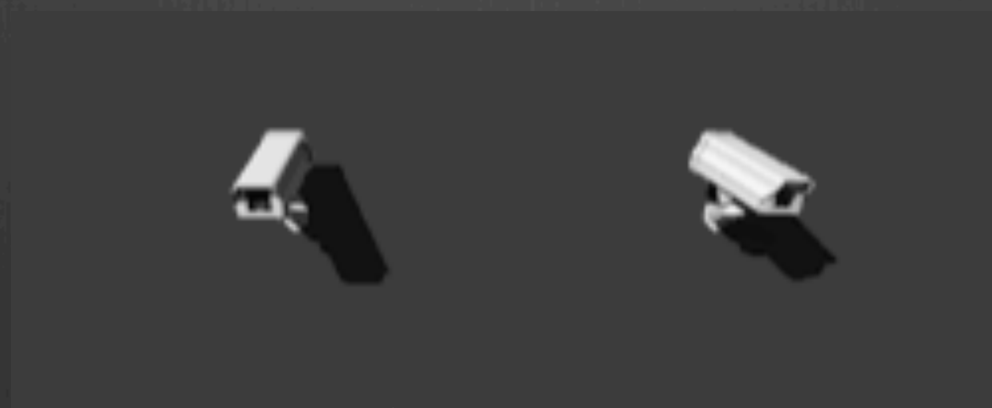
Getting Personal with Differential Privacy

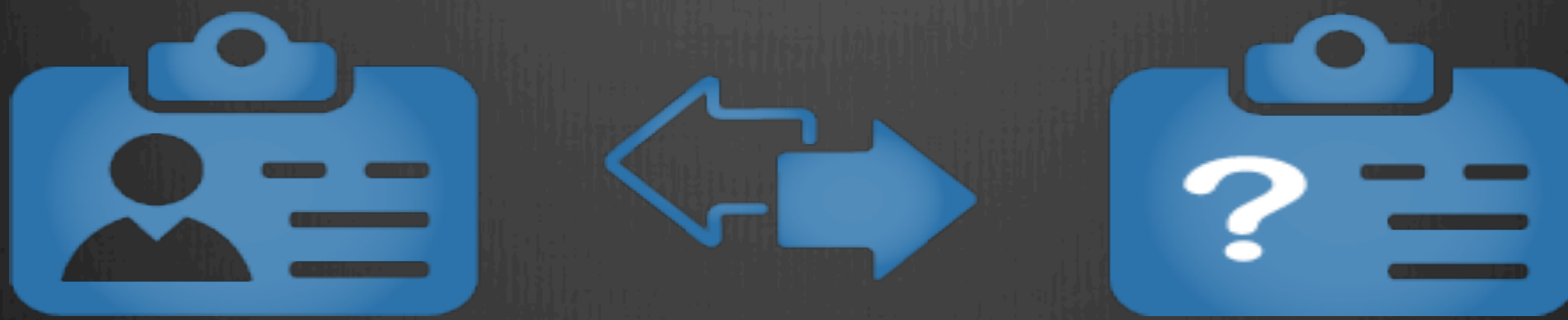


David Sands

Chalmers University of Technology
Sweden

Just how big are Dave's feet?

How Can We Ensure Privacy?



Anonymisation

Massachusetts Group Insurance Commission released "anonymized" health records on state employees

Photo:media.masslive.com

Anonymisation Fail

Netflix released viewer data for half a million subscribers a $1M competition to build the best movie-recommender system

Candidate definition of Absolute Privacy [Dalenius '77]:

Whatever you learn about an individual from a database could already have been learned without access to the database

The 1991 Romanian Mititei Survey

3.6 cm

61% Pork
12% Beef

8.94 cm

Does the 1991 “Mici” Survey Protect Dave’s Privacy?

Dave’s feet are 3x longer than the average 1991 mici

Anonymisation Cannot Guarantee Privacy

On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy, [Dwork & Naor 2010]

Differential Privacy
[Dwork & McSherry '06]

A **quantified** definition of privacy for a **noisy** statistical query:

> quantify the *difference* in what might be learned about any individual from a database with or without said individual

Privacy Preserving Database Queries

- What is privacy for database queries?

