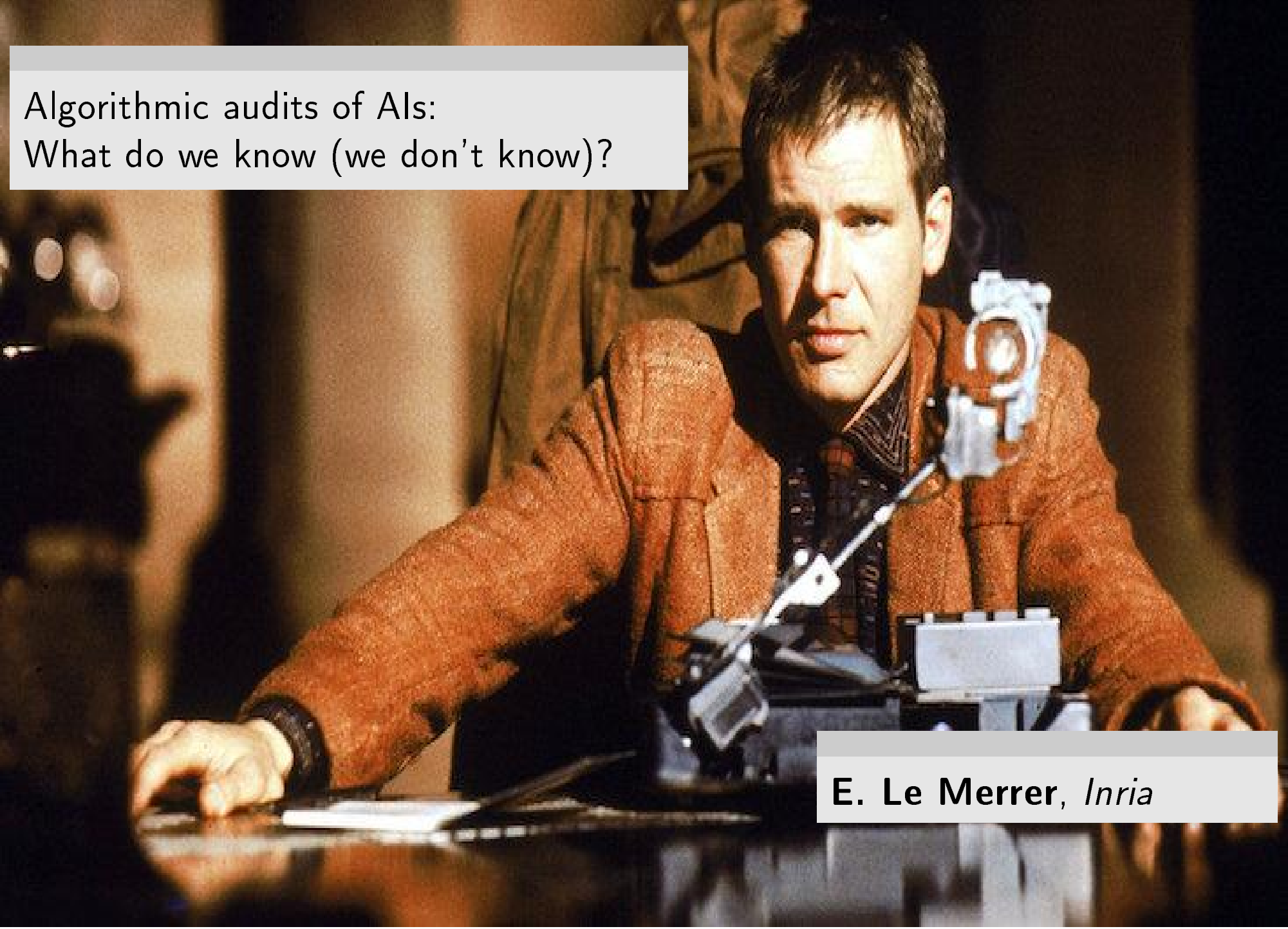
Algorithmic audits of AIs: What do we know (we don't know)?



E. Le Merrer, *Inria*

Blade Runner: the Voight-Kampff test

Is the remote entity a replicant ?
Essentially: investigation on questions/answers (inputs/ouputs)

Today: ChatGPT or student?

Sung Kim

Dec 11, 2022 · 4 min read · ✦ Member-only · Listen

How to Detect OpenAI's ChatGPT Output

How to detect if the student used OpenAI's ChatGPT to complete an assignment

On November 30, 2022, OpenAI released 'ChatGPT' AI system (https://openai.com/blog/chatgpt/), which is a universal writer's assistant that can generate a variety of output, including school assignments. The output (e.g., essays) provided by ChatGPT is so good, if I was a student, I would be using ChatGPT to complete most of my school assignment with minor revisions.

Today: ChatGPT or student?

Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan
vinu@umd.edu

Aounon Kumar
aounon@umd.edu

Sriram Balasubramanian
sriramb@umd.edu

Wenxiao Wang
wwx@umd.edu

Soheil Feizi
sfeizi@umd.edu

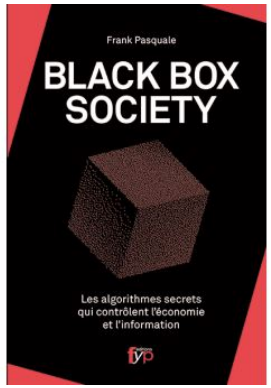
Department of Computer Science
University of Maryland

Abstract

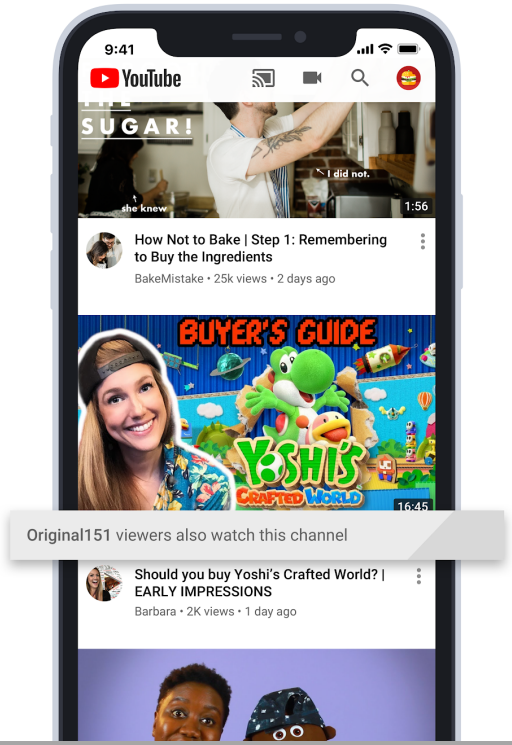
The rapid progress of Large Language Models (LLMs) has made them capable of performing astonishingly well on various tasks including document completion and question answering. The unregulated use of these models, however, can potentially lead to malicious consequences such as plagiarism, generating fake news, spamming, etc. Therefore, reliable detection of AI-generated text can be critical to ensure the responsible use of LLMs. Recent works attempt to tackle this problem either using certain model signatures present in the generated text outputs or by applying watermarking techniques that imprint specific patterns onto them. In this paper, both empirically and theoretically, we show that these detectors are not reliable in practical scenarios. Empirically, we show that *paraphrasing attacks*, where a light paraphraser is applied on top of the generative text model, can break a whole range of detectors, including the ones using the watermarking schemes as well as neural network-based detectors and zero-shot classifiers. We then provide a theoretical *impossibility result* indicating that for a sufficiently good language model, even the best-possible detector can only perform marginally better than a random classifier. Finally, we show that even LLMs protected by watermarking

No replicants yet, but pervasive decision-making AIs

Why we need audits ?

