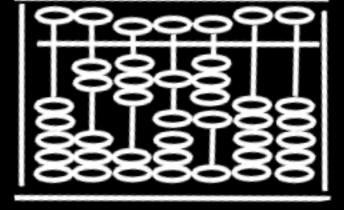
# Data Privacy: Tensions and Opportunities

Katrina Ligett Assistant Professor of Computer Science and Economics Caltech





what's the problem?

T

#### vant to compute on it

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	N
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Υ	Υ
Jennifer Kim	3/1/70	F	135	Ν	Ν
Rachel Waters	9/5/43	F	140	Ν	Ν

marma	008	904	waight	amaka T	LLING CONTROL
Julian Case	化起气性剂	W	11225	W	16
Jame Smith	11/11/日本	F	114ED	196	196
Elliper Japeners	用作用指数	F	11282)	W	W.
Jamillar Kim	3(4)(79)	F	1(385)	196	16
Thesi Presi	12/12/4(3)	)F	54D	196	156

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	Ν
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	Ν	Ν
<b>Rachel Waters</b>	9/5/43	F	140	Ν	Ν



name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	N
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	Ν	Ν
Rachel Waters	9/5/43	F	140	Ν	N

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	N
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Y	Υ
Jennifer Kim	3/1/70	F	135	Ν	Ν
Rachel Waters	9/5/43	F	140	Ν	Ν

public

18%







data privacy	
data privacy day	75,400,000 results
data privacy laws	17,600,000 results
data privacy act	11,100,000 results
data privacy policy	60,400,000 results
data privacy safe harbor	332,000 results
data privacy breaches	1,320,000 results
data privacy legislation	980,000 results
data privacy audit	684,000 results
data privacy through optimal k-anonymization	4,200 results
data privacy laws us	71,900,000 results
	close

**e**<sup>m</sup>

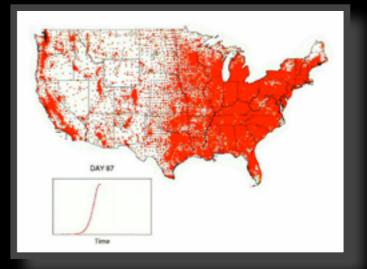
Department of the Treasury-Internal Revenue Service U.S. Individual Income Tax Return 2008

facebook





- Finding statistical correlations
  - Genotype/phenotype associations
  - Correlating medical outcomes with risk factors or events
- Publishing aggregate statistics
- Noticing events/outliers
  - Intrusion detection
  - Disease outbreaks



- Datamining/learning tasks
  - Use customer data to update strategies

See personalized recommendations

Sign in

New customer? Start here

access to the output should not enable one to learn anything about an individual that could not be learned without access

is this possible?

hint: either privacy or usefulness is easy

# what if wanted to do a study about smoking and cancer?

THEFTHE	DOB	904	waight	amaka r	LIND CARTOON
Japhyn Chree	化起气性剂	W	11215	W	16
- Smith	3/5/84	(F	114ED	196	196
	- (5-4) (54)	)F	1100	W	W
		F.	1(385)	196	196
• 6			114ED	196	16

#### what if someone knew Alice is a smoker?

there is a correlation of xxx

access to the output should not enable one to learn anything about an individual that could not be learned without access



name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	N
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Y	Υ
Jennifer Kim	3/1/70	F	135	Ν	Ν
Rachel Waters	9/5/43	F	140	Ν	Ν

public

18%

#### think of output as randomized

public

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Y	Ν
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	Ν	Ν
<b>Rachel Waters</b>	9/5/43	F	140	Ν	Ν

what to promise about output? think of output as randomized promise: if you leave the database, no outcome will change probability by very much 17 18 19 16 public

#### more formally...

- Database D a set of rows, one per person
- Sanitizing algorithm M probabilistically maps
  D to event or object in outcome space



### differential privacy

[DinurNissim03, DworkNissimMcSherrySmith06]

