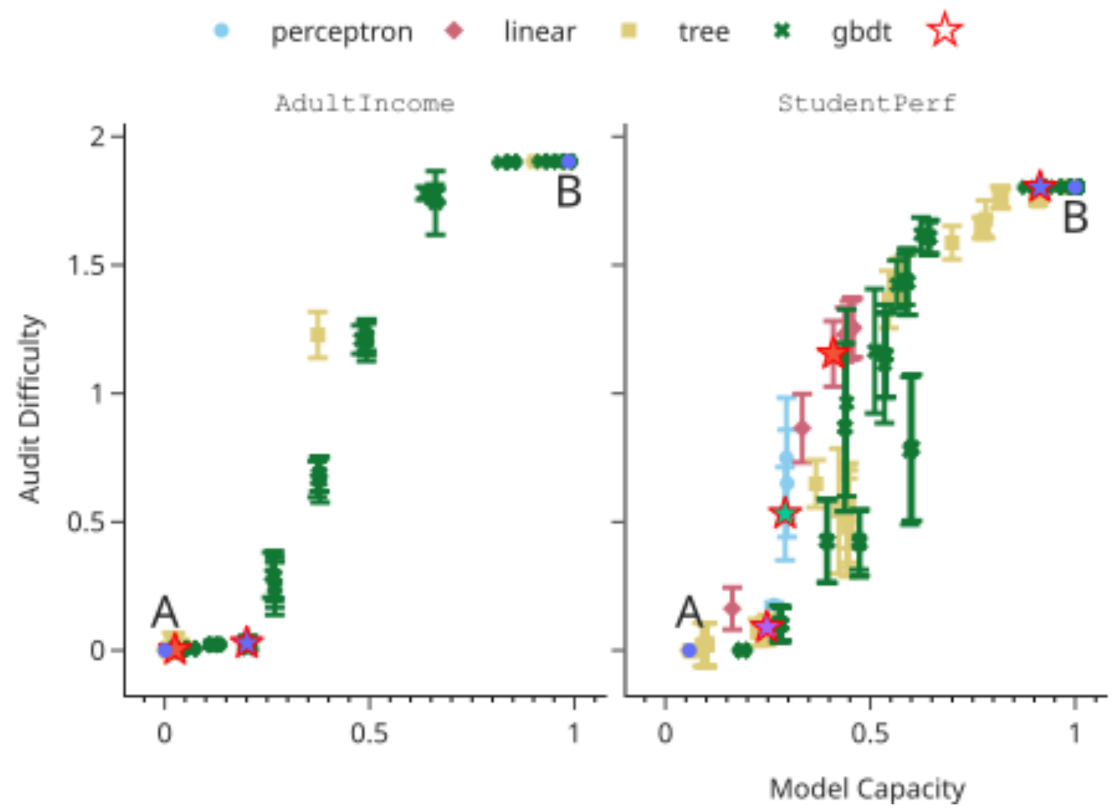


Figure 2: The value of the μ-diameter with respect to the Rademacher complexity of the hypothesis class. Informally: hypothesis class = fixed architeecture + hyper-parameters.

Figure 3: What is the accuracy cost for a platform to evade an audit? Not much. Let $\mathcal{F} = (\mathcal{H}_1, \dots, \mathcal{H}_f)$ be a family of hypothesis classes. *Example: all the decision trees with varying maximum depth.*

- $\mathcal{H}^* \in \mathcal{F}$ with best test accuracy.
- $\mathcal{H}_{\text{evade}} \in \mathcal{F}$ with largest μ-diameter.

$$\mathrm{CostOfExhaustion}(\mathcal{F}) = \mathrm{Accuracy}(\mathcal{H}^*) - \mathrm{Accuracy}(\mathcal{H}_{\text{evade}})$$