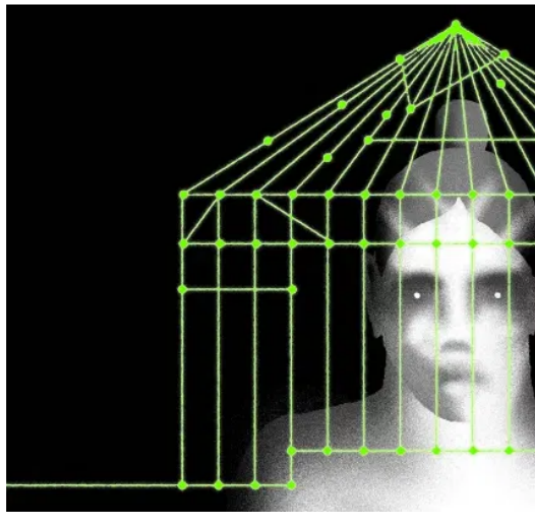


Recommendation



Credit scoring

MIT Technology Review

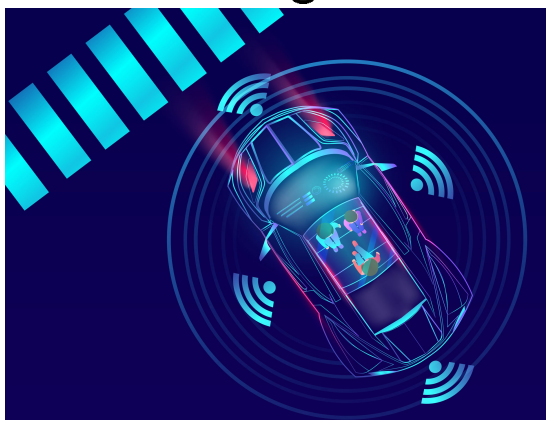


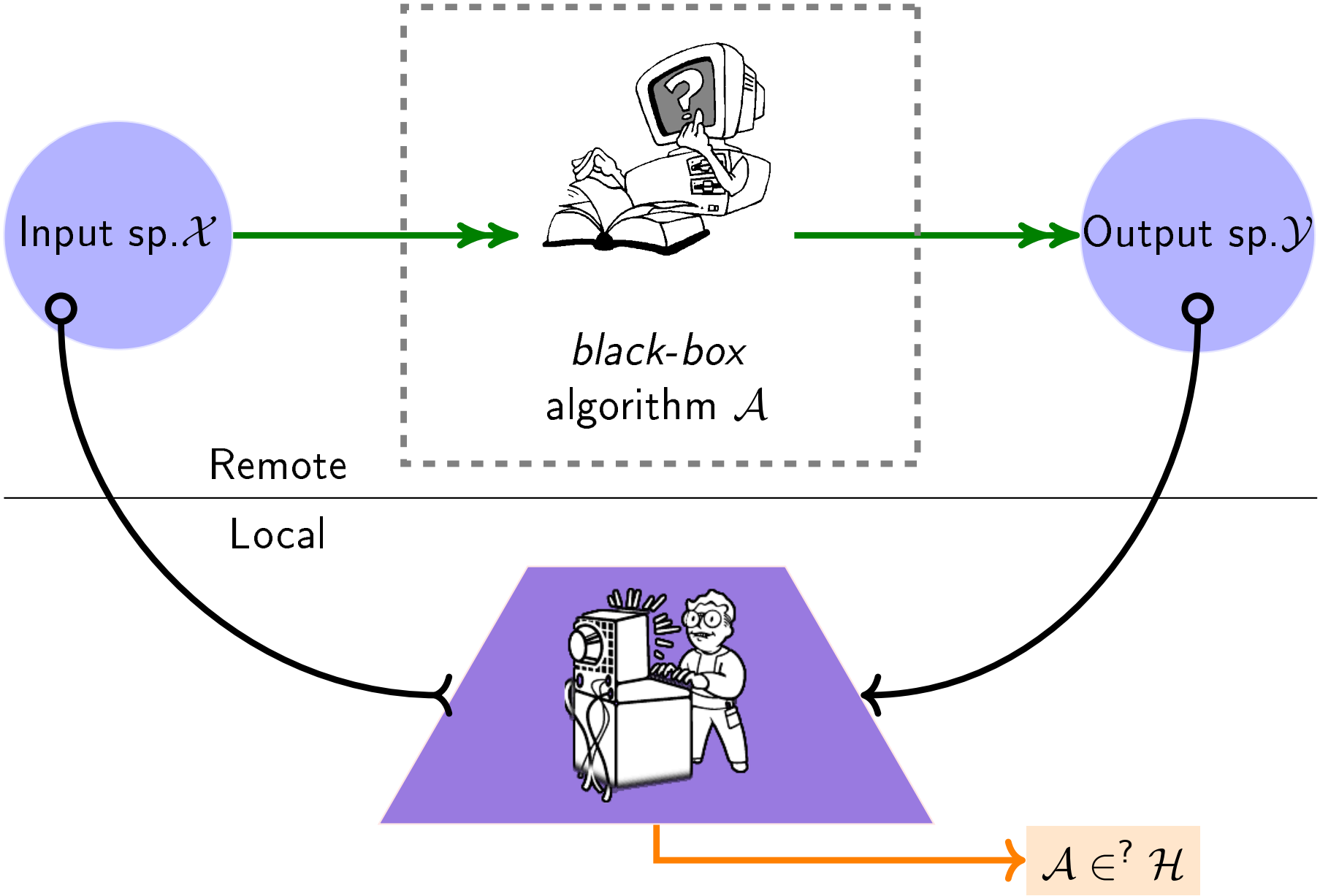
DANIEL ZENDER

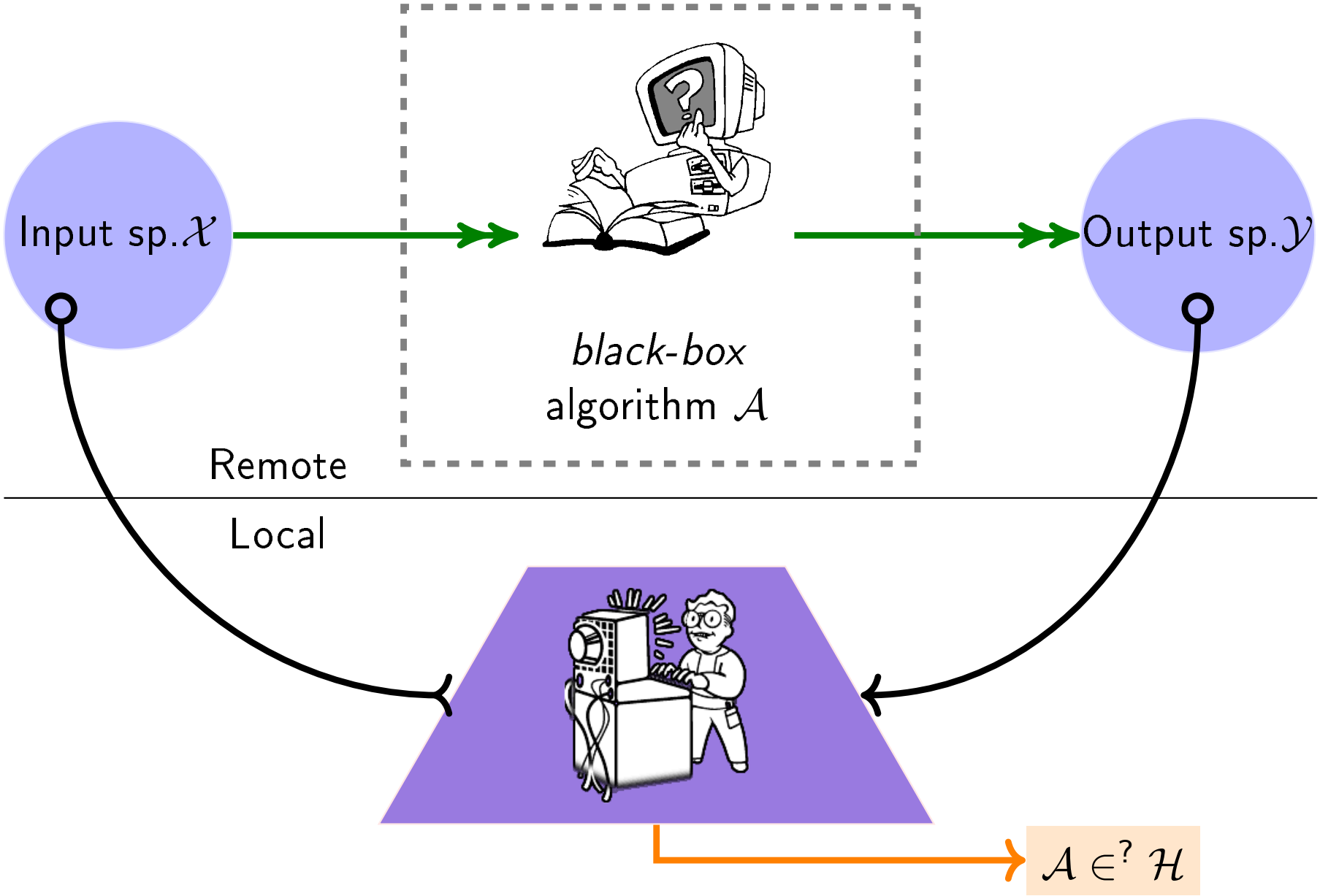
Tech policy / AI Ethics

The coming war on the hidden algorithms that trap people in poverty

Self drinving cars




Input sp.$\mathcal{X}$
Output sp.$\mathcal{Y}$
black-box
algorithm $\mathcal{A}$
Remote
Local
$\mathcal{A} \in^? \mathcal{H}$



and... link to security: information gain, algorithm leak, poisoning

An Input / Output example

Adult Census Income: task to predict whether income exceeds $50K$ / *yrbasedoncensusdata*

Input:

# age	workclass	# fnlwgt	education	# education....	marital.sta...	occupation
90	?	77053	HS-grad	9	Widowed	?
82	Private	132870	HS-grad	9	Widowed	Exec-managerial
66	?	186061	Some-college	10	Widowed	?
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct

Output: Boolean (yes/no)

An Input / Output example

Adult Census Income: task to predict whether income exceeds $50K$ / *yrbasedoncensusdata*

Input:

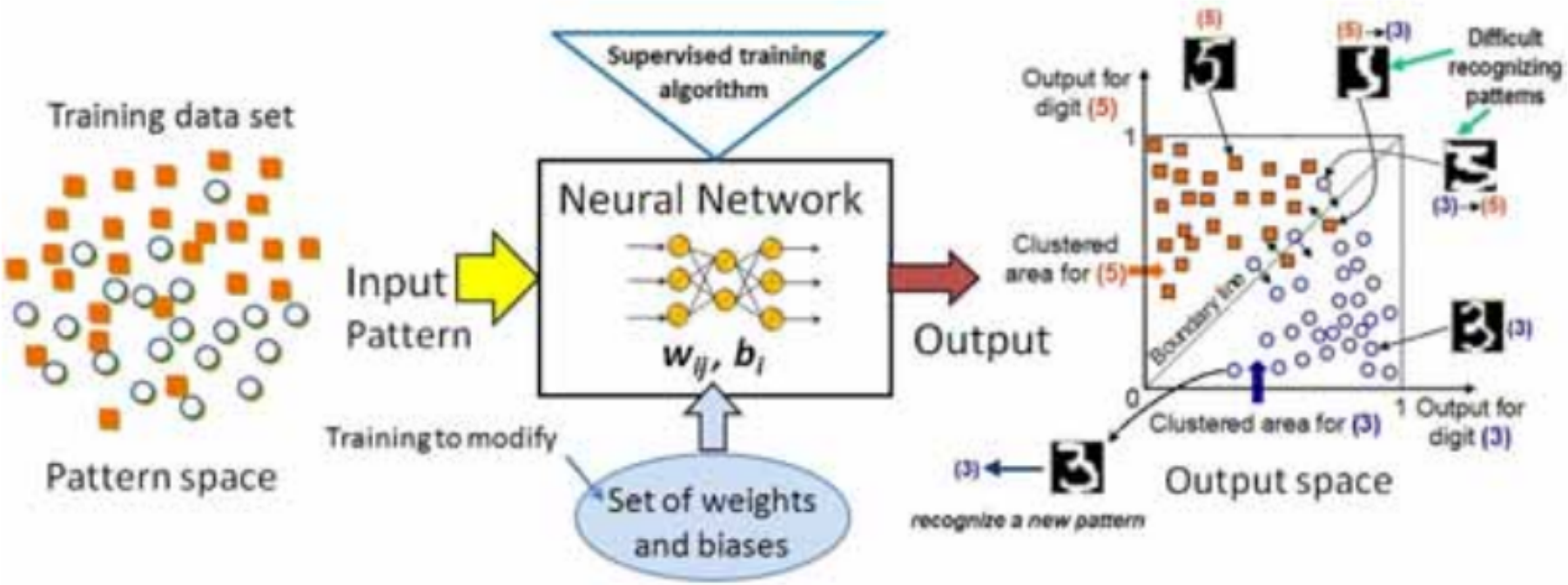
# age	workclass	# fnlwgt	education	# education....	marital.sta...	occupation
90	?	77053	HS-grad	9	Widowed	?
82	Private	132870	HS-grad	9	Widowed	Exec-managerial
66	?	186061	Some-college	10	Widowed	?
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct

Output: Boolean (yes/no)

Other examples:

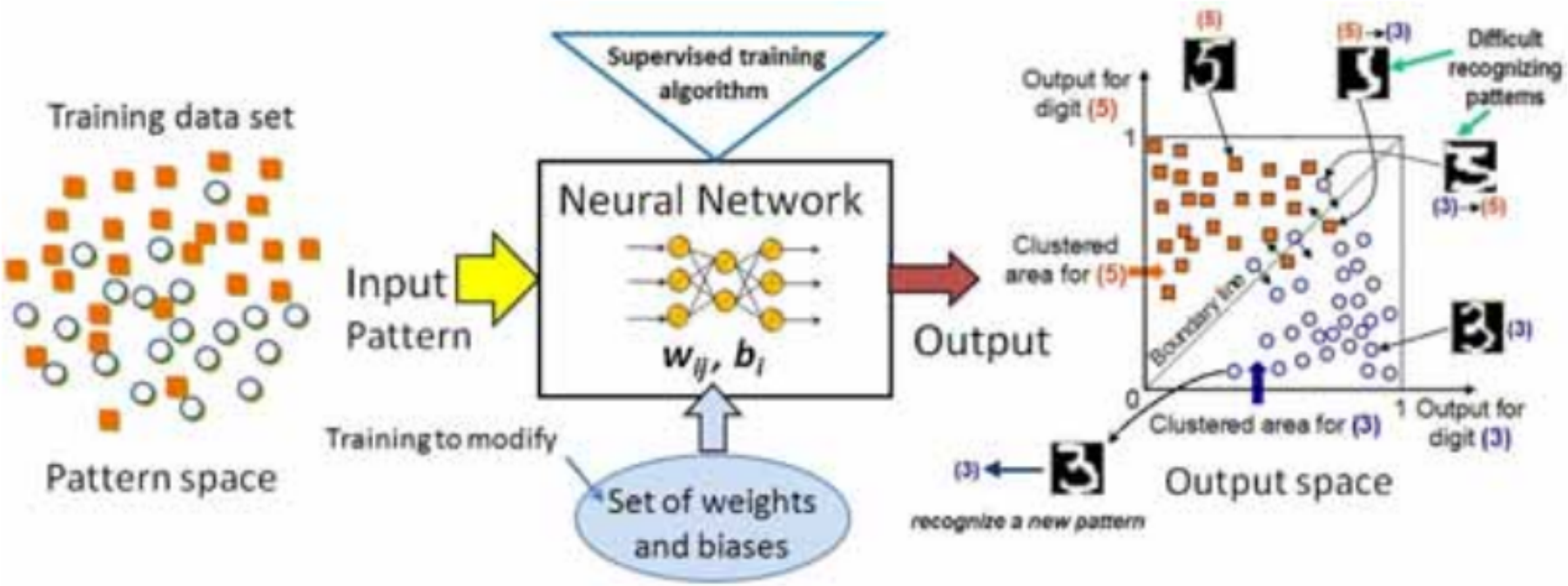
- image (input) $\rightarrow$ label (output)
- user profile $\rightarrow$ item recommended

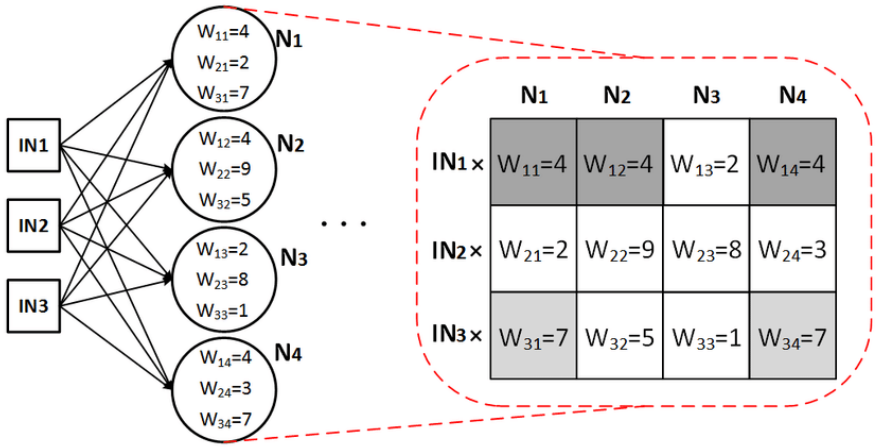
A central notion: boundaries & non native explainability



img: Le Dung et al. 2008.

A central notion: boundaries & non native explainability





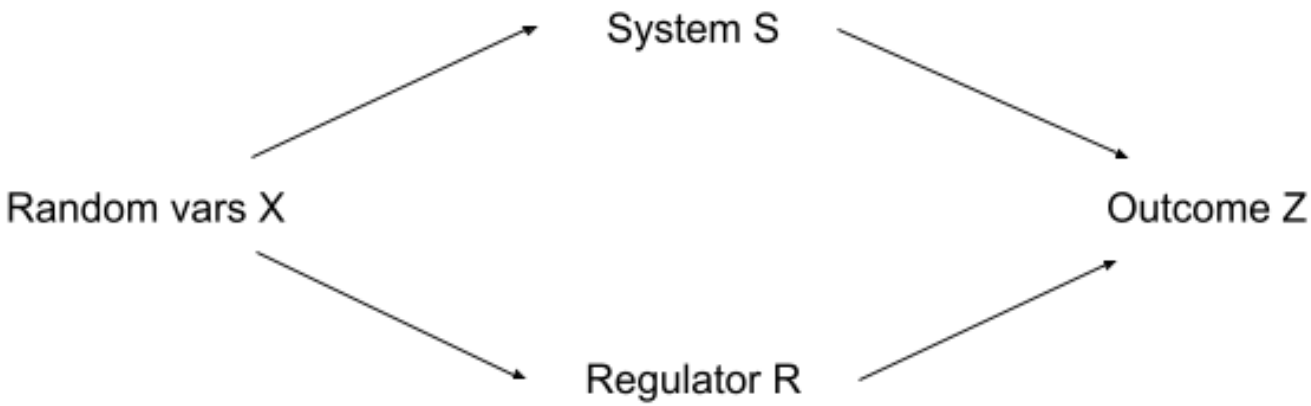
img: Le Dung et al. 2008.

Audits: why not reconstructing the black box?

- From queries/responses $\rightarrow$ copy the black box
- Problem: the good regulator theorem

Audits: why not reconstructing the black box?

- From queries/responses $\rightarrow$ copy the black box
- Problem: the good regulator theorem



- For any optimal regulator $P[R|S]$, Z is a deterministic function of S
- Impractical: gigantic scale of algorithms in black boxes

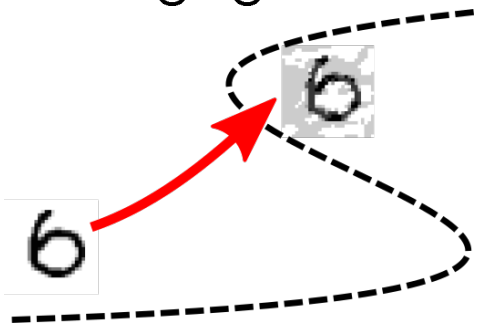
Let's think "local": and audit properties only

Conant & Ashby, 1970. *Fixing The Good Regulator Theorem*, by johnswentworth.

Decision boundaries: how to approach them

PB: "fooling" $\mathcal{A}$

Leveraging *adversarial examples*



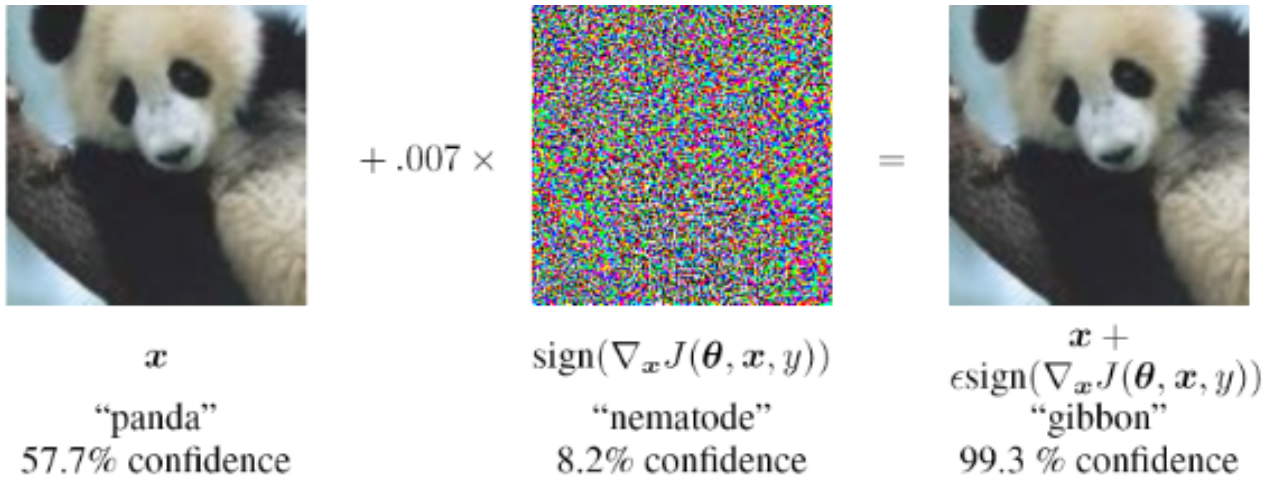
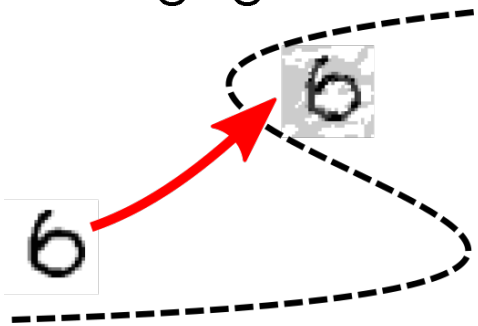


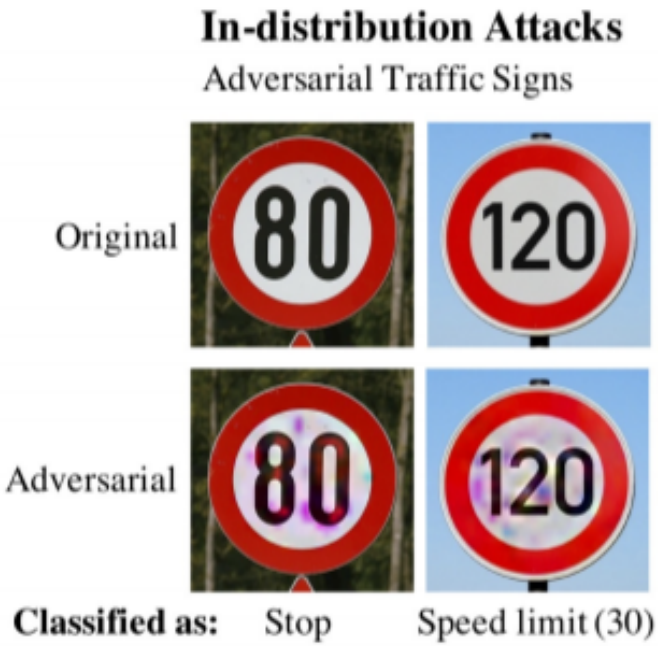
Figure 1: An adversarial image generated by *Fast Gradient Sign Method* [55]: left: a clean image of a panda; middle: perturbation; right: an adversarial image, classified as a gibbon.

