

**MACHINE
LEARNING
APPLICATIONS
AND SUSTAINABLE
DEVELOPMENT**

Vishnu S. Pendyala, Ph.D.
San Jose State University

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

11/12/23

©Vishnu S. Pendyala This work is licensed under a [Creative Commons Attribution-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nd/4.0/)

Video Recording: <https://www.youtube.com/watch?v=GyYWvylJfvg&list=PLLsxQYv4DdJk9HSsby8Whox-JhJ0w7Ldk>

SUSTAINABLE DEVELOPMENT GOALS



This Photo by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/4.0/)

Last Year's November2remember

Machine Learning for Societal Improvement, Modernization and Progress



11/12/23

<https://www.youtube.com/watch?v=RdoguqaM8PA>

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

More Background

Exploring Equity, Diversity, Inclusion, and Social Justice in AI: A Comprehensive Lens



11/12/23

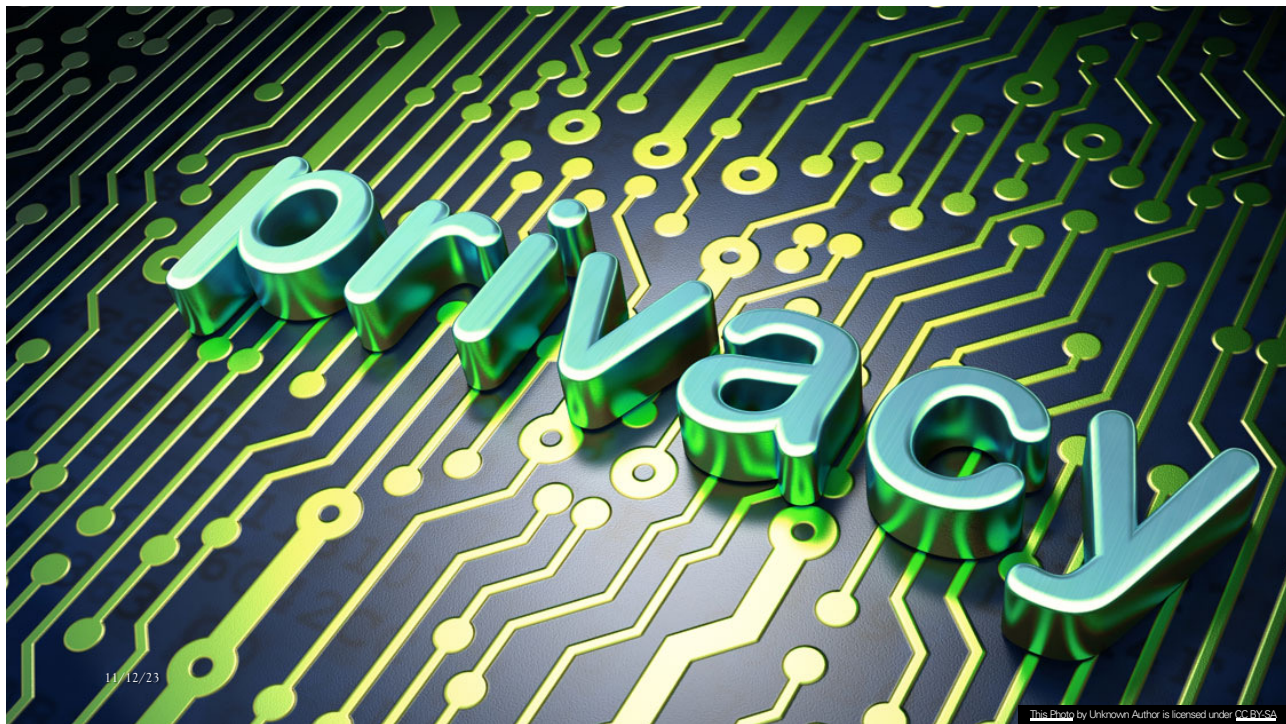
<https://www.youtube.com/watch?v=YKKOE3nwAs>



This Photo by Unknown Author is licensed under [CC BY](#)

Social Sustainability

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.



This Photo by Unknown Author is licensed under [CC BY-SA](#)

The New York Times

How China Uses High-Tech Surveillance to Subdue Minorities

Share full article



202

"The system sifted through billions of records, then displayed details of her education, family ties, links to an earlier case and recent visits to a hotel and an internet cafe."

"It contained facial recognition records and ID scans for about 2.5 million people, mostly in Urumqi, a city with a population of about 3.5 million."



How China Turned a City Into a Prison

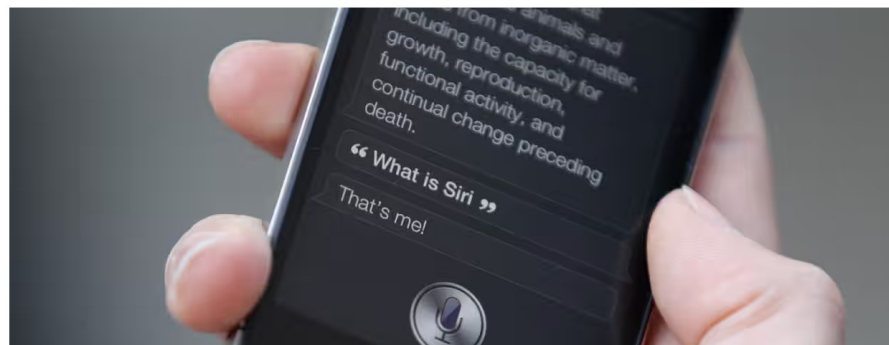
Children are interrogated. Neighbors become informants. Mosques are monitored. Cameras are everywhere.

