Privacy protection, redefined.

Gerome Miklau

CEO / Co-Founder, Tumult Labs
Professor, UMass Amherst

Presented to:
NAS Location Data / Governance Frameworks

Date:
June 8, 2022

Data custodians

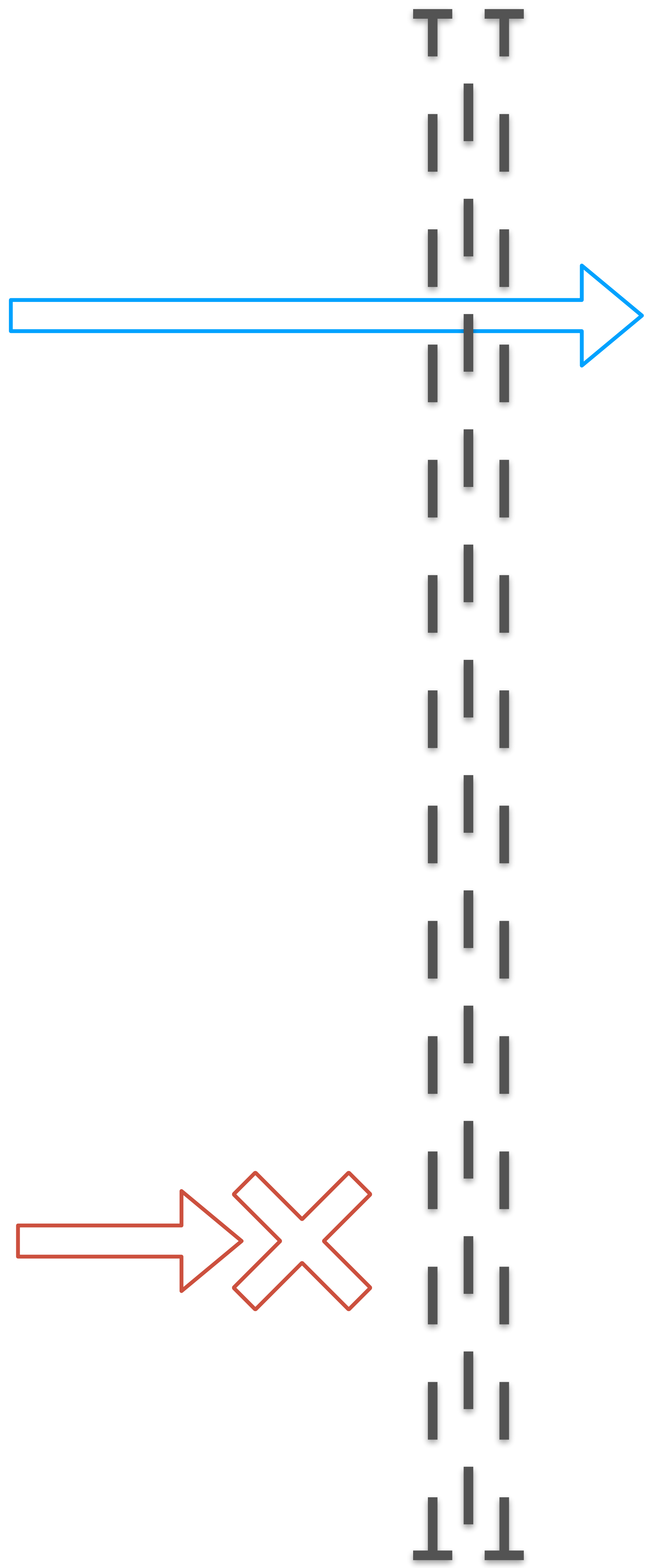
cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records

Data custodians need a privacy "filter"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records



Share insights about groups

Desired Insight

The median weekday drop-off frequency on 59th Street during morning rush hour is 145

