Statistics 414 From Predictive AI to Generative AI
Introduction to Digital Marketing and Synthetic Advertisement Data

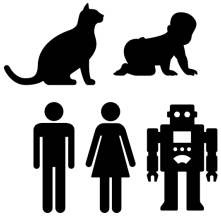
Chi-Hua Wang
chihuawang@ucla.edu

Department of Statistics & Data Science, UCLA

April 04, 2024

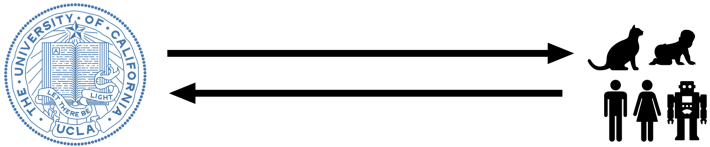
How was UCLA before you got here?

How is UCLA since you've been here?



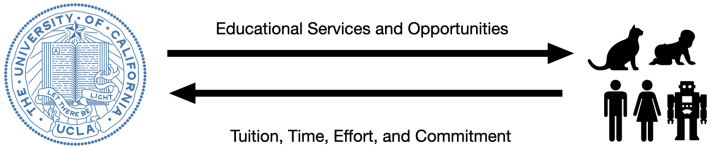
You are here!

What do you expect to gain from UCLA?

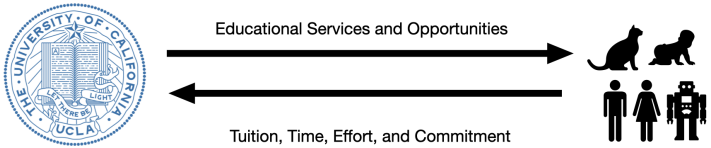


Every Collaboration is a form of Exchange

UCLA use Educational Services and Opportunities
to exchange
Student's Tuition, Time, Effort, and Commitment

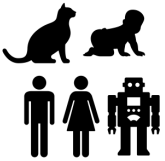


Brand and Consumer



Exchange between UCLA and You

Educational Services and Opportunities



Tuition, Time, Effort, and Commitment

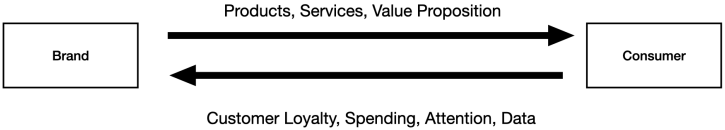
Products, Services, Value Proposition

Brand **Consumer**

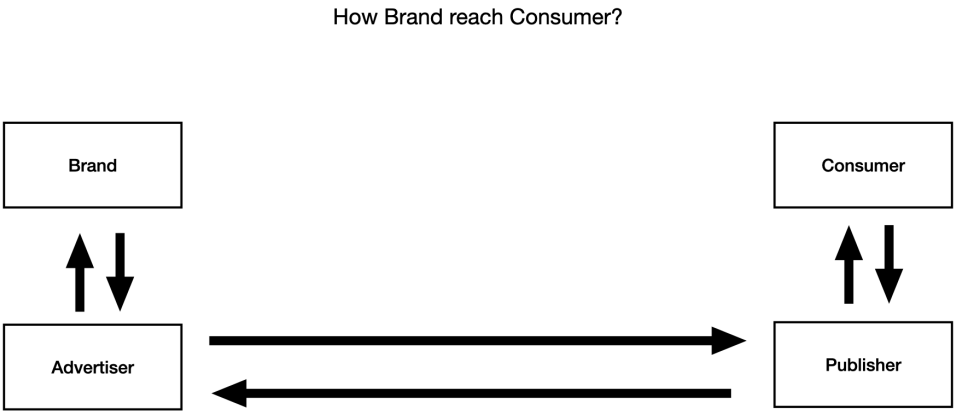
Customer Loyalty, Spending, Attention, Data

Exchange between Brand and Consumer

Brand use Products, Services, Value Proposition
to exchange
Consumer's Customer Loyalty, Spending, Attention, Data



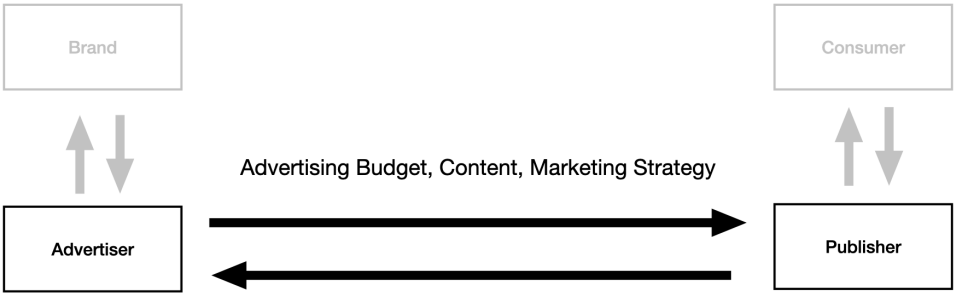
How Brand reach Consumer? Via Advertisement on Content



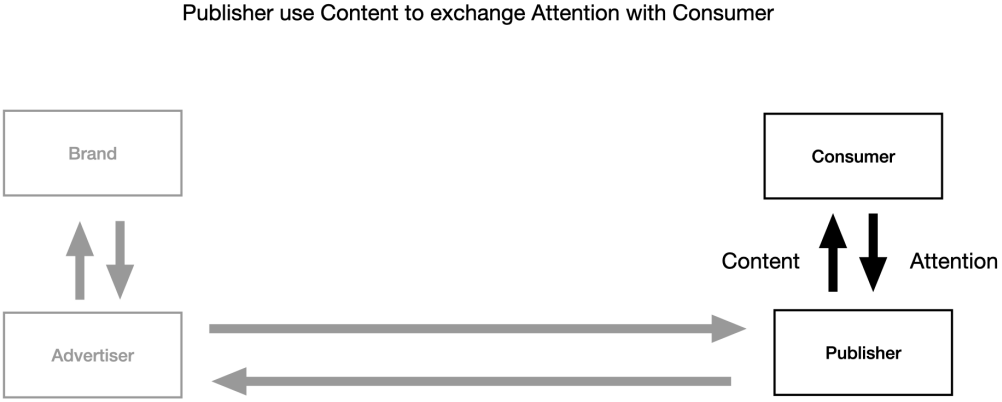
Brand use Marketing Expense to Exchange Customer Awareness with Advertiser

