# Your Data Are a Fingerprint

*Why Anonymization is Not Anonymous (and how Statistics can protect you)*

Dylan Spicker

January 31, 2025

A tremendous amount of data are collected by researchers, companies, governments, and similar institutions.

# Pillar #1

Learning from data is crucial for our understanding of the world.

## Pillar #1

Learning from data is crucial for our understanding of the world.

## Pillar #2

Protecting individual privacy is necessary as data become more abundant.

Imagine *someone* trying to learn your private information using information based on your data.

This is analogous to a detective searching a crime scene.

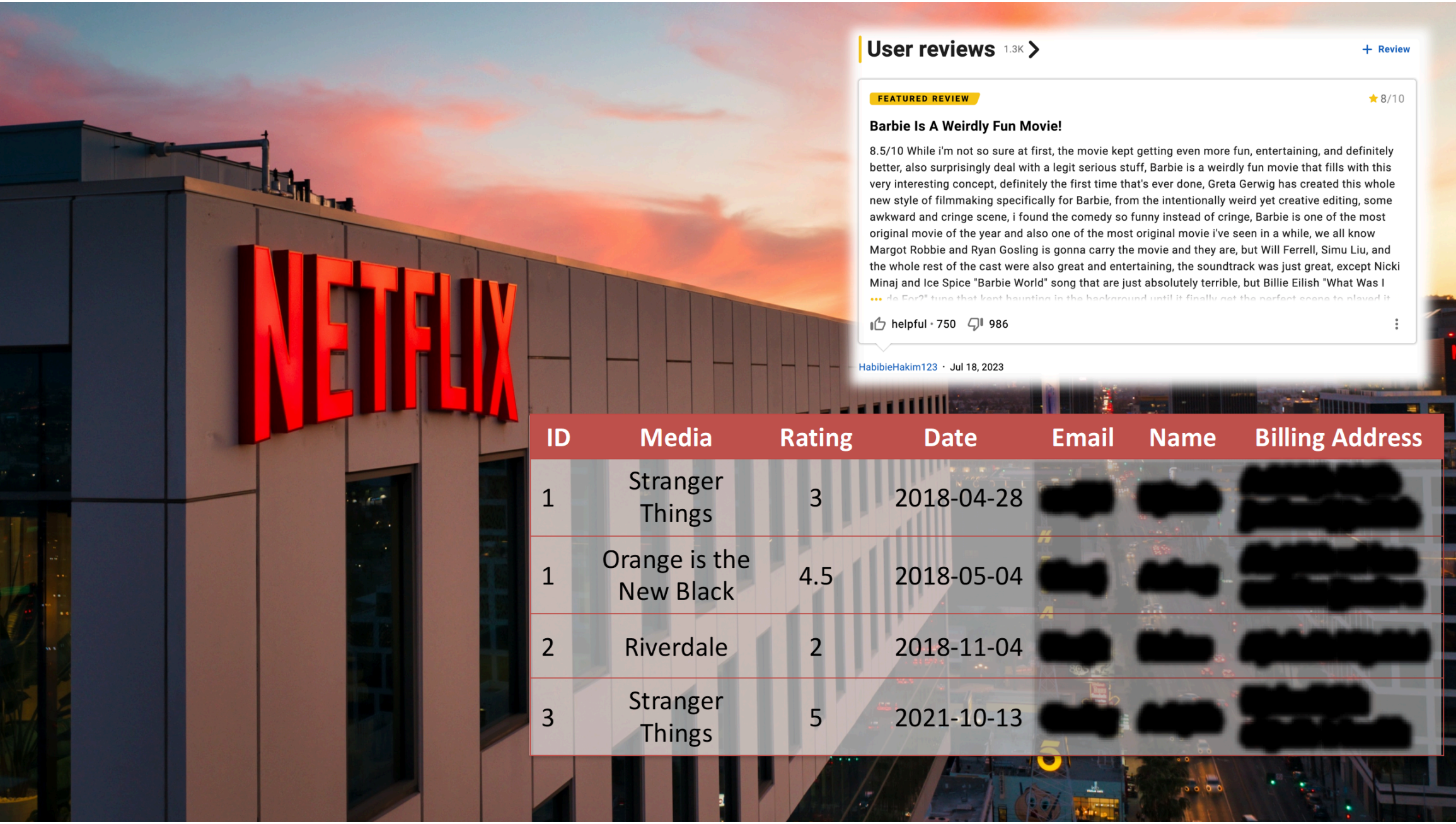
# Caught "Red-Handed": Direct Data Releases

If data are directly released this constitutes a complete privacy violation.