    1. Introducing Differential Privacy

- How to build tools which make it easy to program data analyses while respecting privacy

    2. Building-blocks for DP mechanisms

- Outline of our approach:

    3. Personalised Differential Privacy

Differential Privacy

- A measure of the extent to which anyone can blend into the crowd

- A measure of the plausible deniability of the claim: “I’m not even in that database”

ε -Differentially Private query:

For any dataset and any individual ,

the chance of getting answer A on

vs

the chance of getting answer A on +

differ by at most a factor of exp(ε)

ε -Differentially Private query:

For any dataset  and any individual ,

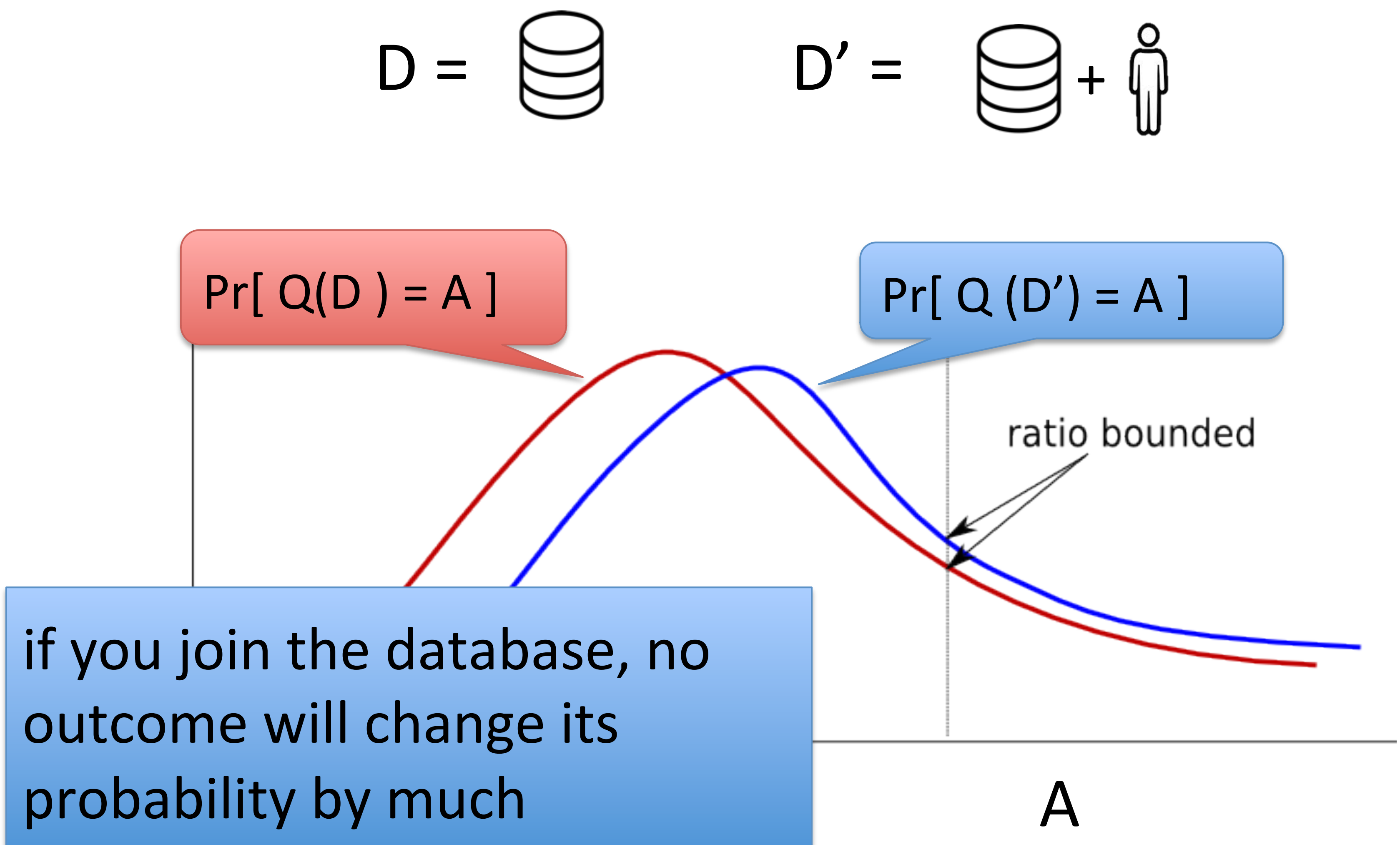
the chance of getting answer A on 

vs

the chance of getting answer A on  + 

differ by at most a factor of $1 \pm \varepsilon$

Differential Privacy



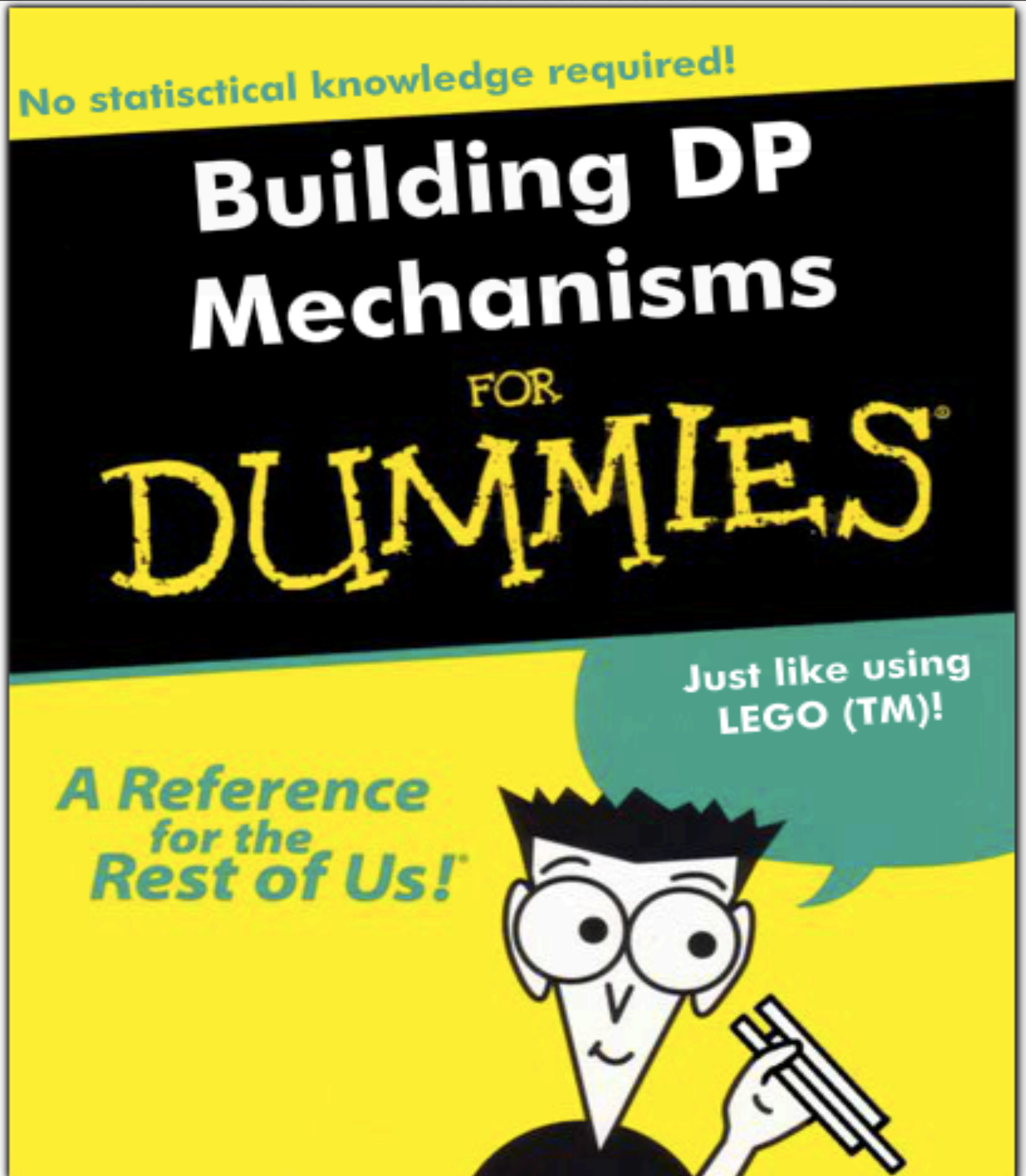
Designing Differentially Private Mechanisms

This talk:

Intro to DP

+

dynamic enforcement method for DP by information flow tracking


No statisctical knowledge required!
Building DP Mechanisms
FOR
DUMMIES
Just like using LEGO (TM)!
A Reference for the Rest of Us!