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

"There have been countless instances of recordings featuring private discussions between doctors and patients, business deals, seemingly criminal dealings, sexual encounters and so on. These recordings are accompanied by user data showing location, contact details, and app data."

Apple contractors 'regularly hear confidential details' on Siri recordings

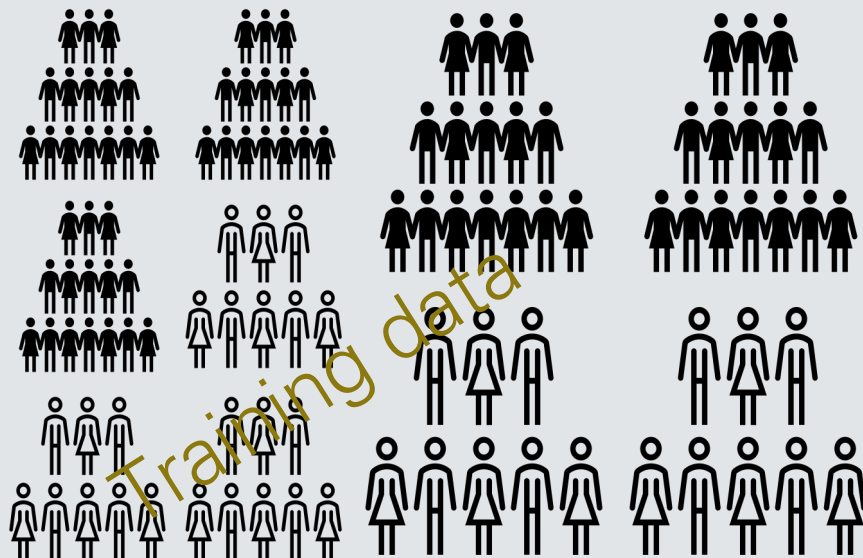
Workers hear drug deals, medical details and people having sex, says whistleblower



Source: <https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings>

Machine Learning: From Generalization to Personalization

Analyze many



Apply to one



person X is
90% of the time
polite, 83%
magnanimous, ...

Are they good or
bad?

©Vishnu S. Pendyala This work is licensed under a [Creative Commons Attribution-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nd/4.0/)

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

But what if...

Analyze many datasets



Identify one
or more



person X has numb
fingers, is looking
for single men, has
some weird
information
needs...

©Vishnu S. Pendyala This work is licensed under a [Creative Commons Attribution-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nd/4.0/)

There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga.”

...

“No. 3505202 asks about “depression and medical leave.” No. 7268042 types “fear that spouse contemplating cheating.”

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

 Share full article  

By **Michael Barbaro** and **Tom Zeller Jr.**

Aug. 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on

To cite this presentation: Pendyala, V.S. (2023) “Machine Learning Applications and Sustainable Development” University of Bolton, UK, November to Remember Webinar series.

Barth-Jones, D. (2012). The're-identification'of governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (July 2012).

"At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then graduate student Sweeney started hunting for the Governor’s hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor’s health records (which included diagnoses and prescriptions) to his office."

11/12/23

"computer science got us into this mess, can computer science get us out of it?"

Sweeney, Latanya. "**k-anonymity**: A model for protecting privacy." *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002): 557-570.

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.

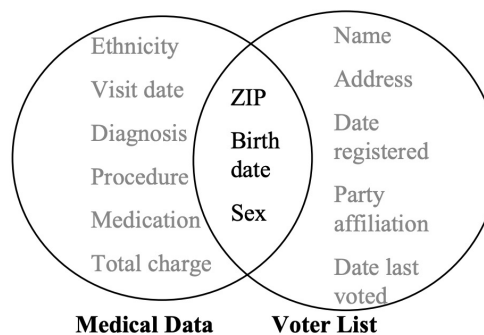


Figure 1 Linking to re-identify data

11/12/23

©Vishnu S. Pendyala This work is licensed under a [Creative Commons Attribution-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nd/4.0/)

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

k-Anonymity

A dataset D satisfies k -Anonymity for a value of k if:

For each row $r_1 \in D$, there exist at least $k - 1$ other rows $r_2 \dots r_k \in D$ such that

$$\Pi_{qi(D)} r_1 = \Pi_{qi(D)} r_2, \dots, \Pi_{qi(D)} r_1 = \Pi_{qi(D)} r_k$$

where $qi(D)$ are the quasi-identifiers of D , and $\Pi_{qi(D)} r$ represents the columns of r containing quasi-identifiers (i.e. the projection of the quasi-identifiers).

11/12/23

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

This Photo by Unknown Author is licensed under CC BY

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

K-Anonymity: Example for K=3

Name	Age	City	Occupation
Alice	25	New York	Engineer
Bob	30	Los Angeles	Doctor
Carol	28	Chicago	Teacher
Dave	27	New York	Engineer
Elaine	35	Los Angeles	Lawyer
Frank	33	Chicago	Architect
Grace	29	New York	Nurse
Henry	26	Los Angeles	Programmer
Ingrid	32	Chicago	Artist
Jack	24	New York	Engineer



Name	Age	City	Occupation
Group 1	25-30	New York	Engineer
Group 2	25-30	Los Angeles	Doctor
Group 3	25-30	Chicago	Teacher
Group 1	25-30	New York	Engineer
Group 2	31-35	Los Angeles	Lawyer
Group 3	31-35	Chicago	Architect
Group 1	25-30	New York	Nurse
Group 2	25-30	Los Angeles	Nurse
Group 3	31-35	Chicago	Artist
Group 1	25-30	New York	Engineer

K-Anonymity: Issues and challenges

- Generalizing using k -Anonymity may involve removing significant information from the data
- Vulnerability to re-identification through background knowledge
- Checking for K-anonymity naively is quadratic in complexity: $O(n^2)$
- Generalizing datasets in NP-hard!