Advertiser use Ad Budget to exchange Ad Space with Publisher

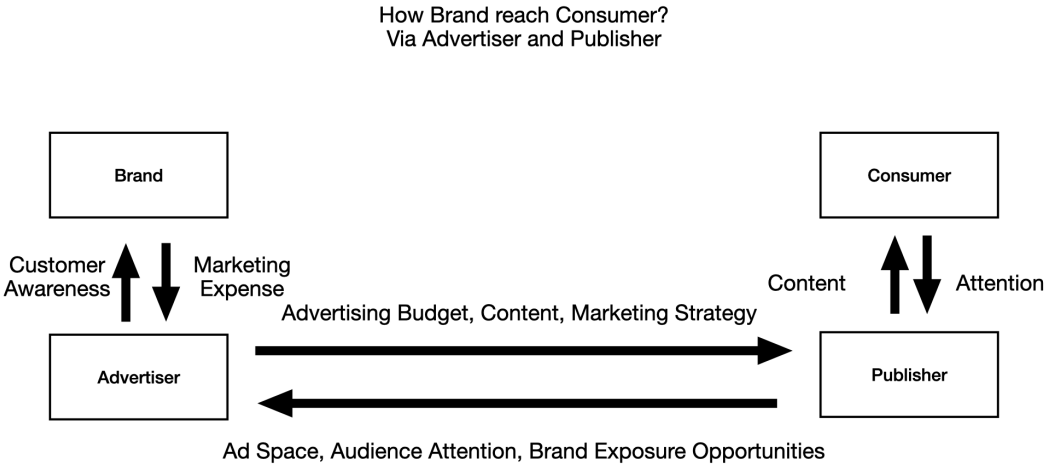
Advertiser use Advertising Budget, Content, Marketing Strategy
to exchange
Publisher's Ad Space, Audience Attention, Brand Exposure Opportunities



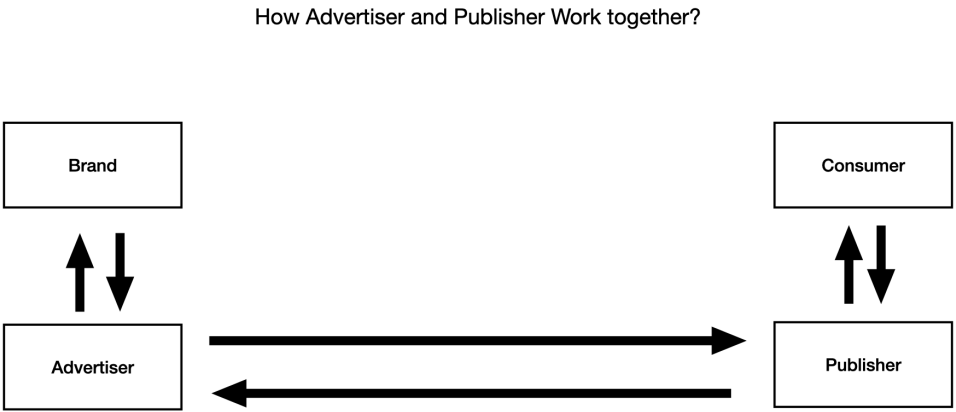
Publisher use Content to exchange Attention with Consumer



Exchange Network between Brand, Advertiser, Publisher, Consumer



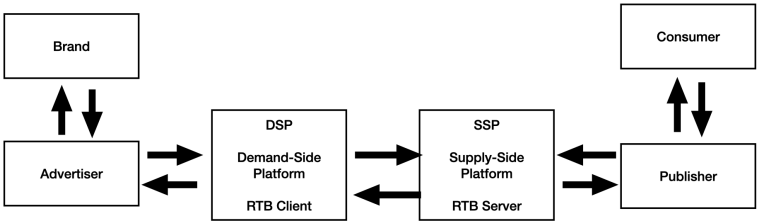
How Advertiser and Publisher work together?



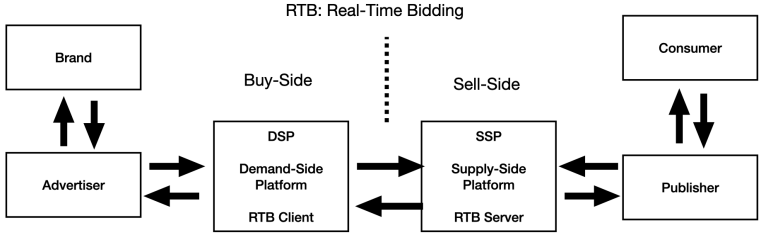
Via Demand-Side Platform and Supply-Side Platform

How Advertiser and Publisher Work together?

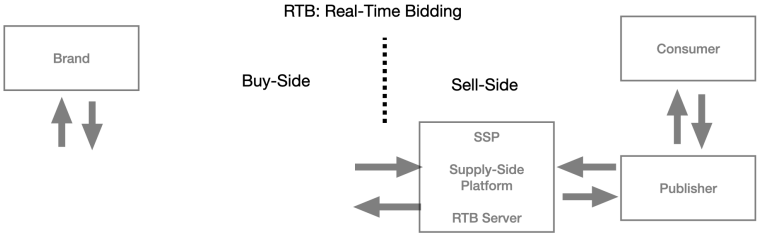
Via RTB: Real-Time Bidding Advertisement



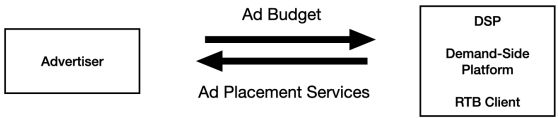
Real-Time Bidding: Buy and Sell the Advertisement Space



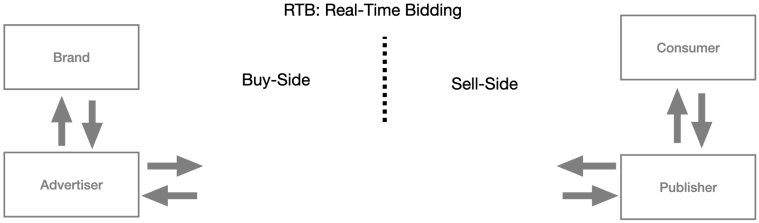
Advertiser use Ad Budget to exchange Ad Placement Service with DSP



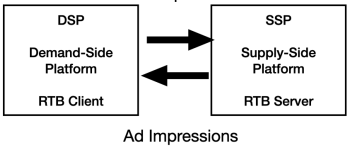
Advertise use Ad Budget to exchange Ad Placement Services with DSP



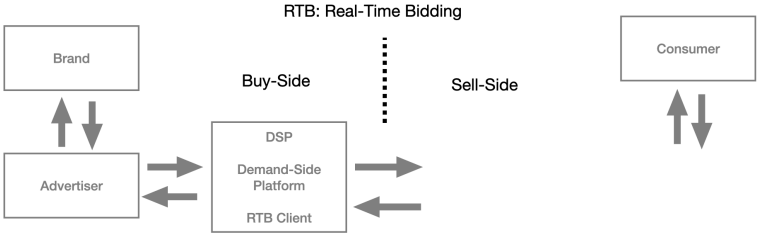
DSP use Ad Spend to exchange Ad Impression with SSP