Decision boundaries: how to approach them

PB: "fooling" $\mathcal{A}$

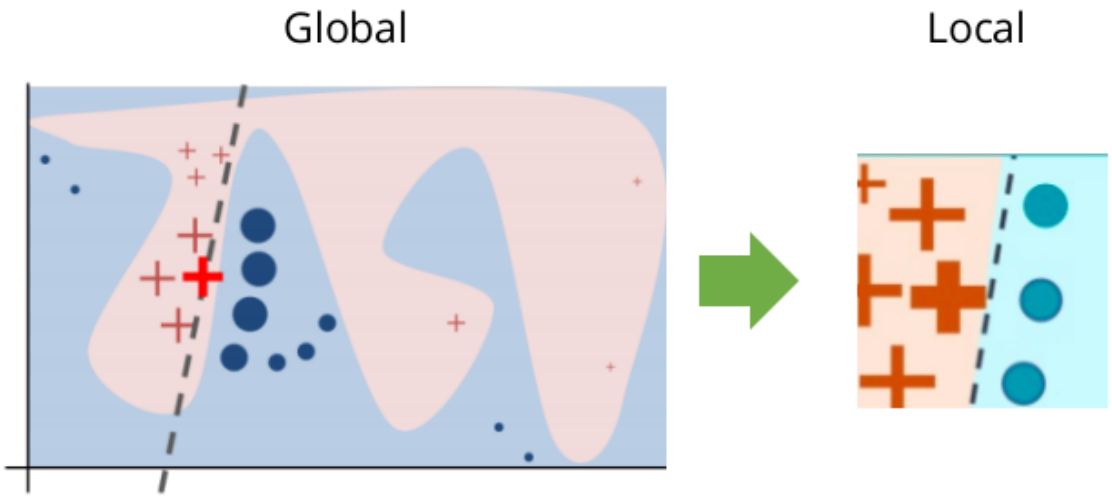
Leveraging *adversarial examples*





Local boundary related explanations: *e.g.*,LIME

PB: explaining $\mathcal{A}$'s decision locally



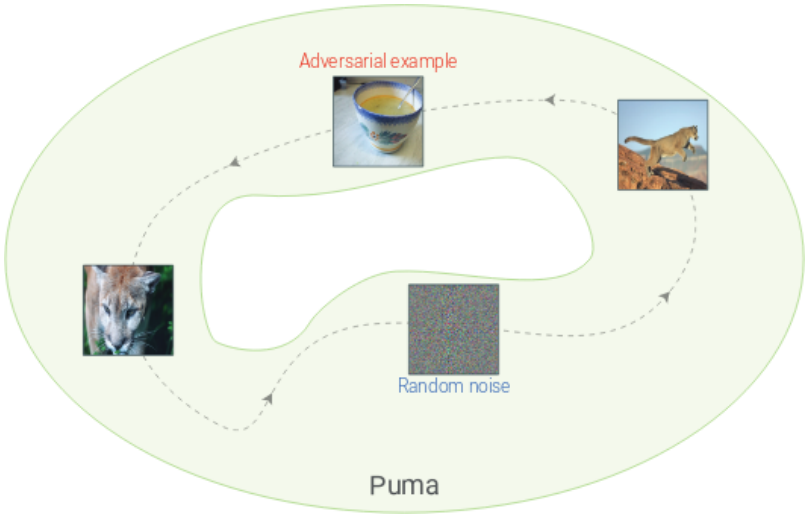
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

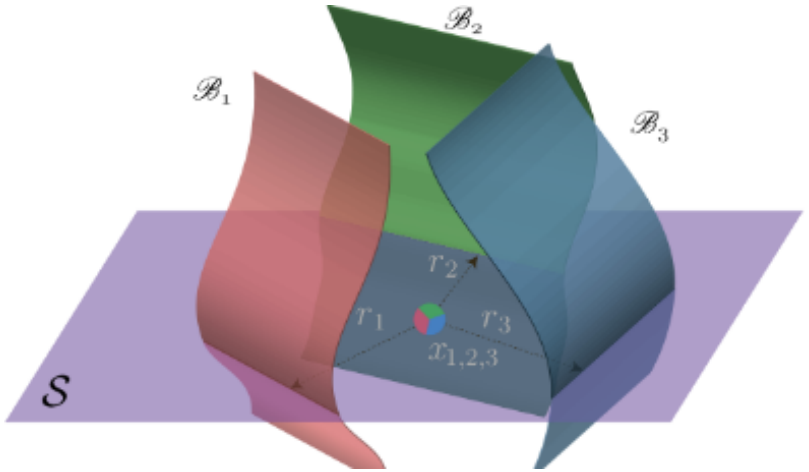
Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier", 2016.

Decision boundaries: what we know

- Fawzi et al. 2017: "classification regions are connected"



- $r(x) = \arg\min_r ||r||_2$ s.t. $\mathcal{A}(x+r) \neq \mathcal{A}(x)$



Let's assume the AI is truthful

Warning: generic assumptions in related work

e.g. demographic parity:

$$\mu_{D_x}(\mathcal{A}) = P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 1) - P_{(x,x_s)\sim D_x}(\mathcal{A}(x) = 1 | x_s = 0)$$

- with D_x the data distribution and x_s a sensitive attribute

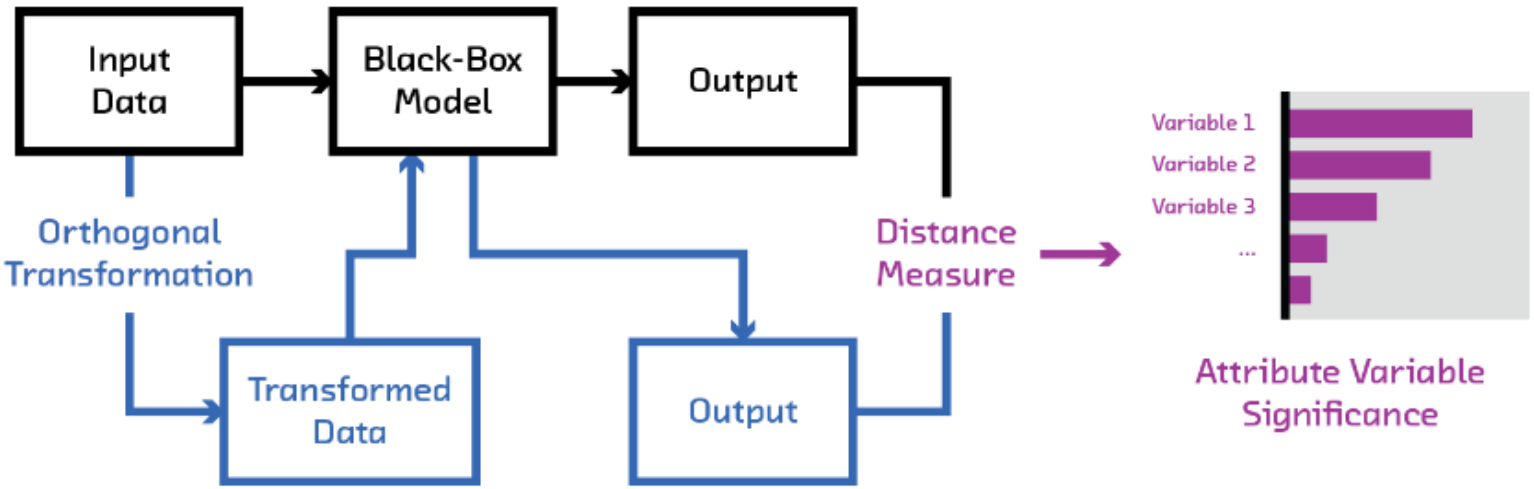
Classic assumptions (e.g. active fairness auditing):

- D_x is know to the auditor
- Events are non negligible: $min(P(x_s = 1), P(x_s = 0)) = \Omega(1)$
- $\mathcal{A}$'s hypothesis class know to the auditor
- + model stable/deterministic in between queries
- ...

Black-box fairness/impact measurement

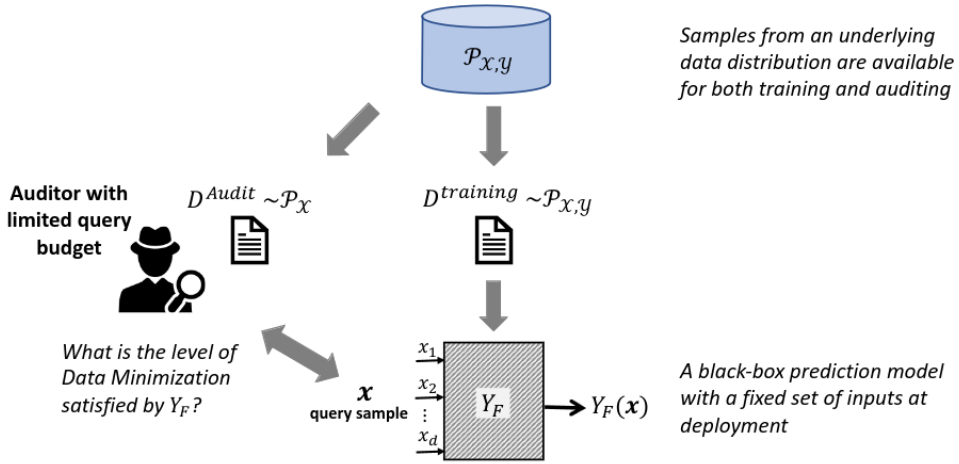
PB: how to assess $\mathcal{A}$'s dependacy on an input feature?
Many, many works, e.g. FairML:

- measure model dependency on inputs by changing them
- small change to a feature changes the output a lot $\Longrightarrow$ model is sensitive to it

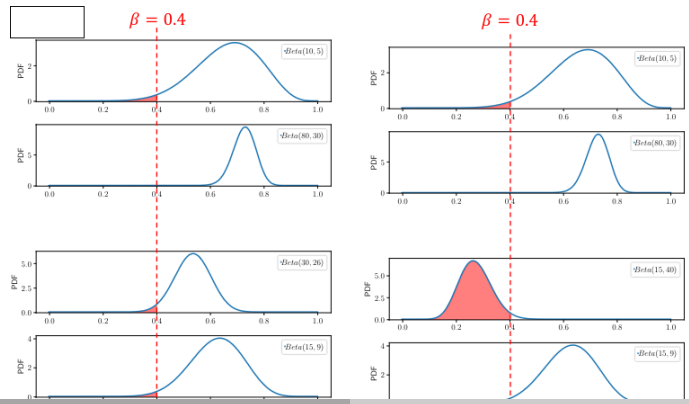


The data minimization principle

PB: how to detect the improper use of an input feature?



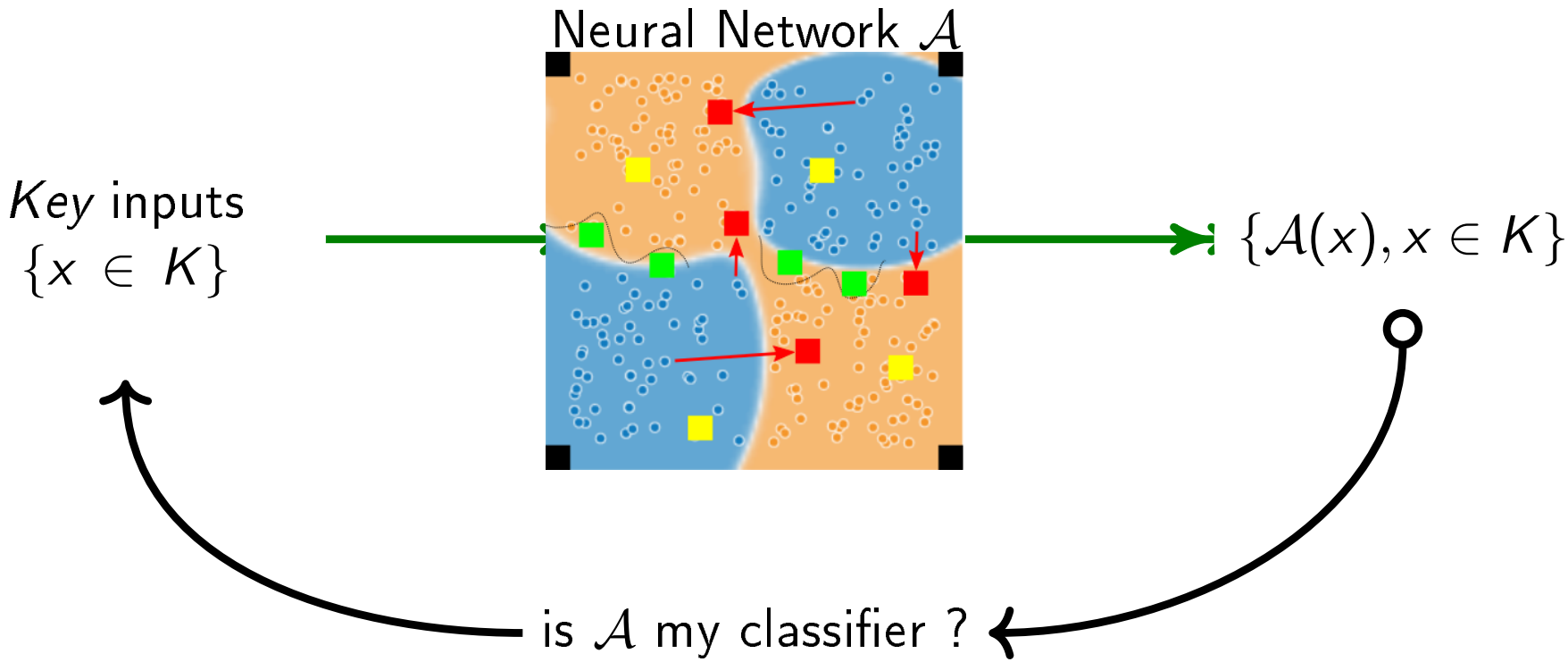
Data minimization guarantee at level β ensures that every input feature used by a prediction model is indeed necessary to reach the predictions made for at least a certain fraction, β, of decisions (predictions).