Protect individuals

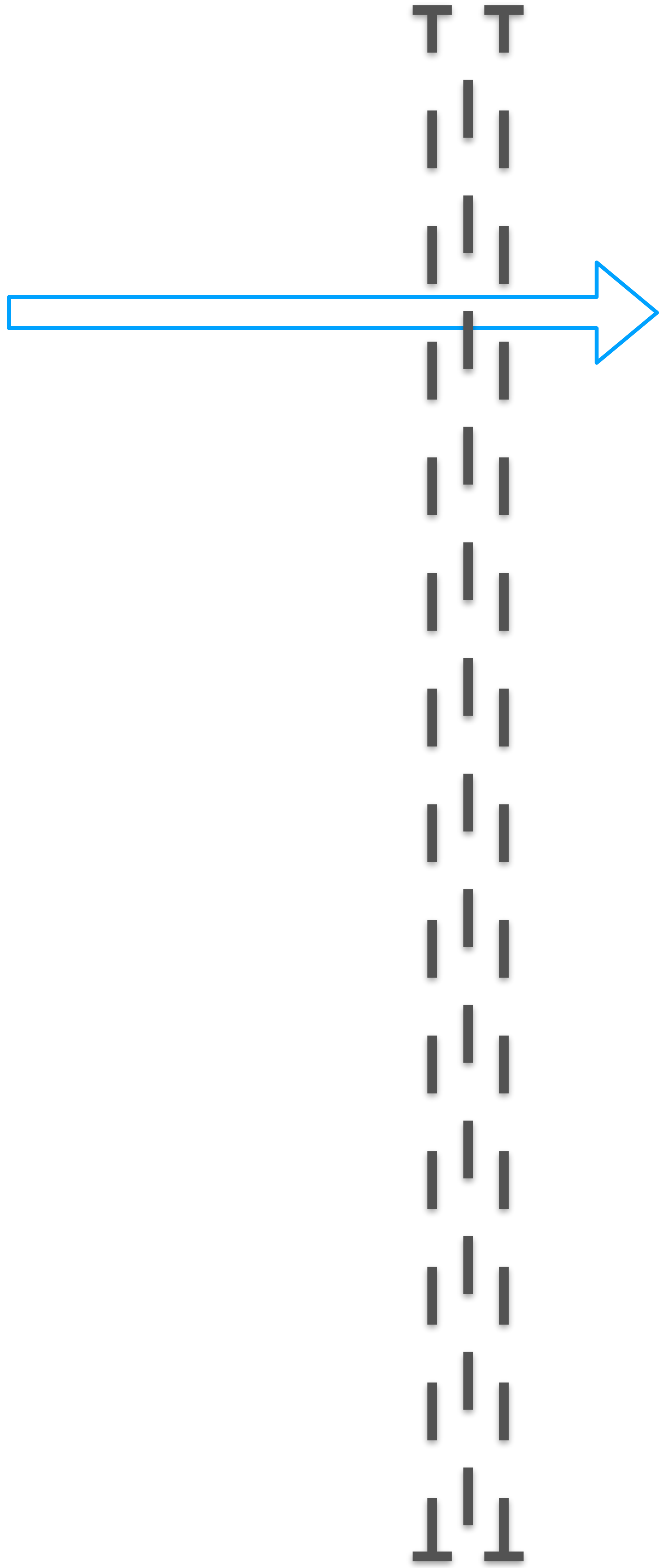
Privacy Violation

Customer x456 traveled from LGA to 59th St and 7th Ave, arriving June 1 at 8:30am

Data custodians need a privacy "filter"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records



- De-identified data

➢ ***Re-identification attack***

A re-identification attacks **uses de-identified data**, in combination with external information sources, to identify individuals and infer their sensitive properties.

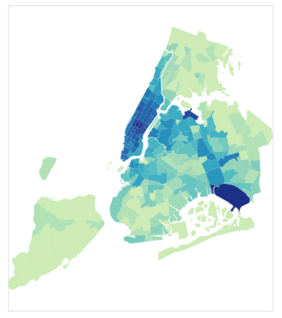
Privacy violation

Data custodians need a privacy "filter"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records