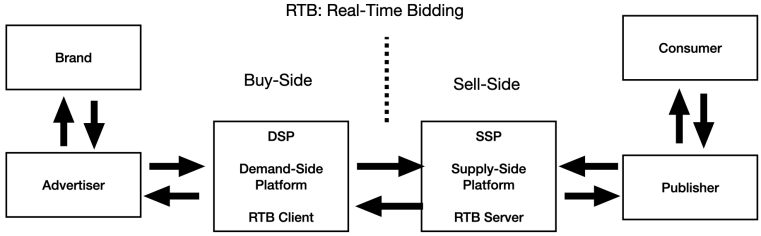
DSP use Ad Spend to exchange Ad Impressions with SSP



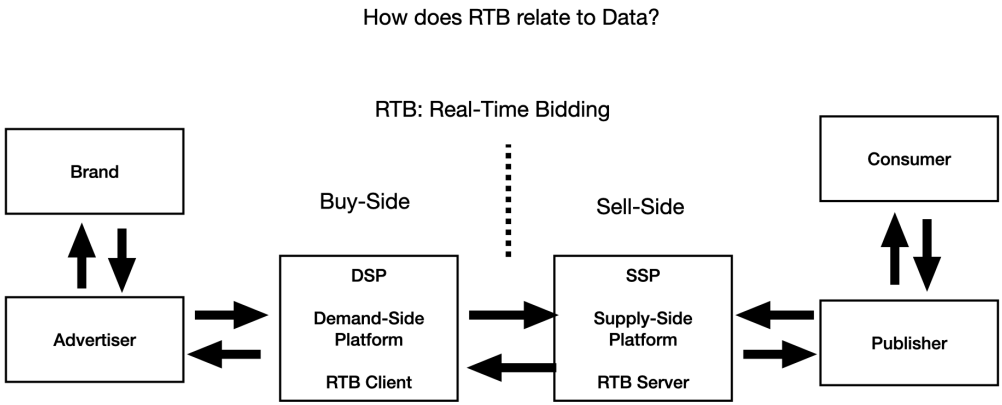
SSP use Ad Earning to exchange Ad Impressions with Publisher



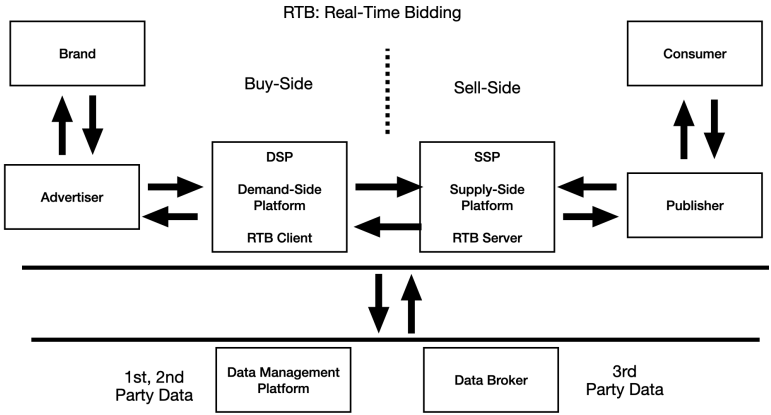
Exchange Network behind Real-Time Bidding



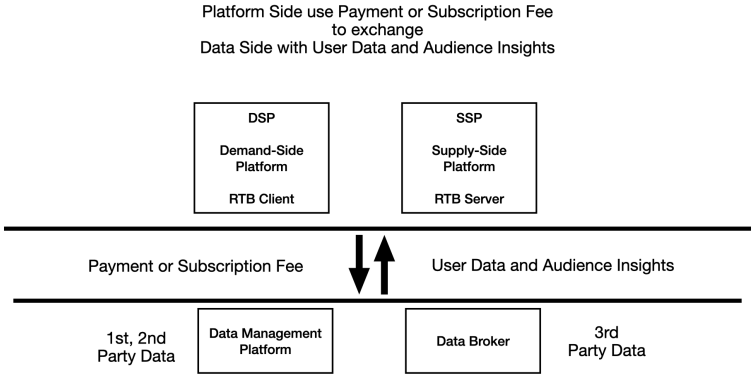
How does Real-Time Bidding relate to Data?



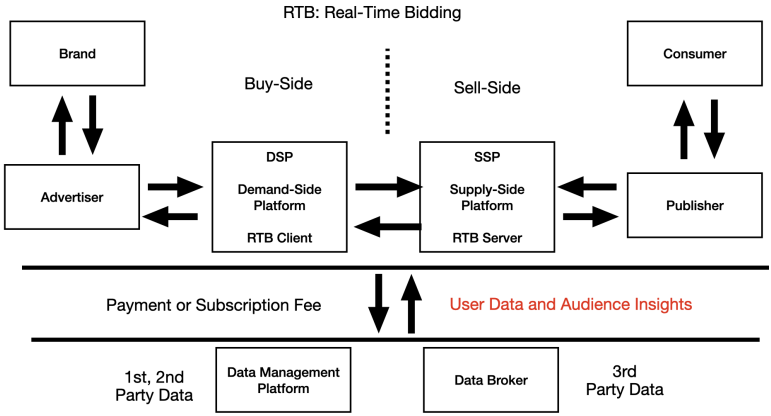
Data Site: Data Management Platform and Data Broker



Platform Side use Payment to Exchange User Data and Audience Insight

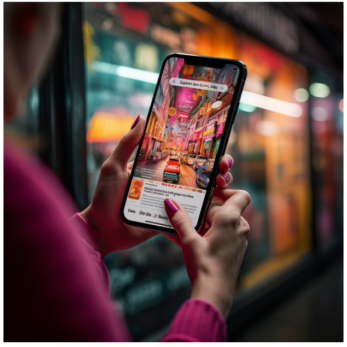


User Data and Audience Insights is Your playground in this class



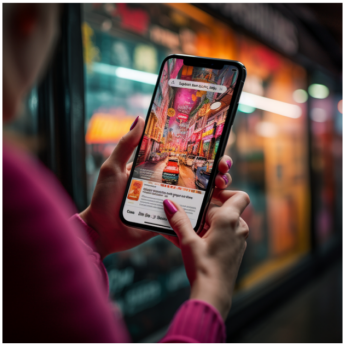
User Data: You are the provider!

User Data and Audience Insights



Advertisement Data: Combine User, Device and Advertisement

User Data and Audience Insights



What is Advertisement Data?

