CRITEO

Advertising on the open internet under privacy constraints

Maxime Vono

Staff Researcher Lead

Privacy-preserving ML

Outline

Context

Novel learning paradigms & challenges

What we do

Outline

Context

Novel learning paradigms & challenges

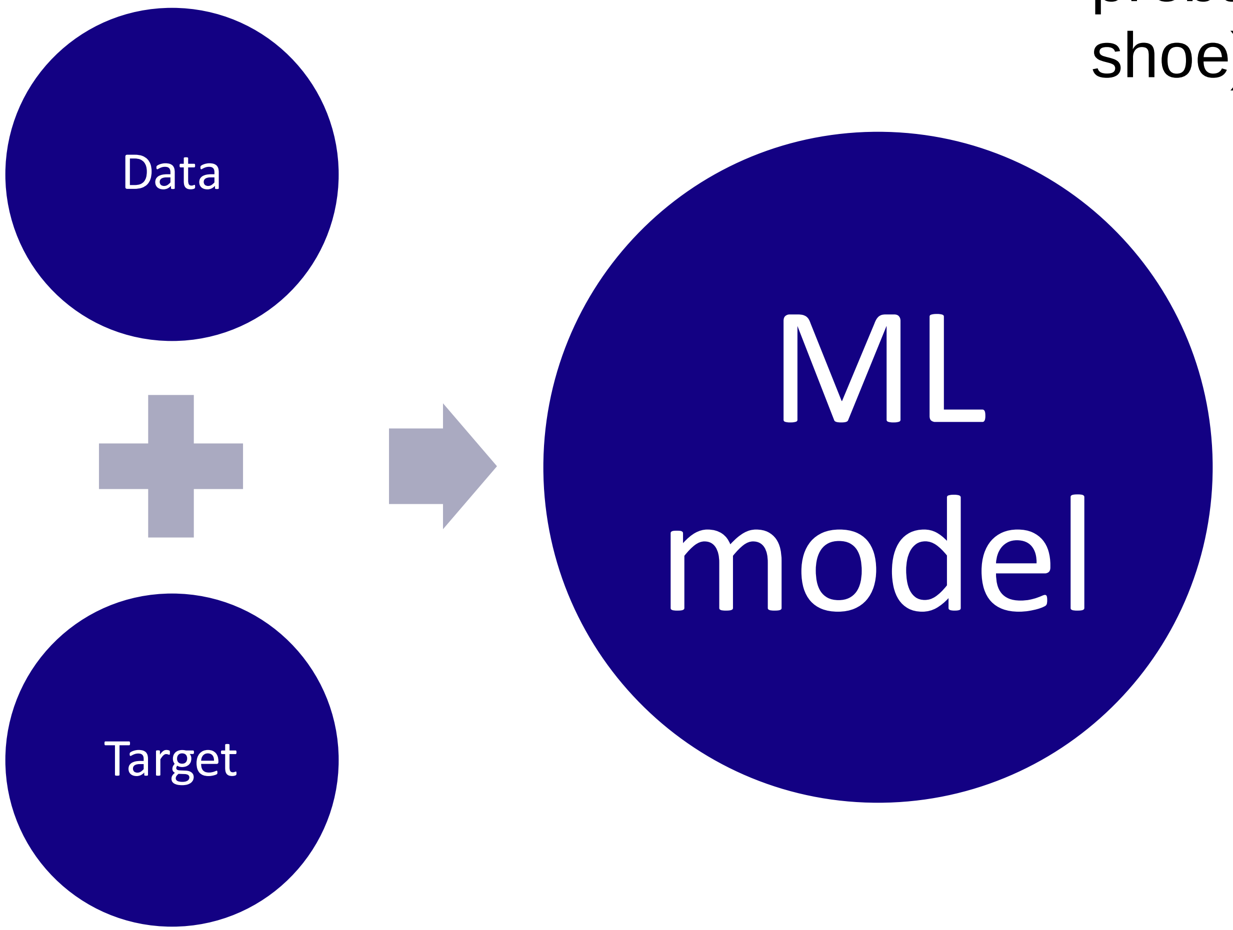
What we do

Supervised Machine Learning - 101

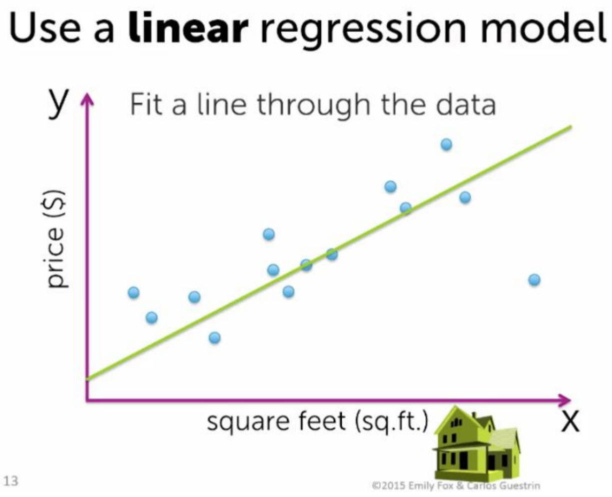
The model learns to predict the target (e.g. probability to buy a Nike shoe) from historical data

