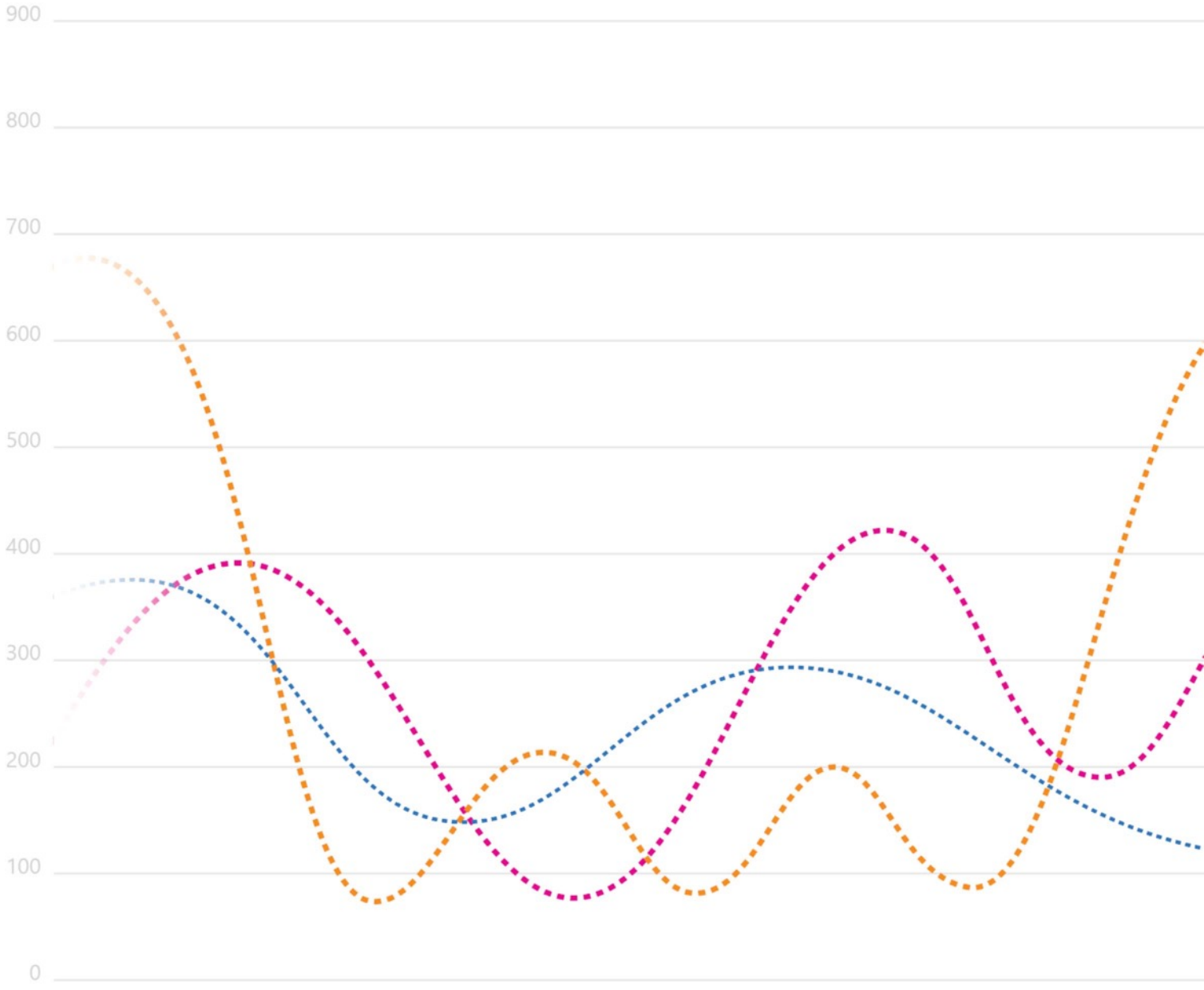
Microsoft

Differential Privacy

Juan M. Lavista Ferres



AI for Good Research Lab

Topics

1. Differential Privacy in Windows 10
2. Why many DP implementations fail?
3. Why I'm still so passionate about DP?

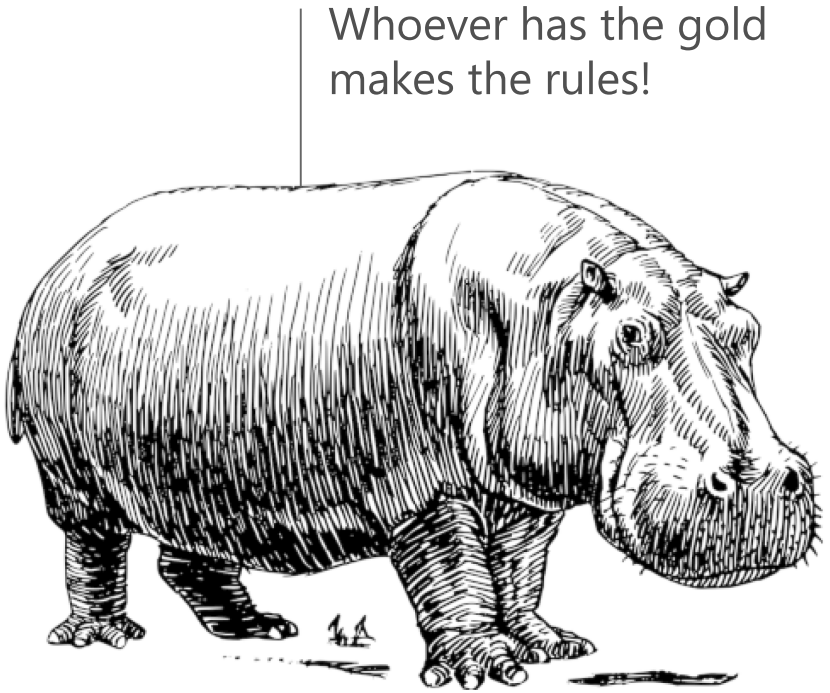
1. Differential Privacy in Windows 10

Differential Privacy in Windows 10

- In Windows 10, we wanted to be as Data Driven as possible
- No more HIPPO decision making (HIPPO = **H**ighest **P**aid **P**erson **O**pinion)
- Objective: try to emulate as best as we could an online service environment.

"If we have data, let's look at data. If all we have are opinions, let's go with mine."

Jim Barksdale, CEO of Netscape, famously said what could be summarized as Data or HiPPO.