	Male	Female
WHO=0	345	1094
WHO=1	214	2439
WHO=2	172	1589



- Reports, analytics, aggregate stats, etc

➢ ***Reconstruction attack***

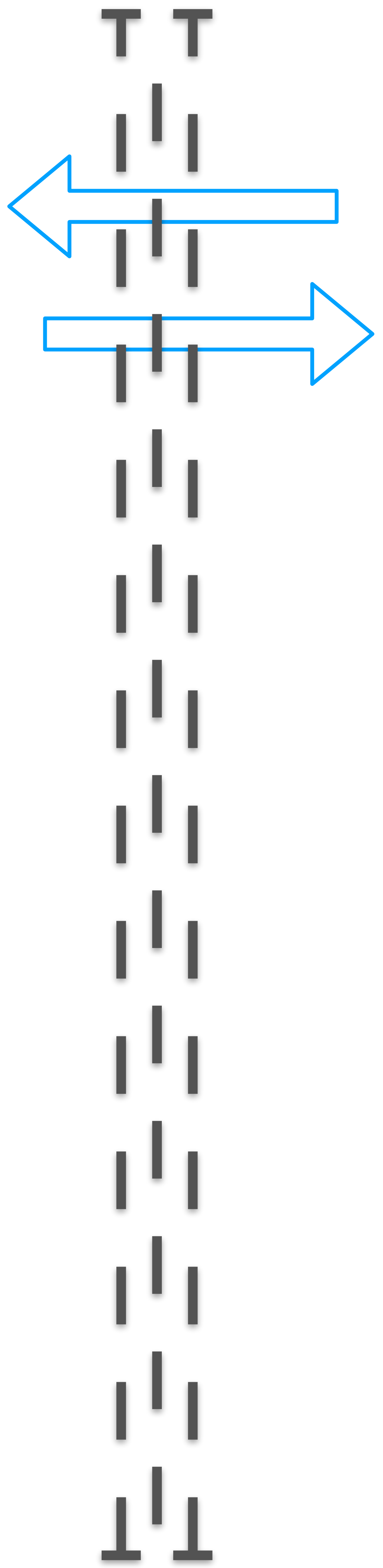
A reconstruction attack **uses a set of aggregate query answers** to reconstruct the set of hidden input records.

Privacy violation

Data custodians need a privacy "filter"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records



- Sharing through a query interface

How many patients in cohort defined by…

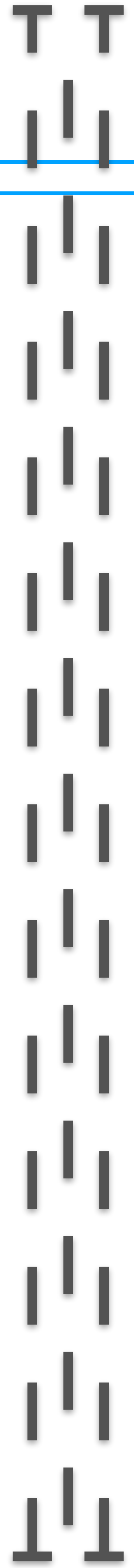
1214

➢ ***Reconstruction attack***

A reconstruction attack **uses a set of aggregate query answers** to reconstruct the set of hidden input records.

Privacy violation

Data custodians need a privacy “filter”



- Sharing synthetic tables

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records

➢ ***Reconstruction attack***

A reconstruction attack **uses a set of aggregate query answers** to reconstruct the set of hidden input records.

➢ ***Membership inference attack***

A membership inference attack occurs when repeated **access to predictions from a machine learning model** reveals sensitive properties of individuals present in the training data.