11/12/23



To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

ZDNET

tomorrow belongs to those who embrace it today



Home / Health

How Strava's "anonymized" fitness tracking data spilled government secrets

Analysis: Strava may "anonymize" the user, but that isn't helpful when that user inadvertently reveals the location of sensitive government facilities.

"That same data also illuminated a scattering of little-known locations in war zones, where US secret facilities and military bases have operations and personnel -- presumably because soldiers and staff are unknowingly uploading their fitness tracking data to Strava."

11/12/23

"...keeping the very existence of military and other intelligence bases classified is impossible"

WIRED BACKCHANNEL BUSINESS CULTURE GEAR MORE ▾ SIGN IN SUBSCRIBE

RYAN SINGEL SECURITY DEC 17, 2009 4:29 PM

Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims

An in-the-closet lesbian mother is suing Netflix for privacy invasion, alleging the movie rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its \$1 million contest to improve its recommendation system. The suit known as *Doe v. Netflix* (.pdf) was filed [...]

11/12/23

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

WIRED BACKCHANNEL BUSINESS CULTURE GEAR MORE ▾ SIGN IN SUBSCRIBE

BRUCE SCHNEIER SECURITY DEC 12, 2007 9:00 PM

Why 'Anonymous' Data Sometimes Isn't

Anonymous data sets are an enormous boon for researchers, but the recent de-anonymization of Netflix customer data shows there are privacy risks as well. Commentary by Bruce Schneier.

Source <https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/>

SAVE

11/12/23

LAST YEAR, NETFLIX published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

FORBES

TECHNOLOGY

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley Contributor
The Firewall Contributor Group @

Follow

Mar 12, 2010, 12:35pm EST

Source: <https://www.forbes.com/sites/firewall/2010/03/12/netflix-settles-privacy-suit-cancels-netflix-prize-two-sequel/?sh=65fbd06b951e>

⌚ This article is more than 10 years old.

On Friday, Netflix [announced](#) on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a \$1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111-125). IEEE.

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

11/12/23

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

Homogeneity Attack

- k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute
- Attackers can exploit the consistency of data across multiple sources or tables and gain insights that were not intended by the data providers.

Demographic Data

ID	Age	Gender	Zip Code
101	25 - 30	Male	90210
102	30 - 35	Female	94102
103	25 - 30	Male	90210
104	40 - 45	Male	60601

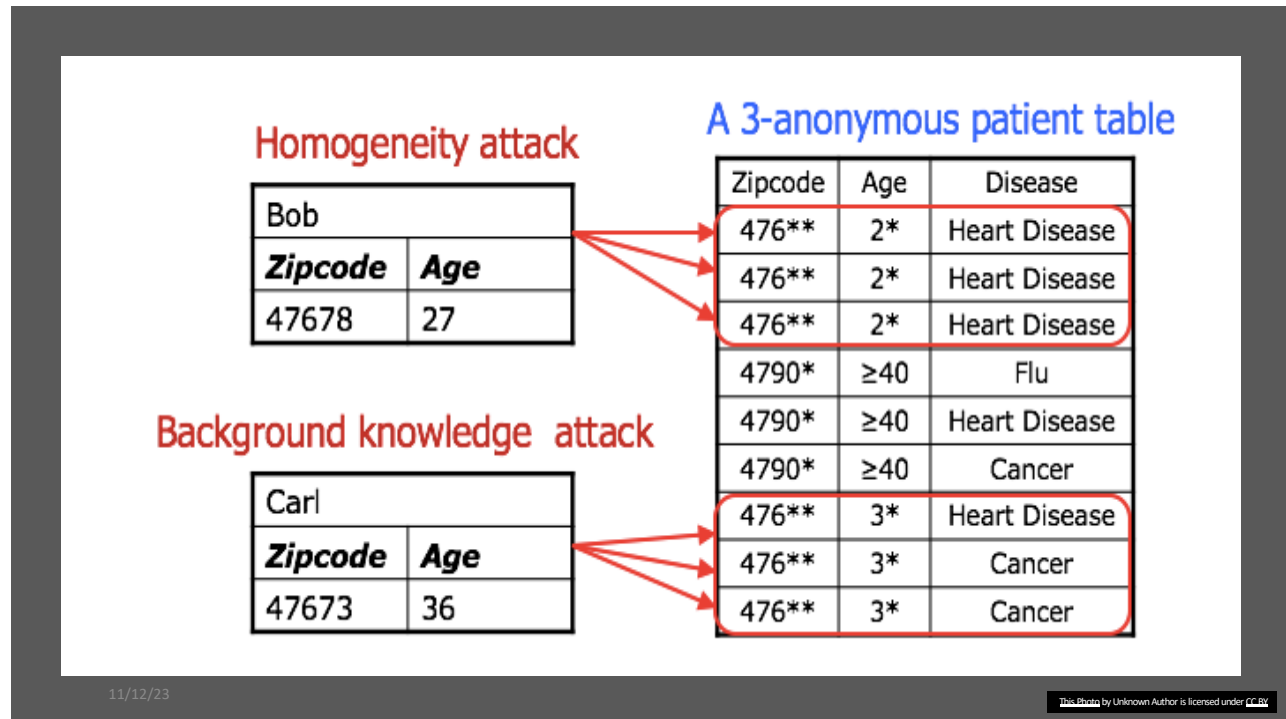
Purchase History

ID	Product	Purchase Date
101	Smartphone	2023-01-25
102	Laptop	2022-12-10
103	Headphones	2023-02-01
104	Smartphone	2023-01-30



*101 and 103 are
the same person!*

11/12/23



To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

L-Diversity

- L-Diversity ensures that for each sensitive attribute value, there is a sufficient diversity of non-sensitive attribute values.
- q^* -block is the set of tuples in T^* whose non-sensitive attribute values generalize to q^*
- A q^* -block is ℓ -diverse if it contains at least ℓ well-represented values for the sensitive attribute S
- A table is ℓ -diverse if every q^* -block is ℓ -diverse
- Provides privacy even when the data publisher does not know what kind of knowledge the adversary possesses
- 1-diversity is always trivially satisfied

Source: Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3-es.