Learning / Training

Historical data
(e.g. : city, number of bedrooms, floor number, …)



Observed value
(flat price)



Source:

CRITEO

Supervised Machine Learning - 101

The model learns to predict the target (e.g. probability to buy a Nike shoe) from historical data

Learning / Training

Historical data
(e.g. : city, number of bedrooms, floor number, …)

Data

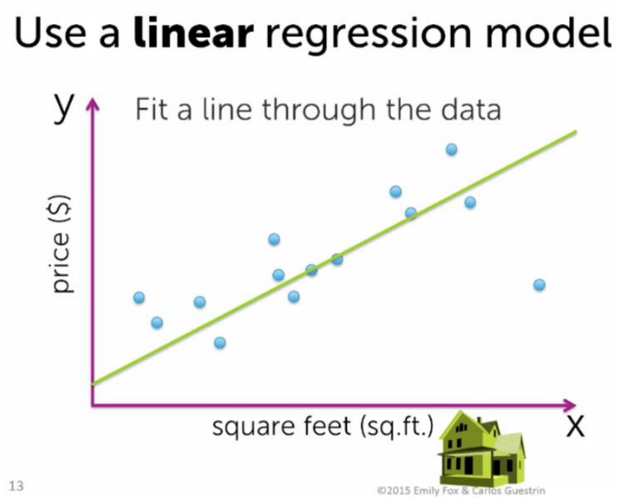
+

Observed value
(flat price)