HIPPO Highest Paid Person's Opinion
Source deanondelivery.com, *by Jafar*

Source https://exp-platform.com/hippo/

Windows 10 Telemetry

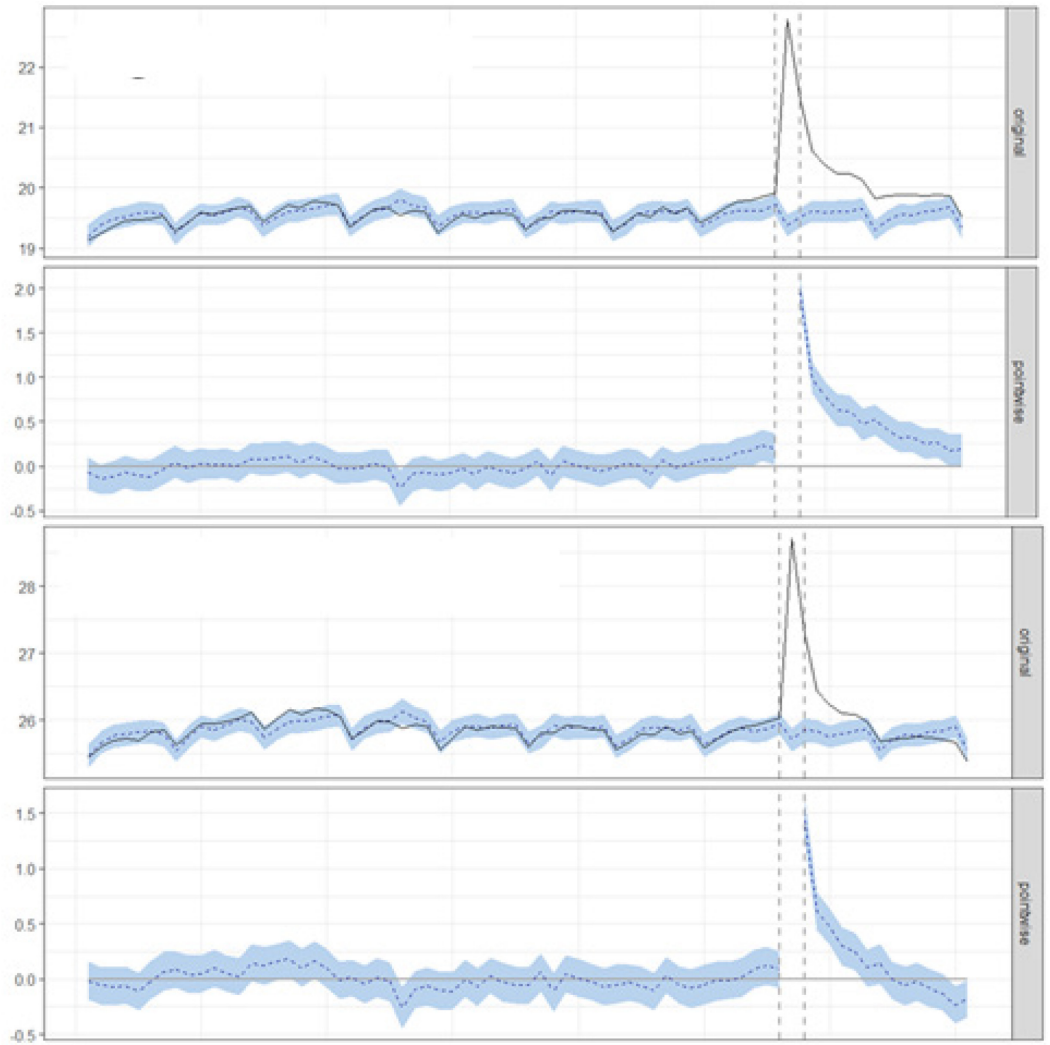
- In Windows 10 we collect a lot of metrics in an anonymous way as part of telemetry
- Users have the option to Opt-Out from telemetry
- There is a tradeoff between data collection and opt-out rate
- We still have 100s of millions of devices that don't opt-out

Problem Information from out-out machines is NOT missing at random

Windows 10 Telemetry

Opt out users effect on metrics

out-out machines is NOT missing at random



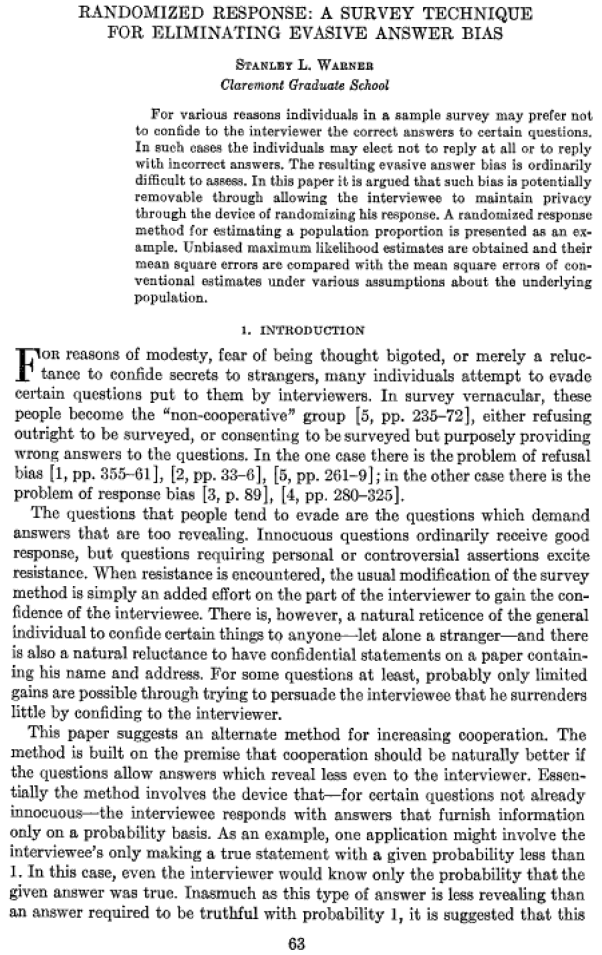
Windows 10 Telemetry

- To help with this problem we needed a solution that could provide us with the signal, but without affecting the privacy of the individuals
- **Solution: use Differential Privacy**