Building Blocks for Differential Privacy

Compositionality principles make it easier to build differentially private mechanisms from components

Sequential Composition

An ε_1-DP query, followed by an ε_2-DP is $(\varepsilon_1 + \varepsilon_2)$-DP [McSherry]

Holds even if Q_2 is chosen using the result of Q_1

Sensitivity (stability)

D =    D' =  + 

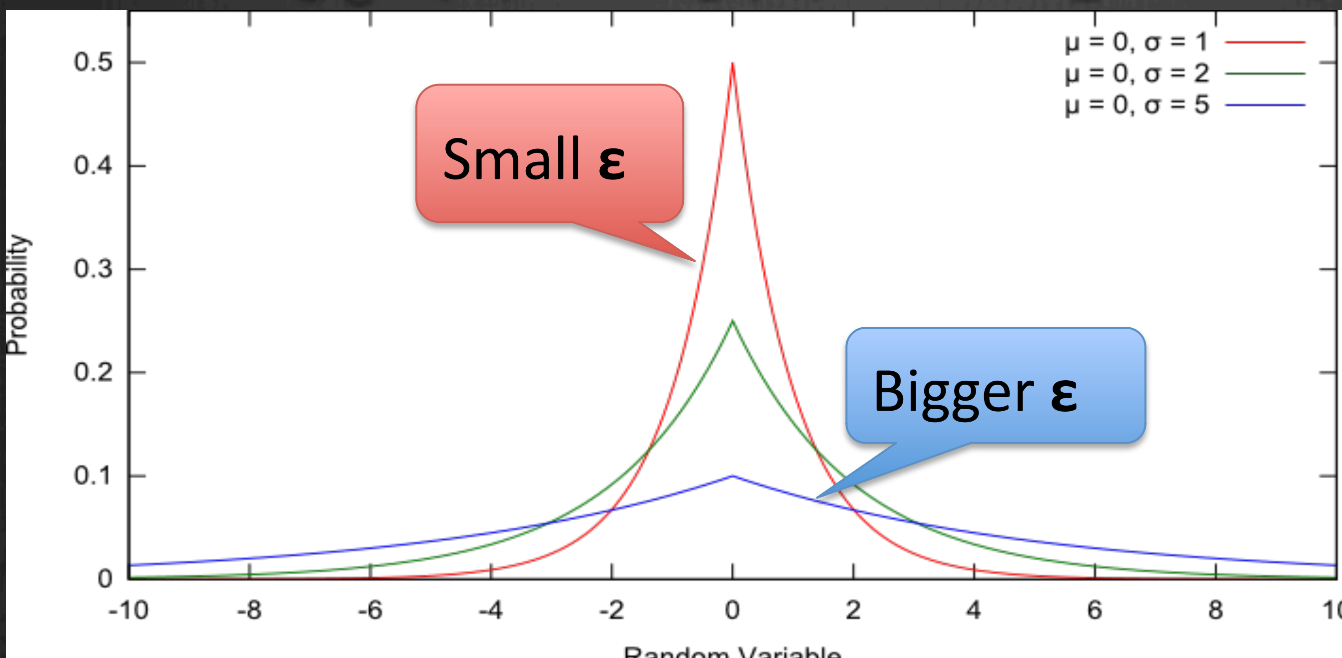
A function **F** has sensitivity **S** if **F(D)** and **F(D')** are different by at most (size) **S**

- count
- select males
- sum

Private Query = Query + Noise

If **Q** has sensitivity **s** then we can compute an ε –differentially private version of **Q**:

$$\mathbf{Q_{\varepsilon}(x) = Q(x) + Laplace(s/\ \varepsilon\ )}$$

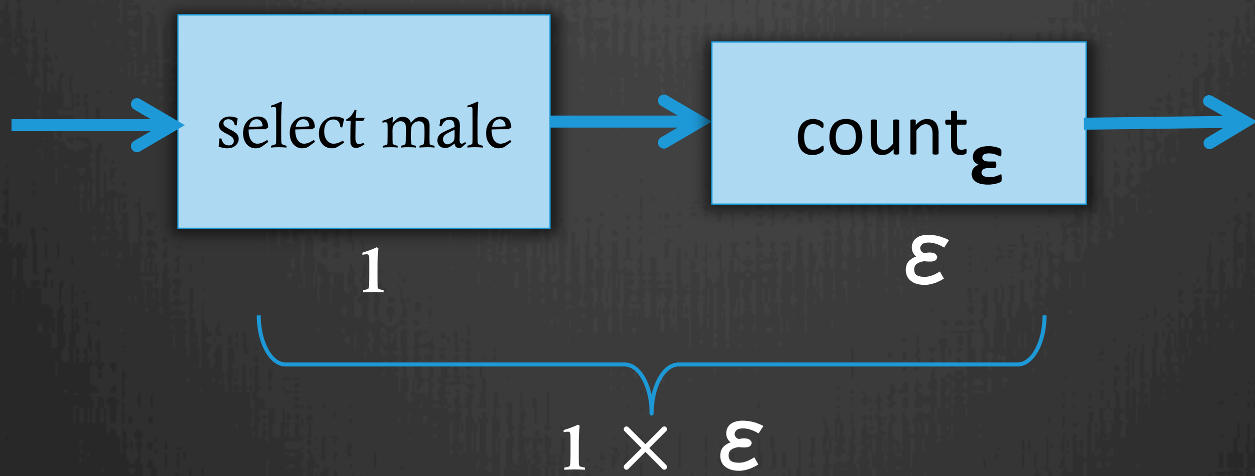


Laplace distribution

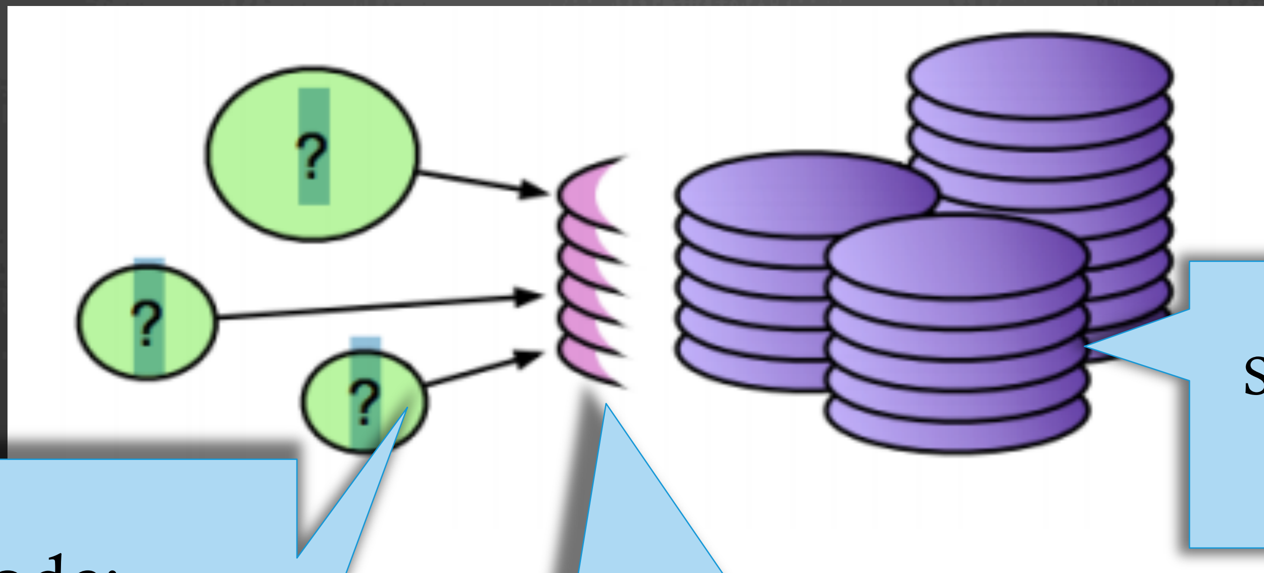
Sensitivity Composition

T has sensitivity **s** and **Q** is ε-DP then

$$\mathbf{Q} \circ \mathbf{T} \text{ is } (\mathbf{s} \times \varepsilon)\text{-DP}$$



PINQ [McSherry]

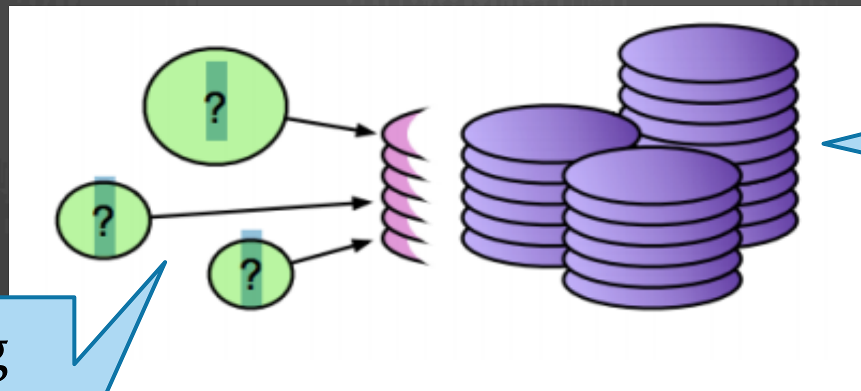


Standard LINQ data store

C# code: Transformations and Queries in a LINQ-like language

API mediating all access to database

PINQ [McSherry]