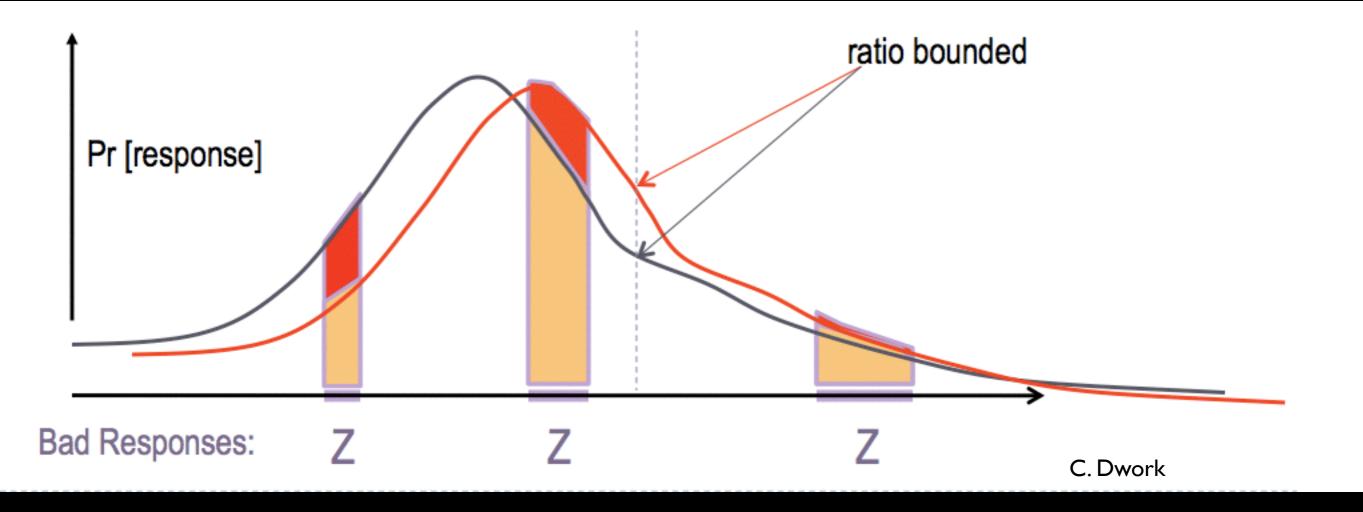
 $\begin{array}{l} \hline {\mathbb E} \mbox{-Differential Privacy for mechanism M:} \\ \mbox{For any two neighboring data sets } D_1, D_2, \\ \mbox{any } {\mathbb C} \in {\rm range}({\mathbb M}), \\ \mbox{Pr}[{\mathbb M}({\mathbb D}_1) \in {\mathbb C}] \leq {\rm e}^{\epsilon} \mbox{ Pr}[{\mathbb M}({\mathbb D}_2) \in {\mathbb C}] \end{array}$ 

16

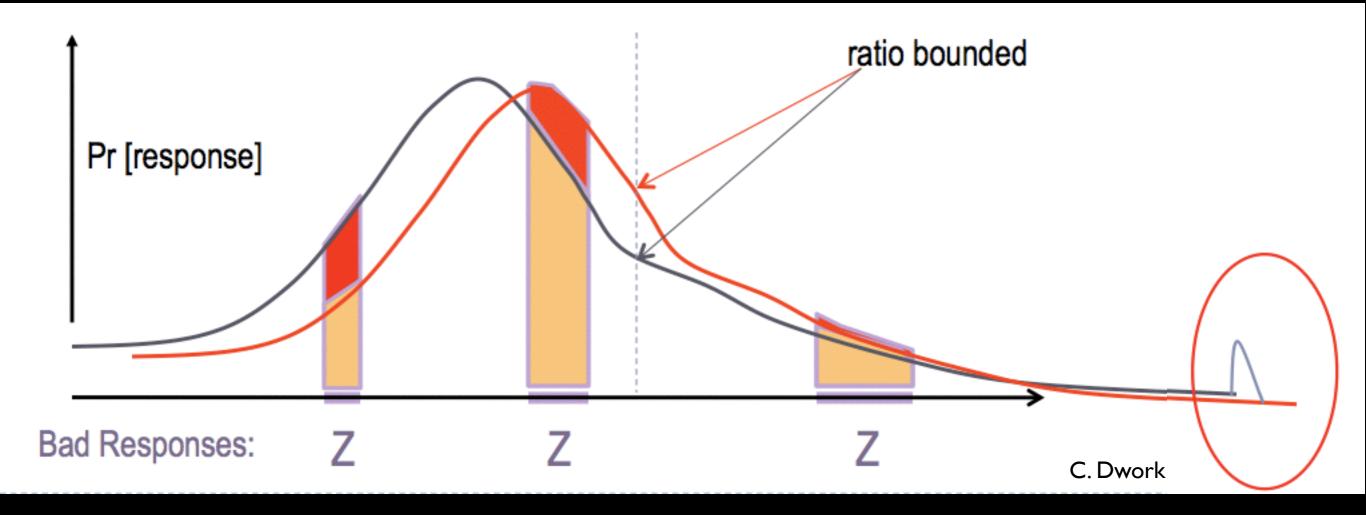
#### differential privacy $Pr[M(D_1) \in C] \le e^{\varepsilon} Pr[M(D_2) \in C]$