  We can't have a trusted curator → Local Differential Privacy (LDP)

  In LDP differential privacy happens before data is being transmitted

  LDP is a generalization from S.L Warner's Randomize response



RANDOMIZED RESPONSE: A SURVEY TECHNIQUE FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
Claremont Graduate School

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

1. INTRODUCTION

FOR reasons of modesty, fear of being thought bigoted, or merely a reluctance to confide secrets to strangers, many individuals attempt to evade certain questions put to them by interviewers. In survey vernacular, these people become the "non-cooperative" group [5, pp. 235–72], either refusing outright to be surveyed, or consenting to be surveyed but purposely providing wrong answers to the questions. In the one case there is the problem of refusal bias [1, pp. 355–61], [2, pp. 33–6], [5, pp. 261–9]; in the other case there is the problem of response bias [3, p. 89], [4, pp. 280–325].

The questions that people tend to evade are the questions which demand answers that are too revealing. Innocuous questions ordinarily receive good response, but questions requiring personal or controversial assertions excite resistance. When resistance is encountered, the usual modification of the survey method is simply an added effort on the part of the interviewer to gain the confidence of the interviewee. There is, however, a natural reticence of the general individual to confide certain things to anyone—let alone a stranger—and there is also a natural reluctance to have confidential statements on a paper containing his name and address. For some questions at least, probably only limited gains are possible through trying to persuade the interviewee that he surrenders little by confiding to the interviewer.

This paper suggests an alternate method for increasing cooperation. The method is built on the premise that cooperation should be naturally better if the questions allow answers which reveal less even to the interviewer. Essentially the method involves the device that—for certain questions not already innocuous—the interviewee responds with answers that furnish information only on a probability basis. As an example, one application might involve the interviewee's only making a true statement with a given probability less than 1. In this case, even the interviewer would know only the probability that the given answer was true. Inasmuch as this type of answer is less revealing than an answer required to be truthful with probability 1, it is suggested that this

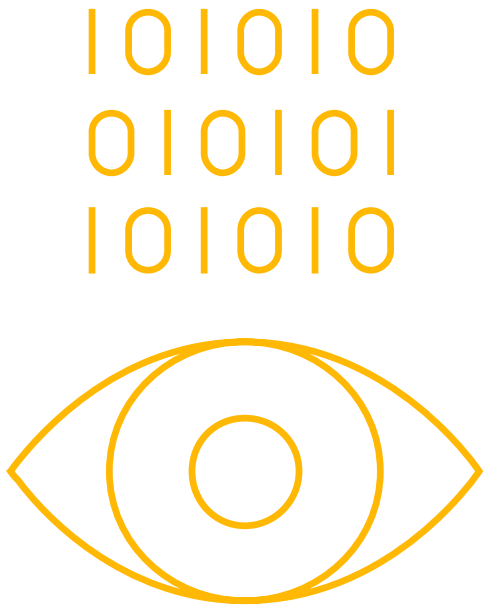
63

Definition 1 ([17, 9, 4]). *A randomized algorithm $\mathcal{A} : \mathcal{V} \to \mathcal{Z}$ is ϵ-locally differentially private (ϵ-LDP) if for any pair of values $v, v' \in \mathcal{V}$ and any subset of output $S \subseteq \mathcal{Z}$, we have that*

$$\mathbf{Pr}[\mathcal{A}(v) \in S] \leq e^{\epsilon} \cdot \mathbf{Pr}[\mathcal{A}(v') \in S] .$$