L-Diversity: “Well-represented”

1. Distinct L-diversity (simplest definition) ensures that at least L distinct values for the sensitive field in each equivalence class exist.
2. Entropy L-diversity (most complex) defines Entropy of an equivalent class E to be the negation of summation of s across the domain of the sensitive attribute of $p(E, s) \cdot \log(p(E, s))$ where $p(E, s)$ is the fraction of records in E that have the sensitive value s .
 - A table has entropy L-diversity when for every equivalent class E , $Entropy(E) \geq \log(L)$.
3. Recursive (c-L)-diversity (compromise definition) ensures the most common value does not appear too often while less common values are ensured to not appear too infrequently.

Source: Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3-es.

To cite this presentation: Pendyala, V.S. (2023) “Machine Learning Applications and Sustainable Development” University of Bolton, UK, November to Remember Webinar series.

L-Diversity Example

4-anonymous inpatient microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

3-diverse inpatient microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Source: Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3-es.

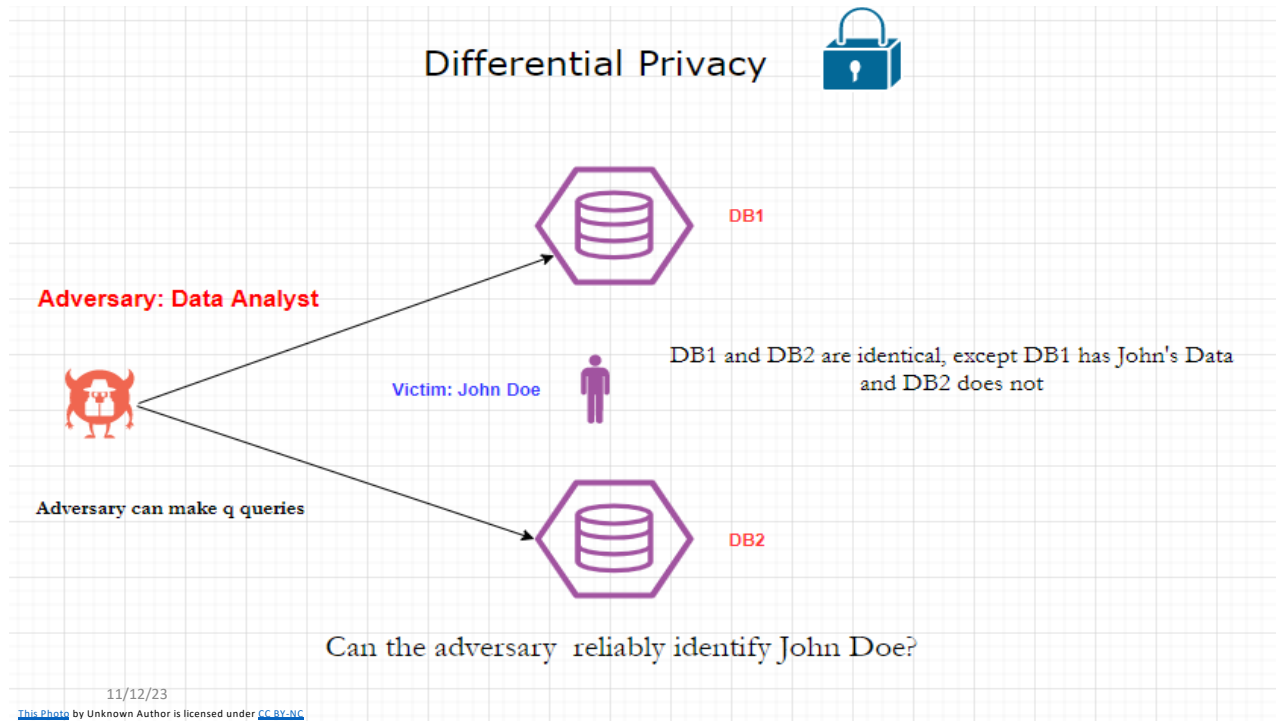
L-Diversity: Limitations

1. Does not provide a quantitative measure of privacy
2. Overgeneralizing data can lead to information (and accuracy) loss
3. Vulnerable to certain forms of background knowledge attacks
4. Less effective in scenarios where data changes over time
5. Computationally expensive, particularly for large datasets as it can involve complex algorithms for generalizing
6. Focuses on a chosen sensitive attribute, but other attributes might still be subject to privacy breaches => The choice of the sensitive attribute can affect the effectiveness of l-diversity.