What if data are anonymized?

| A | B | C | D | E |
|---|---|---|---|---|
| urity_number | first_name | last_name | gender | prescription |
| 03-2310 | Fancy | Thonason | Female | Diclofenac Sodium |
| 11-7096 | Tedd | Hanton | Male | Octinoxate and Oxybenzone |
| 25-6875 | Ashlin | Byatt | Male | Permethrin |
| 48-0157 | Ashby | Pughe | Male | Nafcillin |
| 64-2748 | Avictor | Salerg | Male | Naproxen Sodium |
| 59-3263 | Court | O'Deegan | Male | OCTINOXATE, OXYBENZONE, ZINC OXIDE |
| 64-7089 | Dulce | Tomankiewicz | Female | risperidone |
| 31-5445 | Adamo | Mc Caughan | Male | OPERCULATA FRUIT and MERCURIC IODIDE and SUS SCROFA NASAL ML |
| 90-1403 | Odey | Santus | Male | Levofloxacin |
| 22-8971 | Beck | Bullier | Genderfluid | Menthol |
| 53-2357 | Nita | Castagnone | Female | Acetaminophen |
| 59-7590 | Betta | Vedstra | Female | Oxycodone Hydrochloride and Aspirin |
| 09-5952 | Sharona | Pallaske | Female | Salicylic acid |
| 17-7088 | Wash | M'Quhan | Male | levalbuterol hydrochloride |

Data anonymization is the process of removing clear personal identifiers from a specific dataset.




## User reviews 1.3K

+ Review

FEATURED REVIEW

8/10

**Barbie Is A Weirdly Fun Movie!**

8.5/10 While i'm not so sure at first, the movie kept getting even more fun, entertaining, and definitely better, also surprisingly deal with a legit serious stuff, Barbie is a weirdly fun movie that fills with this very interesting concept, definitely the first time that's ever done, Greta Gerwig has created this whole new style of filmmaking specifically for Barbie, from the intentionally weird yet creative editing, some awkward and cringe scene, i found the comedy so funny instead of cringe, Barbie is one of the most original movie of the year and also one of the most original movie i've seen in a while, we all know Margot Robbie and Ryan Gosling is gonna carry the movie and they are, but Will Ferrell, Simu Liu, and the whole rest of the cast were also great and entertaining, the soundtrack was just great, except Nicki Minaj and Ice Spice "Barbie World" song that are just absolutely terrible, but Billie Eilish "What Was I ... de For?" tune that kept haunting in the background until it finally get the perfect scene to played it