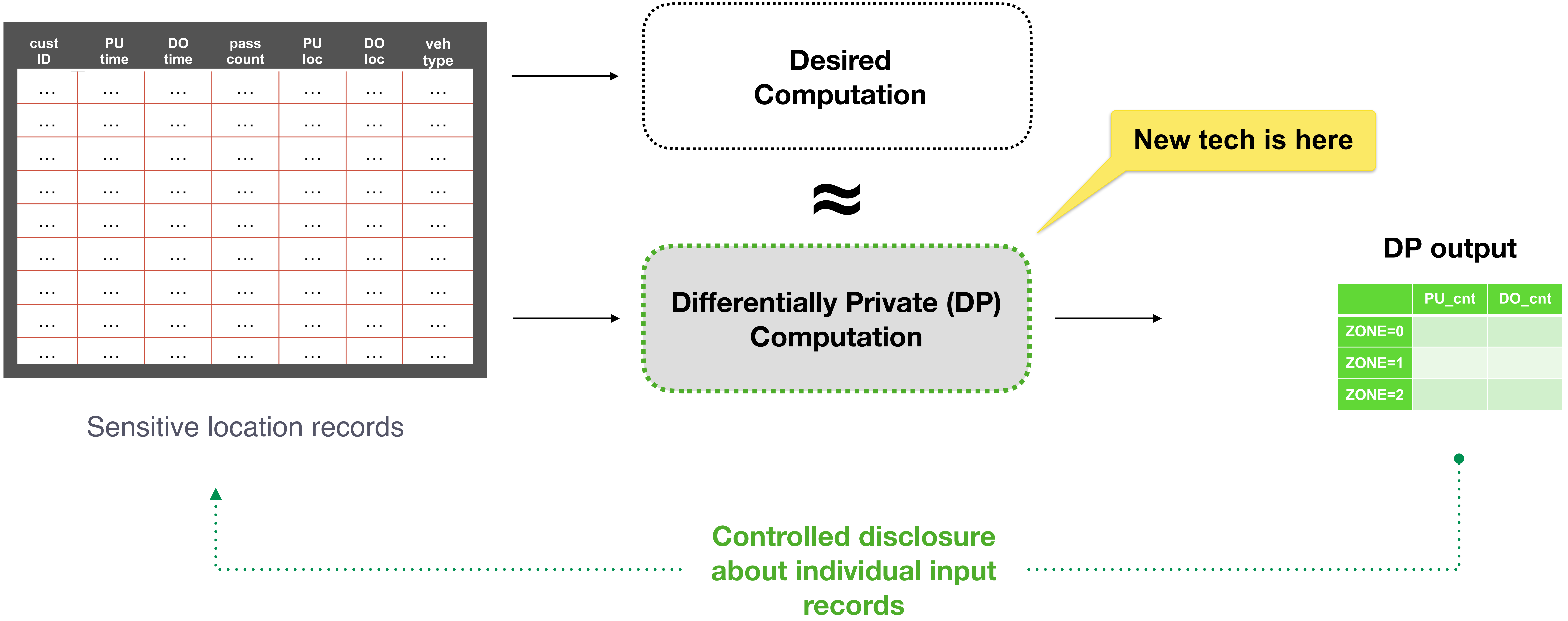
Privacy violation

Data custodians need a reliable privacy “filter”

Differential privacy
a standard for computations on data
that limits the personal information that could be revealed by the output.