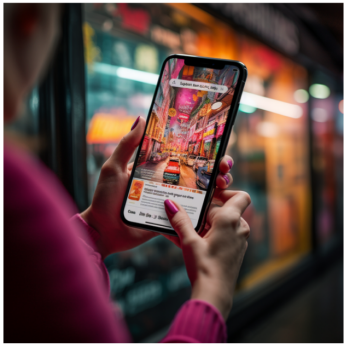
User, Device and Advertisement

Feature = User Feature + Device Feature +Advertisement Feature

User Feature	Age, Gender, Residence (Province, City ID, City Level)
Device Feature	Device series, group, EMUI version, model, size
Advertisement Feature	Network Type, Ad Task ID, Ad Material ID, Advertiser ID, Ad Tag, Creative Material Type, Interaction Material Type

Click-Through Rate: Are you interested in our Brand?

User Data and Audience Insights



CTR Prediction - 2022 DIGIX Global AI Challenge

Click Through Rate Prediction
CTR Prediction Data

DIGIX CTR Predicition Challenge 2022 [Xia22]

CTR Prediction Data

Group columns by the number of unique values

CTR Prediction - 2022 DIGIX Global AI Challenge

About This Data

Ads click-through rate (CTR)

Available at Kaggle

File name: train_data_ads.csv

35 Columns; 7,675,517 Rows

Small Peak on this Data

Group 1	# of Category	Group 2	#	Group 3	#	Group 4	#
site_id	1	net_type	6	app_second_class	20	spread_app_id	116
label	2	series_group	7	series_dev	27	device_name	256
app_score	3	u_feedLifeCycle	8	emui_dev	27	city	341
gender	3	age	8	residence	35	adv_prim_id	545
inter_type_cd	4	creat_type_cd	9	hispace_app_tags	43		
city_rank	4	u_refreshTimes	10	slot_id	60		
Group 5	#	Group 6	#	Group 7	#		
device_size	1547	task_id	11209	ad_click_list_v001	108720		
ad_close_list_v003	1715	adv_id	12615	u_newsCatInterestsST	187576		
ad_close_list_v002	2701	ad_click_list_v003	61102	log_id	1178633		
ad_close_list_v001	3883	user_id	65297				
pt_d	5436	ad_click_list_v002	95376				

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

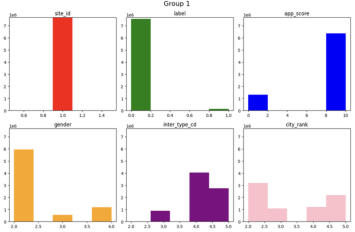
28

Group 1: Columns with unique value less than 5

CTR Prediction Data

Group 1 - Feature with small number of Categories

tab.csv



Group 1	# of Category
site_id	1
label	2
app_score	3
gender	3
inter_type_cd	4
city_rank	4

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

Group 1	Field Description
site_id	Media ID
label	Click
app_score	App score
gender	Gender
inter_type_cd	Interaction type of the material
city_rank	Permanent residence (city level).

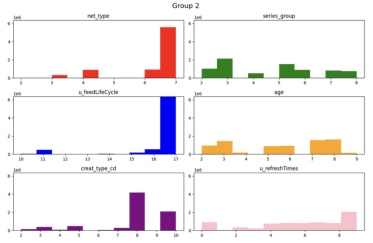
Ad Feature
Ad Feature
Device Feature
User Feature
Ad Feature
User Feature

Group 2: Columns with unique value between 5 to 10

CTR Prediction Data

tab.csv

Group 2 - Major User Feature and Ad Device Feature with small Categories



Group 2	#
net_type	6
series_group	7
u_feedLifeCycle	8
age	8
creat_type_cd	9
u_refreshTimes	10

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

Group 2	Field Description
net_type	Network status where the action occurred
series_group	Device series group
u_feedLifeCycle	User engagement on news feeds.
age	Age
creat_type_cd	Creative type ID of the material
u_refreshTimes	Average number of valid news feeds updates per day.

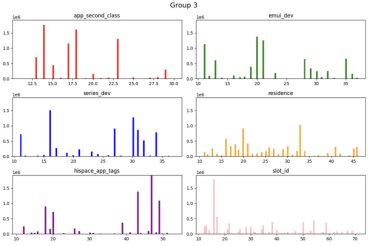
Ad Feature
Device Feature
User Feature
User Feature
Ad Feature
User Feature

Group 3: Columns with unique value between 10 to 100

CTR Prediction Data

tab.csv

Group 3 - Device and Ad feature with medium number of categories



Group 3	#
app_second_class	20
series_dev	27
emui_dev	27
residence	35
hispace_app_tags	43
slot_id	60

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

Group 3	Field Description
app_second_class	Second-level category of the application corresponding to the advertising task
series_dev	Device series
emui_dev	EMUI version number
residence	Permanent residence (province).
hispace_app_tags	Tags of the application corresponding to the advertising task
slot_id	Ad slot ID

Ad Feature
Device Feature
Device Feature
User Feature
Ad Feature
Ad Feature

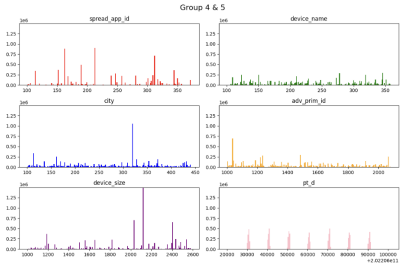
31

Group 4,5: Columns with unique value between 100 to 1000

CTR Prediction Data

tab.csv

Group 4&5 - Device and Ad feature with large number of categories



Group 4	#
spread_app_id	116
device_name	256
city	341
adv_prim_id	545
Group 5	**Datatype**
device_size	1547
pt_d	5436

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

Group 4	Field Description
spread_app_id	Application ID for the advertising task
device_name	User's mobile device model
city	Permanent residence (city ID).
adv_prim_id	Advertiser ID corresponding to the advertising task
Group 5	**Datatype**
device_size	User's mobile device size
pt_d	Timestamp

Ad Feature
Device Feature
User Feature
Ad Feature
Device Feature
Ad Feature

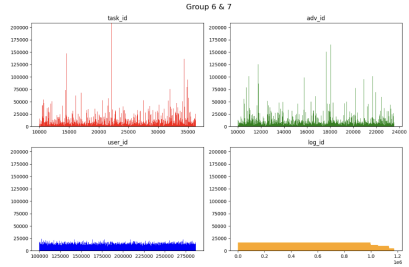
32

Group 6, 7: Columns with unique value over 10,000

CTR Prediction Data

Group 6&7 - id-level Columns

tab.csv



Group 6	Datatype
task_id	11209
adv_id	12615
user_id	65297
Group 7	**Datatype**
log_id	1176633

Group	# of Unique Value	
1	1-5	Less than 5
2	6-10	Between 6 to 10
3	11-100	Less than One hundred
4	100-1000	Less than One Thousand
5	1000-10000	Less than Ten Thousand
6	10000-100000	Less than hundred Thousand
7	100000-1000000	Less than One Million

Group 6	Field Description
task_id	Unique identifier for the advertising task
adv_id	Material ID corresponding to the advertising task
user_id	User ID
Group 7	**Datatype**
log_id	Sample ID

Ad Feature

Ad Feature

User Feature

Ad Feature

Component and Utility of Advertisement Data

What is Advertisement Data?