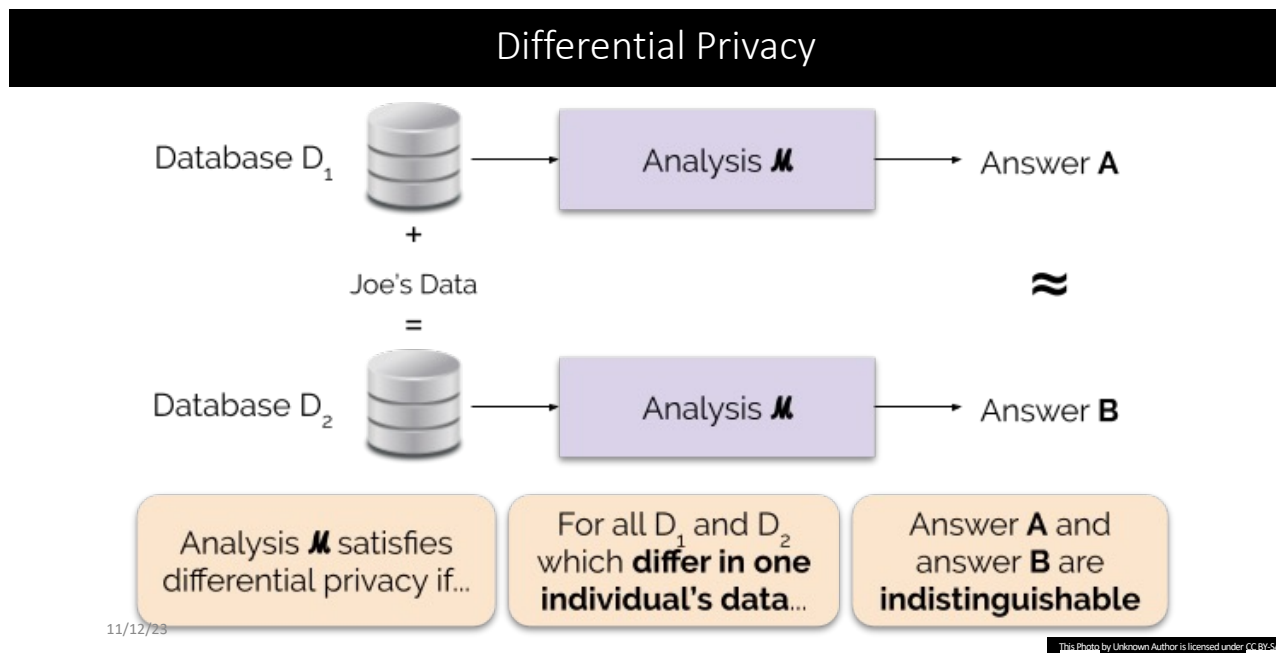
11/12/23

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.





To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.





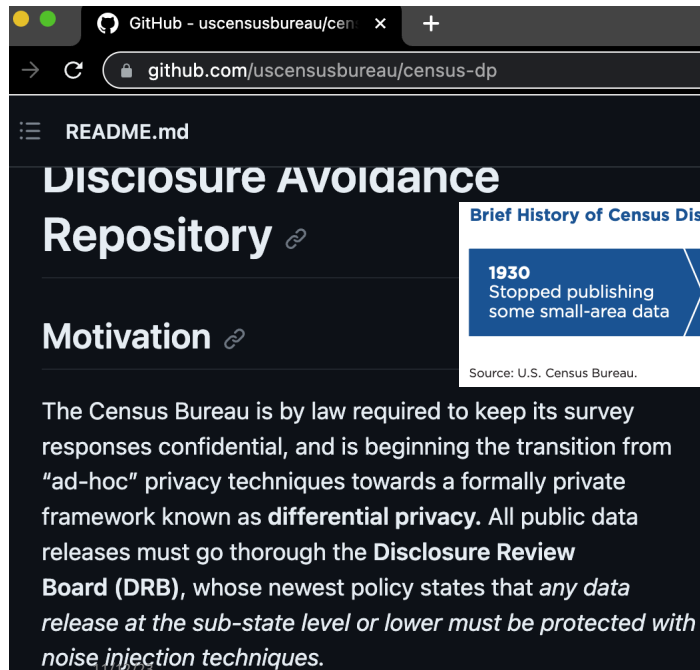
Warner, Stanley L.
"Randomized response: A survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60.309 (1965): 63-69.

Did you ever cheat in an exam?

11/12/23

This Photo by Unknown Author is licensed under CC BY-SA

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.



GitHub - uscensusbureau/cen- x +

github.com/uscensusbureau/census-dp

README.md

Disclosure Avoidance Repository

Motivation

The Census Bureau is by law required to keep its survey responses confidential, and is beginning the transition from "ad-hoc" privacy techniques towards a formally private framework known as **differential privacy**. All public data releases must go through the **Disclosure Review Board (DRB)**, whose newest policy states that *any data release at the sub-state level or lower must be protected with noise injection techniques*.

Source: U.S. Census Bureau.

Brief History of Census Disclosure Avoidance Protection

Year	Technique
1930	Stopped publishing some small-area data
1970	Whole-table suppression
1990	Data swapping, etc.
2020	Differential privacy

Here's how you know

United States[®]
Census
Bureau

SEARCH ME

Why the Census Bureau Chose Differential Privacy

Differential Privacy: From Data to Algorithms

Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer Berlin Heidelberg, 2006.