Tampering detection of a deployed model

PB: how to detect if $\mathcal{A}$ has changed?

- A white box access initially, then deploy & check



If a decision **change** occurs $\rightarrow$ tampered model !

Le Merrer, Tredan, ISSRE 2019.

Measuring distances between evolving models

PB: how to measure the distance between evolutions of $\mathcal{A}$?

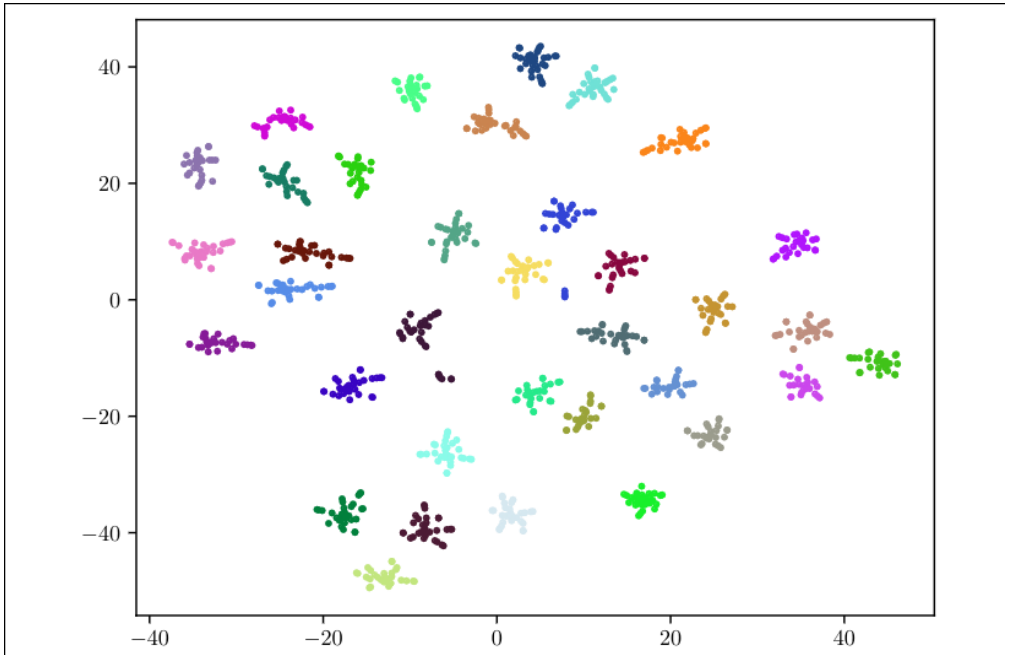


Figure 1: A t-SNE representation of the pairwise distances of 1081 different models: 10 types of variation applied on 35 off-the-shelves vanilla models for ImageNet with different parameters (listed in App. B.2). This work exploits the clear separability (clusters of consistent colors) observed in the decisions of these models. Confusions yet happen (model colors further apart from their cluster), but are under scrutiny for the tracking of false positive identification.

$$dist(\mathcal{A}, \mathcal{A}') = 1 - \frac{\hat{I}(Y_a, Y_{a'})}{\min(\hat{H}(Y_{a'}), \hat{H}(Y_a))} \in [0, 1].$$

Maho et al., ICASSP'22. See also "A zest of lime", ICLR'22

Problem: AIs may lie (like replicants do)

Why? Obvious conflicting interests: users vs providers

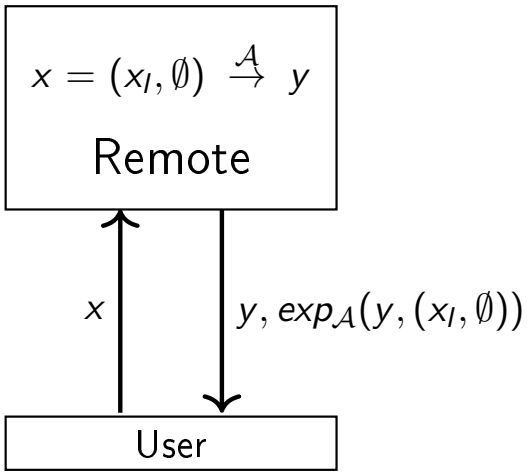
In 1951 American Airlines partnered with IBM to attack the difficult logistical problems of airline reservations and scheduling ($\rightarrow$ SABRE)

Surprisingly, in the face of public scrutiny the company did not deny its manipulations. Speaking before the US Congress, the president of American, Robert L. Crandall, boldly declared that biasing SABRE's search results to the advantage of his own company was in fact his primary aim. He testified that "the preferential display of our flights, and the corresponding increase in our market share, is the competitive raison d'etre for having created the [SABRE] system in the first place" (Petzinger, 1996). We might call this perspective "Crandall's complaint:" Why would you build and operate an expensive algorithm if you can't bias it in your favor?

Sandvig et al., ICA2014.

Or more recently, the Volkswagen "diesel-gate"

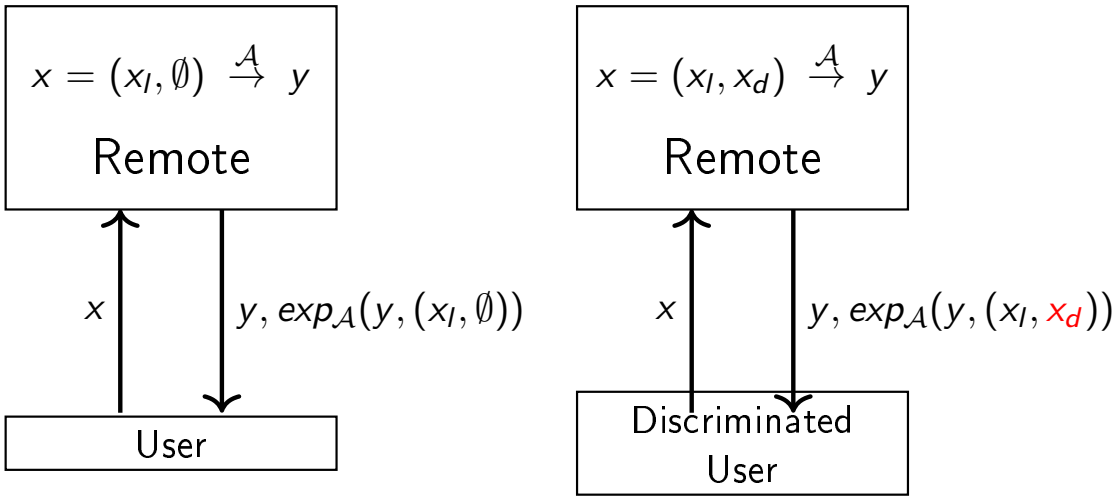
How? The bouncer problem



- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.

Le Merrer, Tredan, Nature Mach. Int. 2020.

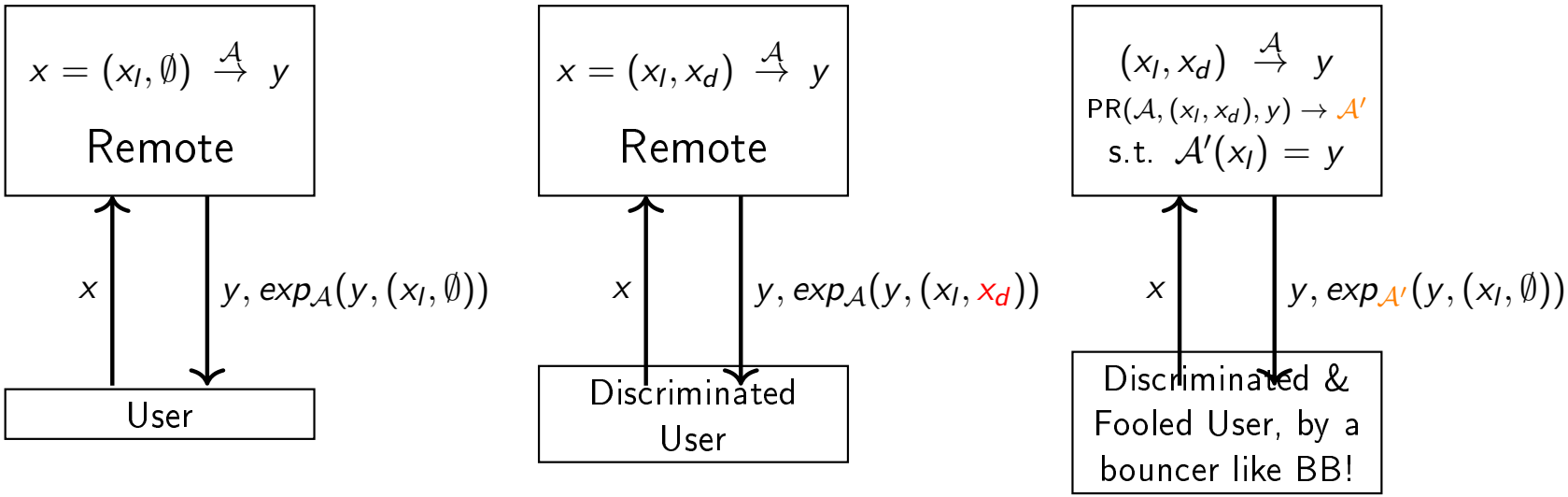
How? The bouncer problem



- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.
- *Intuition*: if decision relies on discriminative variables, explanation will reveal it

Le Merrer, Tredan, Nature Mach. Int. 2020.