name	DOB	sex	weight	smoker	lung cancer				
John Doe	12/1/51	М	185	Y	N				
Jane Smith	3/3/46	F	140	N	N				
Ellen Jones	12455		100	<b>-</b>	γ				
Jennifer Kim	3/1/70	F	135	N	N				
Rachel Waters	9/5/43	F	140	N	N				
			LIBERTY	R DOULNR				20	
							だ。 		

#### differential privacy $Pr[M(D_1) \in C] \le e^{\varepsilon} Pr[M(D_2) \in C]$



#### ( $\varepsilon,\delta$ )-differential privacy Pr[M(D<sub>1</sub>) $\in$ C] $\leq e^{\varepsilon} Pr[M(D_2) \in C]+\delta$



## differential privacy $Pr[M(D_1) \in C] \le e^{\varepsilon} Pr[M(D_2) \in C]$

Is a statistical property of mechanism behavior

- unaffected by auxiliary information
- independent of adversary's computational power

#### differential privacy $Pr[M(D_1) \in C] \le e^{\varepsilon} Pr[M(D_2) \in C]$

promise: if you leave the database, no outcome will change probability by very much

is this achievable?



### if your output is a number...

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	М	185	Υ	Ν
Jane Smith	3/3/46	F	140	Ν	Ν
Ellen Jones	4/24/59	F	160	Υ	Υ
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

add noise with particular shape

public

**8%** 

scale of noise depends on sensitivity of function to compute  $\max_{D1,D2} |f(D_1) - f(D_2)|$ 

for neighboring data sets  $D_1$ ,  $D_2$ 

- measures how much one person can affect output
- sensitivity is 1 for counting queries that count number of rows satisfying a predicate

#### more concrete

name	DOB	sex	weight	sm
John Doe	12/1/51	М	185	Y
Jane Smith	3/3/46	F	140	Ν
Ellen Jones	4/24/59	F	160	Y
Jennifer Kim	3/1/70	F	135	Ν
Rachel Waters	9/5/43	F	140	Ν

what fraction over age 50? what fraction smoke and have lung cancer? what fraction of males over 150 lbs? Ν

 $\bullet \bullet \bullet$ 

11421710	DOB	904	waight	ancika r	lung cancer
Julian Case	化卸气推开	W	11225	W.	16
Jame Smith	3/3/88	F	1148D	16	16
Elliper Japensen	用作用指数	F	11222	W.	W
Jamilar Kim	3(4)(99)	F	1(385)	16	16
(Figni Pres)	制作用	F	114D	16	16

#### public

Ν

# Hardt-Ligett-McSherry algorithm

repeat:

 use Exponentially Weighted Sampling to find query poorly served by our current approximation

2. measure it using Additive Noise

3. use this measurement to improve our distribution using <u>Multiplicative Weights</u> update

we can do something useful with individuals' data once we have it... but...

- participation?
- lying about data?
- compensation?
- model harm from privacy loss?
- even that quantity could be revealing...

#### Data Privacy: Tensions and Opportunities Katrina Ligett <u>katrina@caltech.edu</u>