- Mathematical framework for protecting individual privacy in data analysis and sharing.
- Balances privacy and utility (accuracy etc) in data analysis
- Many applications in healthcare, finance, and government
- Neighbor datasets differ in the data of a single individual
- A mechanism M satisfies differential privacy if for all neighboring datasets x and x' , and all possible outputs S ,

$$\frac{P(M(x) = S)}{P(M(x') = S)} \leq e^\epsilon$$

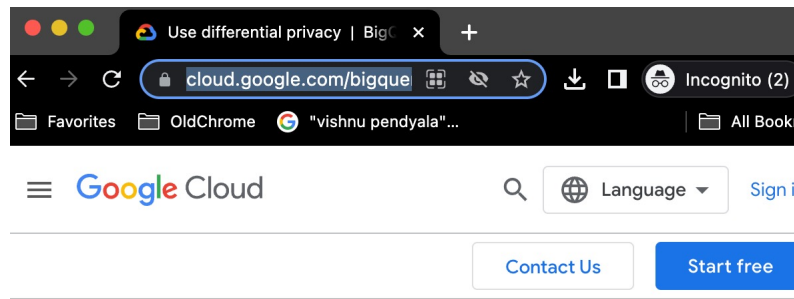
11/12/23

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

Differential Privacy

- Let M be average, x be the scores of students
- $x - x' = \text{score of student, John}$
- If we are told that $M(x) = 75$ and $M(x') = 80$, John's score can be easily guessed => bad privacy
- We need to add noise into the averaging mechanism, so that $M(x)$ and $M(x')$ differ only slightly
- ϵ is called (tunable) privacy parameter or privacy budget, typically ranges between 0.9 and 9
- Larger ϵ => less privacy; $\epsilon = 0$ is ideal
- Laplacian noise is commonly used, but Gaussian can also be used
- Sequential Composition: If M_1 is ϵ_1 -differentially private and M_2 is ϵ_2 then $F(M_1, M_2)$ satisfies $\epsilon_1 + \epsilon_2$ differentially private

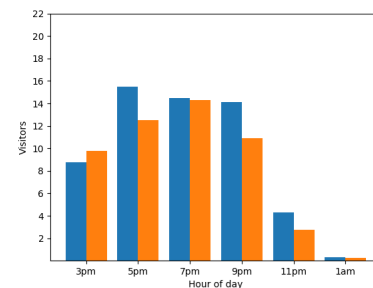
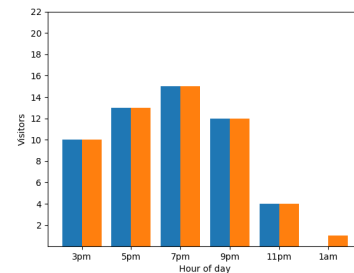
11/12/23



How differential privacy works on queries

The goal of [differential privacy](#) is to mitigate disclosure risk: the risk that someone can learn information about an entity in a dataset. Differential privacy balances the need to safeguard privacy against the need for statistical analytical utility. As privacy increases, statistical analytical utility decreases, and vice versa.

With GoogleSQL for BigQuery, you can transform the results of a query with differentially private aggregations. When the query is executed, it performs the following:



11/12/23

Source: <https://cloud.google.com/bigquery/docs/differential-privacy#:~:text=Differential%20privacy%20is%20commonly%20used,learning%20information%20about%20an%20individual>

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

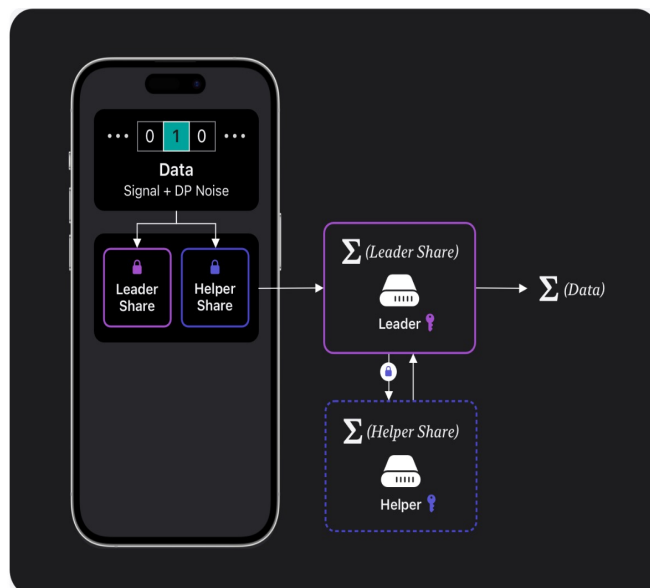


Figure 3: The secure aggregation protocol that enforces the DP assurance. Each binary vector is split into two shares, and each share is encrypted with a different public key. As long as no single entity gets access to both private keys, nobody can see any individual vectors, only the aggregate, which satisfies the desired DP assurance.



"In this article, we share how we apply differential privacy (DP) to learn about the kinds of photos people take at frequently visited locations (iconic scenes) without personally identifiable data leaving their device. This approach is used in several features in Photos, including choosing key photos for [Memories](#), and selecting key photos for locations in Places in iOS 17."

"So you can't learn anything about individual trips, but you can learn a lot about aggregate populations of users and trips."

"So you can't learn anything about individual trips, but you can learn a lot about aggregate populations of users and trips."

"...about a third of all Uber analyst queries Uber's are broad, statistical ones that don't require access to more fine-grained data---and under its differential privacy system, those statistical queries can't be used as cover for more invasive snooping. That would make the system a significant reduction in Uber staff's access to user details,..."

11/12/23

ANDY GREENBERG SECURITY JUL 13, 2017 10:02 AM

Uber's New Tool Lets Its Staff Know Less About You

The controversial ride-sharing service is making a push for "differential privacy," a method that masks users' individual data.

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

Google Chrome's use of Differential Privacy

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova. "Rappor: Randomized aggregatable privacy-preserving ordinal response." *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. 2014.

11/12/23

1. **Signal.** Hash client's value v onto the Bloom filter B of size k using h hash functions.
2. **Permanent randomized response.** For each client's value v and bit $i, 0 \leq i < k$ in B , create a binary reporting value B'_i which equals to

$$B'_i = \begin{cases} 1, & \text{with probability } \frac{1}{2}f \\ 0, & \text{with probability } \frac{1}{2}f \\ B_i, & \text{with probability } 1 - f \end{cases}$$

where f is a user-tunable parameter controlling the level of longitudinal privacy guarantee.