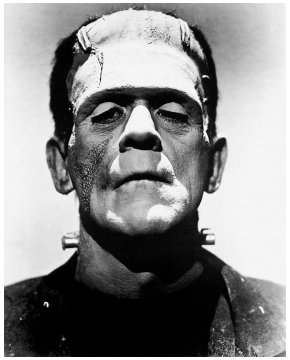
How? The bouncer problem



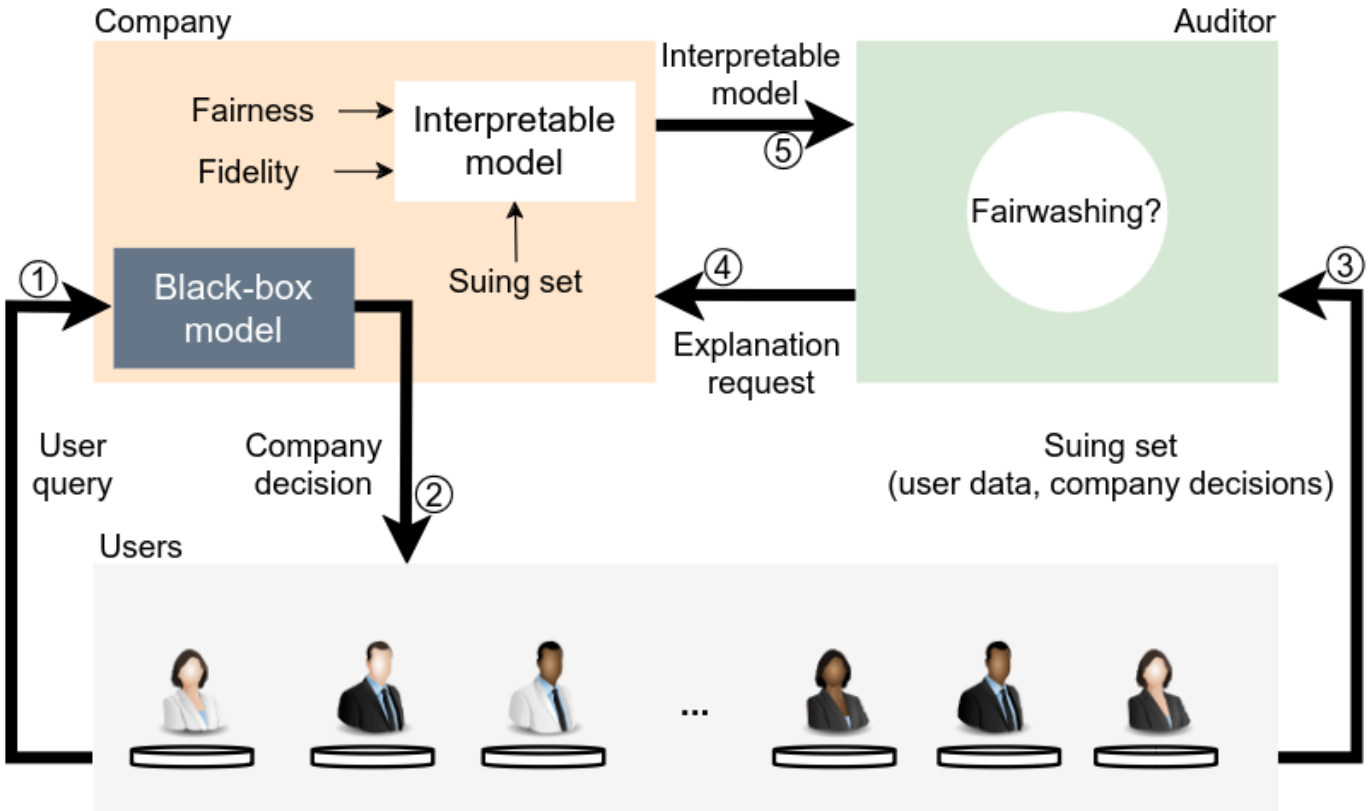
- From *users perspective*: classifier *is a black-box*
  Provide request x, obtain classification y.
- *Intuition*: if decision relies on discriminative variables, explanation will reveal it
- **An attack**: generate a "legit" classifier $\mathcal{A}'$ on the spot, and explain it (like a bouncer would do...)

Le Merrer, Tredan, Nature Mach. Int. 2020.

Bounced! An exemple on Decision Trees



How? (2) Fairwashing



- Rationalization: find **AN interpretable surrogate model** c approximating model b, such that c is fairer than b, to then show it to the auditor.

Shamsabadi et al., NeurIPS'22; Aivodji et al., ICML 2019.

The ideology behind publishing Twitter's source code

A leak. On 31 March, Twitter published parts of the source code that powers its newsfeed. The move came a few days after it was made public that large portions of that code had been leaked on Github already [Gizmodo, 31 Mar].

The 85,797 lines of code contain little new information. Tweets that contain links are less likely to appear in a user's timeline. So are tweets in a language that the system cannot recognize – an obstacle for people whose vernaculars aren't on the radar of Californian engineers. Spaces (Twitter's live podcasting feature) about Ukraine seem to be hidden from view too [Aakash Gupta, 2 Apr].

The most interesting part of the release is the blog post written by Twitter's remaining engineering team. It provides a good high-level overview of how a newsfeed algorithm works.

How (not) to open source. One company led the way in making algorithms public: Twitter. Two years ago, its "Ethics, Transparency and Accountability" team released the code of an image-cropping algorithm and invited auditors to find possible biases [AlgorithmWatch, 2021]. The team was among the first to be fired last year.

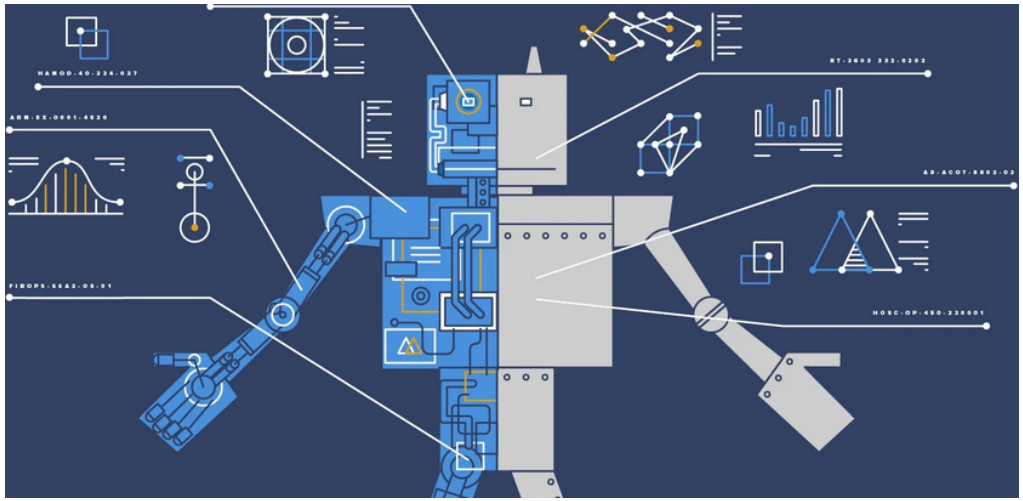
You cannot audit code only by reading it. You need to run it on a computer. On Ukraine, for instance, we only know that Twitter Spaces labeled "UkraineCrisisTopic" undergo the same treatment as items labeled with violence or explicit content. But we don't know how the label is applied or what effects it has. It seems that the code responsible for that task has not even been made public.

Obfuscation. Publishing vast amounts of computer code without instructions can be worse than useless. It allows for claims of transparency while preventing any actual audit. Twitter is not the first

What can an auditor do facing trickery?

- Be stealphy: look like a user
- Make stronger assumptions

Be stealthy: building cases as users with bots

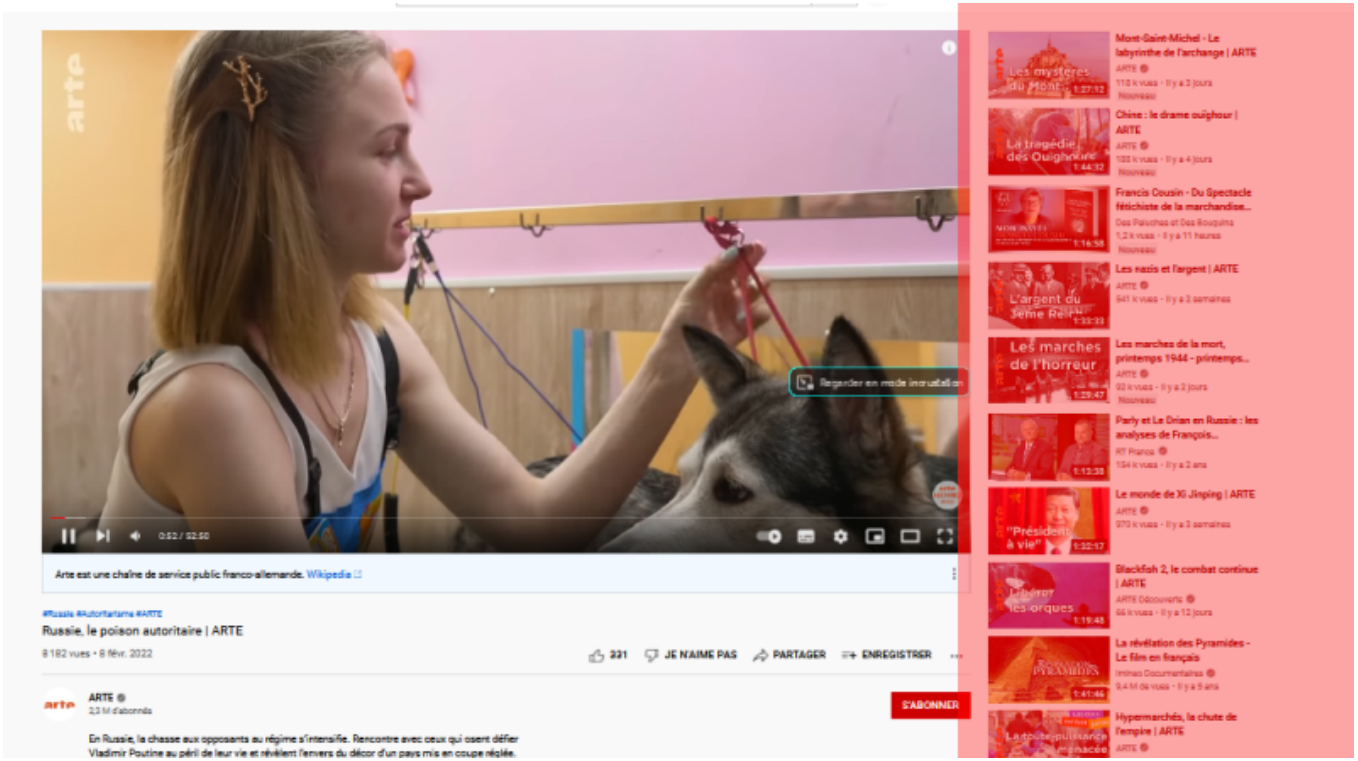


Bots to simulate users: scriptable browsers (Selenium, Pupeteer):

- Bots' homes: stable servers, up during months
- Bots interact: connect/click/watch, and collect results

(Yet, no proof we are not sandboxed... cf diesel-gate)