User, Device and Advertisement

Feature = User Feature + Device Feature +Advertisement Feature

User Feature	Age, Gender, Residence (Province, City ID, City Level)
Device Feature	Device series, group, EMUI version, model, size
Advertisement Feature	Network Type, Ad Task ID, Ad Material ID, Advertiser ID, Ad Tag, Creative Material Type, Interaction Material Type

Utility is about Click Pattern, Conversion Pattern

Label Column	User click the advertisement
Conditional Label Distribution	E.G. The Probability of Age 30-40 iOS user click the advertisement.

Privacy Regulartion on Advertisement Data: Digital Market Act

Digital Market Act

Concurrent Privacy Regulation

Privacy Regulation

2024 Fed News: BigTech DMA Compliance


The Clock is Ticking: Implications of Big Tech Compliance with the EU Digital Markets Act

February 26, 2024
Advisory

Share:

2024 Fed News: Limit the service due to DMA

Apple disables web apps in the EU

iOS firm cites "complex security and privacy concerns" as it changes to comply with Digital Markets Act

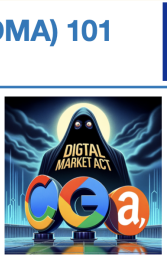
2024 Privacy Landscape

Digital Market Act

Concurrent Privacy Regulation

DMA is a Gatekeeper Regulation

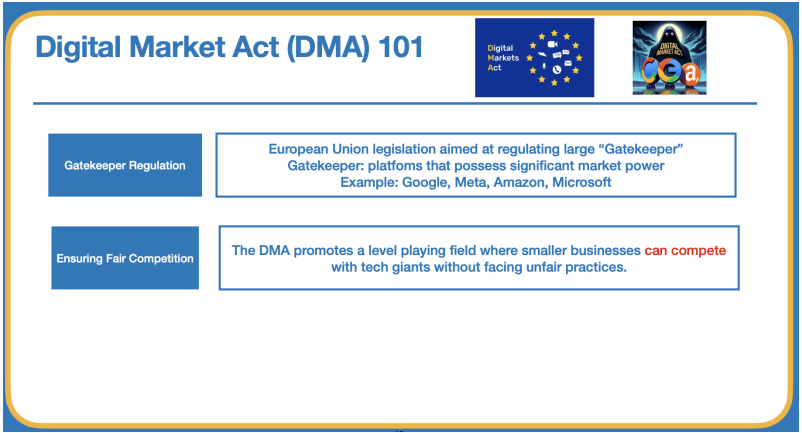
Digital Market Act (DMA) 101



Gatekeeper Regulation

European Union legislation aimed at regulating large “Gatekeeper”
Gatekeeper: platfoms that possess significant market power
Example: Google, Meta, Amazon, Microsoft

DMA wish to ensure fair competition



DMA requires Data Minimization

Digital Market Act (DMA) 101

Gatekeeper Regulation	European Union legislation aimed at regulating large “Gatekeeper” Gatekeeper: platfoms that possess significant market power Example: Google, Meta, Amazon, Microsoft
Ensuring Fair Competition	The DMA promotes a level playing field where smaller businesses can compete with tech giants without facing unfair practices.
Data minimization [1,2]	Companies must only collect and process the minimum amount of data necessary to achieve their purposes. (Compliance with GDPR Article 5)

11

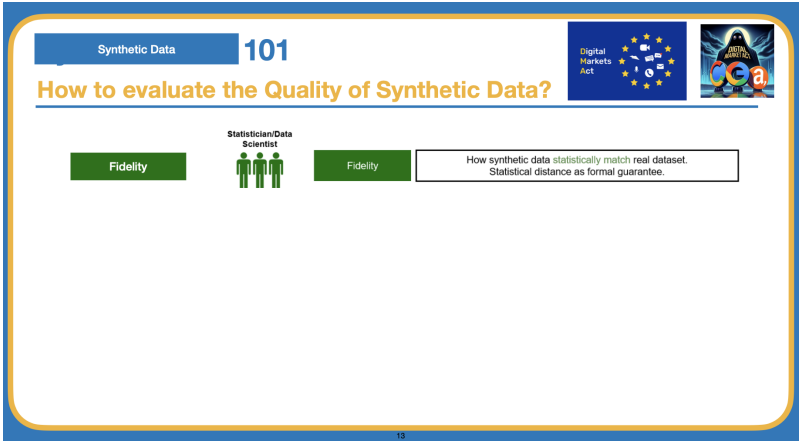
DMA sparks new demand of Synthetic Data

Digital Market Act (DMA) 101

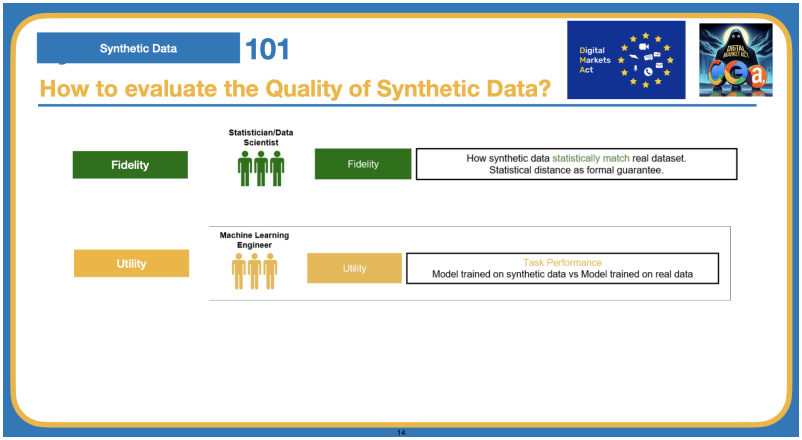
Gatekeeper Regulation	European Union legislation aimed at regulating large “Gatekeeper” Gatekeeper: platfoms that possess significant market power Example: Google, Meta, Amazon, Microsoft
Ensuring Fair Competition	The DMA promotes a level playing field where smaller businesses can compete with tech giants without facing unfair practices.
Data minimization [1.2]	Companies must only collect and process the minimum amount of data necessary to achieve their purposes. (Compliance with GDPR Article 5)