Standard LINQ data store

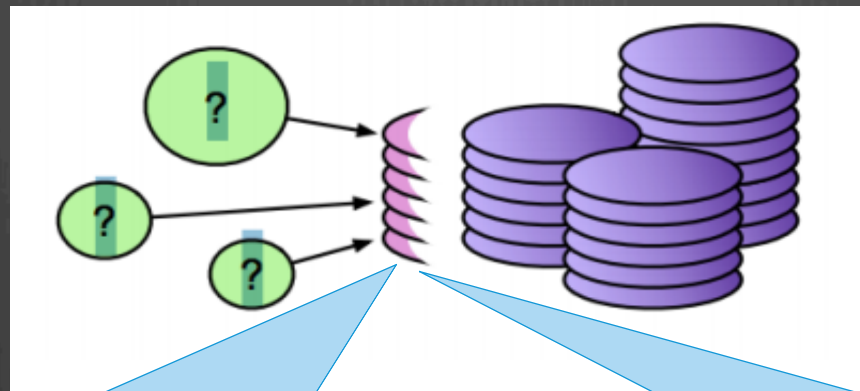
C# code using Transformations and Queries in a LINQ-like language

```
var data = new PINQueryable<SearchRecord>(... ...);

var users = from record in data
            where record.Query == argv[0]
            groupby record.IPAddress;

Console.WriteLine(argv[0] + ":" + users.Count(0.1));
```

PINQ [McSherry]



Data

- A Global Privacy Budget

- The sensitivity of each intermediate database

Bookkeeping

- Deduct $\varepsilon \times \mathbf{s}$ from budget if ε query is applied to a table with sensitivity $\mathbf{s}$
- Deny query whenever the budget is insufficient

Problem 1: Wasteful Global Budget

Detailed, multi-dimensional survey of people with blood type *AB-negative*

Budget Exhausted

Marketing study of all adults

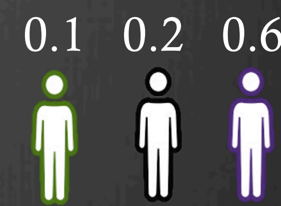
Problem 2: Continuous Data

Detailed, multi-dimensional survey of people with blood type *AB-negative*

New data input to the database

Personalised Differential Privacy (PDP)

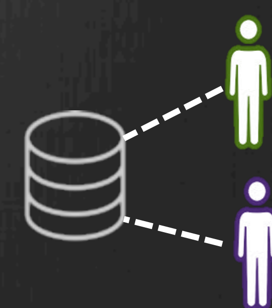
1. Generalise DP: each individual has their own personal “epsilon”



2. Show that PDP has its own composition principles

3. Implement PDP by tracking exact *provenance* of every record

[Ebadi, DS, Schneider, POPL 2015]



Personal (Big-Epsilon) Differential Privacy

Let E be a function from individual records to $\mathbb{R}^{\geq 0}$

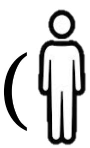
Query Q provides **E**-Differential Privacy

if for all  and all 

the chance of getting answer A on 

vs

the chance of getting answer A on  + 

differ by at most a factor of $1 \pm$ **E**()

PDP generalises DP

$$\begin{tikzcd} \text{Personalised DP} \arrow[r, bend left, "E \mapsto \max(\mathrm{range}(E))"] & \text{DP} \arrow[l, bend left, "\varepsilon \mapsto \lambda x.\, \varepsilon"] \end{tikzcd}$$

If Q is E-DP then Q is ε-DP for $\varepsilon = \max(\mathrm{range}(E))$

PDP composition principles

Sequential composition

An E_1-DP query, followed by an E_2-DP query is E-DP

where $E(y) = E_1(y) + E_2(y)$

PDP composition principles

Sensitivity composition

If Q is E-DP then $Q \circ F$ is E'-DP

where $E'(z) = \text{sensitivity}(F) \times E(z)$

PDP composition principles

"Computing the (noisy) average income of adult smokers is 0-differentially private for Jimmy, aged 10."

Selection: select_P removes elements not in P

Selection composition principle:

Q is E-DP then $Q \circ \text{select}_P$ is E'-DP

where $E'(x) = \text{if } x \in P \text{ then } E(x) \text{ else } 0$

Union-preserving functions

$$F(A \cup B) = F(A) \cup F(B)$$

E.g. select, project, rename, map…

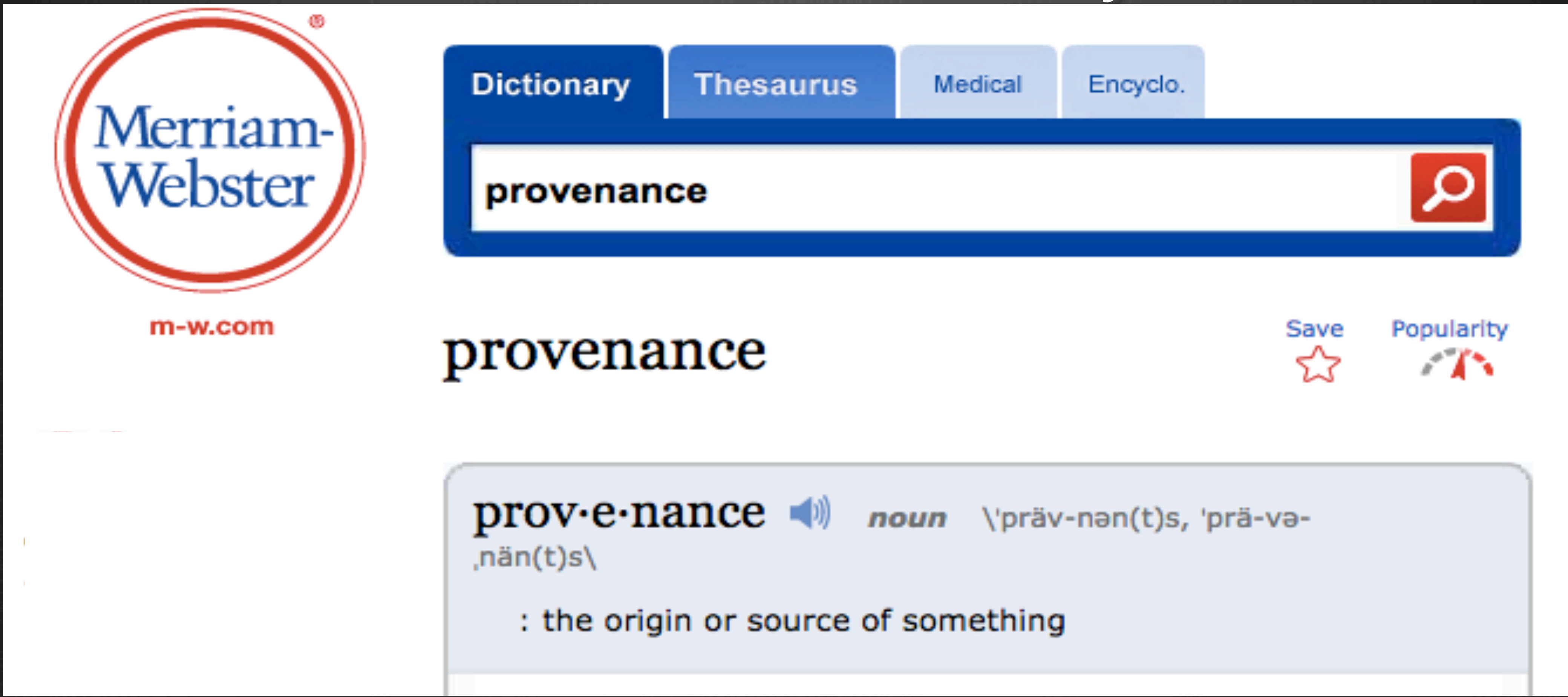
Union-preserving functions

If Q is ε -DP then $Q \circ F$ is E-DP,

where $E(x) = \varepsilon \times \text{size}(F\{x\})$

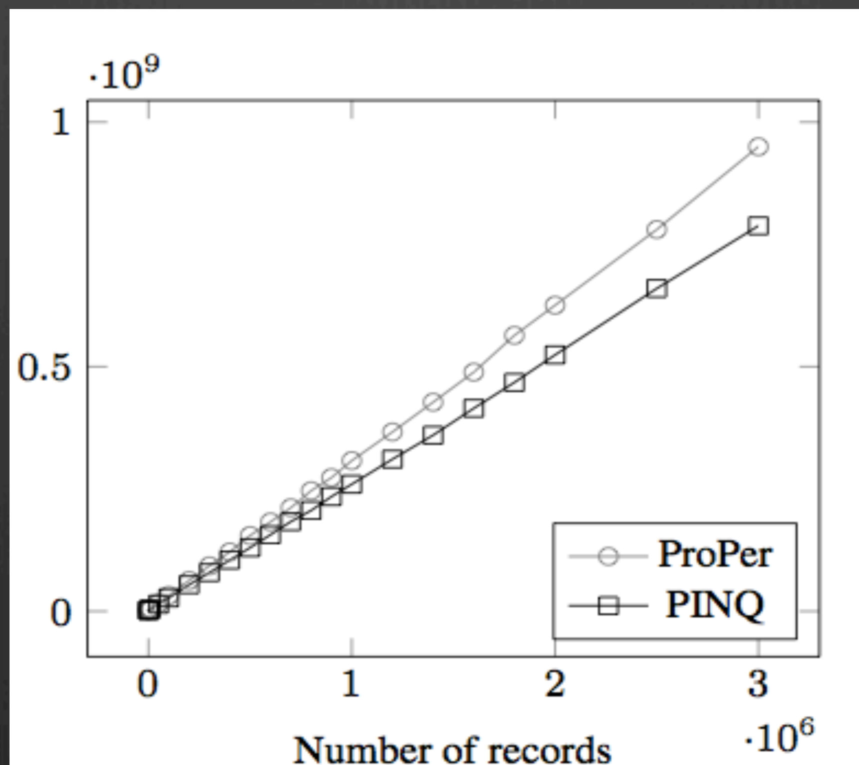
F magnifies the privacy cost of Q for Bob by $|F\{Bob\}|$