Be stealthy: building cases as users with bots



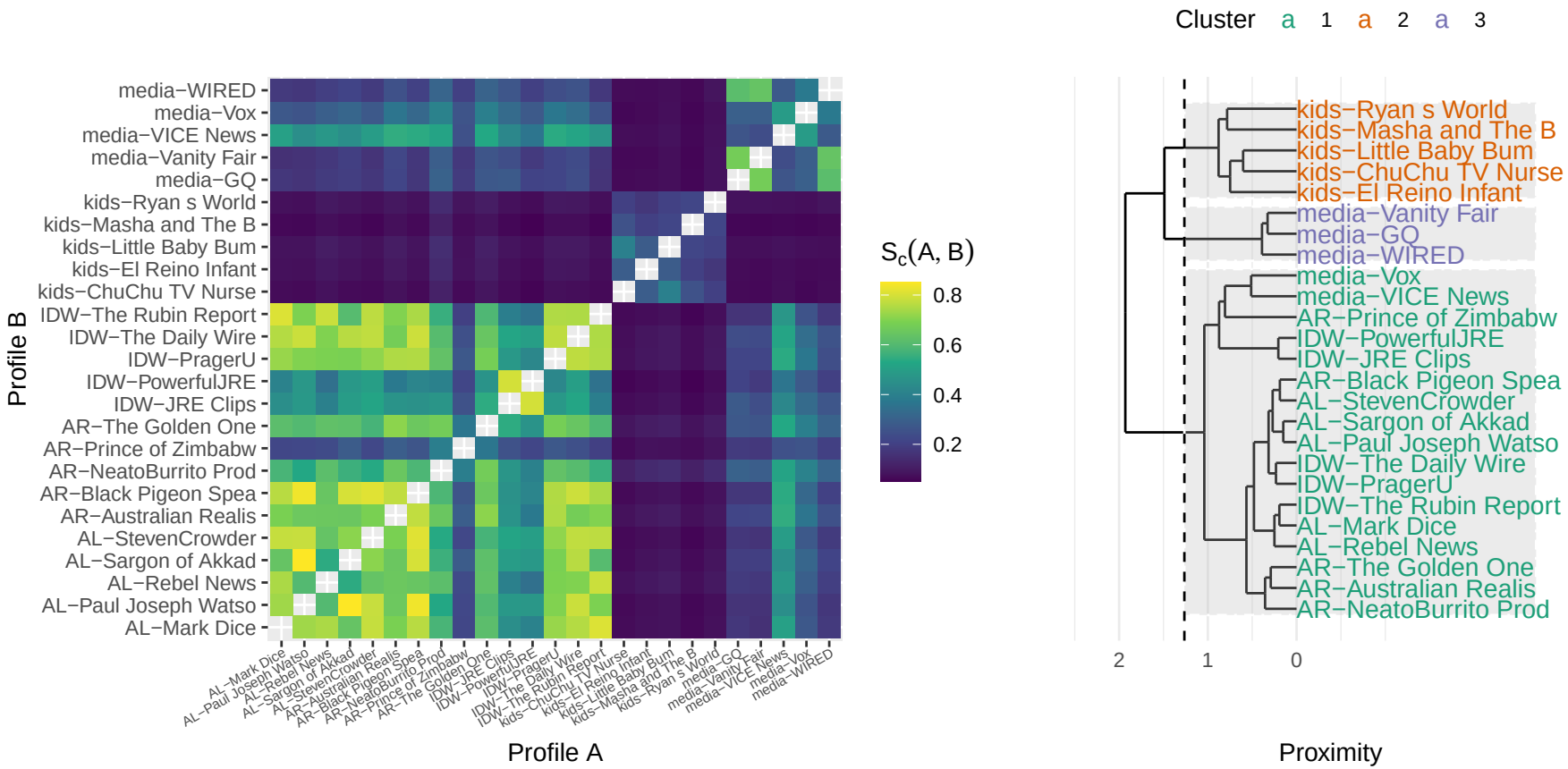
At YouTube:

- In 2018, was accounting for 70% of clicks
- Built to optimize user time on the platform
  - 2016 academic paper listing guidelines

Be stealthy: building cases as users with bots

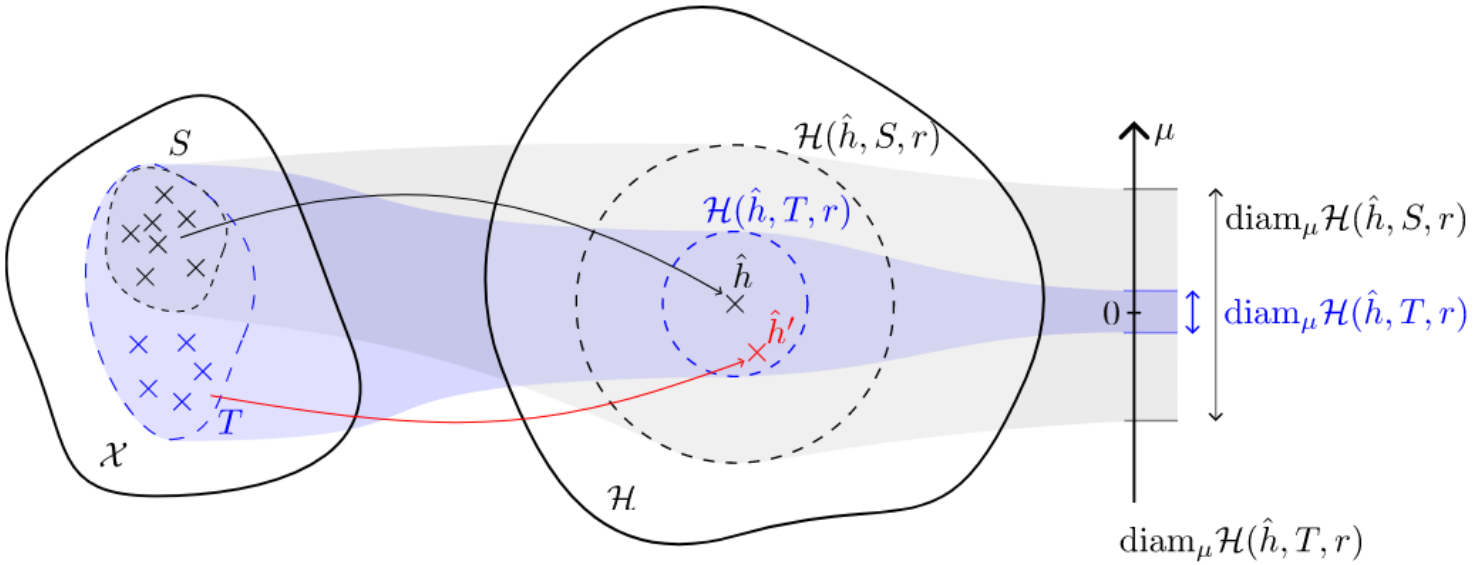
PB: how to measure filter bubbles?

5438 users simulated, watching 5 videos in a row (10.6M recos collected)



Make some assumptions: active fairness auditing

PB: constrain $\mathcal{A}$ to stay consistent with its previous answers

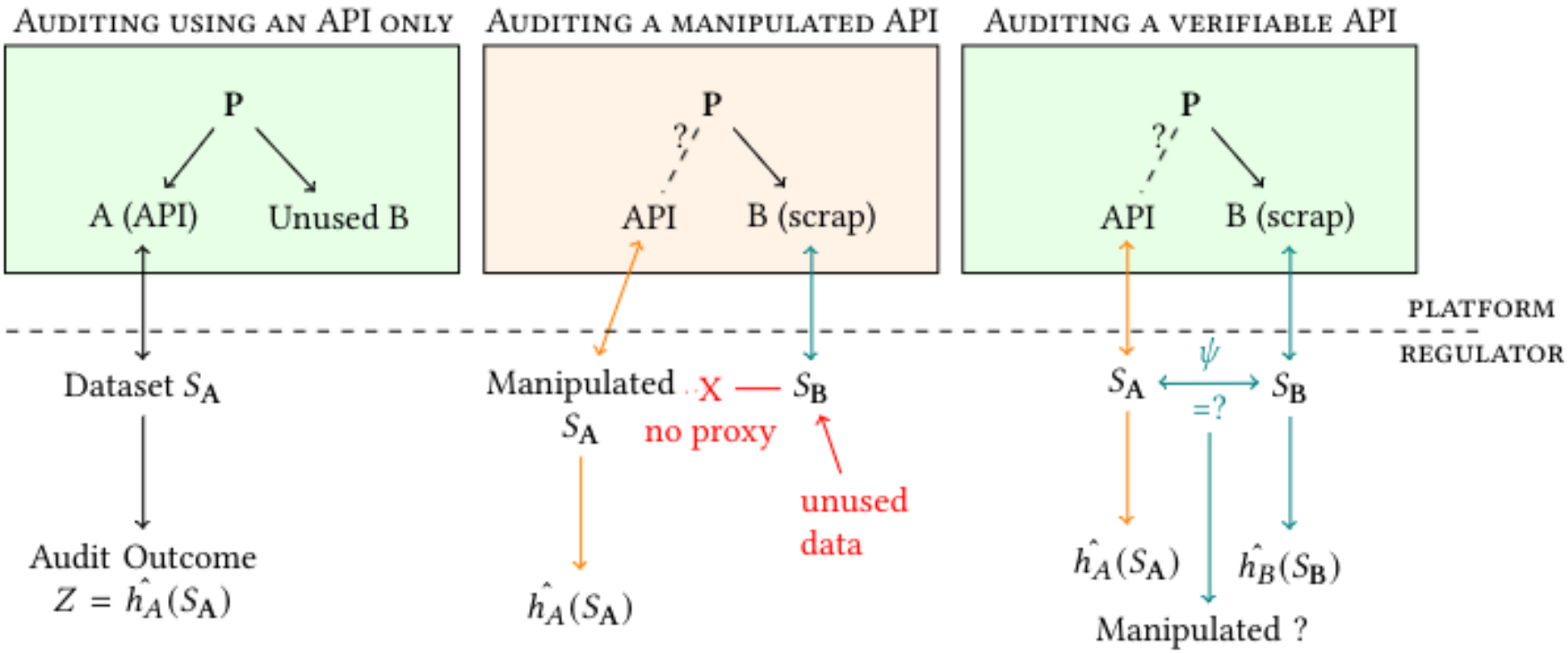


- A.F.A. goal: ensure estimate within ϵ of $\mu(\mathcal{A}_{manipulated})$
- Model cannot change answers once given = certificate
- The auditor wants to crafts queries that constrain the model the most
- Issues: hypothesis class assumed / not robust to lies