This sparks the demand of Synthetic Data

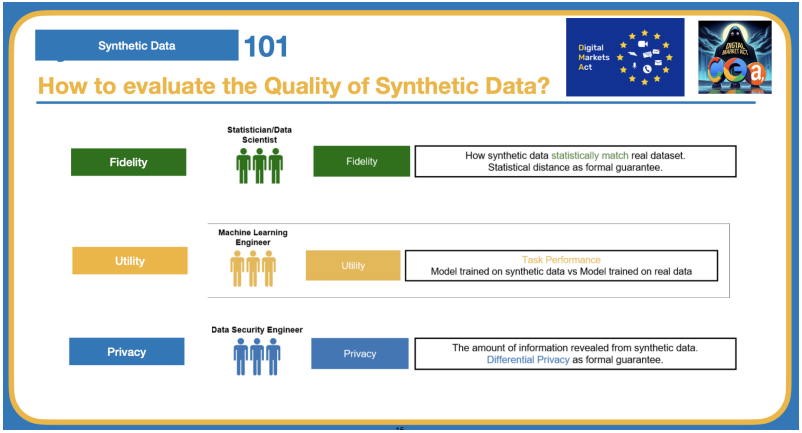
Review Synthetic Data Quality



Review Synthetic Data Quality



Review Synthetic Data Quality



Synthetic Data Quality in DMA context

DMA and Synthetic Tabular Data

How to interpret Synthetic Data Quality into DMA context?

Data Sharing and Portability

Algorithm Transparency

Privacy and Competition

Synthetic Data

Fidelity

Utility

Privacy

19

DMA requirement 1: Data Sharing and Portability

Data Sharing and Portability

Algorithm Transparency

Privacy and Competition

DMA includes provisions for data portability and interoperability Gatekeepers might be required to share certain types of data with competitors or third parties.

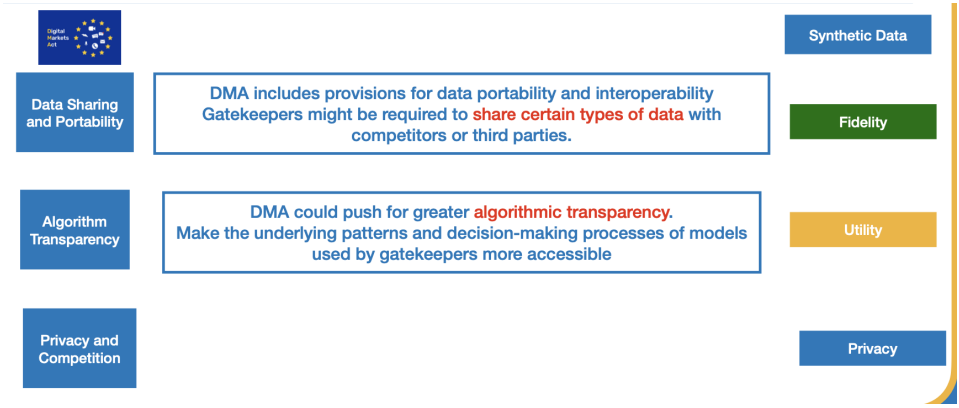
Synthetic Data

Fidelity

Utility

Privacy

DMA requirement 2: Algorithm Transparency



DMA requirement 3: Privacy and Compliant

Data Sharing and Portability

DMA includes provisions for data portability and interoperability Gatekeepers might be required to share certain types of data with competitors or third parties.

Algorithm Transparency

DMA could push for greater algorithmic transparency. Make the underlying patterns and decision-making processes of models used by gatekeepers more accessible

Privacy and Competition

The DMA, in conjunction with data privacy regulations like the GDPR, emphasizes privacy protection.

Synthetic Data

Fidelity

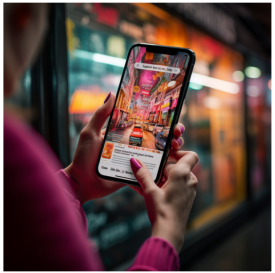
Utility

Privacy

Recall: You are the Data Provider!

How much of your Data been collected every second?

A LOT



Recall: We are the Data Provider!

How much of your Data been collected every second?

A LOT every second

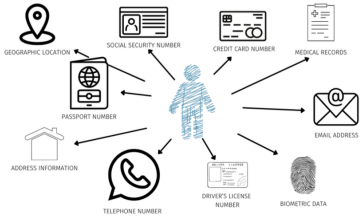
We even provide sensitive information!

How much of your Data been collected every second?

And Some Data are very Sensitive ....

PII: Personal Identifiable Information

Personal Identifiable Information (PII)



Personal Identifiable Information (PII)

64

Differential Privacy make sure PII not leak

We need Differential Privacy to avoid leak PII

Different Privacy is the Legal Standard!

Personal Identifiable Information (PII)

Any information connected to a specific individual that can be used to uncover that individual's identity

HARVARD UNIVERSITY

The Differential Privacy Guarantee

It is mathematically guaranteed that the recipient of a data release generated by a differentially private analysis will make essentially the same inferences about any single individual's private information, whether or not that individual's private information is included in the input to the analysis.

The differential privacy guarantee can be understood in reference to other privacy concepts, such as opt-out and protection of personally identifiable information (PII):

- Differential privacy protects an individual's information essentially as if their data were not used in the analysis at all (i.e., as though the individual opted out and the information was not used).
- Differential privacy ensures that using an individual's data will not reveal essentially any PII that is specific to them. Here, *specific* refers to information that cannot be inferred about an individual unless their information is used in the analysis. Information specific to an individual would be considered PII under a variety of interpretations.[89]

65

Different Privacy as Foundation of Data Privacy

Differential Privacy

Elements of Differential Privacy

Cynthia Dwork

2006

HARVARD UNIVERSITY

Differential Privacy

$$\Pr[\mathscr{A}(\mathbb{D}) \in \mathbb{S}] \leq e^{\epsilon} \Pr\left[\mathscr{A}(\mathbb{D}') \in \mathbb{S}\right] + \delta$$

Randomized Algorithm $\mathscr{A}$

Privacy Parameter ϵ

Neighboring Datasets $\mathbb{D}, \mathbb{D}'$

Subset of Outputs $\mathbb{S} \subset \text{Range}(\mathscr{A})$

Failing Probability δ

66

Share Data Content via Differential Privacy

Share Data Content Via Differential Privacy

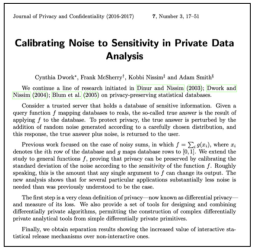
Different Data Content to Share, Different Privacy Mechanism

To Share	For Example...	Randomized Algorithm $\mathscr{A}$
Aggregated Statistics	Mean Function	$M(x) = f(x) + Z$
Machine Learning Model	DP-SGD	$\mathscr{M}(\theta) = \theta - \eta\left(\frac{1}{\lvert B\rvert}\sum_{z\in B}\text{clip}_C\left(\nabla_\theta \ell(\theta, z)\right) + N\left(0, \sigma^2 I\right)\right)$
Synthetic Tabular Data	TabDDPM	$\mathscr{M}(\mathbf{T}) = \tilde{T}$ where $\tilde{T} \sim \mathscr{G}(\mathbf{T})$
Generated Text	Language Model	?????