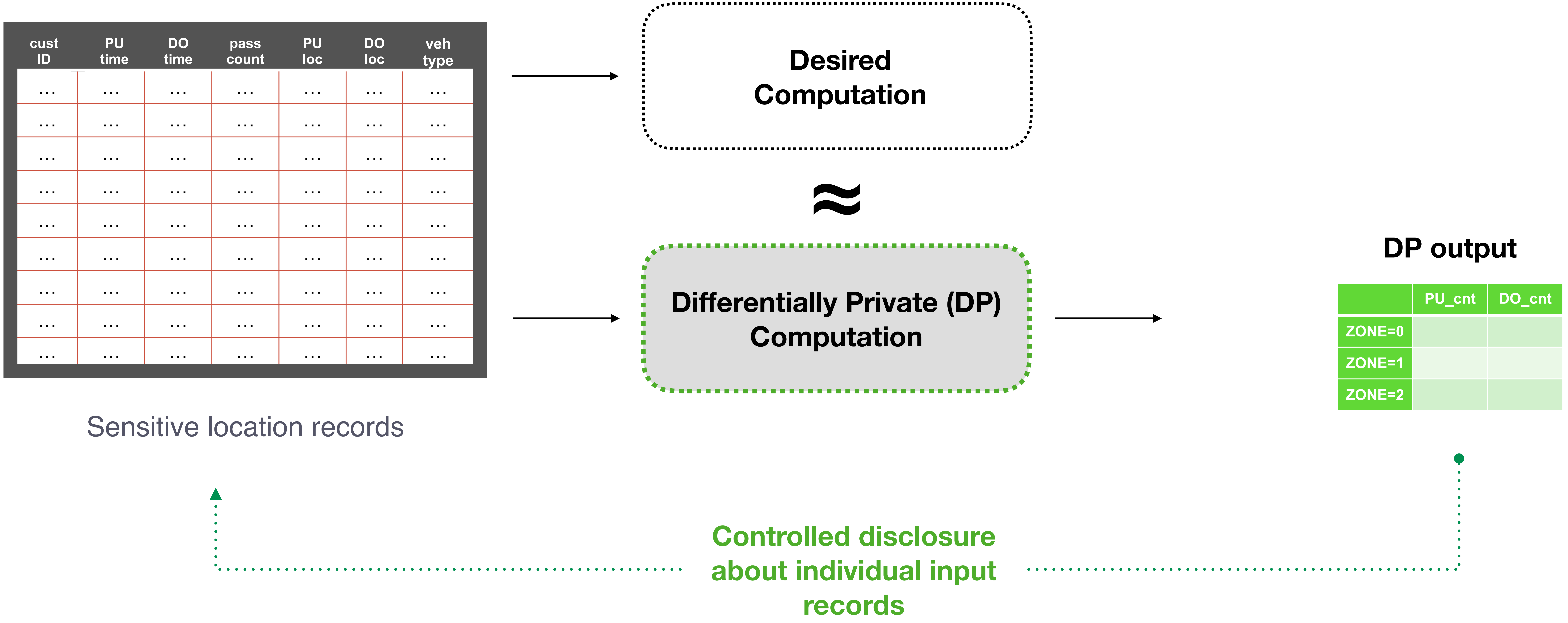
Differential privacy

a standard for computations on data
that limits the personal information that could be revealed by the output.



The differential privacy guarantee

- Every individual protected.
- Every attribute protected.
- The guarantee holds, regardless of compute power or knowledge of potential attacker.
- Resists current and future attacks
- Ahead of regulation



Differential privacy
a standard for computations on data
that limits the personal information that could be revealed by the output.

