CSCI 699: Privacy Preserving Machine Learning - Week 12

Incentives and Privacy

Sai Praneeth Karimireddy, Nov 22 2024.

Some disclaimers

- I am not a lawyer
- Mainly for discussions and inspiring technical questions
- Largely US-centric

Who “owns” data?

- Google News scrapes news outlets and aggregates them
- Google gets eyeballs and displays ads
- News websites lose out on audience and revenue

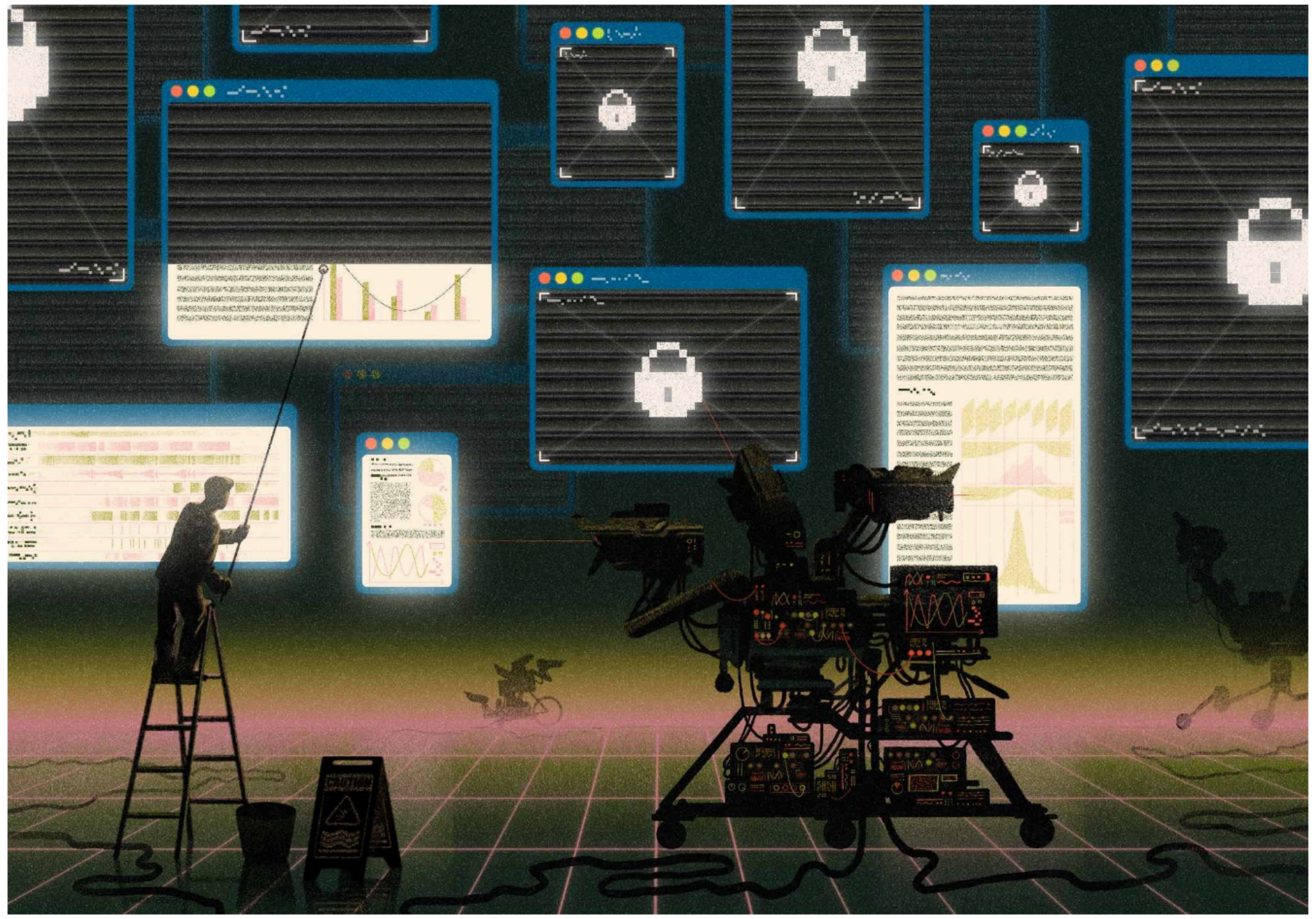
Deal reached in feud between California news outlets and Google: $250 million to support journalism but no new law

California news publishers want Google and other platforms to pay for the articles distributed on the platform. Above, a talk about Google News in 2018. (Jeff Chiu / Associated Press)

Some pushback

- “robots.txt” i.e. Robots Exclusion Protocol describes restrictions on who can access what on a website
- bot traffic jumped by 10x+ over past few years.
- 25% of high-quality websites blocked crawling in 2023-24 alone. So we can no longer replicate ChatGPT data.
- OpenAI and Anthropic seem to be ignoring this.

The data that powers AI is disappearing fast



Raven Jiang / New York Times

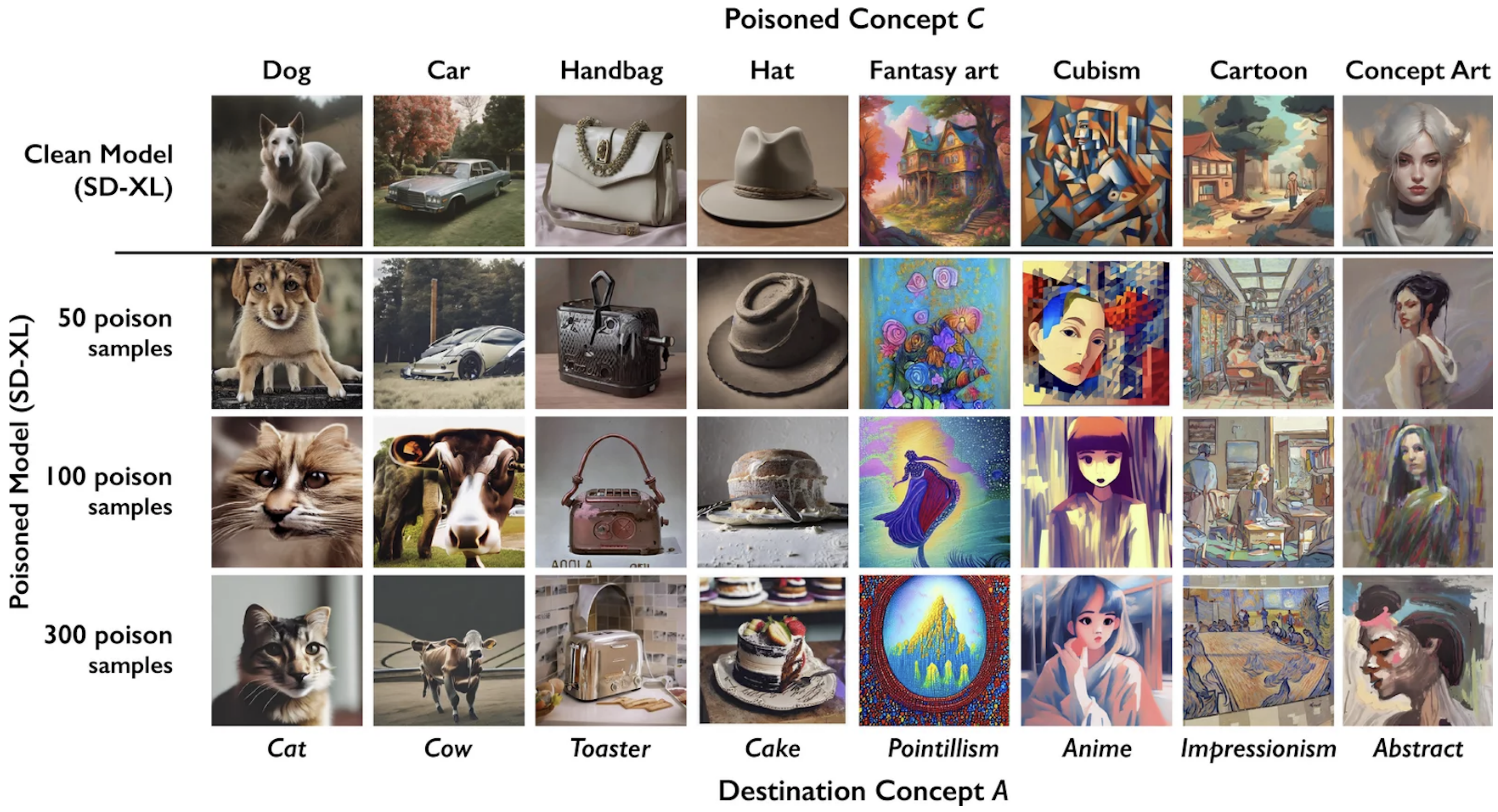
Exclusive: Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says

By **Katie Paul**

June 21, 2024 10:32 AM PDT · Updated 5 months ago

Some pushback

- Nightshade: Generate adversarial data points
- Undetectable to human
- But spoils the model when trained on it



Where does “ownership” stem from?