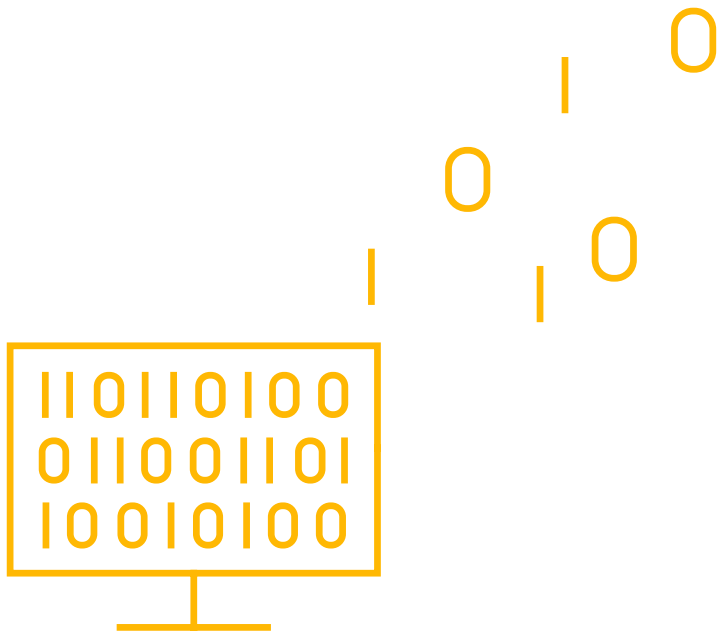
Windows 10 Telemetry

- For a lot of LDP models (for example in surveys) you collect the data once, and this works very well
- In telemetry, we systematically collect data many times across the lifetime of a device
- We do have a privacy leakage problem, given that every time we collect telemetry, we lose some privacy



Windows 10 Telemetry

- Given T is the number of times we have collected telemetry from a device
- and we have an epsilon differentially private mechanism
- Worst case $e^{T \cdot \varepsilon}$
- When T is big, this is too big of a privacy loss



Windows 10 Telemetry

Solution ?

1 Bit algorithm

$$Y_i = A(x_i) = \begin{cases} 1, & \text{with probability } \frac{1}{e^\epsilon + 1} + \frac{x_i}{M} \cdot \frac{e^\epsilon - 1}{e^\epsilon + 1}; \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Mean}(x_1, x_2, \ldots, x_N) \approx \frac{M}{N} \cdot \sum_{i=1}^{N} \frac{Y_i \cdot (e^\epsilon + 1) - 1}{e^\epsilon - 1}.$$

Memoisation (e^epsilon)

Problem Metrics like usage (real number) change every day

Solution discretize the numbers into buckets

Collecting Telemetry Data Privately

Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin
Microsoft Research
{bolind, jakul, yekhanin}@microsoft.com

Abstract

The collection and analysis of telemetry data from user's devices is routinely performed by many software companies. Telemetry collection leads to improved user experience but poses significant risks to users' privacy. Locally differentially private (LDP) algorithms have recently emerged as the main tool that allows data collectors to estimate various population statistics, while preserving privacy. The guarantees provided by such algorithms are typically very strong for a single round of telemetry collection, but degrade rapidly when telemetry is collected regularly. In particular, existing LDP algorithms are not suitable for repeated collection of counter data such as daily app usage statistics.

In this paper, we develop new LDP mechanisms geared towards repeated collection of counter data, with formal privacy guarantees even after being executed for an arbitrarily long period of time. For two basic analytical tasks, mean estimation and histogram estimation, our LDP mechanisms for repeated data collection provide estimates with comparable or even the same accuracy as existing single-round LDP collection mechanisms. We conduct empirical evaluation on real-world counter datasets to verify our theoretical results.

Our mechanisms have been deployed by Microsoft to collect telemetry across millions of devices.

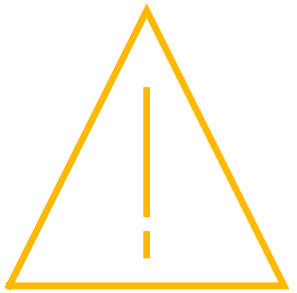
Windows 10 Differential Privacy API

- Randomized Response (single bit report)
- Counts (IE/Edge MAU/DAU, Domain Joined, etc.)
- Histograms (Linear, Exponential, and Custom)
- Mean (System Uptime and App Usage in Seconds)

```
TraceLoggingWrite(g_hProvider,
    "AggregateEvent",
    TraceLoggingKeyword(MICROSOFT_KEYWORD_MEASURES),
    TraceLoggingEventTag(MICROSOFT_EVENTTAG_AGGREGATE),
    UtcAggregationParameters(60,
5, UTC_EVENT_AGGREGATION_MODE_KEEP_RANDOM),
    TraceLoggingValue(1, "key"),
    UtcDiffpValue_NumericMean(L"DiffpVal",
        100000,
        4500000,
        UTC_FIELD_AGGREGATION_MODE_SUM,
        DIFFP_DEFAULT_EPSILON,
        DIFFP_DEFAULT_PERTURBATION_FACTOR));
```