Case 1: Share Data Analytics Result [Dwo+16]

Recommended Reading 1

DP for Queried Statistics



Journal of Privacy and Confidentiality (2016-2017) 7, Number 3, 17–51

Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork*, Frank McSherry†, Kobbi Nissim‡ and Adam Smith§

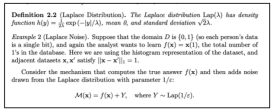
We continue a line of research initiated in Dinur and Nissim (2003); Dwork and Nissim (2004); Blum et al. (2005) on privacy-preserving statistical databases.

Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called true answer is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the sensitivity of the function f. Roughly speaking, this is the amount that any single argument to f can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean definition of privacy—now known as differential privacy—and measure of its loss. We also provide a set of tools for designing and combining differentially private algorithms, permitting the construction of complex differentially private analytical tools from simple differentially private primitives.

Finally, we obtain separation results showing the increased value of interactive statistical release mechanisms over non-interactive ones.



Definition 2.2 (Laplace Distribution). *The Laplace distribution* $\mathrm{Lap}(\lambda)$ *has density function* $h(y) = \frac{1}{2\lambda}\exp(-|y|/\lambda)$*, mean 0, and standard deviation* $\sqrt{2}\lambda$*.*

Example 2 (Laplace Noise). Suppose that the domain D is $\{0, 1\}$ (so each person's data is a single bit), and again the analyst wants to learn $f(\mathbf{x}) = \mathbf{x}(1)$, the total number of 1's in the database. Here we are using the histogram representation of the dataset, and adjacent datasets $\mathbf{x}, \mathbf{x}'$ satisfy $\|\mathbf{x} - \mathbf{x}'\|_1 = 1$.

Consider the mechanism that computes the true answer $f(\mathbf{x})$ and then adds noise drawn from the Laplace distribution with parameter $1/\varepsilon$:

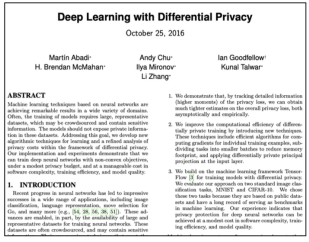
$$\mathcal{M}(\mathbf{x}) = f(\mathbf{x}) + Y, \quad \text{where } Y \sim \mathrm{Lap}(1/\varepsilon).$$

- Foundation of Privacy-Preserving Data Analysis

Case 2: Share Machine Learning Result [Aba+16]

Recommended Reading 2

DP for Trained Model



Deep Learning with Differential Privacy

October 25, 2016

Martin Abadi · Andy Chu · Ian Goodfellow ·
H. Brendan McMahan · Ilya Mironov · Kunal Talwar ·
Li Zhang ·

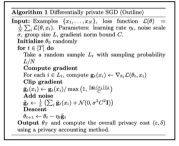
ABSTRACT

Machine learning techniques based on neural networks are achieving remarkable results in a wide variety of domains. Often, the training of models requires large, representative datasets, which may be crowdsourced and contain sensitive information. The models should not expose private information in these datasets. Addressing this goal, we develop new algorithmic techniques for learning and a refined analysis of privacy costs within the framework of differential privacy. Our implementation and experiments demonstrate that we can train deep neural networks with non-convex objectives, under a modest privacy budget, and at a manageable cost in software complexity, training efficiency, and model quality.

1. INTRODUCTION

Recent progress in neural networks has led to impressive breakthroughs in many machine learning tasks including classification, language representation, move selection for Go, and many more [e.g., 54, 28, 56, 38, 51]). These advances are enabled, in part, by the availability of large and representative datasets for training neural networks. These datasets are often crowdsourced, and may contain sensitive

1. We demonstrate that, by tracking detailed information (higher moments) of the privacy loss, we can obtain much tighter estimates on the overall privacy loss, both asymptotically and empirically.
2. We improve the computational efficiency of differentially private training by introducing new techniques. These techniques include efficient algorithms for computing gradients for per-example updates, and dividing tasks into smaller batches to reduce memory footprint, and applying differentially private principal projection at the input layer.
3. We build on the machine learning framework TensorFlow [1] for training models with differential privacy. We evaluate our approach on two standard image classification tasks, MNIST and CIFAR-10. We chose these two tasks because they are based on public datasets and have a long record of serving as benchmarks in machine learning. Our experience indicates that privacy protection for deep neural networks can be achieved at a modest cost in software complexity, training efficiency, and model quality.



Algorithm 1 Differentially private SGD (Outline)

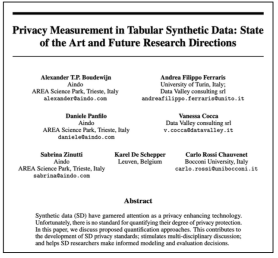
Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.
Initialize θ_0 randomly
for $t \in [T]$ **do**
  Take a random sample L_t with sampling probability L/N
  Compute gradient
  For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$
  Clip gradient
  $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$
  Add noise
  $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
  Descent
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$
Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

- Foundation of Differential Private Machine Learning

Case 3: Share Generator and Synthetic Data [Bou+23]

Recommended Reading 3

DP for Generator



Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions

Alexander T.P. Boudewijn
Aindo
AREA Science Park, Trieste, Italy
alexander@aindo.com

Andrea Filippo Ferraris
University of Turin, Italy;
Data Valley consulting srl
andreafilippo.ferraris@unito.it

Daniele Panfilo
Aindo
AREA Science Park, Trieste, Italy
daniele@aindo.com

Vanessa Cocca
Data Valley consulting srl
v.cocca@datavalley.it

Sabrina Zinutti
Aindo
AREA Science Park, Trieste, Italy
sabrina@aindo.com

Karel De Schepper
Leuven, Belgium

Carlo Rossi Chauvenet
Bocconi University, Italy
carlo.rossi@unibocconi.it

Abstract

Synthetic data (SD) have garnered attention as a privacy enhancing technology. Unfortunately, there is no standard for quantifying their degree of privacy protection. In this paper, we discuss proposed quantification approaches. This contributes to the development of SD privacy standards; stimulates multi-disciplinary discussion; and helps SD researchers make informed modeling and evaluation decisions.