- **Copyright**: Protects expression of *creative* and *original* output (e.g. images or texts).
- **Patent**: Protects innovative processes or methods
- **Trademark**: Protects branding elements
- **Contracts**: agreements you enter when accessing services (e.g *Terms of Service)*

Copyright and GenAI

Copyright and AI

- taken in 2011 by a crested macaque named Naruto, using British photographer David Slater's unattended camera in Indonesia.
- Was uploaded to Wikimedia Commons image library in 2014
- In 2018 US court ruled that it is in public domain because “non-human” cannot hold copyright

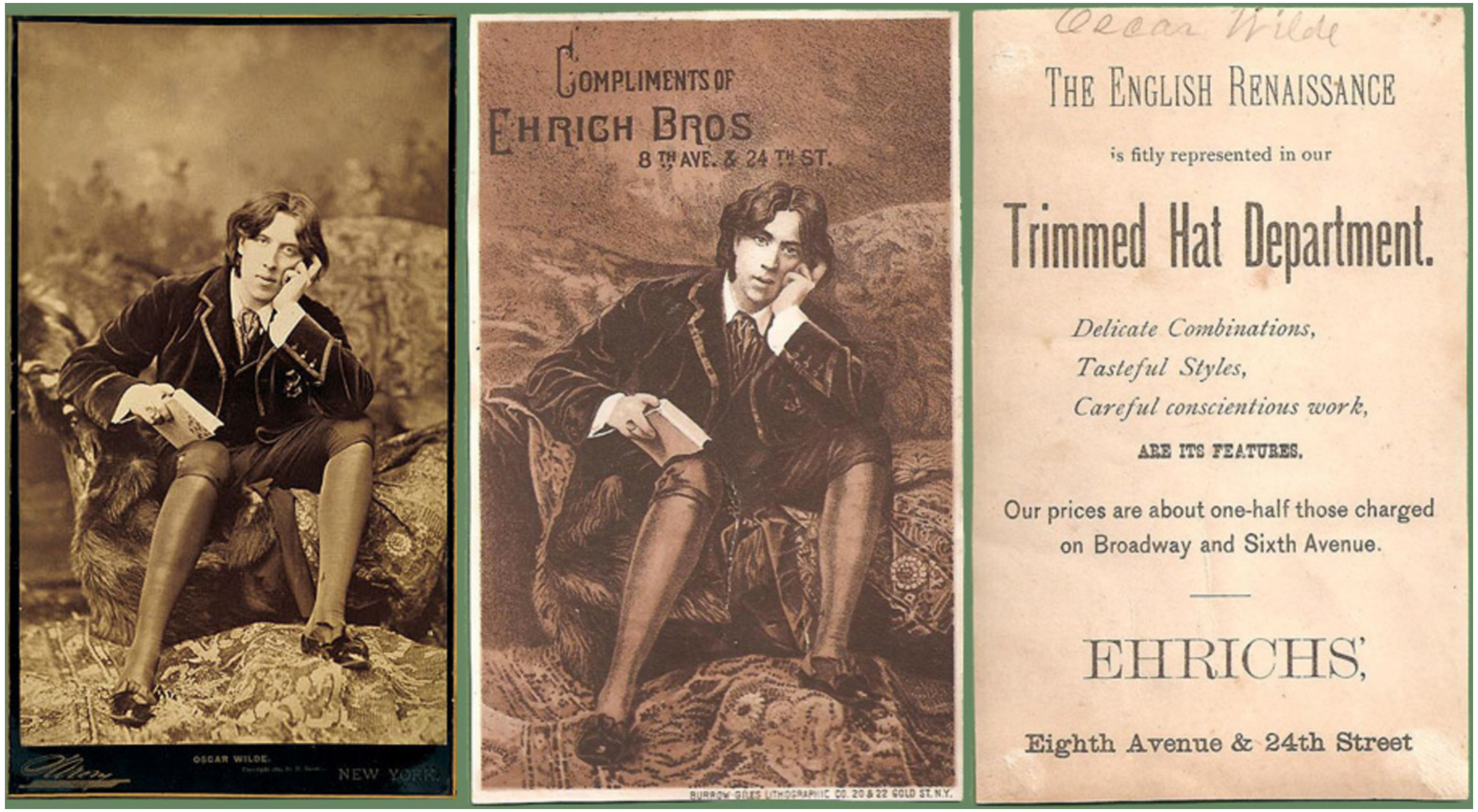
Copyright and AI

- AI-assisted comic book Zarya of the Dawn by author Kris Kashtanova

- US Copyright Office revoked copyright of images after it was found that they were generated by mid journey.

- *"We conclude that Ms. Kashtanova is the author of the Work's text as well as the selection, coordination, and arrangement of the Work's written and visual elements"*

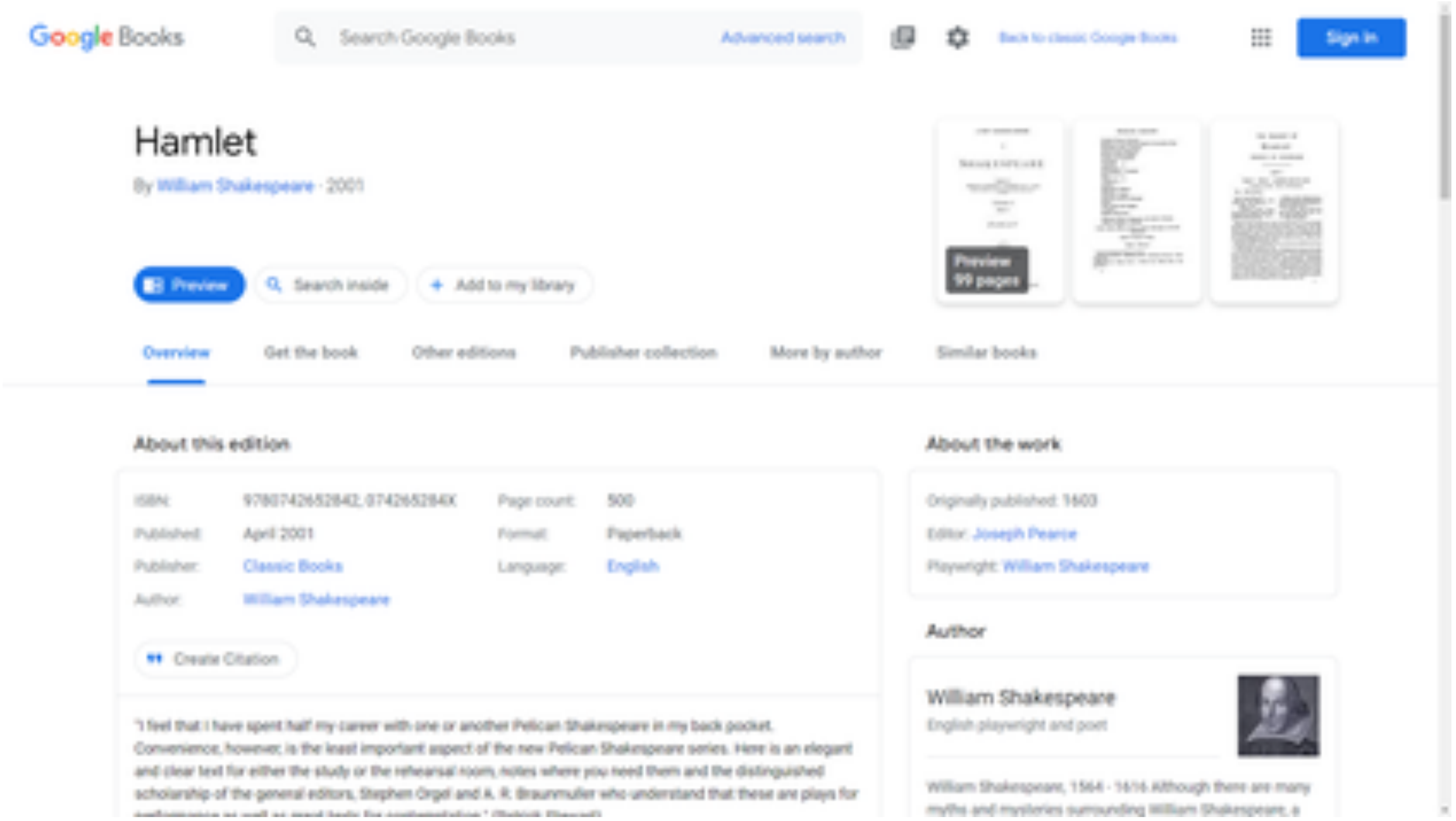
Copyright and AI



- Burrow-Giles Lithographic Co. v Sarony (1884)
- Does a photographer own the copyright of the picture they took?
- Photo has two parts:
    - Human participation: *creative decisions*
    - Tool use

What constitutes a copyright breach?