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records

First key difference: randomness

Differentially Private (DP) Computation

Some "noise" in output

DP analytics output

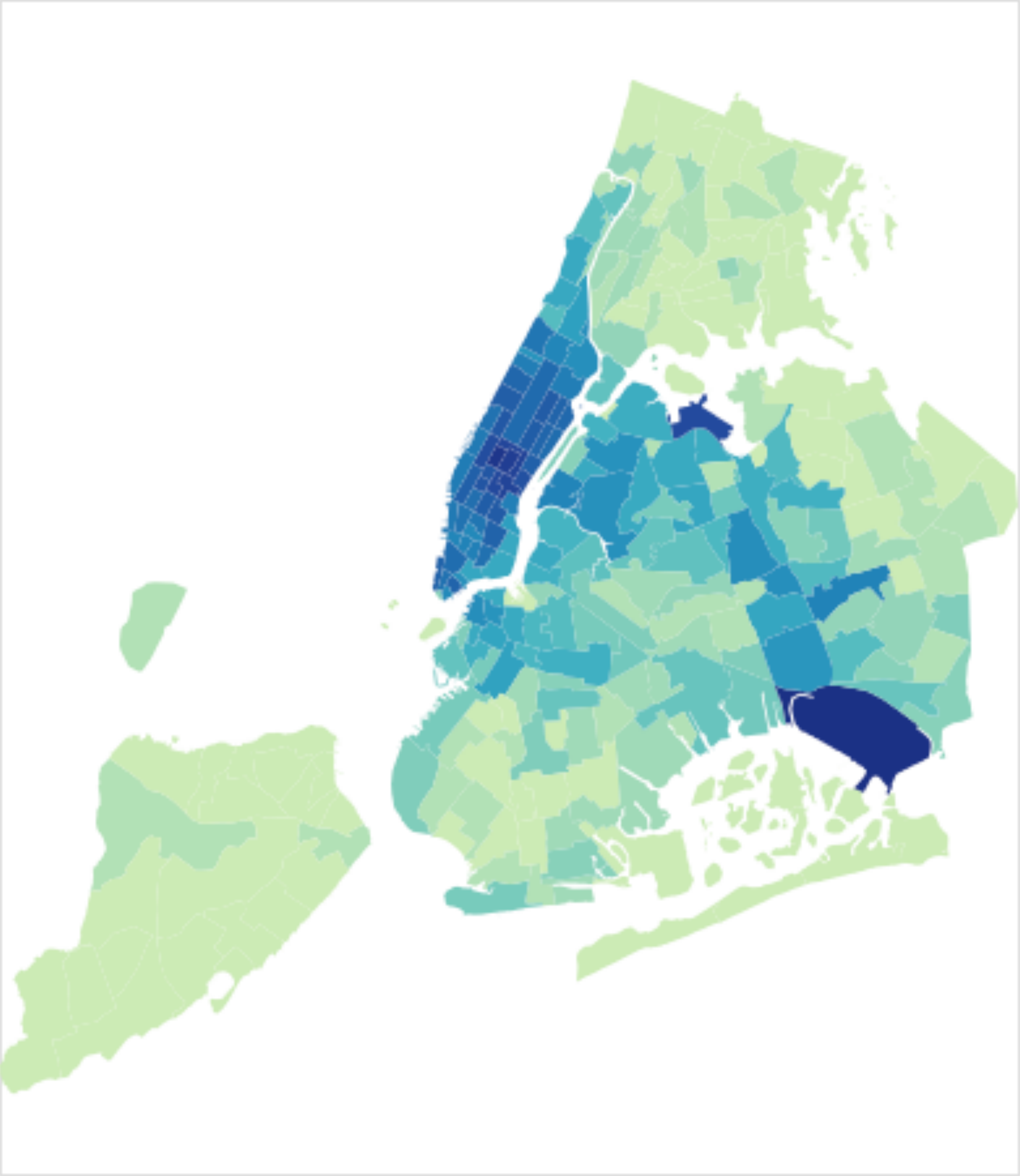
	PU_cnt	DO_cnt
ZONE=0		
ZONE=1		
ZONE=2		

Controlled disclosure about individual input records

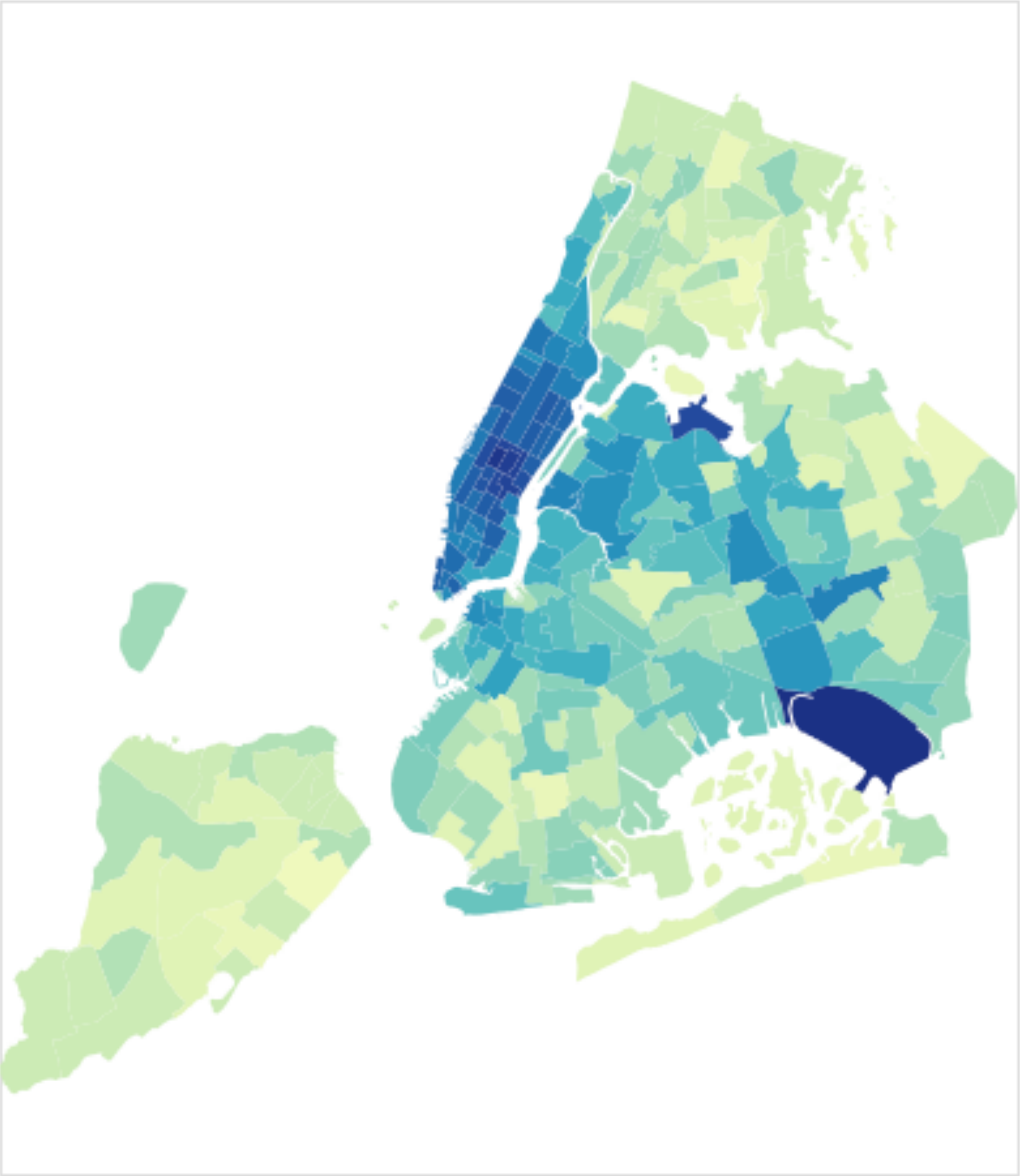
Pickup frequency by taxi zone

New York City taxi/passenger data; 3.17 million records

Original data



Differentially private, epsilon = 1.0



Differential privacy

a standard for computations on data
that limits the personal information that could be revealed by the output.

Second key difference: privacy loss parameter, epsilon

Bound on "privacy loss"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records