Target

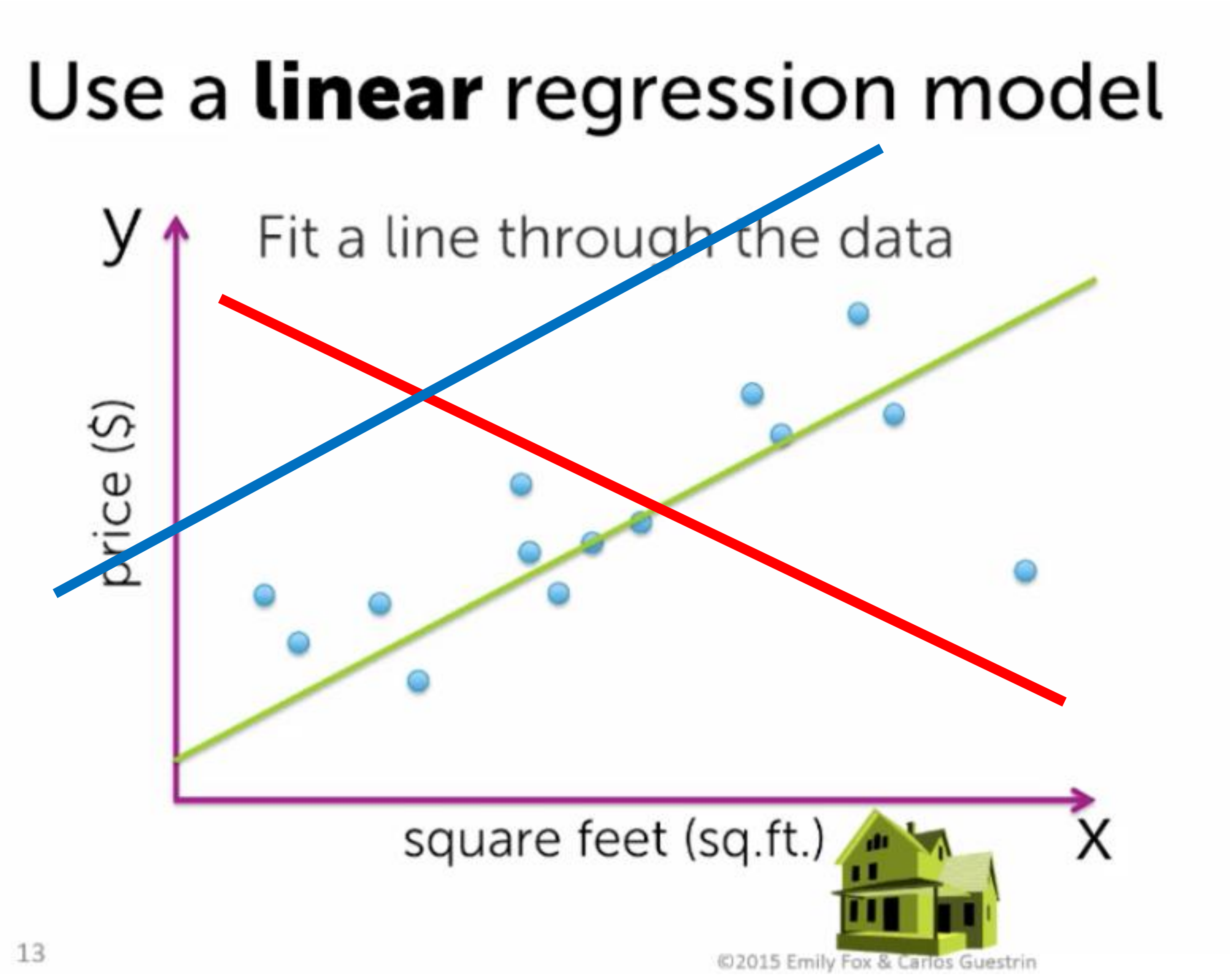
→

ML model
=
parameters



Source:

Supervised Machine Learning - 101



ML model = straight line

2 parameters :

- **Slope**
- **Intercept**

Source:

CRITEO

Supervised Machine Learning - 101

- What's my flat price?

Prediction / Inference

Available data

Trained model

In Criteo context, data = publisher & advertiser data

In Criteo context, value = Click / Visit / Sale

Source:

Data scarcity

Regulators

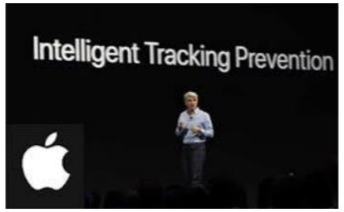
GDPR

CALIFORNIA CONSUMER PRIVACY ACT

CCPA

Vendors

Privacy Sandbox



The user privacy context

Government & Regulators

GDPR & CCPA lead the way in regulating privacy and data protection to prevent **True Privacy risks**

Gatekeeper Restrictions

Browsers place limitations on **third-party cookies, pretexting "posing greater risks than first-party cookies"**

Evolution of web advertising

2017 Apple ITP limits third-party cookie tracking in Safari

2018 Global Data Protection Regulation (GDPR) rolls out in the EU

Firefox blocks cross-site tracking

2019 Microsoft Edge introduces tracking prevention

2020 California Consumer Privacy Act (CCPA) goes into effect

2021 Apple IDFA opt-in requirement rolls out

2022 Firefox rolled out Total Cookie Protection (TCP)

We are here

2025 Google is expected to deprecate third-party cookies in Chrome

Ad performance measurement post-3PC: a fragmented landscape

Different approaches from various browsers, **discussed at the W3C, and meant to converge**

Attribution Reporting API Private Learning API	PAM – Private Ad Measurement SKAN – Stored Kit Ad Network	Ad Selection API	IPA

Many ways of ensuring Privacy !