Original document	a) Exact match	b) Near-duplicate match	c) Semantically similar
Mrs Dursley had a sister called Lily Potter. She and her husband James Potter had a son called Harry Potter. They lived far from the Dursleys and did not speak to them much. They did not get along.	Mrs Dursley had a sister called Lily Potter. She and her husband James Potter had a son called Harry Potter. They lived far from the Dursleys and did not speak to them much. They did not get along.	Mrs Dursley had a sibling named Lily Potter. She and her spouse James Potter had a child named Harry Potter. They lived far from the Dursleys and did not speak to them much. They did not get along.	Mrs. Dursley's sister went by the name Lily Potter. Alongside her spouse James Potter, they parented a son named Harry Potter. They resided at a considerable distance from the Dursleys and seldom engaged in conversation. Their relationship was strained.



- Decided based on:
  - Similarity: what constitutes similar - closeness, length of match, etc.
  - Use: Market replacement? Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015)

How to protect against copyright breach?

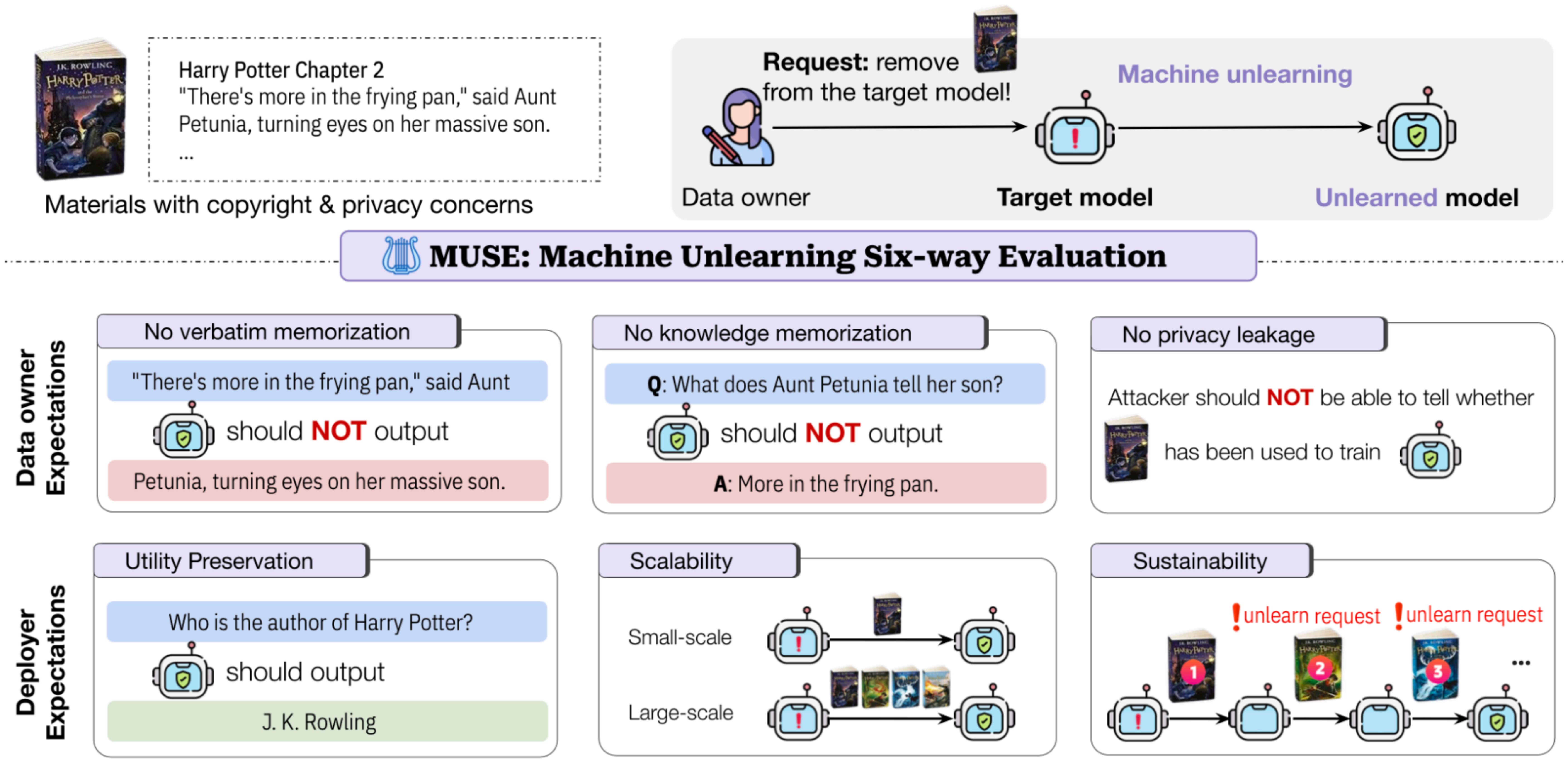
(a) Target copyrighted character: Mario

(b) Target copyrighted character: Batman

Figure 1: Examples of copyrighted characters generated by the open-source Playground v2.5 model (Li et al., 2024a) and proprietary DALL·E 3 model. The figures show Mario (a) and Batman (b), which can be generated with their names directly included in the prompt (*character name anchoring*, though **DALL·E 3 rejects the generation with its built-in guardrails** with messages like, “I can’t generate an image of Mario/Batman due to content policy restrictions”) or without their names using relevant keywords (*indirect anchoring*, still possible for DALL·E 3 despite its guardrails).

[He et al. 24]

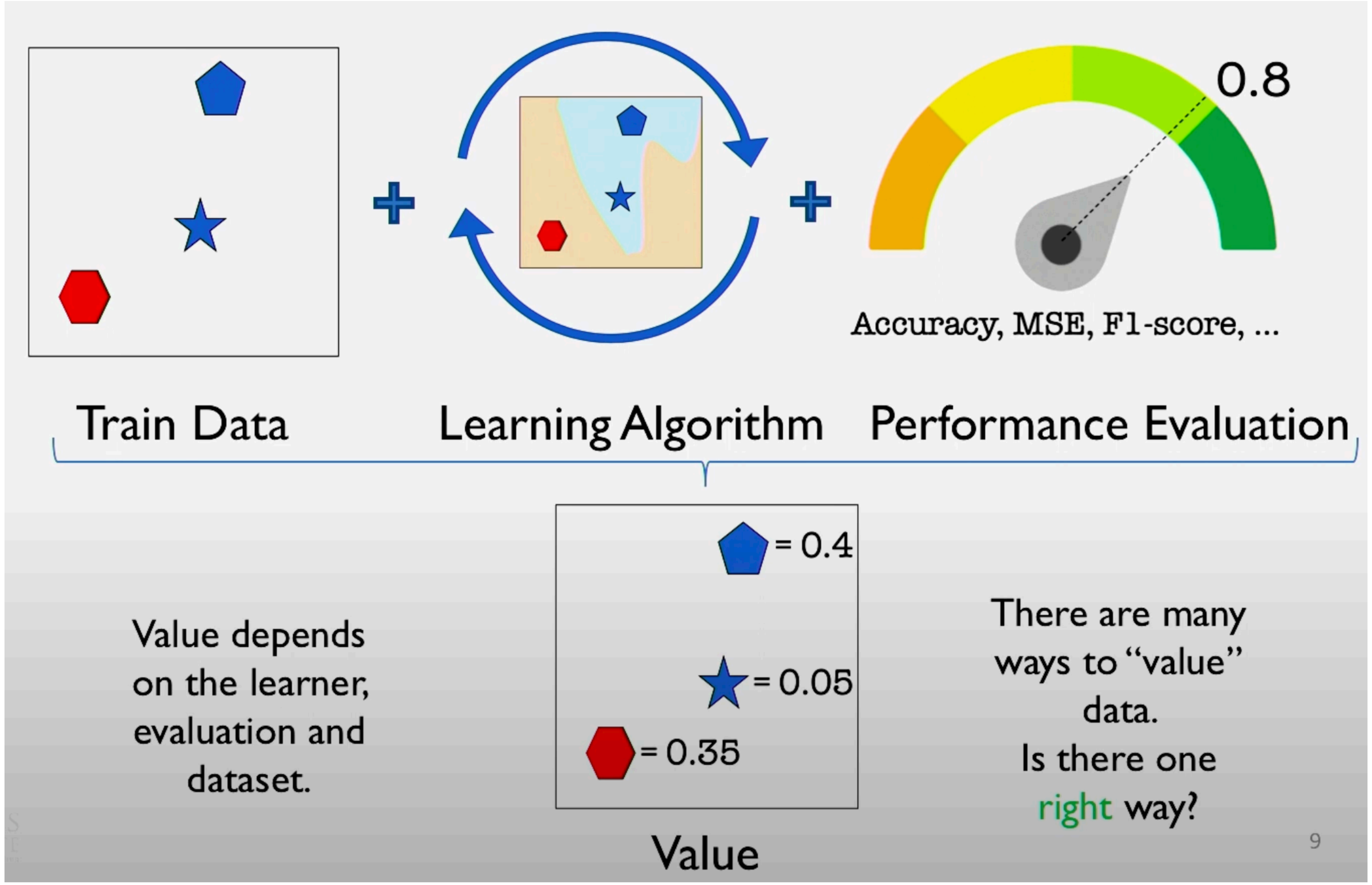
How to protect against copyright breach?