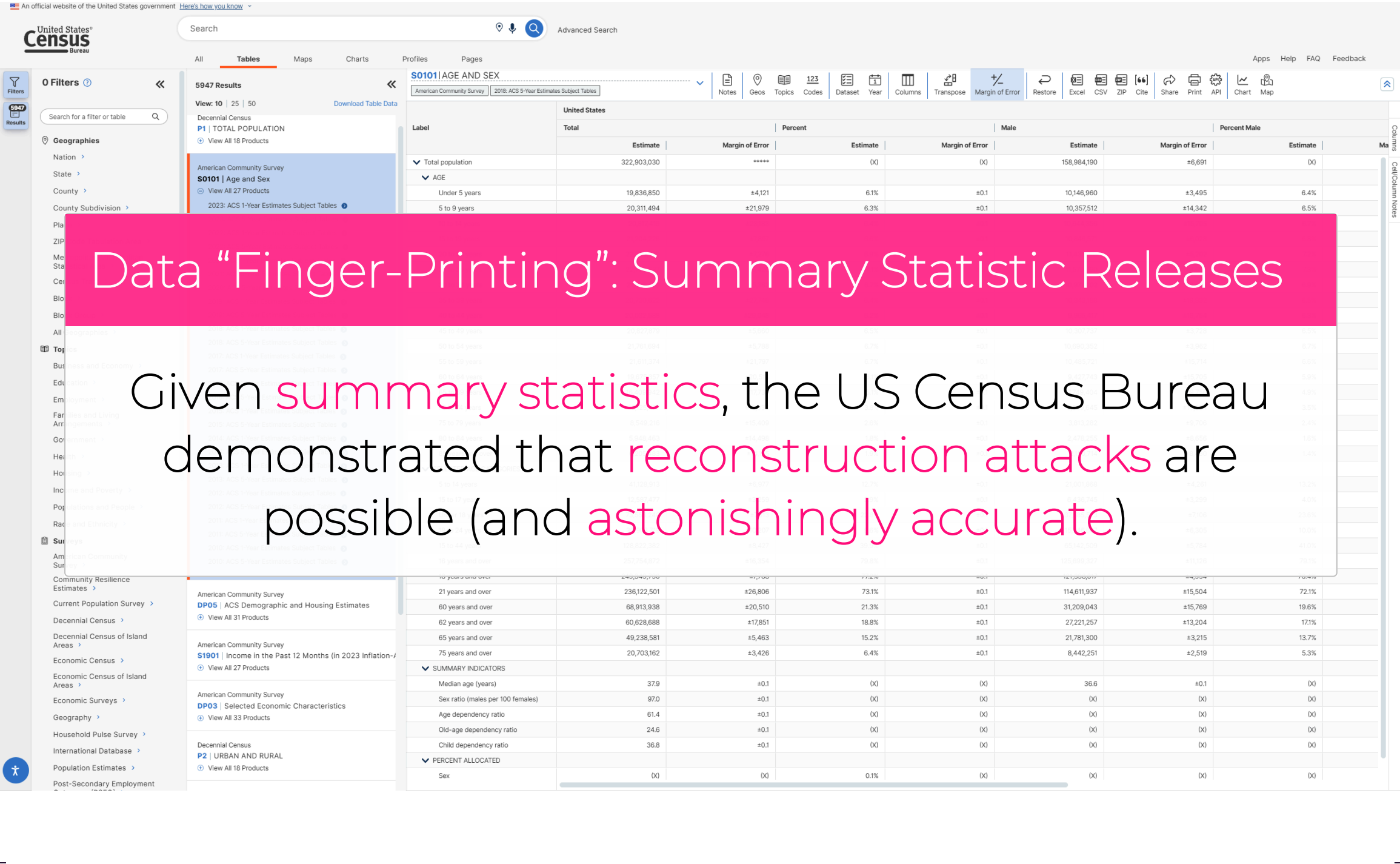
helpful · 750 986

HabibieHakim123 · Jul 18, 2023

| ID | Media | Rating | Date | Email | Name | Billing Address |
|---|---|---|---|---|---|---|
| 1 | Stranger Things | 3 | 2018-04-28 | [redacted] | [redacted] | [redacted] |
| 1 | Orange is the New Black | 4.5 | 2018-05-04 | [redacted] | [redacted] | [redacted] |
| 2 | Riverdale | 2 | 2018-11-04 | [redacted] | [redacted] | [redacted] |
| 3 | Stranger Things | 5 | 2021-10-13 | [redacted] | [redacted] | [redacted] |

Sufficient amounts of non-personally identifying information can act as personal identifiers.

Summary statistics like means, counts, or standard deviations can be reported to describe the overall data.




# Data "Finger-Printing": Summary Statistic Releases

Given summary statistics, the US Census Bureau demonstrated that reconstruction attacks are possible (and astonishingly accurate).

You leave trace amounts of personal information in summary statistics.

What about using advanced statistical procedures?

Received: 14 September 2023 | Revised: 18 October 2023 | Accepted: 12 November 2023
DOI: 10.1002/sta4.641

ORIGINAL ARTICLE

**Differentially private outcome-weighted lea
dynamic treatment regime estimation**

Dylan Spicker[1] | Erica E. M. Moodie[2] | Susan M. Shortre

[1]Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, New Brunswick, Canada
[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada
[3]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA
[4]Department of Biostatistics, University of Washington, Seattle, WA, USA

**Correspondence**
Dylan Spicker, Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, NB, Canada.
Email: dylan.spicker@unb.ca

**Funding information**
Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873. This research was undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program. Additionally, Dylan Spicker was supported by the Canadian Statistical Sciences Institute (CANSSI) and STATLAB and Erica E.M. Moodie by a chercheur de mérite career award from the FRQ, Santé. The STAR*D study was supported by NIMH Contract #N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is CT00021528.

Precision medicine is a framework for develo
mendations that seeks to determine the optima
all of the relevant, observable patient-level cha
cine relies on highly sensitive, patient-level data
is of great importance. Dynamic treatment regi
of precision medicine in a longitudinal setting.
a family of techniques for estimating optimal
OWL techniques leverage support vector mach
form estimation. SVMs perform classification b
the data known as support vectors. The classifi
requires direct access to the support vectors. Th
mated with OWL requires the release of patien
sample. As a result, the classification rules from S
lation for those individuals whose data compri
violation is a major concern, particularly in ligh
medical data that are used in DTR estimation. D
mathematical framework for ensuring the privac
able guarantees on the likelihood that individual
an adversary. We provide the first investigation
of DTRs and provide a differentially private OW
allowing us to quantify the cost of privacy in t
estimators.

