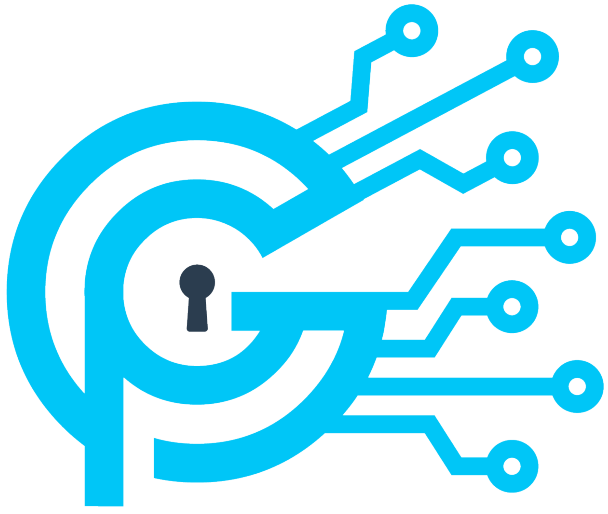
Demystifying Differential Privacy



GHOST PII

Capnion, Inc. March 2022

Who is the speaker?

30 second resume:

- an ancient metro Saint Louis townie
- grew up in University City
- University City High School 2003
- B.A. Washington University 2007 (econ and math)
- Ph.D. University of Michigan 2013 (math)
- a few years of data science
- founded data privacy tech company Capnion, Inc.

Slides URL:
`https://github.com/capnion/random/blob/master/acm_capnion_dpriv.pdf`

Agenda

overview

- goals
- stylized example and setting

differential privacy definitions

- heuristic
- mathematical
- comparison in our example

practical

- pros and cons
- outliers
- privacy budgets

Goals

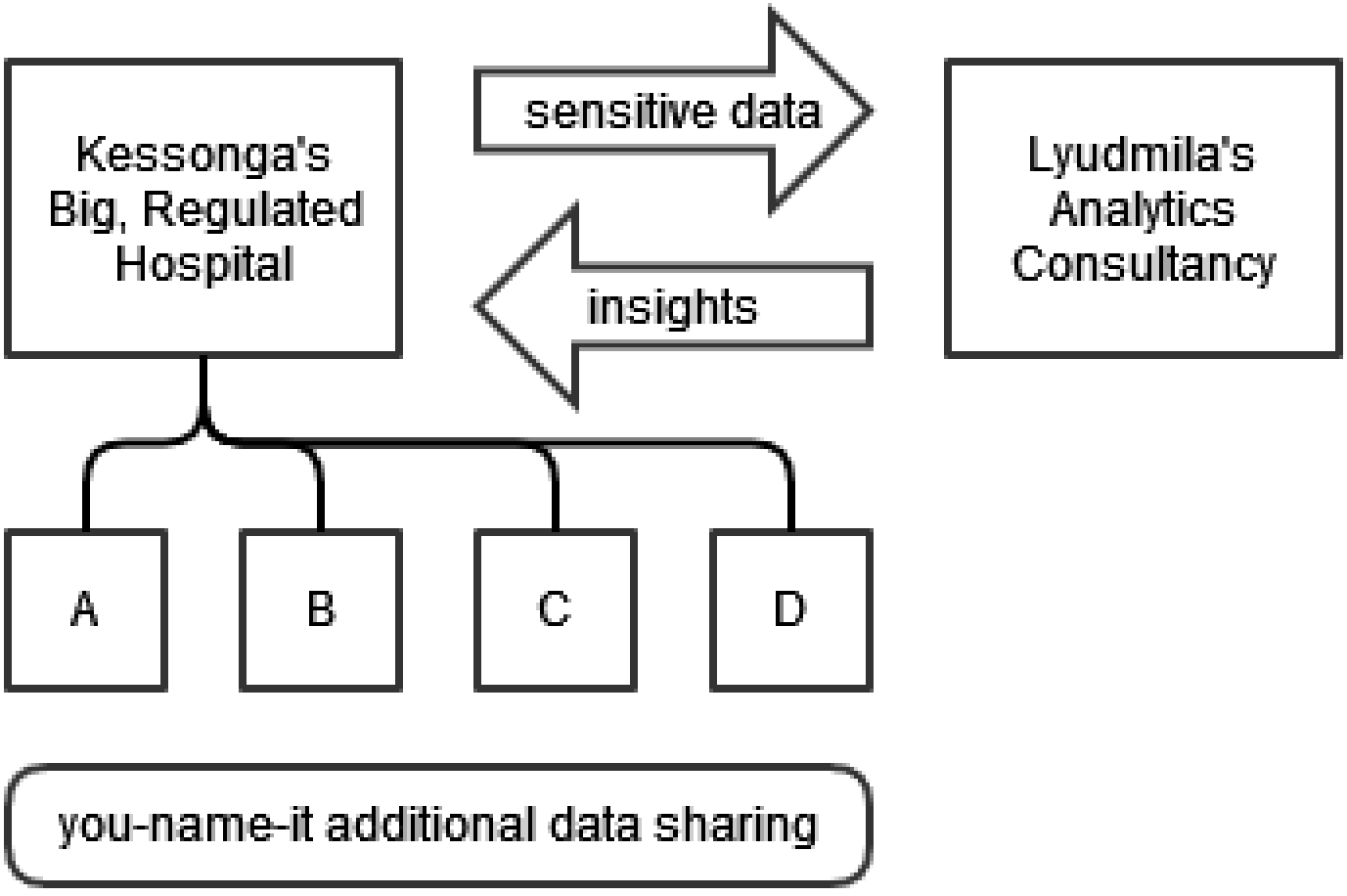
Harmonize two types of definition:

"Everyone in the dataset has plausible deniability"

vs.

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S],$$

Example Setting



anonymization is stronger than deidentification

Data With Added Noise

Patient ID	Longitude	Latitude	Height	Weight	Age	Has [DX ?]
1	38.69917	-90.3855	61.76584	176.0291	24.31459	Y
2	38.74275	-90.5305	62.17394	181.0755	12.78603	N
3	38.6937	-90.4123	62.13901	190.3331	40.78684	N
4	38.59847	-90.5228	56.1521	194.3795	30.9224	N
5	38.82691	-90.3496	64.99869	165.1797	26.77323	Y
6	38.65063	-90.6067	60.87851	211.6166	39.85179	N
7	38.75332	-90.4875	59.26405	240.6255	32.93192	Y
8	38.62144	-90.3807	60.72137	159.6717	43.57262	N
9	38.50983	-90.4078	55.37783	192.1354	55.4374	N
10	38.72783	-90.365	54.98859	213.6395	39.32518	N

Differential Privacy: A Heuristic Definition

"Everyone in the dataset has plausible deniability"

Add random noise to the data.

- maybe my record now looks more like someone else's
- maybe someone else's record looks like me
- adding noise is easy, fine-tuning is harder

My Cover Story

Patient ID	Longitude	Latitude	Height	Weight	Age	Has [DX ?]
1	38.69917	-90.3855	61.76584	176.0291	24.31459	Y
2	38.74275	-90.5305	62.17394	181.0755	12.78603	N
3	38.6937	-90.4123	62.13901	190.3331	40.78684	N
4	38.59847	-90.5228	56.1521	194.3795	30.9224	N
5	38.82691	-90.3496	64.99869	165.1797	26.77323	Y
6	38.65063	-90.6067	60.87851	211.6166	39.85179	N
7	38.75332	-90.4875	59.26405	240.6255	32.93192	Y
8	38.62144	-90.3807	60.72137	159.6717	43.57262	N
9	38.50983	-90.4078	55.37783	192.1354	55.4374	N
10	38.72783	-90.365	54.98859	213.6395	39.32518	N

People that resemble me are showing up randomly all over.

Mathematical Definition on Wikipedia

Definition breakdown via our example

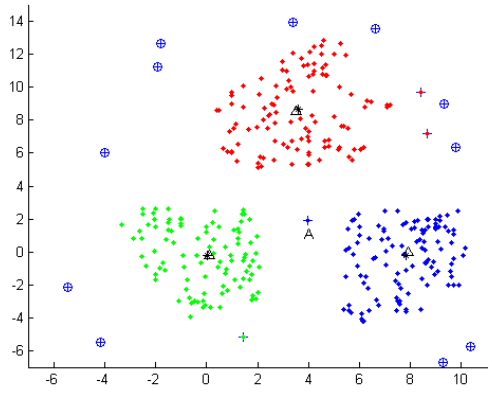
- Kessonga's hospital applies diff. priv. noise $\mathcal{A}$.
- Example S: Is there someone just like me in the data?
- If yes, it is just as likely (with ϵ confidence) a coincidence.

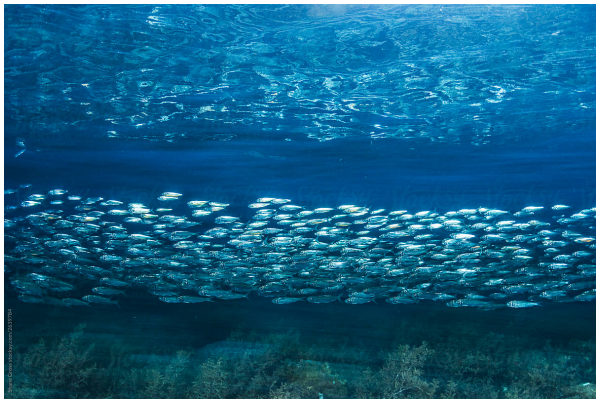
$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S],$$

If you found somebody that resembled me in your data, that is **just as likely** (up to a bit of error ϵ) because of the differential privacy noise and not because I was included in the original, pre - differential privacy dataset.

Pros and Cons

Good Data: continuous, lots of data, dense





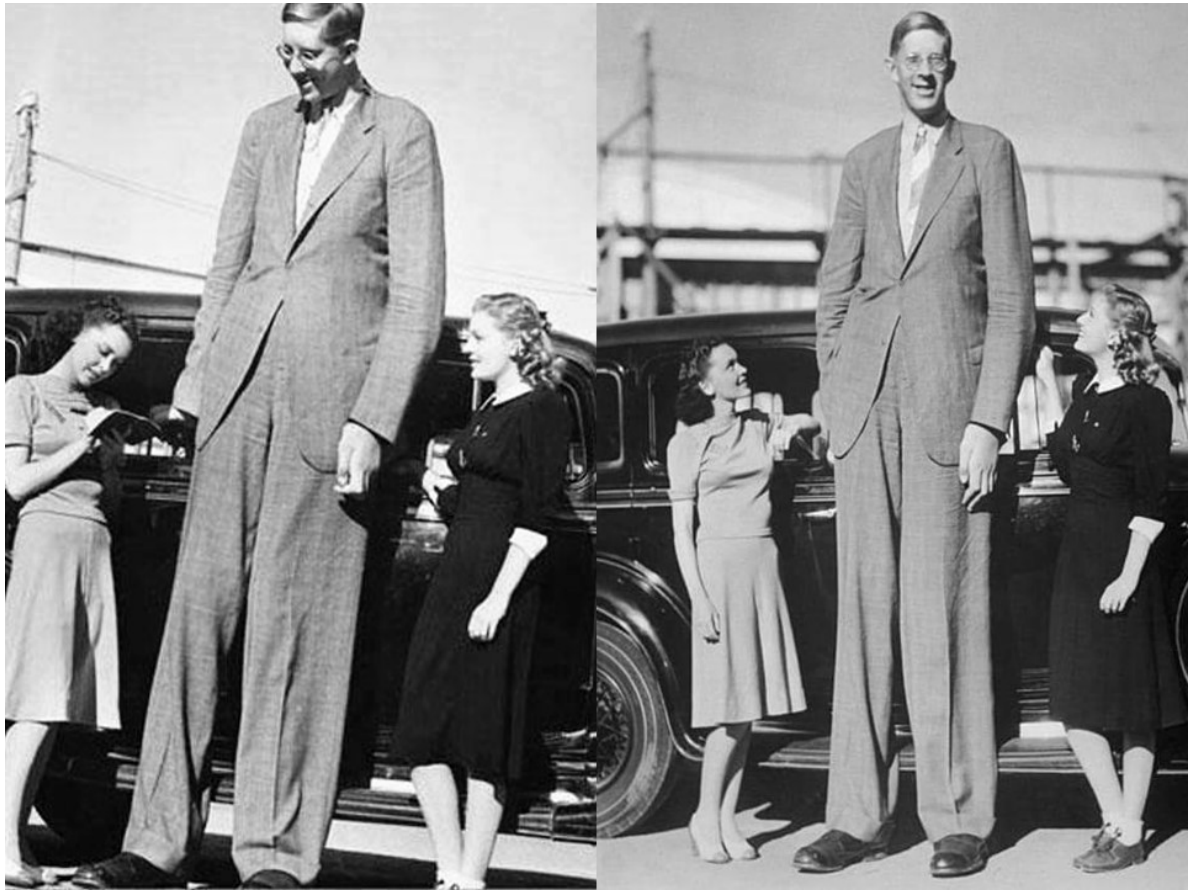
Bad Data: lots of outliers, outcomes driven by outliers

You need to be able to accept some level of random error.

- Bad: an audit on lots of money
- Better: hypothesis testing on a treatment effect

Pro: Data with added noise can be handled as normal.

Outlier Intuition and Example



standard noise might not protect the world's tallest (107 in.)

Tradeoffs

It's easy to guess at which record comes from the world's tallest man, and his exclusion changes mean height and weight significantly.

						Means without Patient 1		
Longitude	Latitude	Height	Weight	Age	Has [DX ?]	Height	Weight	Age
38.69917	-90.3855	107.1684	350.0291	24.31459	Y	58.8462	193.4733	31.99335
38.74275	-90.5305	62.17394	181.0755	12.78603	N	Means without Patient 20		
38.6937	-90.4123	62.13901	190.3331	40.78684	N	Height	Weight	Age
38.59847	-90.5228	56.1521	194.3795	30.9224	N	61.51476	199.9965	32.32686
38.82691	-90.3496	64.99869	165.1797	26.77323	Y			
38.65063	-90.6067	60.87851	211.6166	39.85179	N	All		
38.75332	-90.4875	59.26405	240.6255	32.93192	Y	Height	Weight	Age
38.62144	-90.3807	60.72137	159.6717	43.57262	N	61.26231	201.3011	31.60941

Two basic questions that get advanced quickly...

- Can I just add more noise or data? How much?
- Can I exclude outliers? How and how often?

Tradeoffs

Basic: It's the ϵ in the definition. Smaller epsilon = more privacy.

It is important that repeating a differential private query can be less private with each iteration if new noise is used.

`BoundedMean.quick_result()` takes a List of integer/ float as an input and returns the mean of the list values.

In [4]:
```
# calculates mean applying differential privacy
def private_mean(privacy_budget: float) -> float:
    x = BoundedMean(privacy_budget, 0, 1, 100)
    return x.quick_result(list(df["carrots_eaten"]))
```

As you can see, the value of the private mean varies compared to the mean calculated using non-private statistical methods.

This difference in value corresponds to the privacy that is actually preserved for individual records in it.

In [5]:
```
print("Mean: ", mean_carrots())
print("Private Mean: ", private_mean(0.8))
```

```
Mean:  53.01648351648352
Private Mean:  71.27272727272728
```

(from PyDP carrots tutorial)

(add noise then compute mean) provides privacy and so does (compute mean and then add noise)

GHOST PII

Questions and Conversation

Any questions?

acmueller@capnion.com

https://twitter.com/capnion

https://www.linkedin.com/in/
alexander-c-mueller-phd-0272a6108/

GHOST PII