Yan, Zhang, ICML 2022. Img: Augustin

APIs: really? + spotting inconsistencies

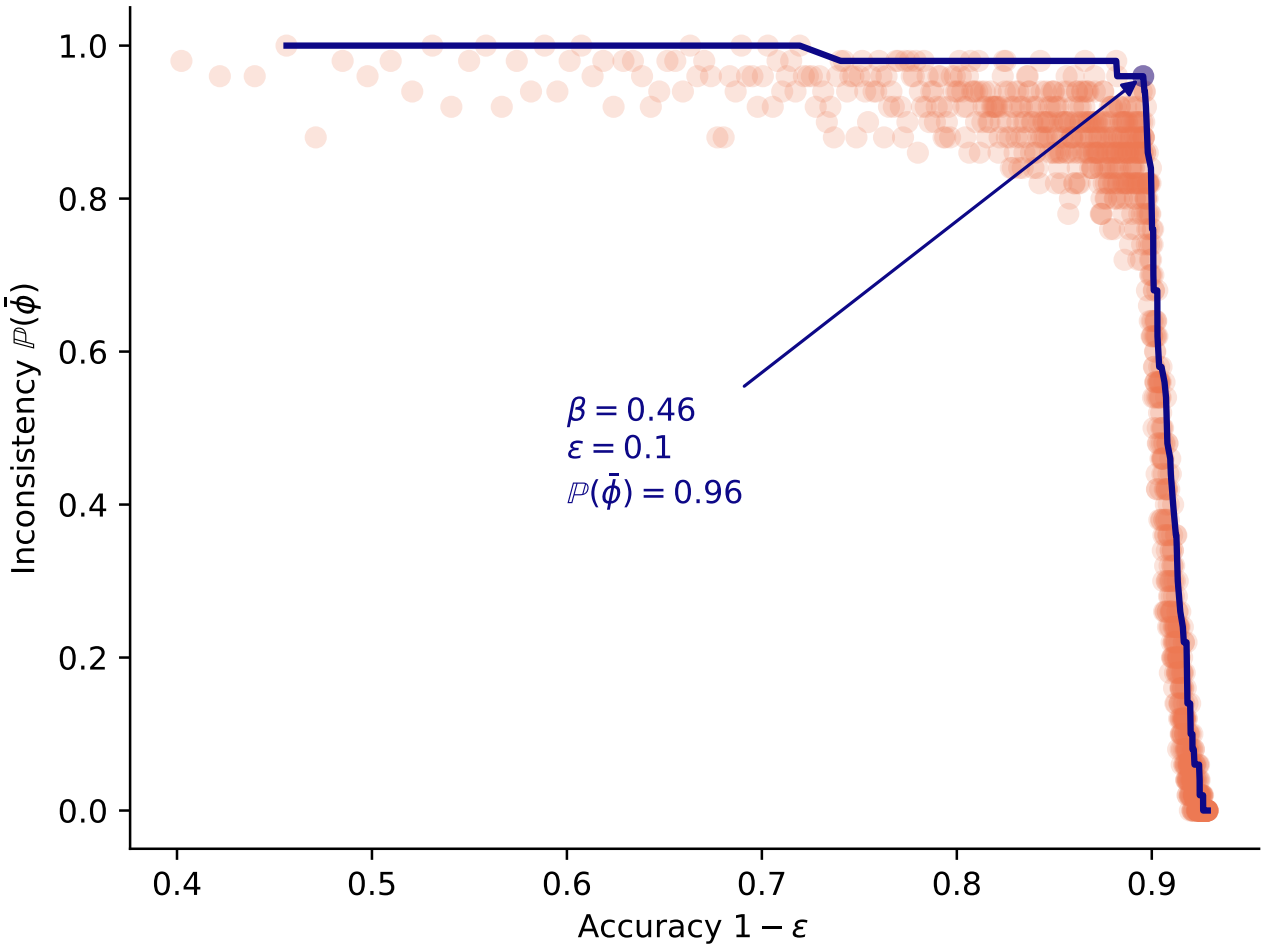
PB: acknowledging fairwashing, are APIs useful anyway?



Compare observations from several sources to spot inconsistencies

J. Garcia-Bourrée et al., submitted.

APIs: really? + spotting inconsistencies

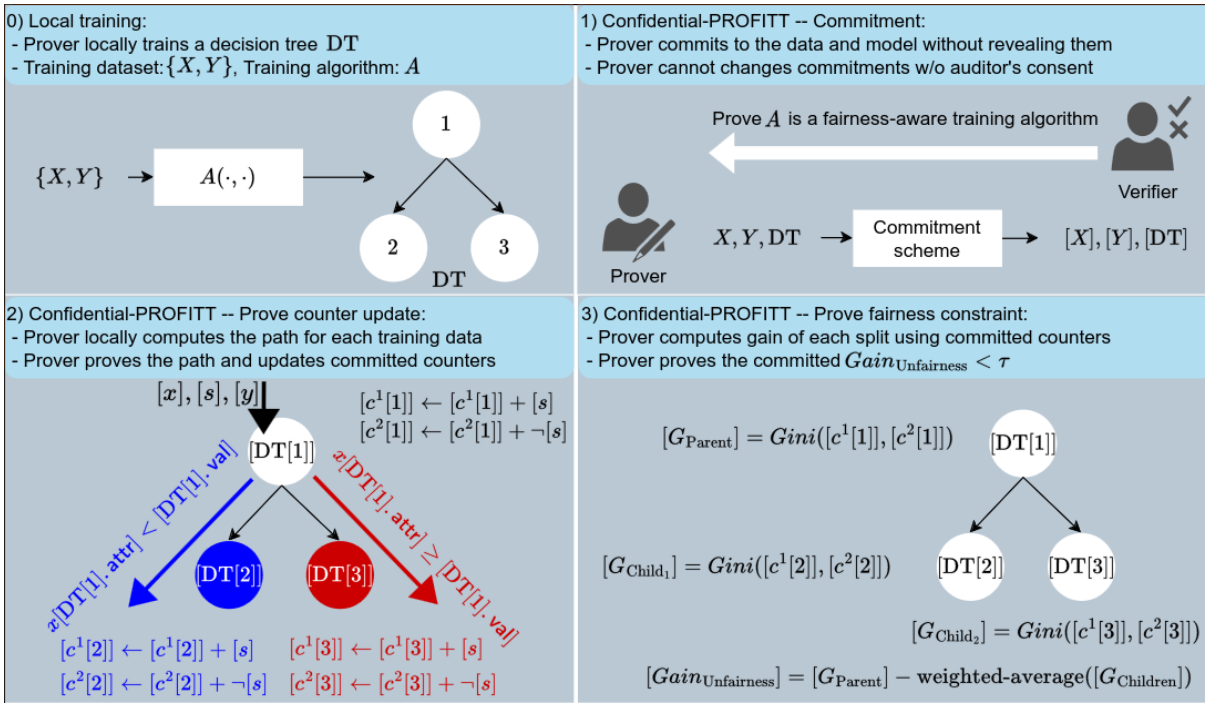


Estimating economic disparity while also checking for manipulation (inconsistencies between answers from A and B) under a fixed audit budget. A Pareto frontier appears: the higher the estimation accuracy, the harder it is to spot inconsistencies

Towards leveraging cryto proofs

PB: means for $\mathcal{A}$ to prove fairness in training?

- ZK proofs: give a proof without reveiling information
- ZK-friendly fair decision tree learning algorithm: fairness verified efficiently w.o. repeating the entire training process



But: not proving the use of the proven model in the deployment!

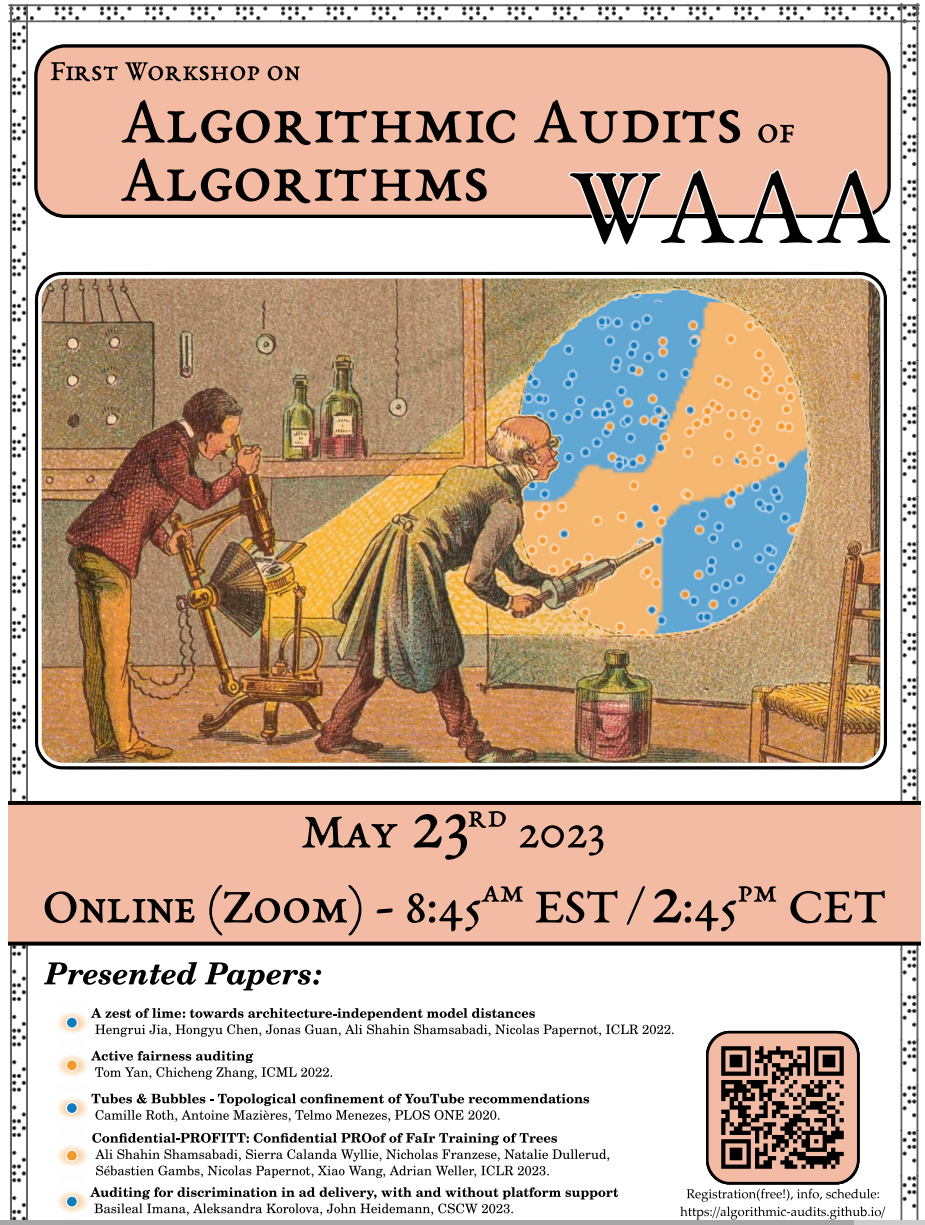
Shamsabadi et al., ICLR 2023.

The long road to efficient black-box audits...

- Societal push: scandals, Chat-GPT on "pause", DSA, AI-act:

  Prop. résol. Européenne mars 2023, 68: Souhaite que soit généralisée l'évaluation par des tiers de la conformité des systèmes d'IA
- Research has not yet provided practical manipulation-proof audit algorithms
  - Dimensionnality of inputs, vs need of bounding query budget
  - Need for assumptions (are black box audits realistic in practice?)
  - Many impossibility theorems yet to come?
- Hope
  - Laws with more enforcement (auditors with *correct* assumptions)
  - Collaborative user-audits? (many users instead of bots)

The end



Thanks to Gilles, Augustin, Jade, Thibault, Teddy and François!

erwan.le-merrer@inria.fr

https://algorithmic-audits.github.io