Since 2017, Windows 10 standard telemetry collection library, Universal Telemetry Client, has included support for differential privacy. Developers sending telemetry simply need to add some flags to the call, and the result will be randomized before sending to the server.

2. Why many DP implementations fail?



2 2. Lessons learned from failed implementations of DP

Majority of DP projects fail (some even before starting)

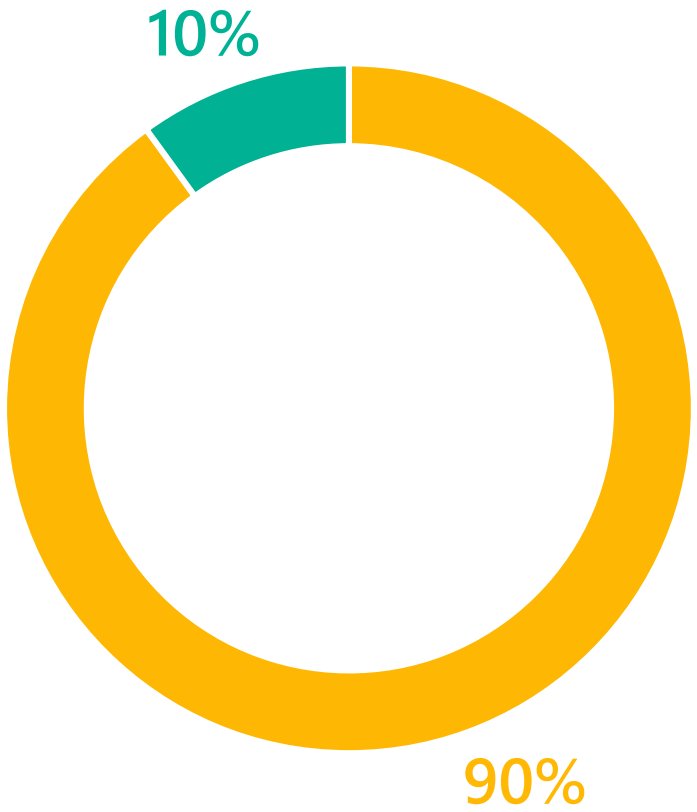
Common theme?

We have a marketing problem

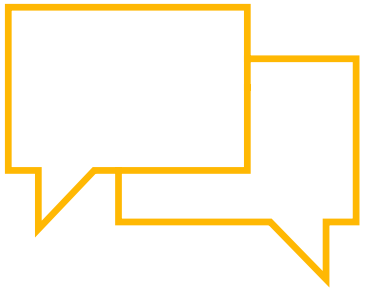
1. We have an awareness problem

We have a problem of awareness; people don't know what differential privacy is…

In a survey, over 90% of software engineers were not aware of what differential privacy is



2. We have a branding problem



People that have never heard about Differential Privacy, think DP is about having different levels of privacy depending on who the individual is.

I believe it means depending on who a person is they're subjected to a different level of privacy

It is the amount of privacy people think you have a right to depending on social status, occupation or maybe an event that occurred to you.

I don't know what it is. I think it may mean that someone who is more high in status may have more privacy than someone with less status.

Differential privacy is based on the individual. It takes into factors such as job, education, criminal background, current employment status, etc.. From these factors, a score is given to determine the amount of privacy one person is allowed to have.

I believe it means depending on who a person is they're subjected to a different level of privacy

Differential Pricing $

3. We are selling a product to solve a problem our customers don't know they have

- The majority of our customers think that by just sharing anonymized data they solve the privacy problems.
- In **Healthcare** they recognize they have a de-anonymization problem, however, they make it clear that de-anonymization is prohibited by law

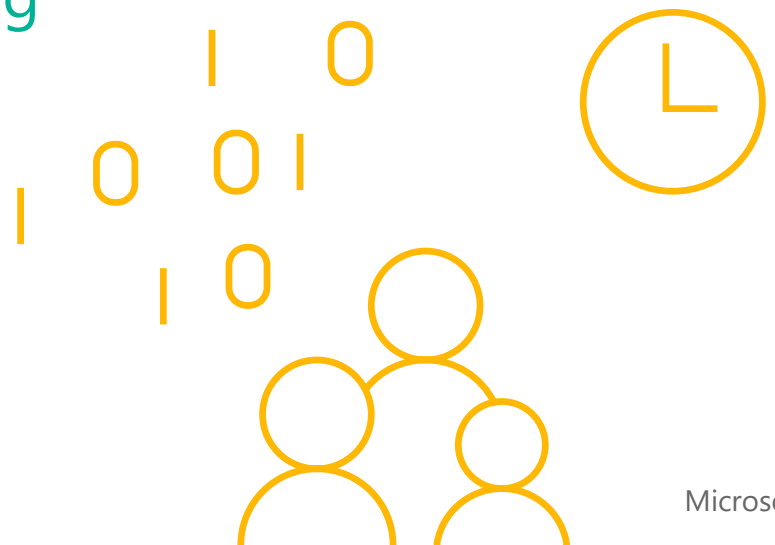
The Public Health Service Act (Section 308 (d)) provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.

Any effort to determine the identity of any reported case is prohibited by this law.

4. When they learn about differential privacy, they have irrational expectations about the product

- Customers think DP is a magic box that can solve all their problems
- Common frustration: Global models they can query, but data is not accessible in a raw format

Data Scientist spend 80% of their time wrangling with data, and 20% of their time complaining about wrangling with data.



5. Once they learn that they have a problem, it is not entirely clear to them that DP can solve it.

AKA The Epsilon Problem

For Neighboring datasets: Two datasets D and D', such that D' can be obtained by changing one single tuple in D

A randomized algorithm M satisfies ε-differential privacy, if for any two neighboring datasets D and D' and for any output R,

Where the probability space in each case is over the coin flips of the mechanism M

$$\frac{\text{Prob (R | D)}}{\text{Prob (R | D')}} <= \text{e^epsilon}$$

3. Why I'm still so passionate about DP?

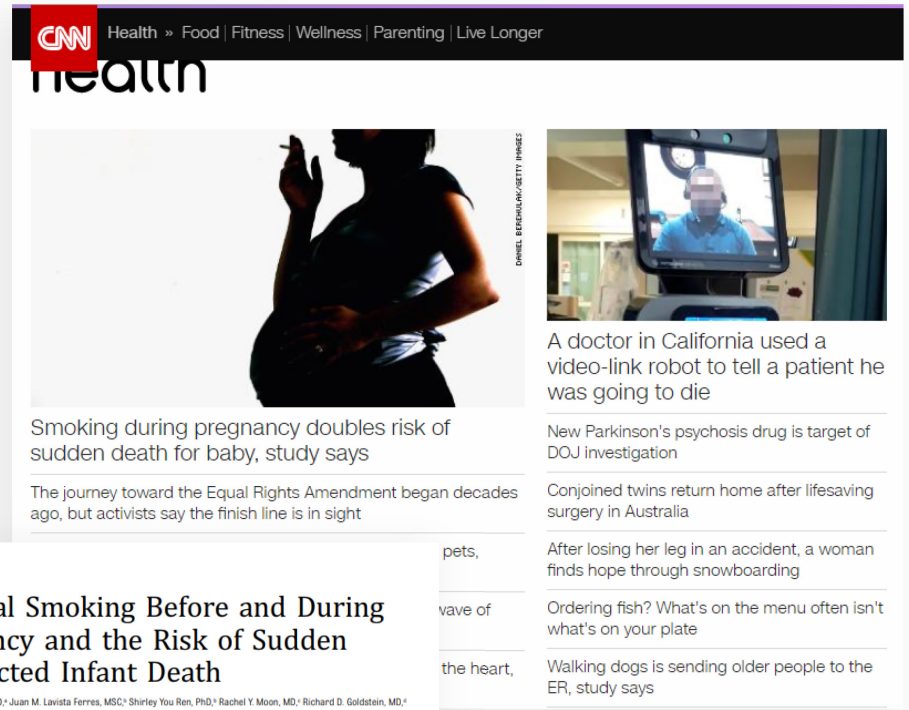
A colleague lost a child to SIDS, and he was doing an amazing campaign raising awareness about SIDS

SIDS = Sudden Infant Death Syndrome, main cause of death of babies from 1 month to 1 year old

~4000 children die every year of SIDS in the US, and we don't know why (**2 buses full of children** every week for a full year)

3. Why I'm still so passionate about DP?

- I wanted to help with Research
- Data = CDC Dataset **(every kid that was born in the US, and 1-year cohort of those that died)**
- Working with top SIDS researchers in the world



Maternal Smoking Before and During Pregnancy and the Risk of Sudden Unexpected Infant Death

Tatiana M. Anderson, PhD,[a] Juan M. Lavista Ferres, MSC,[b] Shirley You Ren, PhD,[b] Rachel Y. Moon, MD,[c] Richard D. Goldstein, MD,[d] Jan-Marino Ramirez, PhD,[e,f] Edwin A. Mitchell, FRACP[g]

abstract

OBJECTIVES: Maternal smoking during pregnancy is an established risk factor for sudden unexpected infant death (SUID). Here, we aim to investigate the effects of maternal prepregnancy smoking, reduction during pregnancy, and smoking during pregnancy on SUID rates.

METHODS: We analyzed the Centers for Disease Control and Prevention Birth Cohort Linked Birth/Infant Death Data Set (2007–2011: 20 685 463 births and 19 127 SUIDs). SUID was defined as deaths at <1 year of age with *International Classification of Diseases, 10th Revision* codes R95 (sudden infant death syndrome), R99 (ill-defined or unknown cause), or W75 (accidental suffocation or strangulation in bed).

RESULTS: SUID risk more than doubled (adjusted odds ratio [aOR] = 2.44; 95% confidence interval [CI] 2.31–2.57) with any maternal smoking during pregnancy and increased twofold between no smoking and smoking 1 cigarette daily throughout pregnancy. For 1 to 20 cigarettes per day, the probability of SUID increased linearly, with each additional cigarette smoked per day increasing the odds by 0.07 from 1 to 20 cigarettes; beyond 20 cigarettes, the relationship plateaued. Mothers who quit or reduced their smoking decreased their odds compared with those who continued smoking (reduced: aOR = 0.88, 95% CI 0.79–0.98; quit: aOR = 0.77, 95% CI 0.67–0.87). If we assume causality, 22% of SUIDs in the United States can be directly attributed to maternal smoking during pregnancy.

CONCLUSIONS: These data support the need for smoking cessation before pregnancy. If no women smoked in pregnancy, SUID rates in the United States could be reduced substantially.

[a]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, Washington; [b]Microsoft, Redmond, Washington; [c]Department of Pediatrics, School of Medicine, University of Virginia, Charlottesville, Virginia; [d]Boston Children's Hospital and Harvard Medical School, Harvard University, Boston, Massachusetts; [e]Department of Neurological Surgery and Pediatrics, School of Medicine, University of Washington, Seattle, Washington; and [f]Department of Paediatrics: Child and Youth Health, The University of Auckland, Auckland, New Zealand

WHAT'S KNOWN ON THIS SUBJECT: Approximately 3500 infants <1 year old die suddenly and unexpectedly each year in the United States. Previous research has revealed that maternal smoking during pregnancy is a known risk factor for sudden unexpected infant death (SUID).

3. Why I'm still so passionate about DP?

Challenge?

- Many hypothesis required PII Data
- Manual process
  - Write the scripts
  - Submit them to a trusted curator (human)
  - If approved, After 3 months, ($900), you could run this script
- Research doesn't work if every query takes 3 months to run

Opening Datasets

- Our job needs to be about opening datasets for research
- The world has a huge amount of data locked because of privacy
- Opening datasets can help save lives
- Differential privacy can be an amazing tool for opening these datasets while preserving the privacy of the individuals

Microsoft

Thank you.

AI for Good Research Lab