**KEYWORDS**
differential privacy, dynamic treatment regimes, indivi
support vector machines

## 1 | INTRODUCTION

Health data are, by their nature, sensitive information. We assume that patients expect that their sensitive information will remain protected and secure once it has been collected. The expectation is that no one outside of those who were explicitly granted access (such as the researchers receiving informed consent) should be able to reliably learn about the sensitive information that has been collected on each individual. This idea is broadly referred to as privacy. Our concern is in the study of privacy as it relates to precision medicine. Precision medicine

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Authors. *Stat* published by John Wiley & Sons Ltd.

# Circumstantial Evidence: Advanced Statistical Modelling

Even advanced statistical modelling is susceptible to data leakage.

The severity depends on the underlying analysis.

# What can we do?

Differential Privacy: A rigorous mathematical standard for privacy, that, if achieved, provides provable guarantees for individual protection.

| Family | Income |
|---|---|
| 1 | 60,000 |
| 2 | 58,000 |
| 3 | 63,000 |
| 4 | 95,000 |
| 5 | 61,000 |
| 6 | 12,000 |
| 7 | 63,000 |
| 8 | 125,000 |
| 9 | 57,000 |
| 10 | 55,000 |

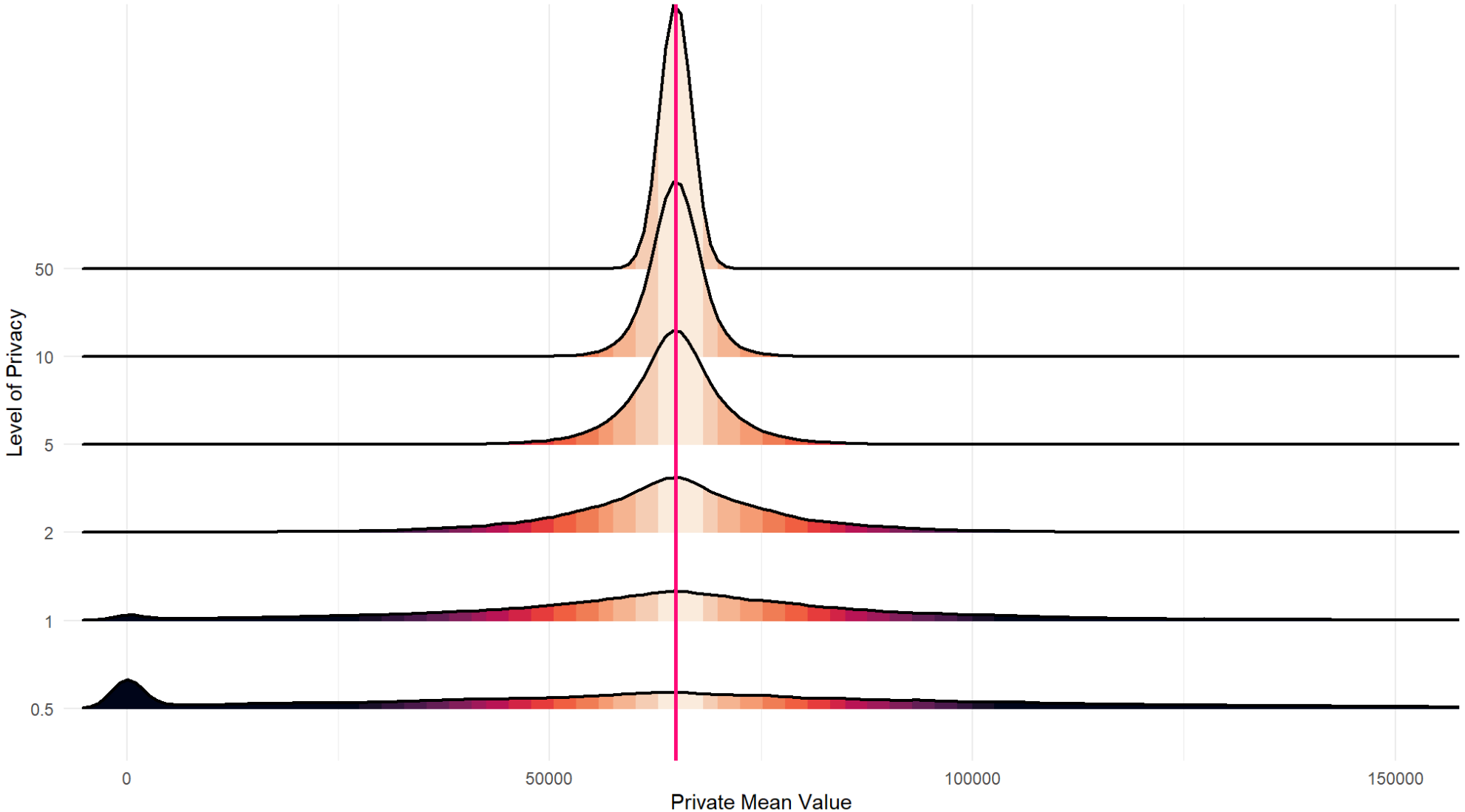
$$\text{Mean} = \frac{60 + 58 + 63 + 95 + 61 + 12 + 63 + 125 + 57}{10}$$