Differentially Private (DP) Computation

ϵ

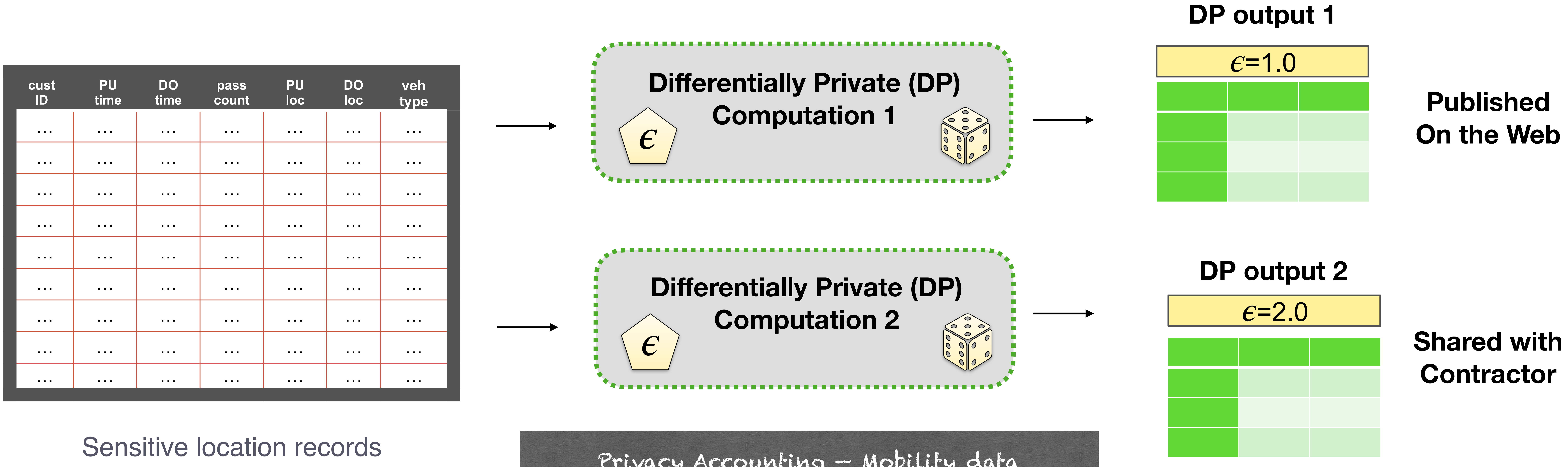
DP output

ϵ=1.0

	PU_cnt	DO_cnt
ZONE=0		
ZONE=1		
ZONE=2		

Controlled disclosure about individual input records

Managing **cumulative** privacy loss



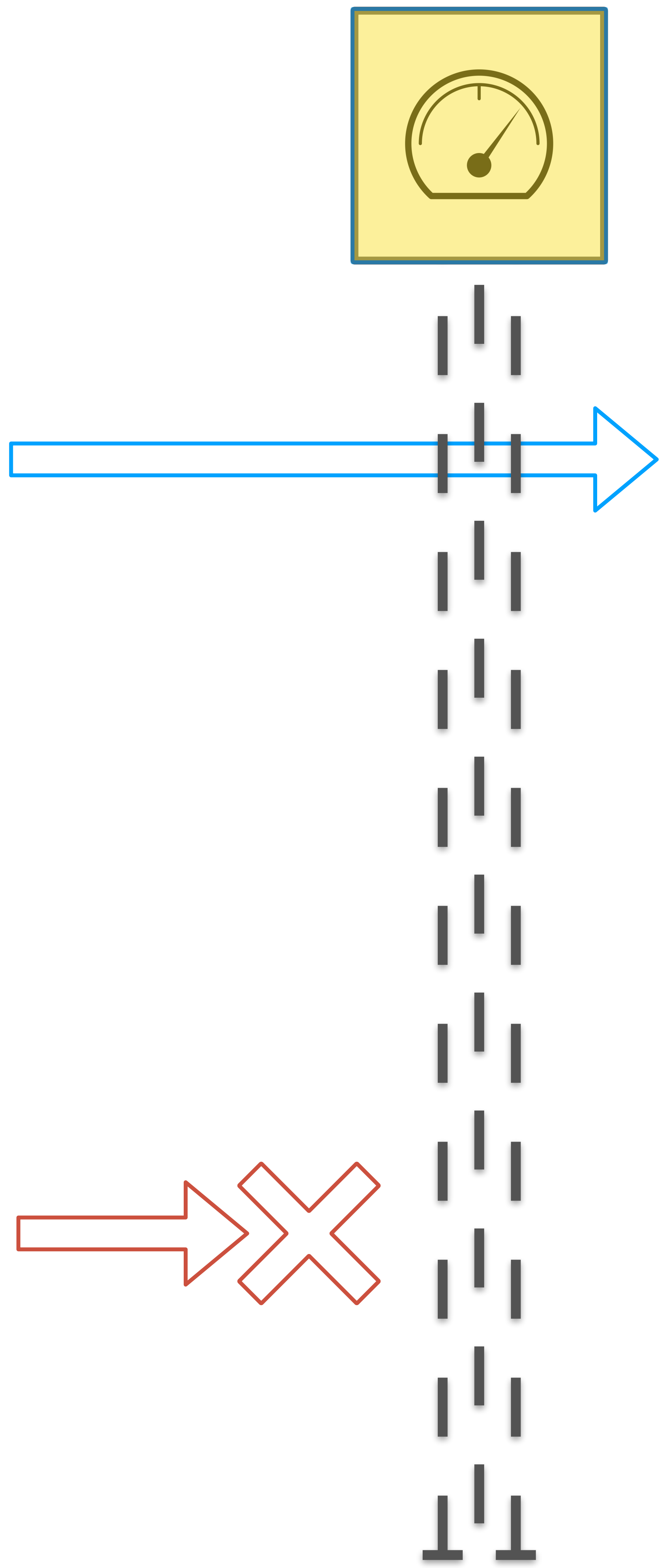
Privacy Accounting – Mobility data

Release	Recipient	Privacy Loss
Analytics 1	Web Pub	ϵ=1.0
Analytics 2	Contractor	ϵ=2.0
TOTAL PRIVACY LOSS		ϵ=3.0

Differential privacy gives
Data custodians a reliable, metered privacy "filter"

cust ID	PU time	DO time	pass count	PU loc	DO loc	veh type
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Sensitive location records



Share insights about groups

Protect individuals

TUMULT LABS

Try our platform

www.tmlt.io/connect

Free trial available; open source soon!

Thank you

Gerome Miklau

Contact: miklau@tmlt.io