Definition 4.1. (Differential Privacy, [44]) *A randomized algorithm $\mathcal{M}$ is (ε, δ)-**differentially private** ((ε, δ)-DP) if for all $S \subseteq A(P)$:*

$$\mathbb{P}\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(D') \in S\right] + \delta \quad (1)$$

for all databases D, D' such that $\exists d \in D : D' = D \setminus \{d\}$.

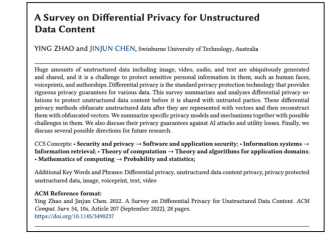
Generators are data releasing systems and can thus be DP: suppose we have two real datasets D and D' with $D' = D \setminus \{d\}$. Then generator $\mathcal{G}$ is DP if a data controller with access to $D \sim \mathcal{G}$ cannot infer whether $\mathcal{G}$ was trained on D or D' (Appendix B, Figure 3). Appendix B.2 details approaches to train generators with built-in mechanisms to guarantee output data is DP. Importantly, in this context, DP is a *property of generators*, and not of the synthetic data they may produce.

- Ongoing Topic of Differential Private Generative Modeling

Case 4: Share AI-generated Text [ZC22]

Recommended Reading 4

DP for Text ?



A Survey on Differential Privacy for Unstructured Data Content

YING ZHAO and JINJUN CHEN, Swinburne University of Technology, Australia

Huge amounts of unstructured data including image, video, audio, and text are ubiquitously generated and shared, and it is a challenge to protect sensitive personal information in them, such as human faces, voiceprints, and authorships. Differential privacy is the standard privacy protection technology that provides rigorous privacy guarantees for various data. This survey summarizes and analyzes differential privacy solutions to protect unstructured data content before it is shared with untrusted parties. These differential privacy methods obfuscate unstructured data after they are represented with vectors and then reconstruct them with obfuscated vectors. We summarize specific privacy models and mechanisms together with possible challenges in them. We also discuss their privacy guarantees against AI attacks and utility losses. Finally, we discuss several possible directions for future research.

CCS Concepts: • **Security and privacy** → **Software and application security**; • **Information systems** → **Information retrieval**; • **Theory of computation** → **Theory and algorithms for application domains**; • **Mathematics of computing** → **Probability and statistics**;

Additional Key Words and Phrases: Differential privacy, unstructured data content privacy, privacy protected unstructured data, image, voiceprint, text, video

ACM Reference format:
Ying Zhao and Jinjun Chen. 2022. A Survey on Differential Privacy for Unstructured Data Content. *ACM Comput. Surv.* 54, 10s, Article 207 (September 2022), 28 pages.
https://doi.org/10.1145/3490237

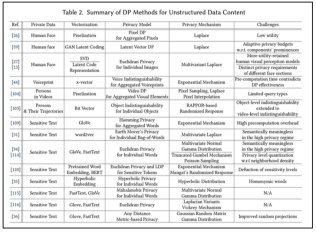


Table 2. Summary of DP Methods for Unstructured Data Content

Ref.	Private Data	Vectorization	Privacy Model	Privacy Mechanism	Challenges
[26]	Human Face	Pixelization	Pixel DP for Aggregated Pixels	Laplace	Low utility
[39]	Human face	GAN Latent Coding	Latent Vector DP	Laplace	Adaptive privacy budgets w.r.t. different latent dimensions
[27], [33]	Human Face	SVD Latent "Code" Representation	Euclidean Privacy for Individual Images	Multivariate Laplace	More utility-oriented human visual perception models
					Distinct privacy requirements of different face sections
[46]	Voiceprint	x-vector	Voice Indistinguishability for Aggregated Voiceprints	Exponential Mechanism	Pre-computation time overtaken DP effectiveness
[104]	Persons in Videos	Pixelization	Video DP for Aggregated Visual Elements	Pixel Sampling, Laplace Pixel Interpolation	Limited query types
[103]	Persons & Their Trajectories	Bit Vector	Object Indistinguishability for Individual Objects	RAPPOR-based Randomized Response	Object-level indistinguishability extended to video-level indistinguishability
[109]	Sensitive Text	Glove	Hamming Privacy for Aggregated Words	Exponential Mechanism	High precomputation overhead
[31]	Sensitive Text	word2vec	Earth Mover's Privacy for Individual Bag-of-Words	Multivariate Laplace	Semantically meaningless in the high privacy regime
[58], [114]	Sensitive Text	GloVe, FastText	Euclidean Privacy for Individual Words	Multivariate Normal Gamma Distribution, Truncated Gumbel Mechanism	Semantically meaningless in the high privacy regime
					Privacy level-quantization w.r.t neighborhood density
[108]	Sensitive Text	Pretrained Word Embedding, BERT	Euclidean Privacy and LDP for Sensitive Tokens	Exponential Mechanism, Mangat's Randomized Response	Deflection of sensitivity levels
[35]	Sensitive Text	Hyperbolic Embedding	Hyperbolic Privacy for Individual Words	Hyperbolic Distribution	Homonymic words
[113]	Sensitive Text	FastText, GloVe	Euclidean Privacy for Individual Words	Multivariate Normal Gamma Distribution	N/A
[116]	Sensitive Text	Glove, FastText	Euclidean Privacy	Laplacian Variants Vickrey Mechanism	N/A
[36]	Sensitive Text	Glove, FastText	Any Distance Metric-based Privacy	Gaussian Random Matrix Gamma Distribution	Improved random projections

- Text Privacy is to protect Authorship Privacy (i.e. who wrote the text), but Different Privacy fail to do it.
- Authorship Privacy is different from previous data content to protect PII (Personal Identifiable Information)

71

References I

[Aba+16] Martin Abadi et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[Bou+23] Alexander Theodorus Petrus Boudewijn et al. "Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions". In: *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*. 2023.

[Dwo+16] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Journal of Privacy and Confidentiality* 7.3 (2016), pp. 17–51.

[Hub23] HubSpot. *Programmatic Ads 101: The Plain-English Guide to Programmatic Advertising*. Accessed: 2024-03-27. 2023. URL: https://blog.hubspot.com/agency/programmatic-advertising-glossary.

[Xia22] Xiaojiu1414. *CTR Prediction - 2022 DIGIX Global AI Challenge*. https://www.kaggle.com/datasets/xiaojiu1414/digix-global-ai-challenge. Accessed: 2024-03-27. 2022.

[ZC22] Ying Zhao and Jinjun Chen. "A survey on differential privacy for unstructured data content". In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–28.