[Shi et al. 2024]

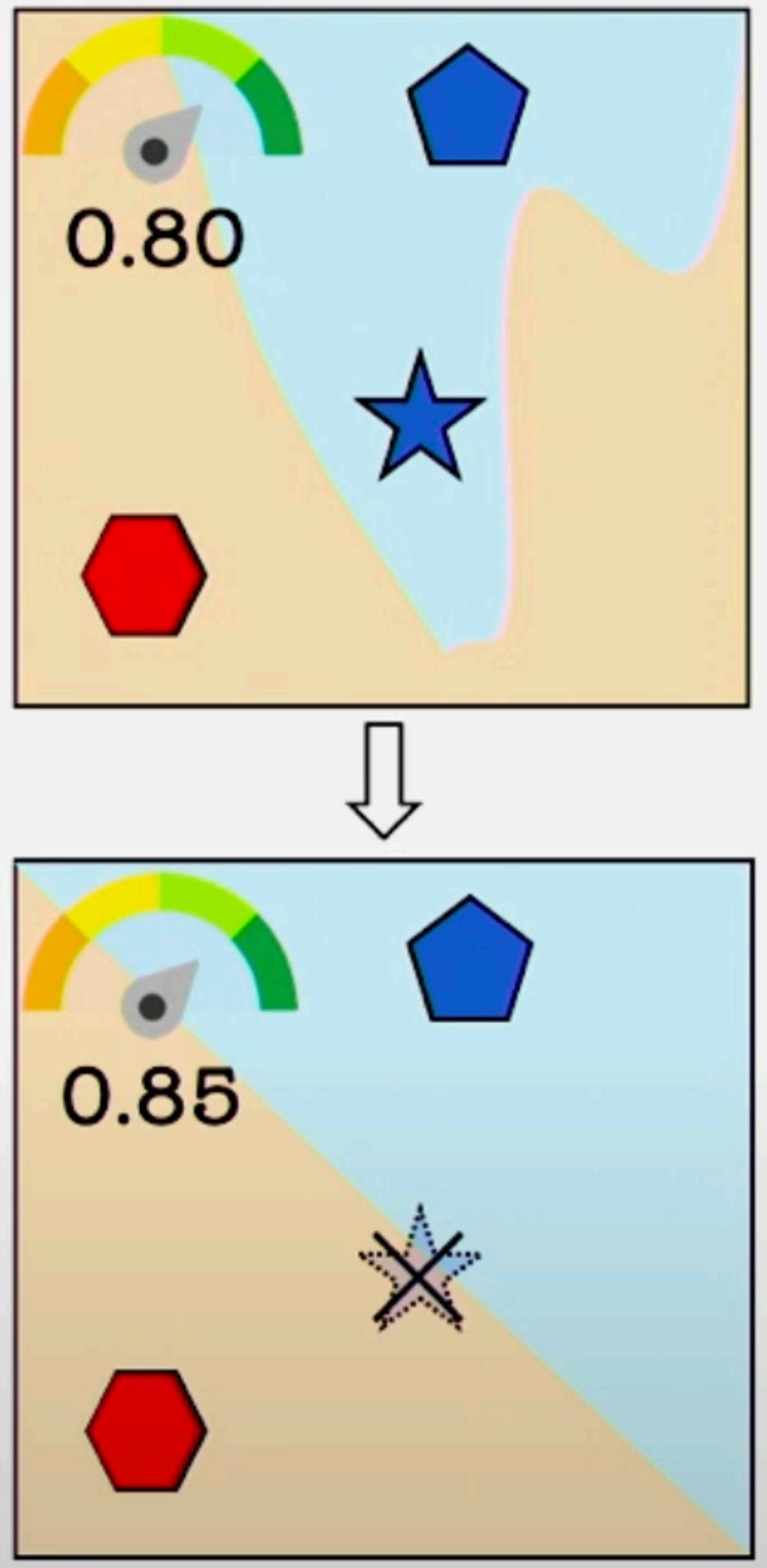
Data Valuation

Ingredients of data valuation



Leave one out (LOO)

<u>Example</u>: value (★) = 0.80 – 0.85 = -0.05



Widely used in statistics and ML.
Many variations and approximations:
leverage score, influence score, …

Does LOO capture the importance
of specific data?

Data Shapley Values: properties

1. <u>Null Element</u>: If adding ★ to any part of data never changes the learned model's performance:

value(★) = 0

2. <u>Symmetry</u>: If adding ★ or ⬟ to any part of data always results in the same performance:

value(⬟) = value(★)

3. <u>Decompostable:</u> In ML, performance metric can be the sum of performance on individual tasks (e.g. individual test)

$$\sum_i L(\text{classifier}(x_i^{\text{test}}), y_i^{\text{test}})$$

Add/remove the task ⟺ add/remove value(★) for that task.

[Ghorbani et al. 2019]

Data Shapley Values: properties

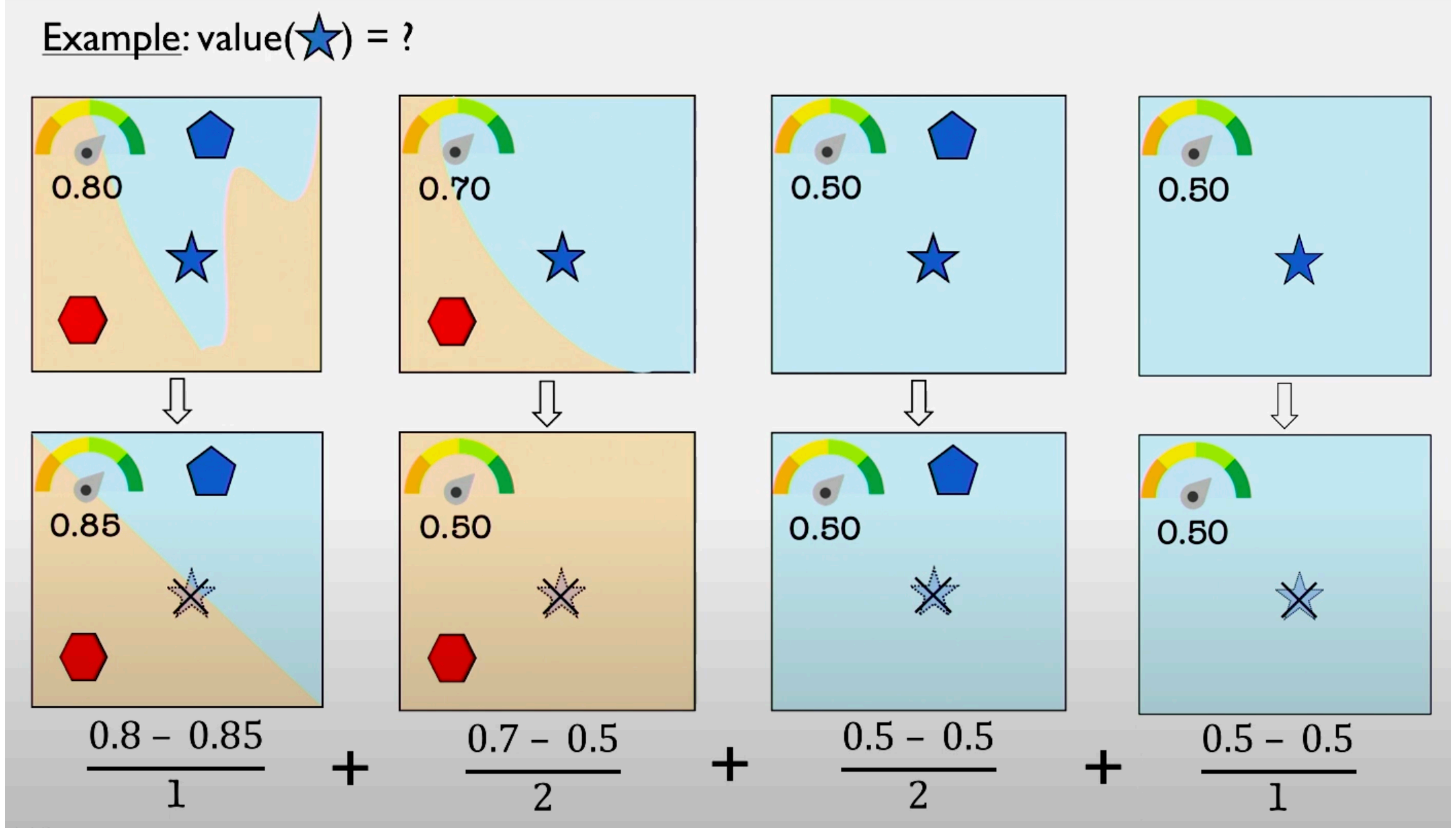
Theorem. The only data value that satisfies these properties is

$$\textit{Value}(\text{data k}) = \sum_{\text{subsets S not containing k}} \frac{\overbrace{\textit{performance}(S \cup k) - \textit{performance}(S)}^{\text{marginal contribution}}}{\underset{\#\text{ of size } |S| \text{ subsets}}{\binom{n-1}{|S|}}}$$

Expected contribution to all possible sizes of train data samples

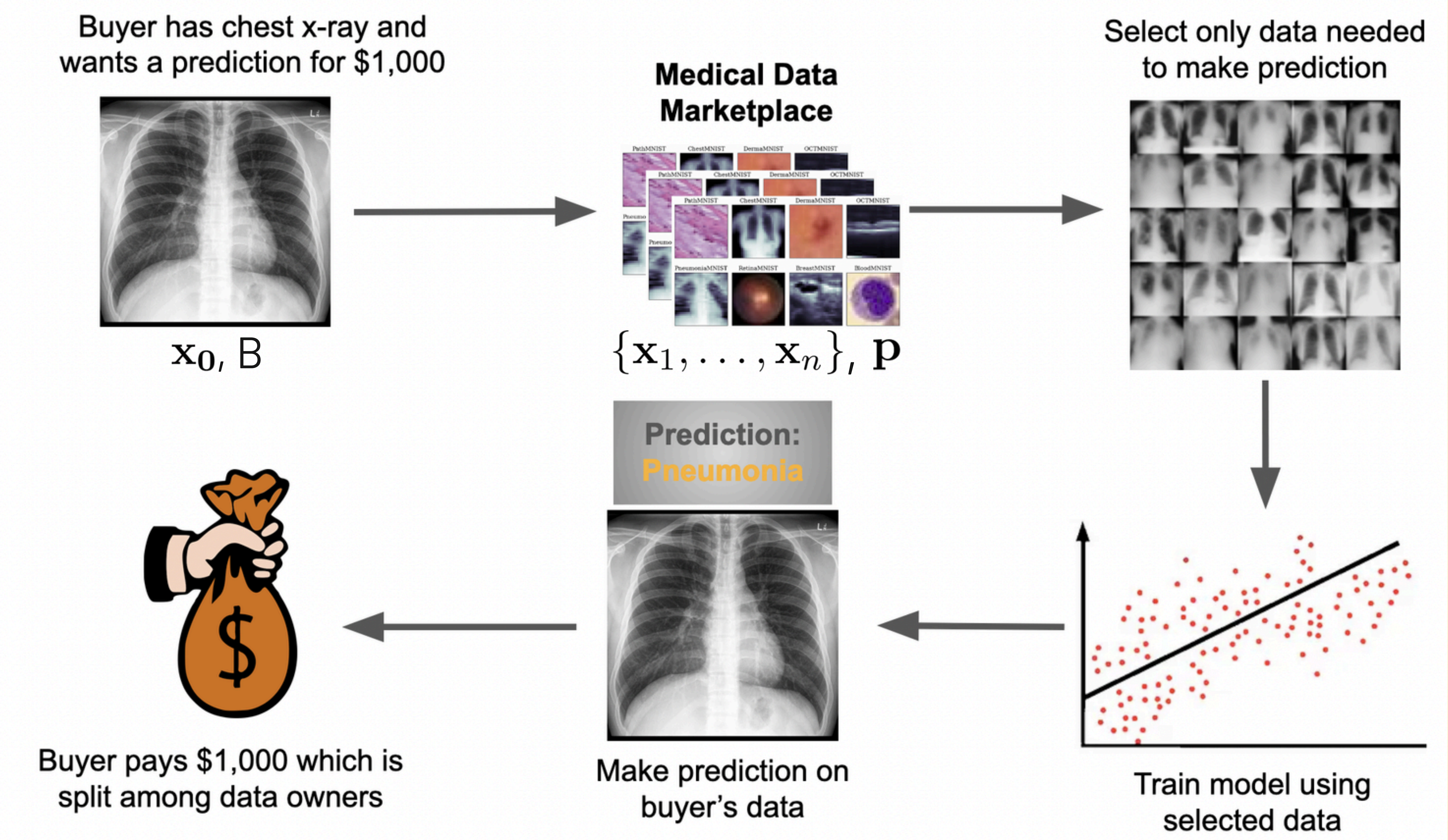
[Ghorbani et al. 2019]

Data Shapley Values



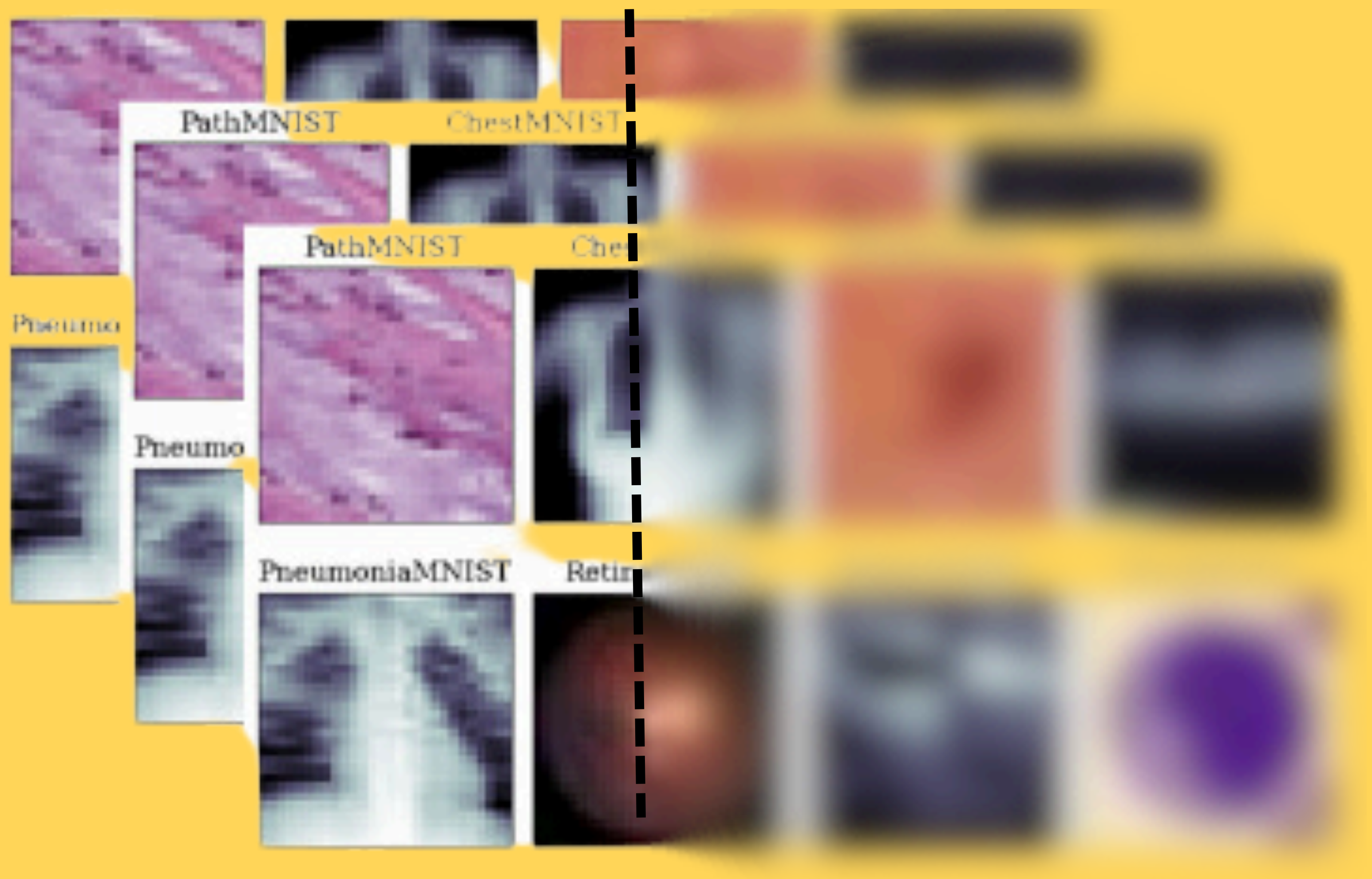
[Ghorbani et al. 2019]

Open data marketplace



- Buyer test data $\mathbf{x_0}$ and budget B
- n sellers' data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with prices $\mathbf{p}$
- Select best data within budget

Open or private?



- But, seller can’t reveal data
  - privacy concerns
  - IP concerns: data can be easily copied
- How can a buyer tell which data is useful for them?
- **Collaborative** data markets.

Collaborative data marketplaces

1

Model value of each seller datapoint

?

2

Collaborative data selection within budget

?

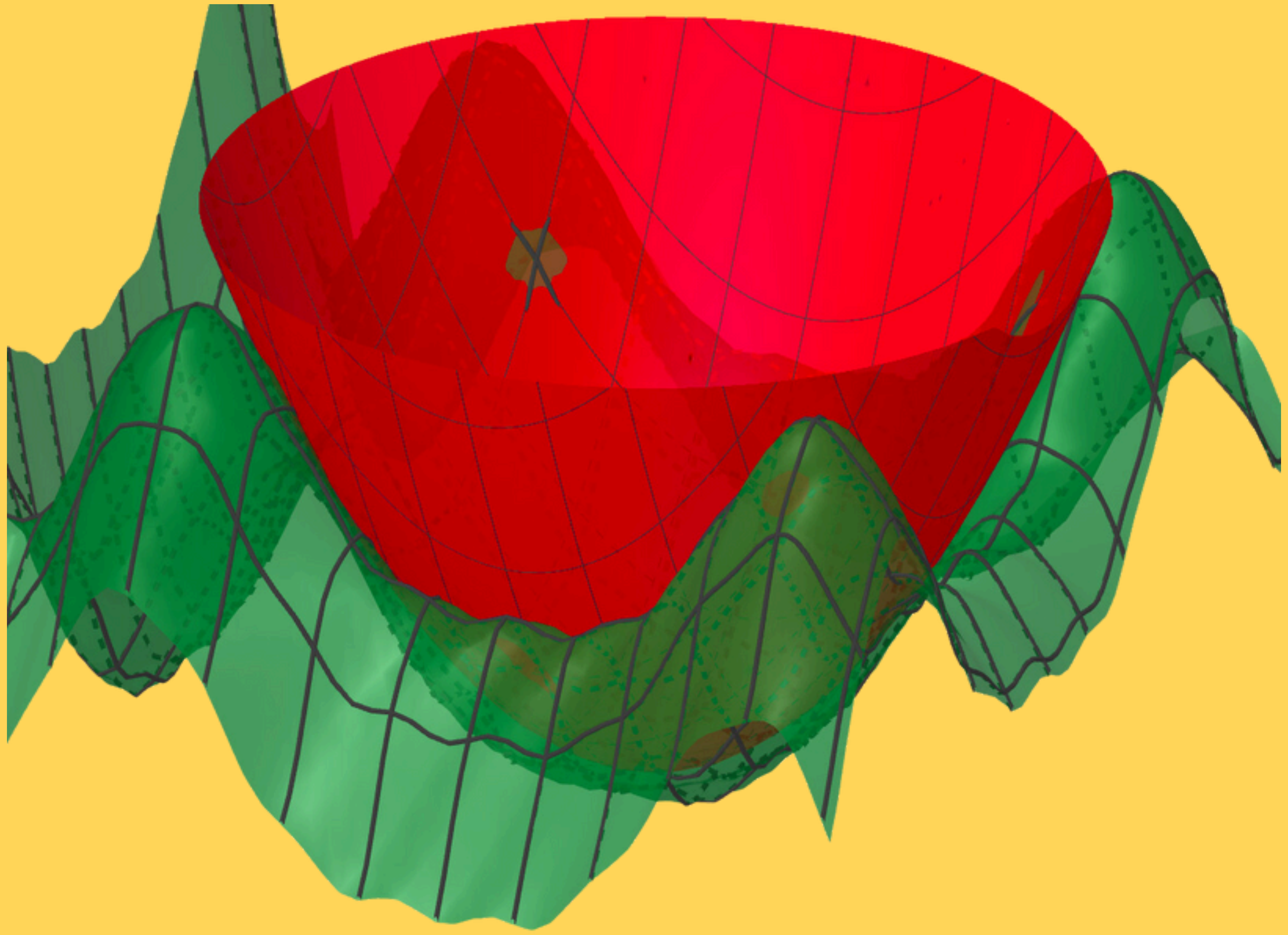
3

Train a model using CL on selected data

?

Collaborative data discovery

Modeling effect of training data on error



- Understanding effect of data on deep learning is hard!
- Construct a linear proxy-model using:
  - Neural Tangent Kernel (NTK)
  - Embeddings

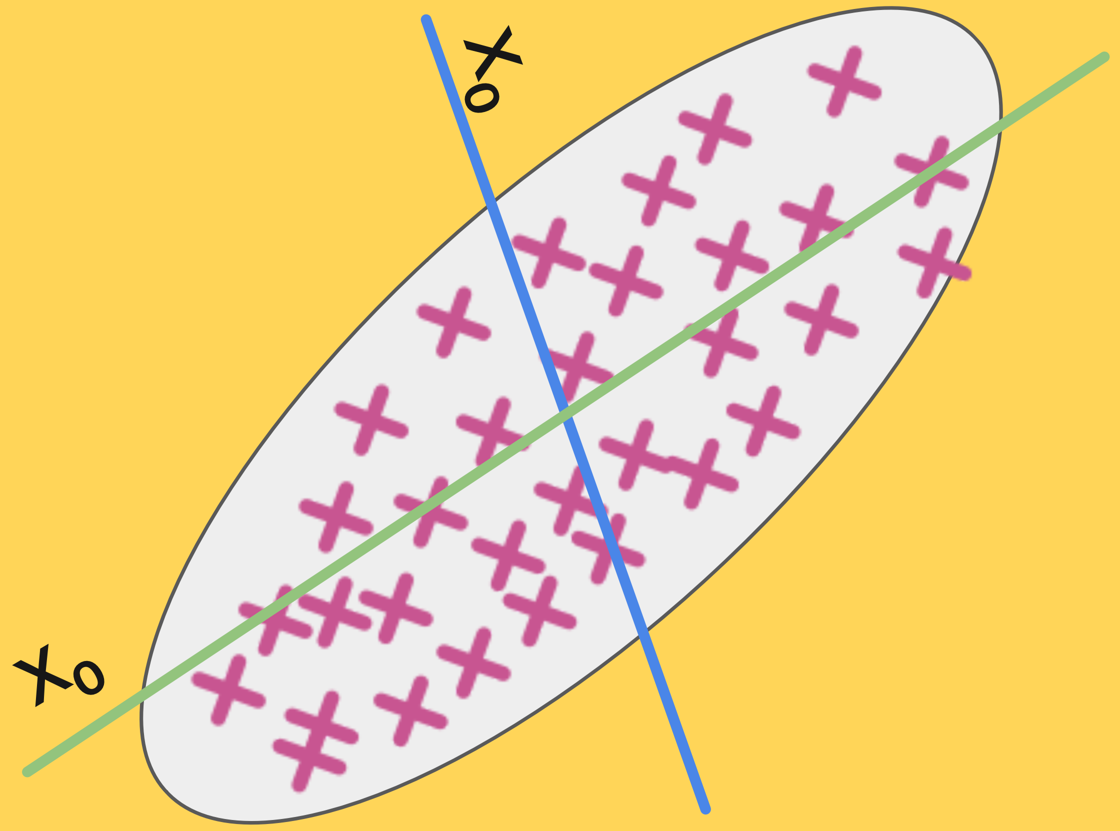
Linear Experiment Design

Estimating error

- Assume $y = \mathbf{x}^\top \theta^* + \text{iid noise}$
- Error on any test $\mathbf{x_0}$ determined by $\mathcal{I}$

$$\mathcal{E}(\mathbf{x}_0) = \mathbf{x}_0^\top \left(\sum_{i=1}^n x_i x_i^\top\right)^{-1} \mathbf{x}_0$$
$$= \mathbf{x}_0^\top \mathcal{I}^{-1} \mathbf{x}_0$$

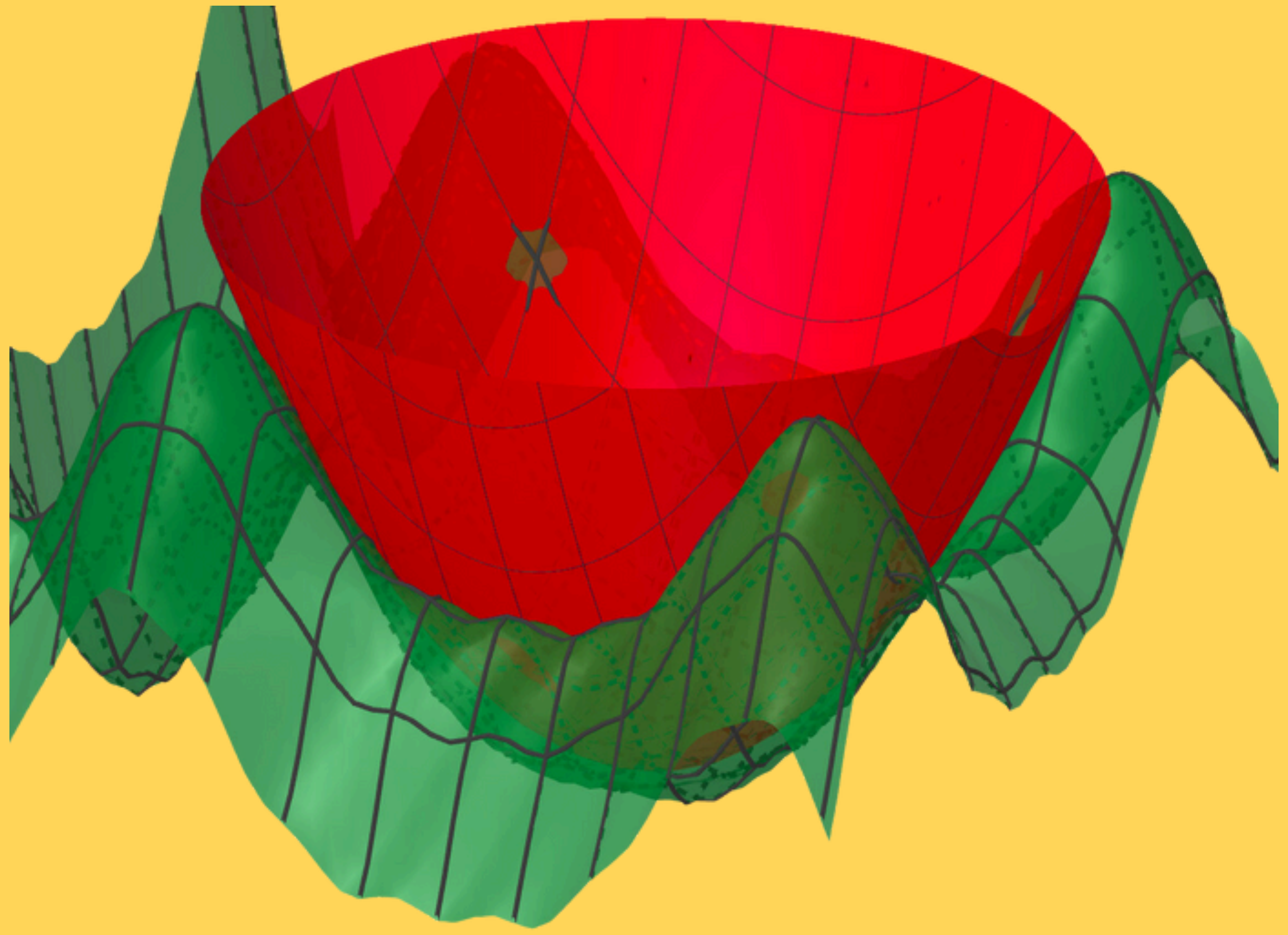
Error determined by information matrix

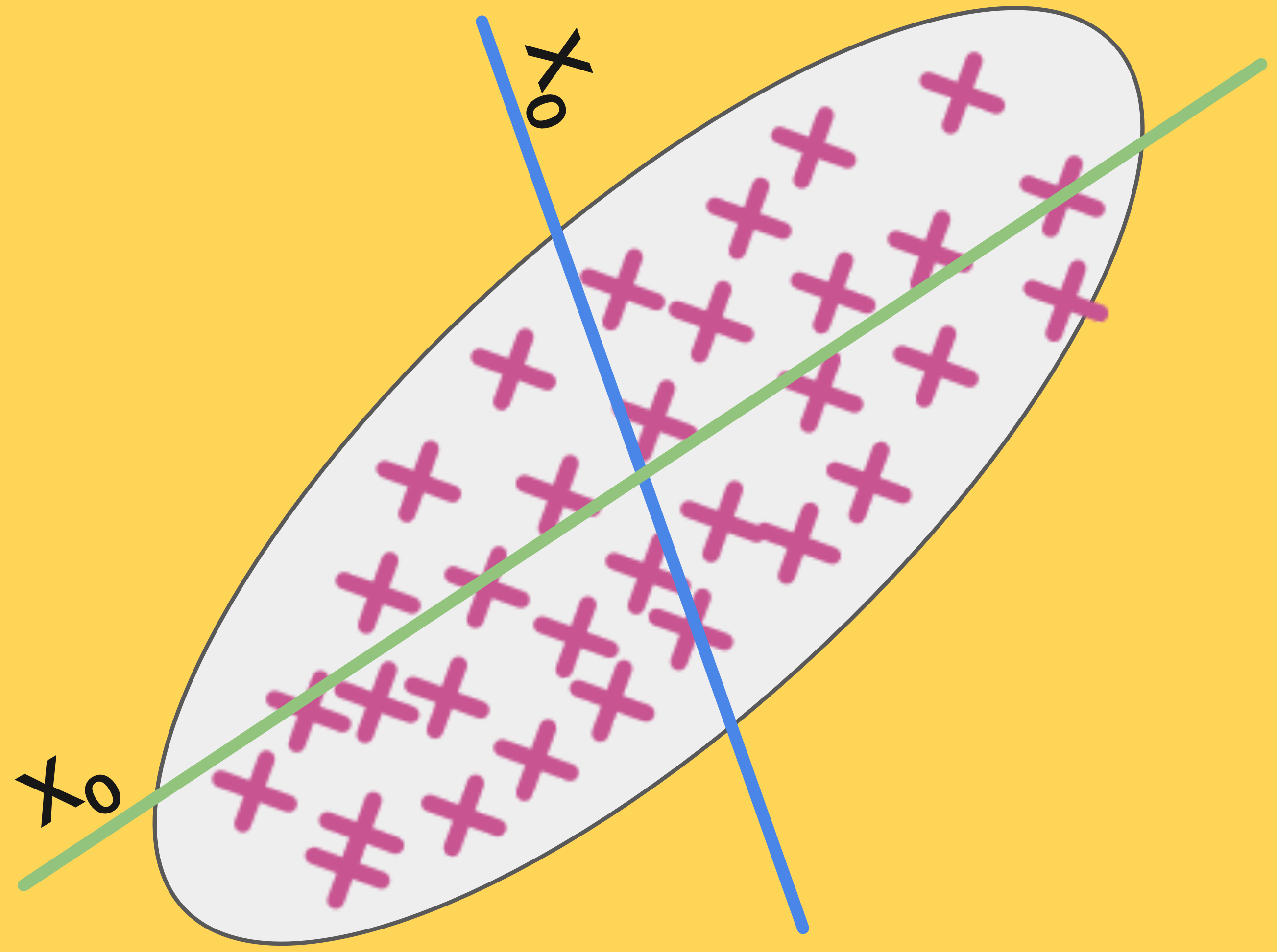


[Wald 1943; Pukelsheim 2006]

Collaborative data discovery

Modeling data value





a. Construct a linear proxy-model using

- Neural Tangent Kernel (NTK)
- Embeddings

b. Use information matrix to estimate error to buyer on $\mathbf{x_0}$

$$\min_{\substack{\mathbf{p}^\top \mathbf{w} \leq B \\ w_j \in \{0,1\}}} \mathbf{x_0}^\top \left(\sum_j w_j \mathbf{x_j} \mathbf{x_j}^\top\right)^\dagger \mathbf{x_0}$$

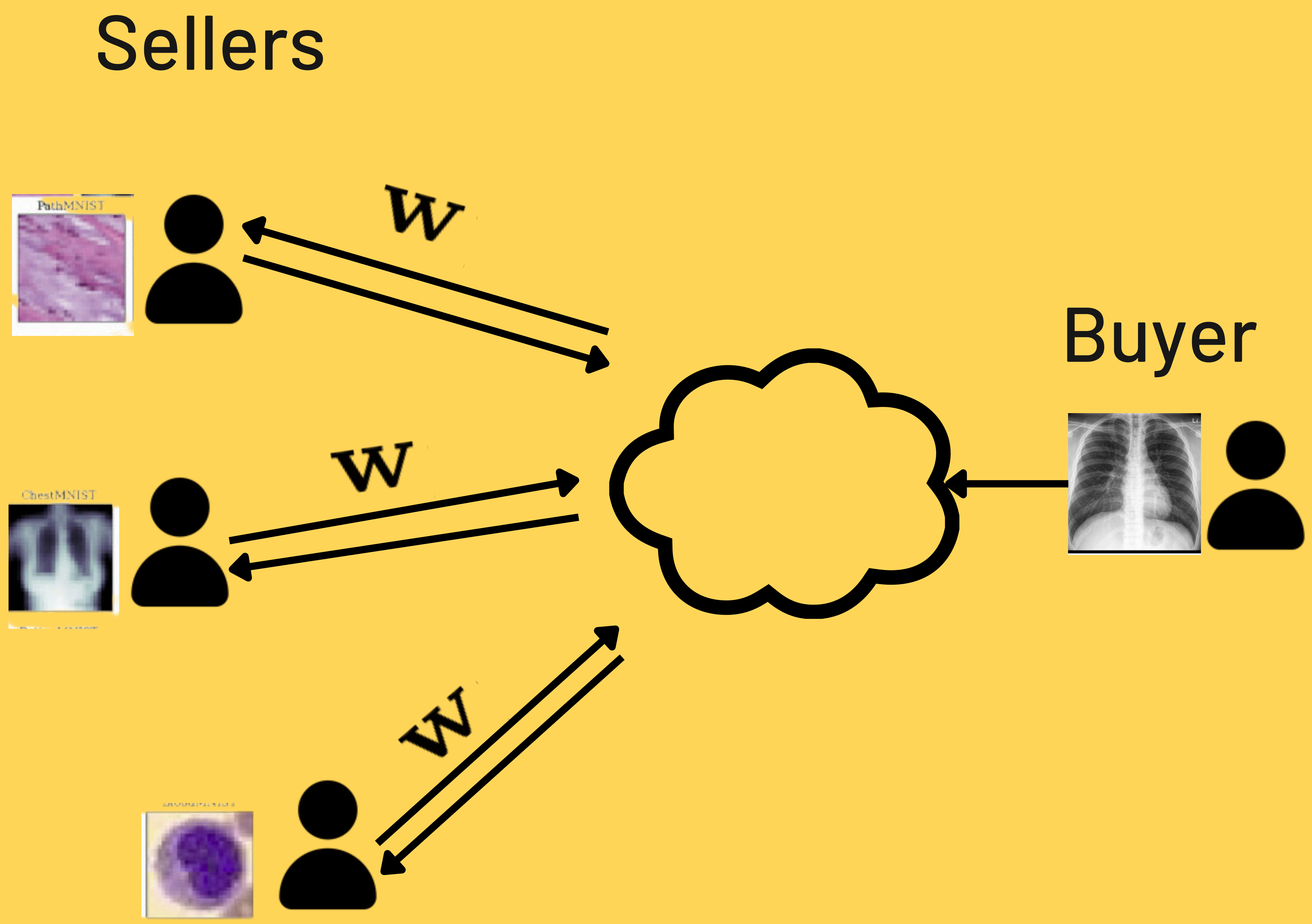
Collaborative data discovery

Iterative, collaborative algorithms

- Minimize error on buyer $\mathbf{x_0}$ within budget

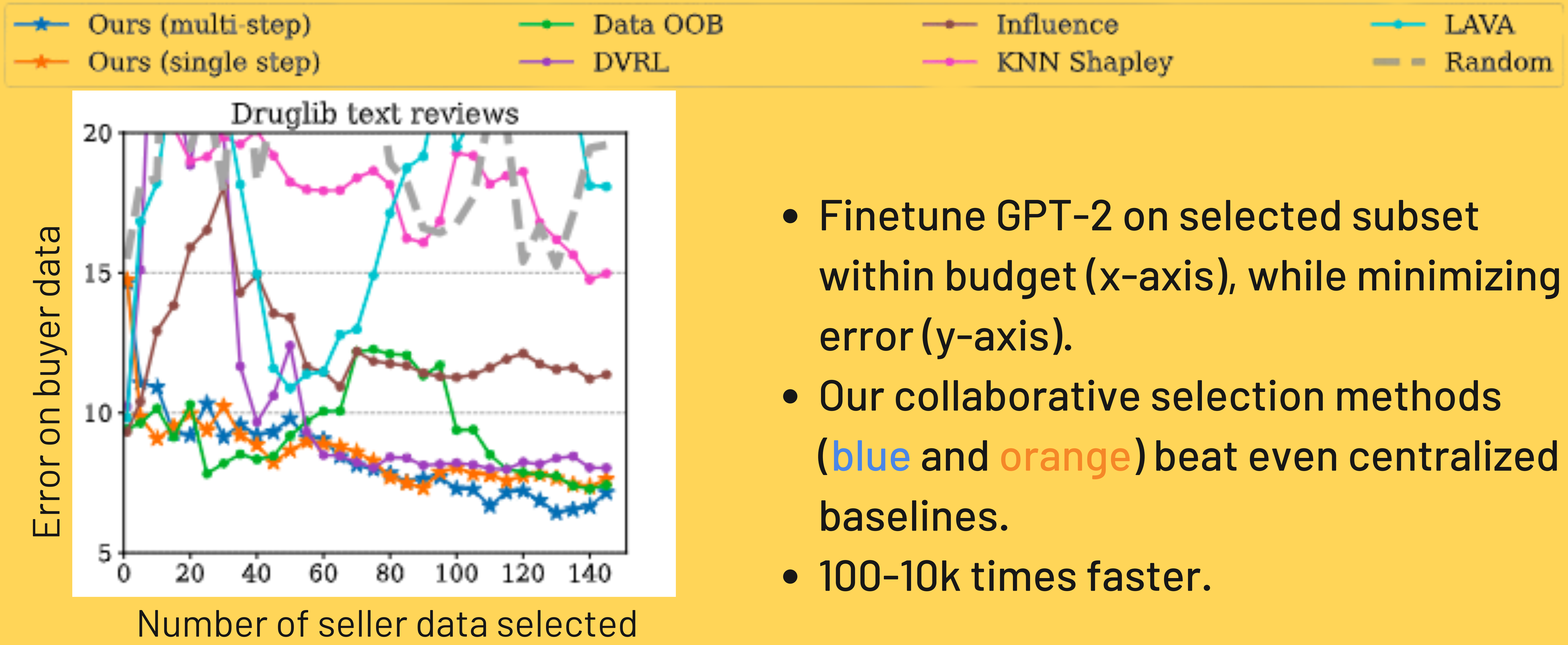
$$\min_{\substack{\mathbf{p}^\top \mathbf{w} \le B \\ \cancel{w_j \in \{0,1\}}}} \mathbf{x_0}^\top \left(\sum_j w_j \mathbf{x_j}\mathbf{x_j}^\top\right)^\dagger \mathbf{x_0}$$

- NP Hard => O(1/B) convex approximation
- Using ***collaborative conditional gradient*** [Frank and Wolfe 1956]



[Lu, Huang, *Karimireddy*, Jordan, Raskar NeurIPS 2024]

Collaborative data discovery
Results



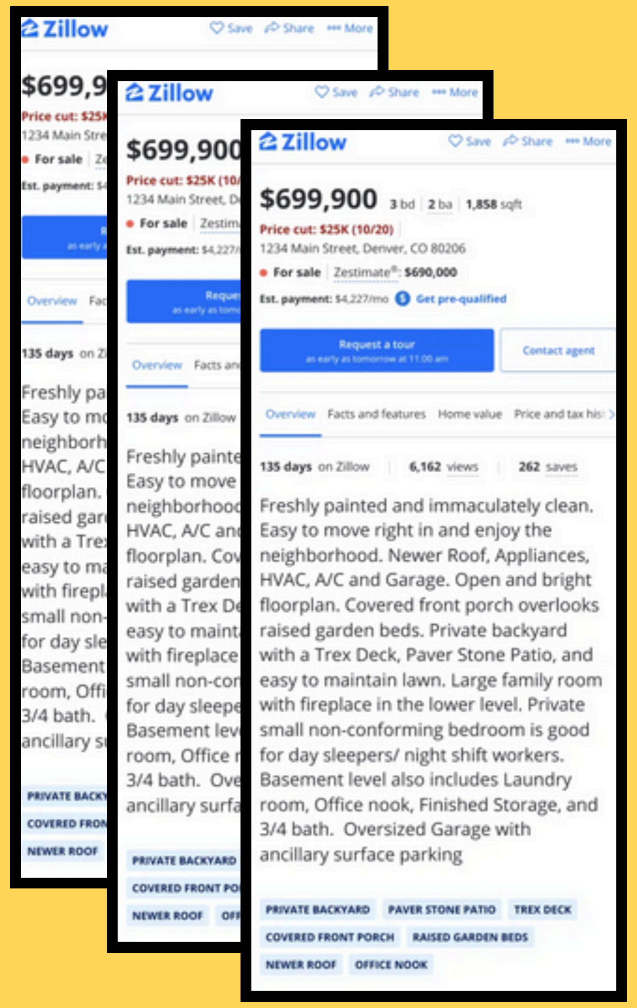
- Finetune GPT-2 on selected subset within budget (x-axis), while minimizing error (y-axis).
- Our collaborative selection methods (blue and orange) beat even centralized baselines.
- 100-10k times faster.

[Lu, Huang, *Karimireddy*, Jordan, Raskar NeurIPS 2024]

Better data understanding
Future work

Prompt
Which house should I buy?

ChatBot



Zillow listings

1. A theory of data utility
   - Beyond linear models
   - Statistically sound
   - Incentive compatible

2. What if data is manipulated or fake? (peer prediction/Bayesian persuasion + ML)

3. Establish authenticity & provenance (watermarks, ZKP)