Table 1. Overview of Key Technical Approaches Essential for PPDSA.

Technique	Description	Value	Limitations
K-anonymity	Transforms a given set of *k* records in such a way that in the published version, each individual is indistinguishable from the others	Reduces the risk of re-identification	Vulnerable to reidentification attack if additional public information is available
Differential Privacy	Adds noise to the original data in such a way that an adversary cannot tell whether any individual's data was or was not included in the original dataset	Provides formal guarantee of privacy by reducing the likelihood of data reconstruction or linkage attacks	Limited to simpler data types; challenge in managing tradeoff between privacy, accuracy, or utility of data
Synthetic Data	Information that is artificially manufactured as an alternative to real-world data	Preserves the overall properties or characteristics of the original dataset	May still disclose privacy-sensitive information contained in the original dataset; difficult to mirror real-world data
Secure Multiparty Computation	Allows multiple parties to jointly perform an agreed computation over their private data, while allowing each party to learn only the final computational output	Increases the ability to compute over distributed datasets without revealing original data	Higher computational and communication costs/burdens, and difficult to scale
Homomorphic Encryption	Allows computing over encrypted data to produce results in an encrypted form	Only authorized users can see original and/or computed data	Higher computational cost and time
Zero-Knowledge Proof	Allows one party to prove to another party that a particular statement is true without revealing privacy-sensitive information	Increases ability to validate information without disclosing sensitive information	Cost and scalability
Trusted Execution Environment	Creates a secure, isolated execution environment parallel to the main operating system to process sensitive data	Allows faster secure analytics on data compared to encryption-based techniques	Introduces other ways sensitive data can leak
Federated Learning	Allows multiple entities to collaborate in building an ML model on distributed data without sharing original data	Minimizes data sharing while training a combined model	Various data reconstruction or inference attacks are still possible; require consistency across datasets held by multiple entities

Do not forget that pseudonymisation also stands for a PET!

https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf

Main privacy mechanisms

ARA	Noise added to data Trusted Server (backend: TEE)
PAM SKAN	Noise added to data Trusted Server (backend: MPC)
Masked Lark Ad Selection	Noise added to data Trusted Server (backend: TEE)
IPA	Noise added to data Trusted Server (backend: MPC)

	Main privacy mechanisms
ARA	Noise added to data Trusted Server (backend: TEE)
PAM SKAN	Noise added to data Trusted Server (backend: MPC)
Masked Lark Ad Selection	Noise added to data Trusted Server (backend: TEE)
IPA	Noise added to data Trusted Server (backend: MPC)

Future-proof AI post-3PC: a missing use-case

Current state (3PC) :

We have access to

- Contextual features
- X-device user features
- X-advertiser user features
- Advertiser-centric features

We don’t have access to

N/A

Future-proof AI post-3PC: a missing use-case

Short-term future state (without 3PC) - 2025:

We have access to

- Contextual features
- 12-bit & noisy advertiser-centric features
- Noisy aggregated or granular labels

We don't have access to

- X-device user features
- X-advertiser user features

We are influencing the future of AI on the open internet

1. **Collaborating closely with Google Chrome** to maximize Privacy Sandbox utility and ensure the use cases of our customers and partners are addressed.

2. **Participating in W3C community and working groups** Web Incubator (WICG) & Private Advertising Technology (PATCG/PATWG).

3. **Sharing with the industry**
   - feedback on Github,
   - online articles detailing our experiments,
   - private session of knowledge sharing,
   - collaborative testing opportunities.

4. **Working with regulators** to define test frameworks, participate to the tests and share feedback on business impacts.

Outline

Context

Novel learning paradigms & challenges

What we do

CRITEO

A minimal obfuscation mechanism to consider: Differential Privacy

docs-and-reports / design-dimensions / **Dimensions-with-General-Agreement.md**

Top

Preview | Code | Blame | 62 lines (34 loc) · 6.4 KB

Raw

reached general agreement that data join could potentially occur off device within a some type of server side architecture. This is conditional on having adequate protections for any data that leaves a device, in line with our security and privacy goals.

Privacy defined at least by Differential Privacy

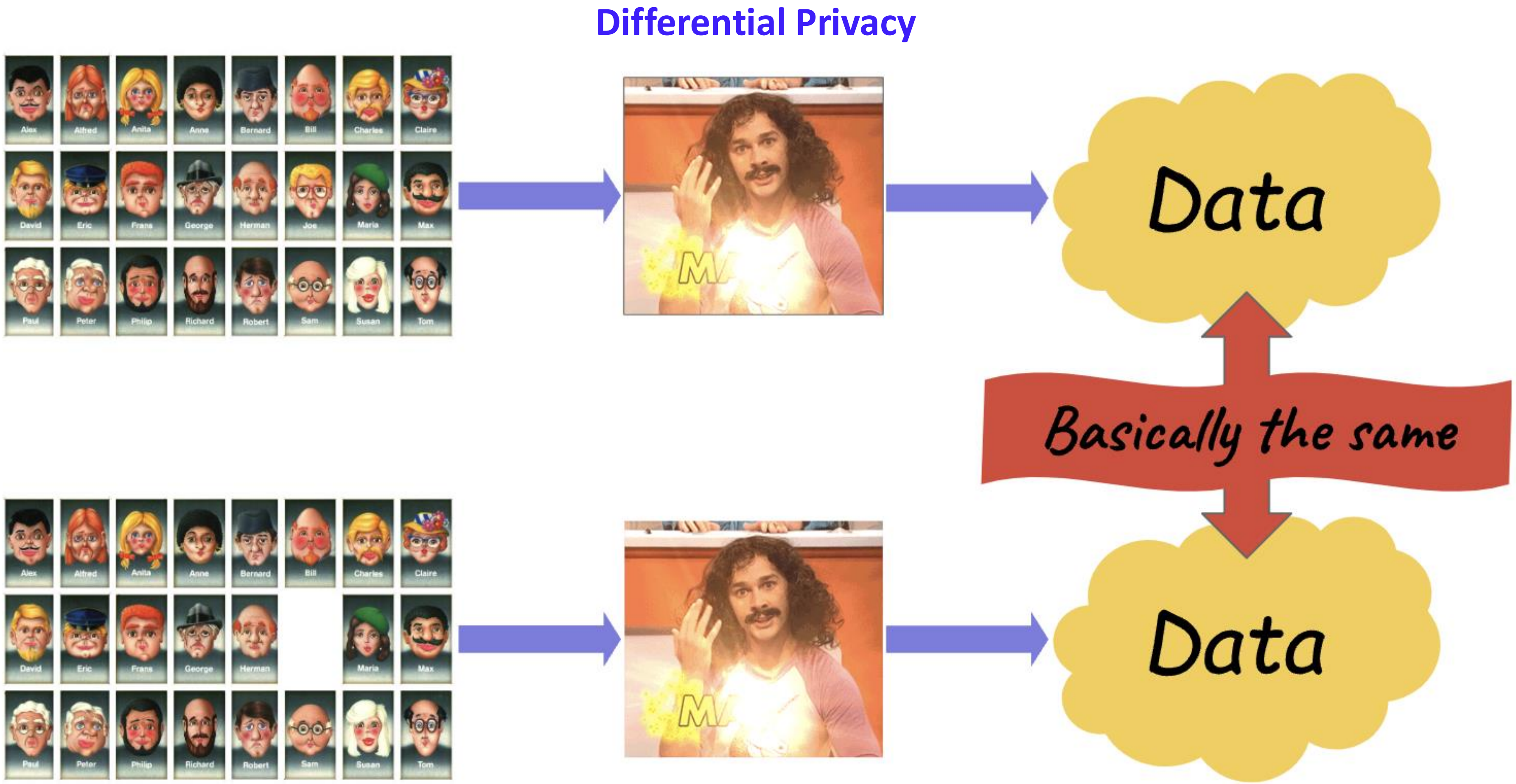
We've explored three main definitions of privacy:

1. Information theoretic
2. K-anonymity
3. Differential privacy

The community group has reached general agreement that the *Private Measurement Technical Specification MVP* should use a definition of privacy based on differential privacy. This does not preclude the use of other privacy definitions in conjunction with differential privacy, however any proposal should aim to provide differential privacy guarantees.

Source: W3C PATCG github repository

The promise



How ? Quézako

Randomness

Differential Privacy

How ? Quézako

Differential Privacy requires **Randomness**

adds **NOISE**

What is noise?

Did a user convert after clicking on a Criteo ad?

What is noise?

How the noise is calibrated/chosen?

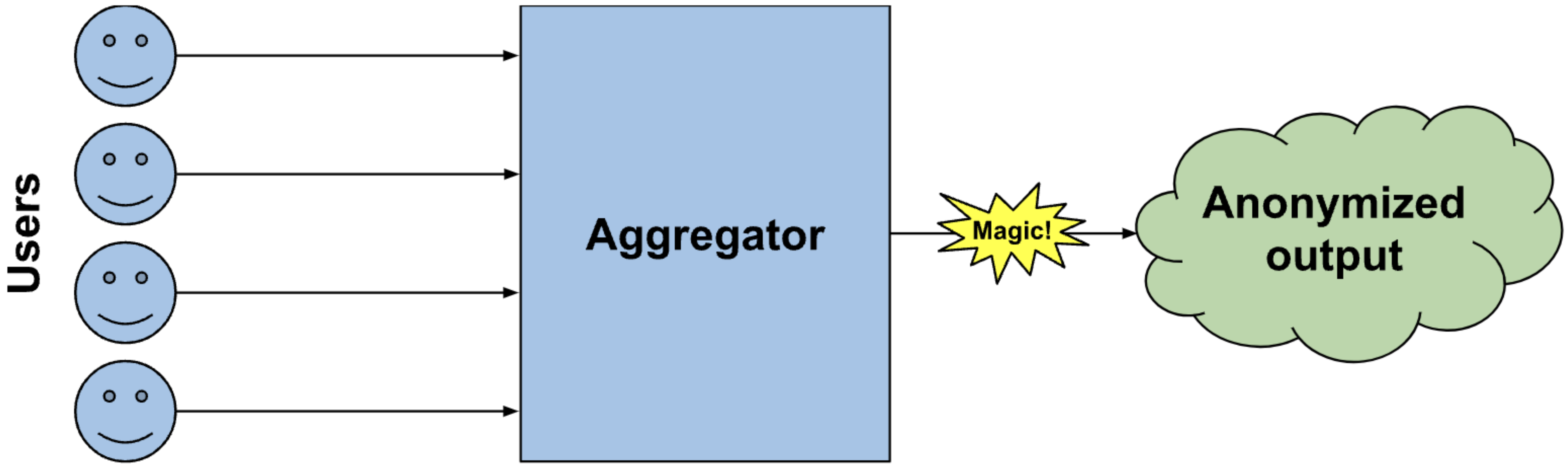
Epsilon

More budget = less privacy = less noise = more performance

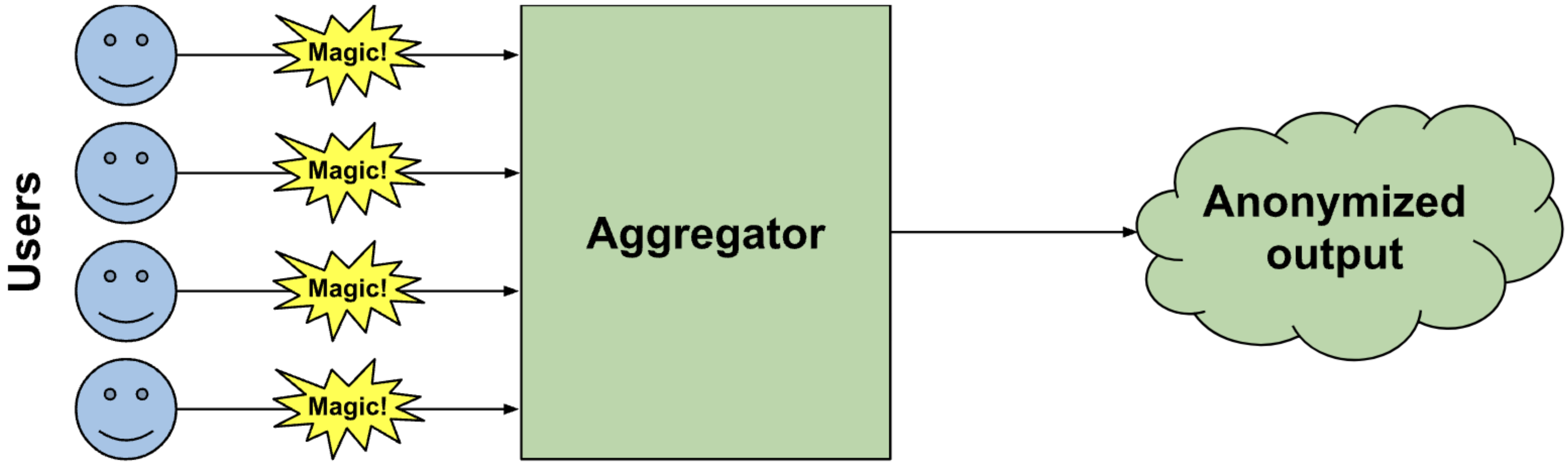
Budget is given by Chrome
Criteo can only optimise the budget planning

Quézako – Global DP

ARA API



Quézako – Local DP



Protected Audiences API

CRITEO

Main AI learning paradigms

Paradigm	Probability to happen	AI Performance	Flexibility	Main cost
Learning on aggregated data	High	Medium	Medium	AI Research
Learning on event-level data	Medium	Low to High	Low to High	AI Research
Learning in a trusted server	Medium	Low to High	Low to High	AI/Platform/Infra Engineering

All learning paradigms involve a specific instance of differential privacy

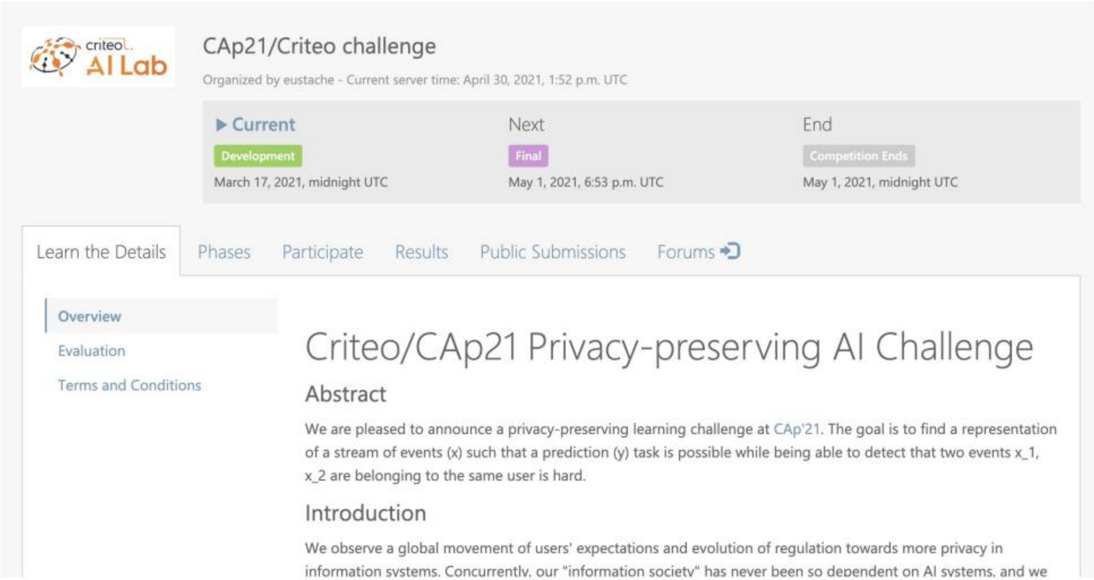
Outline

Context

Novel learning paradigms & challenges

What we do

Open data & AI competitions



Criteo Research Datasets

Terms of use.

We regularly release datasets to ML practitioners and enthusiasts. It is to be noted that Criteo holds the record for releasing the world's largest truly public ML dataset at a healthy 1TB in size and 4B event lines.

All datasets have been anonymized to confirm to privacy standards.

- Criteo Uplift Modeling Dataset (CRITEO-UPLIFT-1)
- Criteo Sponsored Search Conversion Logs
- Criteo Attribution Modeling for Bidding Dataset
- Kaggle Display Advertising dataset
- Criteo 1TB click logs
- Dataset for evaluation of couterfactual algorithms
- Criteo @Hugging Face

Top-tier AI Research

Distribution-Aware Mean Estimation under User-level Local Differential Privacy

Corentin Pla
Criteo AI Lab

Hugo Richard
Criteo AI Lab

Maxime Vono
Criteo AI Lab

Position Paper: Open Research Challenges for Private Advertising Systems under Local Differential Privacy

Matilde Tullii *[2], Solenne Gaucher*[2], Hugo Richard*[1], Eustache Diemert[1], Vianney Perchet[1, 2], Alain Rakotomamonjy[1], Clément Calauzènes[1], and Maxime Vono[1]

[1]Criteo AI Lab, France
[2]ENSAE, Crest, France

Personalised Federated Learning On Heterogeneous Feature Spaces

Alain Rakotomamonjy [* 1] **Maxime Vono** [* 1] **Hamlet Jesse Medina Ruiz** [1] **Liva Ralaivola** [1]

Local Differential Privacy for Regret Minimization in Reinforcement Learning

Evrard Garcelon
Facebook AI Research & CREST, ENSAE
Paris, France
evrard@fb.com

Vianney Perchet
CREST, ENSAE Paris & Criteo AI Lab
Palaiseau, France,
vianney@ensae.fr

Ciara Pike-Burke
Imperial College London
London, United Kingdom
c.pikeburke@gmail.com

Matteo Pirotta
Facebook AI Research
Paris, France
matteo.pirotta@gmail.com

Application to production data and feedbacks to the industry

- **2023/03 –** Alonzo Velasquez (Chrome PM) : https://github.com/WICG/turtledove/issues/435
  - Short term : noisy event-level reporting
  - Long term : learning eventually outsourced to a TEE-based trusted server
- **From 2023/03 to 2023/06 –** Multiple Github issues/presentations of Charlie on event-level label DP : London PATCG slides
- **2023/09 –** Criteo follow-up presentation on ML training using label DP to Chrome + PATCG
- **2024/02 –** Charlie on future of learning : https://github.com/WICG/turtledove/issues/1017
- **2024/04 –** Criteo follow-up presentation on ML training using DP to Boston PATCG
- **2024/06 –** Criteo/Chrome WS

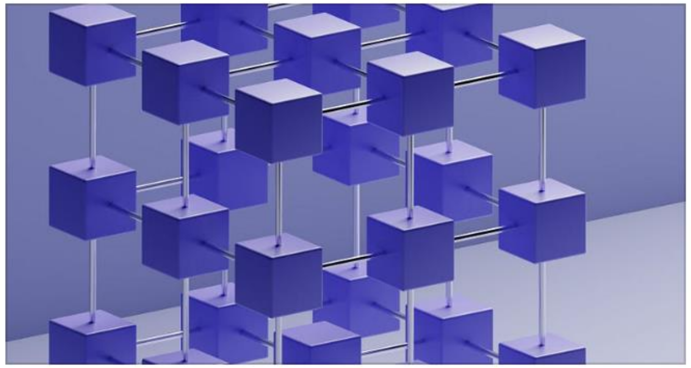
Application to production data and feedbacks to the industry

An Introduction to PETs for Attribution and Reporting

To avoid direct cross-site tracking, several browsers are developing attribution and reporting proposals based on Privacy-Enhancing Techs.

Maxime Vono

Apr 25 · 10 min read



PETs in Advertising: Scenarios for Secure Multi-Party Computation

It aims to deep-dive into the tech details of MPC for ads use cases including private attribution, reporting and campaign optimisation

Maxime Vono

May 4 · 10 min read

Thank you!