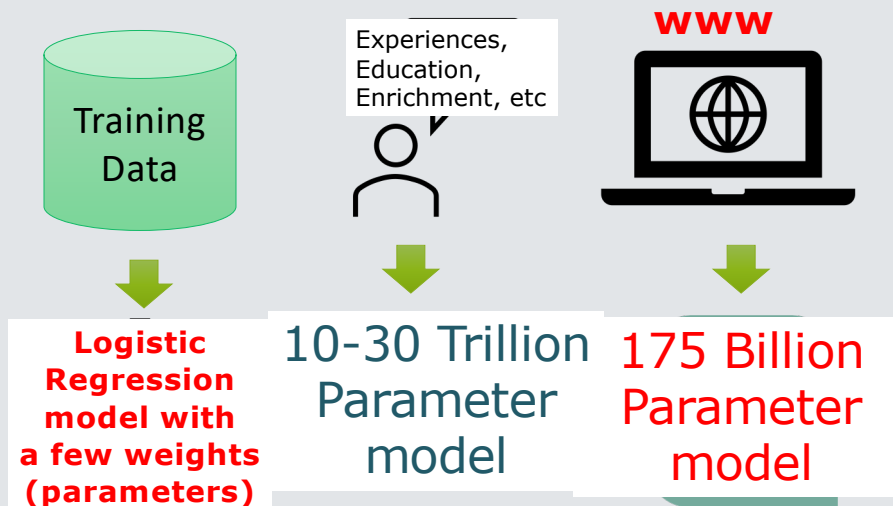
Subsequently, this B' is memoized and reused as the basis for all future reports on this distinct value v .

3. **Instantaneous randomized response.** Allocate a bit array S of size k and initialize to 0. Set each bit i in S with probabilities

$$P(S_i = 1) = \begin{cases} q, & \text{if } B'_i = 1. \\ p, & \text{if } B'_i = 0. \end{cases}$$

4. **Report.** Send the generated report S to the server.

MACHINE LEARNING MODELS ARE ADVANCED SEARCH ENGINES



11/12/23

©Vishnu S. Pendyala This work is licensed under a [Creative Commons Attribution-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nd/4.0/)

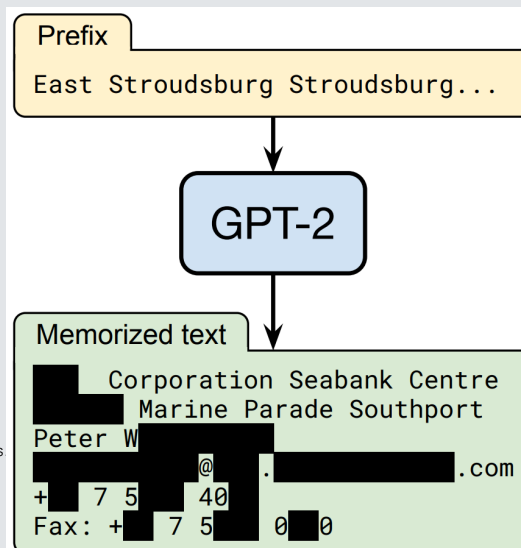
To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.



Source: https://imgs.xkcd.com/comics/predictive_models

11/12/23

Carlini, Nicholas, et al. "Extracting training data from large language models." *30th USENIX Security Symposium (USENIX Security 21)*. 2021.



"Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect

"The following order provisions apply to Edmodo, but they merit the attention of others in the ed tech industry."

...

*Edmodo **must delete models or algorithms** it developed using information collected from kids without school authorization or their parents' verifiable consent."*

11/12/23

Source: <https://www.ftc.gov/business-guidance/blog/2023/05/oh-no-you-dont-edmodo-ftc-sues-ed-tech-company-violating-school-kids-privacy>

An official website of the United States government

[Here's how you know](#) v

 **FEDERAL TRADE COMMISSION**
PROTECTING AMERICA'S CONSUMERS

[Home](#) / [Business Guidance](#) / [Business Blog](#)

Business Blog

Oh no, you don't, Edmodo: FTC sues ed tech company for violating school kids' privacy

By: Lesley Fair | May 22, 2023 | [f](#) [t](#) [in](#)

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

 TRISTAN GREENE

OCT 02, 2023

Researchers find LLMs like ChatGPT output sensitive data even after it's been 'deleted'

According to the scientists, there's no universal method by which data can be deleted from a pretrained large language model.

7979 Total views 5 Total shares Listen to article   3:47

"...even state-of-the-art model [editing](#) methods, such as Rank-One Model Editing "fail to fully delete factual information from LLMs, as facts can still be extracted 38% of the time by whitebox attacks and 29% of the time by blackbox attacks."

11/12/23

Source: <https://cointelegraph.com/news/artificial-intelligence-researchers-llms-chat-gpt-sensitive-data-after-deleted>

Machine Unlearning: "Right to be forgotten"

- Machine unlearning is the process of deliberately removing or updating knowledge or data from AI models and systems.
- Use cases: Privacy protection, rectify unethical behavior of AI models, eliminate outdated or irrelevant data from the models
- The problem is related, but tangential to privacy-preserving ML - here, ϵ must be 0 - in DP, it is usually non-zero
- Naïve approach: Retrain the model from scratch without D_u , which is computationally expensive
- Challenge: Non-deterministic training makes it difficult to understand the impact of select data items
- Goals: Comparable accuracy, minimum time unlearning, provable guarantees, and model agnostic.