Provenance for Personalised Differential Privacy



Provenance for **Per**sonalised Differential Privacy

Our implementation, ProPer, is based on (and subsumes) PINQ with small overhead



ProPer in Action

Budget	ID	Age
1	Mary	24
1	Bob	29
1	Harry	17

Initial budgets associated with the original data

ProPer in Action

	ID	Age
1	Mary	24
1	Bob	29
1	Harry	17

SELECT age
WHERE age $\geq$ 18

Transformation

ProPer in Action

	ID	Age
1	Mary	24
1	Bob	29
1	Harry	17

SELECT age
WHERE age $\geq$ 18

Age
24
29

Table plus the provenance of each record

ProPer in Action

	ID	Age
1	Mary	24
1	Bob	29
1	Harry	17

SELECT age
WHERE age $\geq$ 18

Age
24
29

27.937

Average($\varepsilon = 0.1$)

"Primitive"
DP-query

ProPer in Action

	ID	Age
0.9	Mary	24
0.9	Bob	29
1	Harry	17

Update budgets

SELECT age
WHERE age $\geq$ 18

Age
24
29

27.937

Average(ε $=$ **0.1**)

The Catch

PINQ

deny the query (throw exception)

OK because the budget is not private

ProPer

Not OK! Budget *is* private

Solution

1. Silently drop the records from the query which would otherwise get negative budget

- Not obvious that this is privacy preserving
  - It isn't , in general

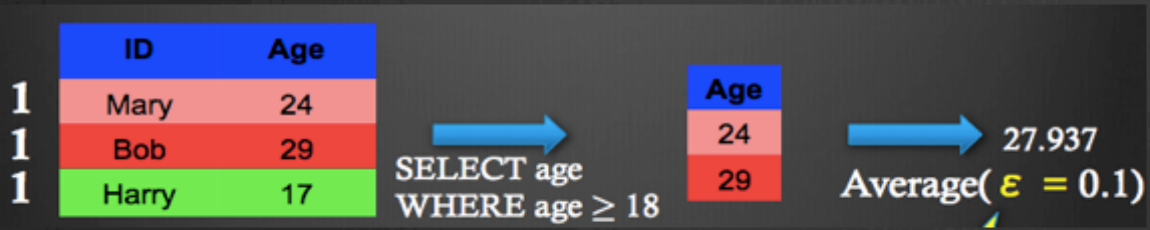
2. Restrict to *unary union-preserving transformations (e.g. map & filter)*
  - *Small change to dataset implies only small change to set of over-budget records*

Conclusions

- Introduced Personalised Differential Privacy
  - more fine-grained budgeting
  - capable of handling interactive queries over data arriving over time
- ProPer provenance-based tracking
  - Implementation subsuming PINQ with small overhead
  - Formal model & proof of correctness (PDP)

End


1/2
1/1


ID Age
1 Mary 24
1 Bob 29
1 Harry 17
SELECT age
WHERE age ≥ 18
Age
24
29
27.937
Average(ε = 0.1)

Further Work

- Permissiveness: Prove more permissive than PINQ
  - requires formal model of PINQ
- Utility: method degenerates to noise; analyst may be unaware
  - Track utility based on analysts prior knowledge

Related Work

- See paper
- Don't see the paper:
  [Xiao & Tao, SIGMOD'06] Personalised version of k-anonymity
- [Alaggan, Gambs, Kermarrec TPDP 2015] Heterogeneous differential privacy
- [Jorgensen, Yu, and Cormode, ICDE 2015] Conservative or liberal? personalized differential privacy.

Related Work

Not cited in the paper: