



# CriteoPrivateAd Dataset



## Why a CriteoPrivateAd dataset?

- Criteo had been investing in Privacy Sandbox long-term vision
  - We saw the current version of the PSB as temporary, which should also allow for some privacy flaws
  - We worked on the next stages, like Bidding and Auction Service, and Private Model Training
- We think that future privacy-preserving advertising standards shall be based on **data-driven analysis**
  - We've released the **CriteoPrivateAd** dataset publicly and anyone can use it
  - Criteo dataset can be used to benchmark the performance of any Private Model Training system
  - It is thus possible to get mathematical results about both privacy and performance/market impact, and use those results for data-driven decisions

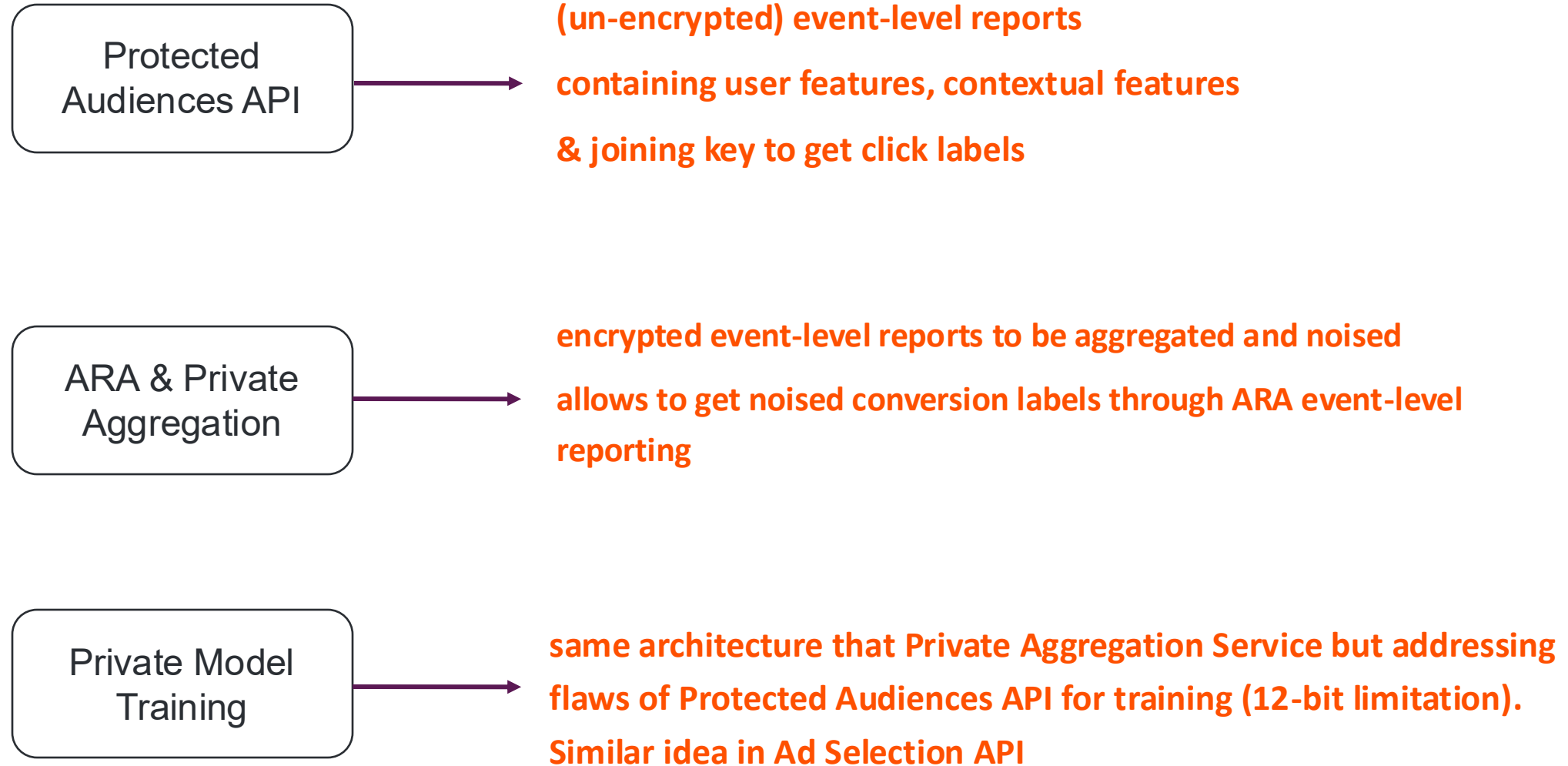
# CriteoPrivateAd dataset - Agenda

- Presentation of the status quo regarding ground truth to assess private advertising systems
- Presentation of the dataset
  - Description
  - Simulation of APIs
  - Use-cases beyond current Privacy Sandbox APIs
- Usefulness for other open web proposals
- Usefulness for the research community (academia)

## CriteoPrivateAd dataset – What is it about?

- Largest real-world anonymised bidding dataset, in terms of number of features, from Criteo production data
- ~ 150 features, 4 business labels (click, sales, ...)
- Additional data to simulate Privacy Sandbox APIs and beyond (user id, conversion delays, ...)
- Objectives:
  - assess the drop of performance associated to the removal of cross-domain user signals and hence illustrating one of the impacts of third-party cookie deprecation on adtech companies
  - design and test private bidding optimisation approaches leveraging both contextual signals and user features to predict click and conversion events
  - design and test the relevancy of answers provided by aggregation APIs for measurement and learning bidding models.

## CriteoPrivateAd dataset – Status quo




## CriteoPrivateAd dataset – Status quo

- Common benchmarks to assess the campaign optimisation use-case are depicted in Table 1
- None of them can be used to evaluate the impact of private model training on Click Through Rate prediction based on browser vendors APIs constraints.
- **[Feature tagging]** No sufficient information about features types (e.g., contextual, cross-domain, single-domain), availability at inference time
- Public challenges as Criteo AdKDD 2021 are great but huge investment for a one-time impact

Data	Rows	Features
KDD Cup Track 2 ( <a href="#">KDD Cup 2012 Track 2</a> )	5M	Search engine context, with queryn ad features, user id and click information.
Criteo-Kaggle Display Advertising Challenge ( <a href="#">Tien et al., 2014</a> )	100K	Over 7 days of live traffic, 39 features hashed and fully undisclosed, click label.
Avazu dataset ( <a href="#">Avazu, 2015</a> )	40M	Over 11 days of live traffic, 9 anonymised features, 6 contextual features, 5 user device features, click label.
Criteo 1TB Click Logs dataset ( <a href="#">Lab, 2023</a> )	4B	Over 7 days of live traffic, 39 features hashed and fully undisclosed, click label. Similar to the dataset in Criteo-Kaggle-Display challenge.
Criteo Attribution Modeling for Bidding Dataset ( <a href="#">Diemert et al., 2017</a> )	16.5M	Over 30 days of live traffic, 9 contextual features, attribution data, user and campaigns ids, click and conversion labels.

Table 1: Public Datasets for Bidding.

# CriteoPrivateAd dataset – Dataset Description

 **Hugging Face**

Models

Datasets

Spaces


Posts

Docs

Enterprise

Pricing

⌵



Datasets:

🔖

criteo

CriteoPrivateAd

📄

📖

like

2

Following

🔖

CRITEO

28

Tasks:

📄

Tabular Classification

📈

Tabular Regression

Size:

10M<n<100M

ArXiv:

📄

arxiv:2502.12103

📄

arxiv:2201.13123

Tags:

criteo

advertising

License:

📄

cc-by-sa-4.0

Dataset card

Data Studio

Files and versions

👤 Community 3

⚙️ Settings

👁️ Dataset Preview

ⓘ

</> API

📄 Embed

📄 Data Studio

Split (1)

train

⌵

▶ The full dataset viewer is not available (click to read why). Only showing a preview of the rows.

id	user_id	display_order	sale_delay_after_display_arra
string	string	int32	sequence
acf2f01356f857ea7aa0cfdb2b6bb987	6b5523637cbd8b972e1850e4cc344b14	8	nul
d5680c5d388ab7a50c766d62e6162cd2	a80b944919728985bf8c66127e1af158	4	nul
6c75e878e5c1f57ab8feb1964587b483	ca5365463da260d415eea465255cb8a6	1	nul

Downloads last month

12,454

View full history

📄 Edit dataset card

⋮

## CriteoPrivateAd dataset – Overview

- 100M displays spanning 30 consecutive days of data
- More than 100 relevant features to learn common bidding models
- Binary labels (click, landed click, visit)
- Integer labels: number of sales attributed to a specific display
- Data retrieved from third-party cookie traffic on Chrome and hence not necessarily representative of Chrome Protected Audiences traffic. These choices have been made in order to have a sufficient number of positive examples enabling to learn relevant bidding models
- Sufficient data to get real-world production performance (+/- 5%)



## CriteoPrivateAd dataset – Risk Mitigation

- Feature values are anonymised:
  - Hashing for categorical features
  - Use of a monotone transformation for continuous features
- Major part of the features' names are not available
- Labels in this dataset have been sampled non-uniformly from our production logs. Such labels could be sub-sampled from the released dataset to meet a specific business key performance indicator relevant for the advertising industry, e.g., a target click-through rate (CTR) representative of online traffic
- Some features used in production have been removed from this dataset without compromising its relevance to estimate bidding model performance

## CriteoPrivateAd dataset – Ensuring production-friendly performance

- **Sub-sampling of negative examples** to ensure that the number of clicks remains sufficient for achieving representative performances, leading to a click-through rate  $CTR \approx 0.36$  and a conversion rate  $CVR \approx 0.02$ , where a conversion here means a positive sale event (i.e., at least one product has been bought by the user)
- Hence, it is important to rescale validation metrics for interpretable results, close to production performance
- We are providing a simple re-scaling formula taking as input:
  - The target business KPI (e.g.  $CTR = 1\%$ )
  - The dataset business KPI (e.g.  $CTR = 36\%$ )

## CriteoPrivateAd dataset – Ensuring production-friendly performance

<b>Task \ Target CTR</b>	<b>0,1%</b>	<b>0,5%</b>	<b>1%</b>
Landed Click   Display	0.170	0.186	0.234
Sales   Landed Click	0.218	0.218	0.218
Sales   Display	0.171	0.187	0.237

Table 2: Rescaled LLH of baseline bidding models for a given target CTR and a given task. Learned from day 1 to 25, validated from day 26 to 30.

## CriteoPrivateAd dataset – Dataset deep-dive

- Features used to train bidding models are **grouped into five buckets** with respect to their logging and inference constraints in Chrome Protected Audiences API
- We chose the latter API as a reference for building this dataset as it stands for the most mature API to train bidding models without third-party cookies
- This feature bucketisation into five groups can be easily relaxed to other browser vendors proposals or private learning frameworks as we are providing the precise semantics of such bucketisation

## CriteoPrivateAd dataset – Dataset deep-dive

- **id:** the id of the row
- **uuid:** user id **consistent over the day**. The same user will have two different user ids for two different days. To clarify the difference between uuid and a device id, note that the data in CriteoPrivateAd comes mainly from the third-party cookie in the Chrome instance, and the rest from aggregated data from the 3rd-party cookie and app data from the same device (cookie id is matched with gaid from apps)
- **campaign id:** the id of the advertising campaign associated to the display, which could be used as a proxy for the advertiser id
- **publisher id:** the id of the publisher on which the display has been made

## CriteoPrivateAd dataset – Dataset deep-dive

- **features\_kv\_bits\_constrained ( $\approx 30$  different features):** single-domain (a domain refers to a website) user features. In Protected Audiences API of Chrome Privacy Sandbox, these features can be encoded in the modelingSignals field, subjected to the 12-bit constraint and available in the key-value server at inference time
- **features\_kv\_bits\_not\_constrained ( $\approx 10$  different features):** all the features derived from Interest Group (IG) name / renderURL, that is all ad features available in reportWin field outside of modelingSignals field in Protected Audiences API setting. These features are available in the key-value server at inference time and do not have any logging constraint (e.g., 12-bit encoding)
- **features\_browser\_bits\_constrained ( $\approx 10$  different features):** all cross-advertiser user features available in generateBid in Protected Audiences API setting. This includes features that can be encoded in recency and joinCount fields. These features can be logged in modelingSignals field but are only available in the browser at inference time
- **features\_ctx\_not\_constrained ( $\approx 10$  different features):** all features only available in the contextual call with no logging constraints. These features stand for contextual ones
- **features\_not\_available ( $\approx 80$  different features):** all cross-domain user features

## CriteoPrivateAd dataset – Dataset deep-dive

- **is clicked:** binary label indicating whether the display has been clicked or not
- **is click landed:** binary label indicating whether the click has been observed on the advertiser website
- **is visit:** binary label indicating whether the user interacted with the advertiser website after a landed click (at least one advertiser event after the landing event)
- **nb sales:** number of sales attributed to the clicked display
- The **is sale** label has to be created with  $\text{binarise}(\text{is visit} * \text{nb sales})$

## CriteoPrivateAd dataset – Dataset deep-dive

- **is clicked:** binary label indicating whether the display has been clicked or not
- **is click landed:** binary label indicating whether the click has been observed on the advertiser website
- **is visit:** binary label indicating whether the user interacted with the advertiser website after a landed click (at least one advertiser event after the landing event)
- **nb sales:** number of sales attributed to the clicked display
- The **is sale** label has to be created with  $\text{binarise}(\text{is visit} * \text{nb sales})$

Additional fields allowing to emulate custom report windows in ARA event-level and aggregated reports



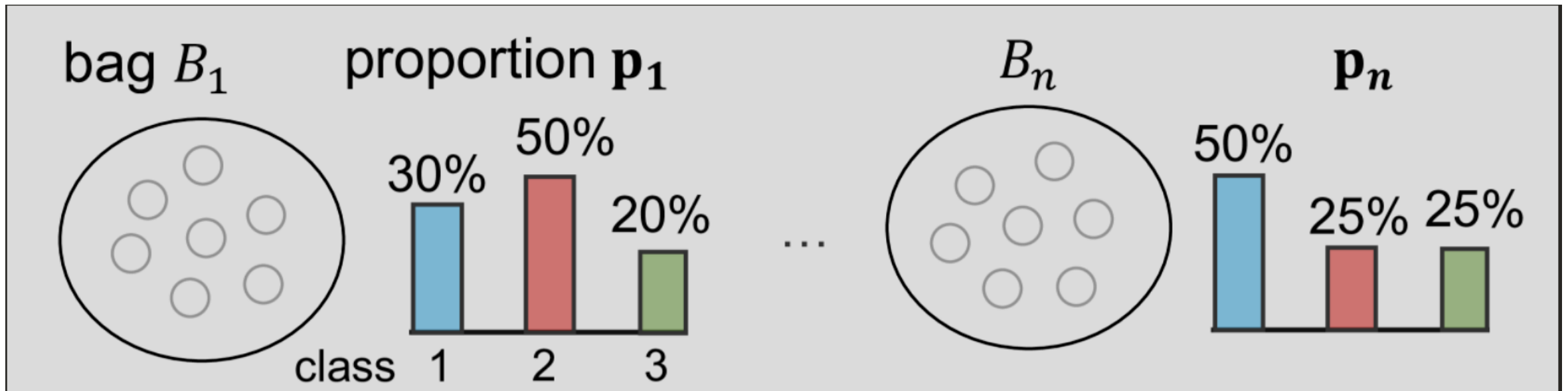
## CriteoPrivateAd dataset – Offline performance metrics

- **Log-likelihood:** standard machine learning metric
- **calibration:**  $\# \text{ predictions} / \# \text{ true positive events}$
- **/ ! \ Classical classification metrics (e.g. AUC or F1-score) could lead to misleading conclusions on performance**

# CriteoPrivateAd dataset – Simulating Aggregation APIs

- **Two constraints:**
  - Reporting delays
  - Global & local differential privacy

Can be used to learn bidding models from "noisy label proportions"



# CriteoPrivateAd dataset – Private Model Training

- **Two constraints:**
  - Training outsourced to a trusted server (backend: TEE)
  - Global differential privacy

**Noise is added inside the training process or afterwards (to the learnt model weights)**

## CriteoPrivateAd dataset – Usefulness for other open web proposals

- User-level differential privacy currently researched in W3C work (cf. Cookie Monster), and is also a request from Chrome
- Features' tagging versatile enough to split contextual and user features
- Can also be used as a ground truth for B&A (size of the model to be sent to the TEE)

## CriteoPrivateAd dataset – Usefulness for the research community

- Differential Privacy is the workhorse approach to design private machine learning methods
- Learning from (noisy) label proportions – cf Aggregation APIs
- Private Model Training (DP-SGD) is the most used approach for both traditional machine learning and more advanced (deep learning and generative AI)

## CriteoPrivateAd dataset – Key take aways

- This dataset is also intended to strengthen and accelerate collaboration with W3C PATWG & other industry proposals
- We plan on communicating later this year on training performance of learning from an Aggregation Service for bidding optimisation



**Any other question?**

## CriteoPrivateAd dataset – User-level versus Display-level DP

- This dataset being sampled at the display level, it can be naturally used to design and test private bidding model training strategies based on display- level differential privacy (also coined item-level DP in the academic literature)
- Albeit we are providing a user identifier, this dataset cannot be used directly, without involving some bias, to assess the relevancy of user-level DP frameworks.
- To tackle this issue, we are providing two ways to use this dataset in a user-level DP context:
  - CriteoPrivateAd might be representative to what is sent to a private aggregation server
  - CriteoPrivateAd could be aligned to the true user distribution



## CriteoPrivateAd dataset – User-level versus Display-level DP