$$= 6$$

$$\text{Private Mean} = \text{Mean} + \text{Random Noise}$$

- $3994.80$ produces: $64,900 + 3994.80 = 68,894.80$
- $-4987.57$ produces: $64,900 - 4987.57 = 59,912.43$
- $-12943.21$ produces: $64,900 - 12,943.21 = 51,956.79$
- $35374.81$ produces: $64,900 + 35,374.81 = 100,274.81$

The specific value of the private mean depends on the random noise. It may be close or far from the truth.


Level of Privacy
50
10
5
2
1
0.5
0
50000
100000
150000
Private Mean Value

Received: 14 September 2023 | Revised: 18 October 2023 | Accepted: 12 November 2023
DOI: 10.1002/sta4.641

ORIGINAL ARTICLE

**Differentially private outcome-weighted lea… dynamic treatment regime estimation**

Dylan Spicker[1] | Erica E. M. Moodie[2] | Susan M. Shortre…[3,4]

[1]Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, New Brunswick, Canada
[2]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada
[3]Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA
[4]Department of Biostatistics, University of Washington, Seattle, WA, USA

**Correspondence**
Dylan Spicker, Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, NB, Canada.
Email: dylan.spicker@unb.ca

**Funding information**
Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873. This research was undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program. Additionally, Dylan Spicker was supported by the Canadian Statistical Sciences Institute (CANSSI) and STATLAB and Erica E.M. Moodie by a chercheur de mérite career award from the FRQ, Santé. The STAR*D study was supported by NIMH Contract #N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is CT00021528.

Precision medicine is a framework for develo… mendations that seeks to determine the optima… all of the relevant, observable patient-level cha… cine relies on highly sensitive, patient-level data… is of great importance. Dynamic treatment regi… of precision medicine in a longitudinal setting. … a family of techniques for estimating optimal … OWL techniques leverage support vector mach… form estimation. SVMs perform classification b… the data known as support vectors. The classifi… requires direct access to the support vectors. Th… mated with OWL requires the release of patien… sample. As a result, the classification rules from … lation for those individuals whose data compri… violation is a major concern, particularly in ligh… medical data that are used in DTR estimation. … mathematical framework for ensuring the priva… able guarantees on the likelihood that individual… an adversary. We provide the first investigation … of DTRs and provide a differentially private OW… allowing us to quantify the cost of privacy in t… estimators.

KEYWORDS
differential privacy, dynamic treatment regimes, indivi… support vector machines

## 1 | INTRODUCTION

Health data are, by their nature, sensitive information. We assume that patients expect that their sensitive information will remain protected and secure once it has been collected. The expectation is that no one outside of those who were explicitly granted access (such as the researchers receiving informed consent) should be able to reliably learn about the sensitive information that has been collected on each individual. This idea is broadly referred to as privacy. Our concern is in the study of privacy as it relates to precision medicine. Precision medicine

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2024 The Authors. *Stat* published by John Wiley & Sons Ltd.

# Private Personalized Medicine

Treatment is more impactful for some patients than others.

Our method achieves high accuracy for those patients and provides privacy for all patients.

## Pillar #1

Learning from data is crucial for our understanding of the world.

## Pillar #2

Protecting individual privacy is necessary as data become more abundant.

There is an inherent trade-off between these pillars. Privacy research makes this explicit.

# Thank you.

dylanspicker.com | dylan.spicker@unb.ca