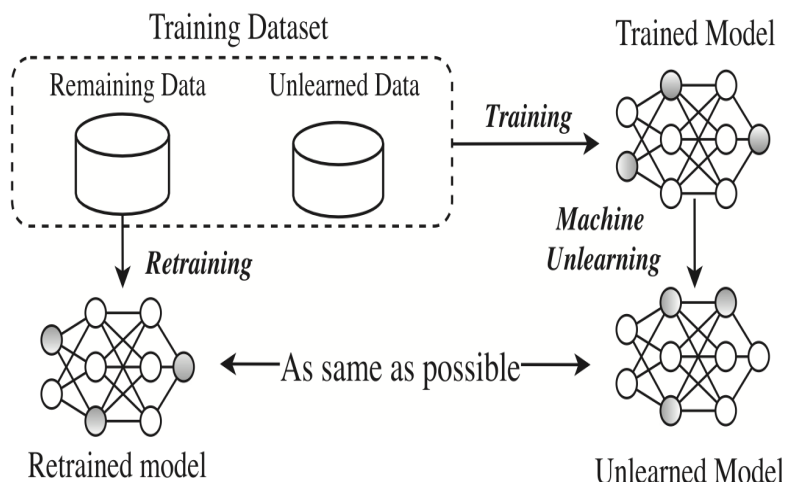
Source: Bourtole, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021, May). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 141-159). IEEE.

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

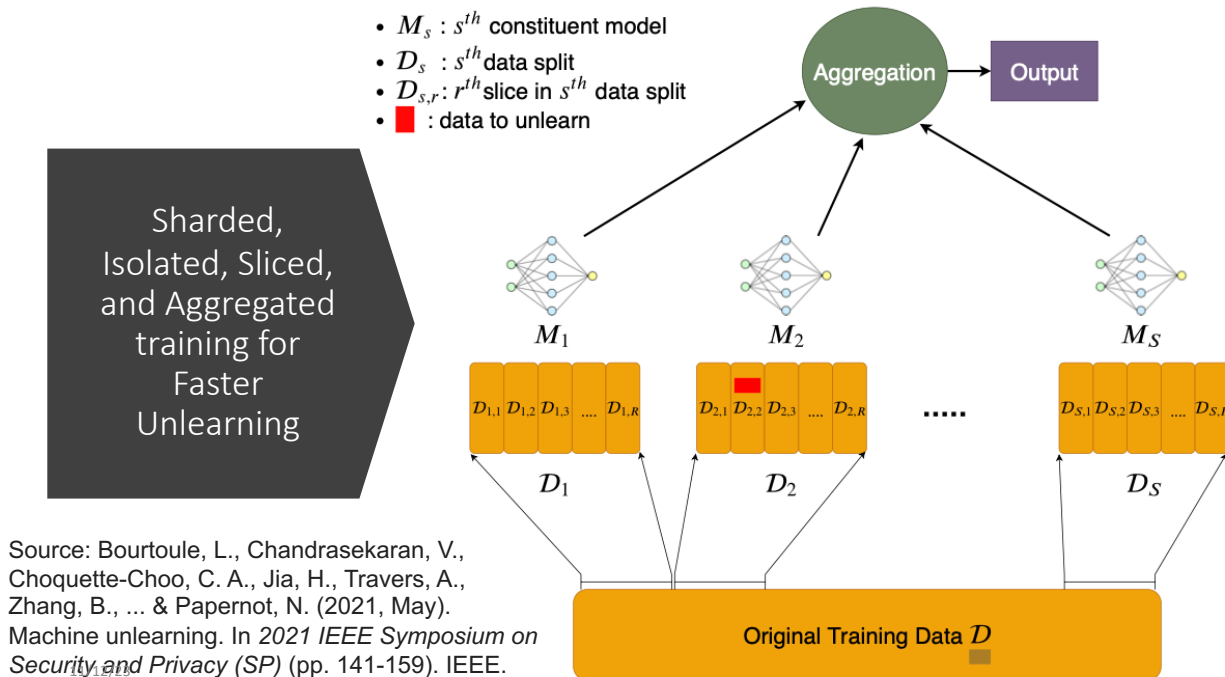
Def: An unlearning process

$U(A(D), D, D_u)$ is defined as a function from a trained model $A(D)$, a training dataset D , and an unlearning dataset D_u to a model w_u , which ensures that the unlearned model w_u performs as though it had never seen the unlearning dataset D_u .

Machine Unlearning



Source: Xu, Heng, et al. "Machine unlearning: A survey." *ACM Computing Surveys* 56.1 (2023): 1-36.



To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.

The big news from the Whitehouse (verbatim)

New Standards for AI Safety and Security

- Require that developers of the most powerful AI systems share their **safety test results** and other critical information with the U.S. government.
- Government will Develop standards, tools, and tests to help ensure that AI systems are **safe, secure, and trustworthy**.
- Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for **detecting AI-generated content and authenticating official content**.
- **Harness AI's** potentially game-changing cyber capabilities **to make software and networks more secure**.

Source: <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>

Protecting Americans' Privacy, Equity and Civil Rights

1. Accelerate the **development** and use of **privacy-preserving** techniques
2. Strengthen privacy-preserving **research** and technologies
3. **Evaluate how agencies collect and use** commercially available information
4. Develop guidelines for federal agencies to **evaluate** the effectiveness of privacy-preserving techniques
5. Address **algorithmic discrimination** through training, technical assistance, and coordination on best practices for investigating and prosecuting civil rights violations related to AI.
6. Ensure **fairness** throughout the criminal justice.
7. Support educators deploying AI-enabled educational tools, such as **personalized tutoring** in schools.

Source: <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/>

To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.





To cite this presentation: Pendyala, V.S. (2023) "Machine Learning Applications and Sustainable Development" University of Bolton, UK, November to Remember Webinar series.