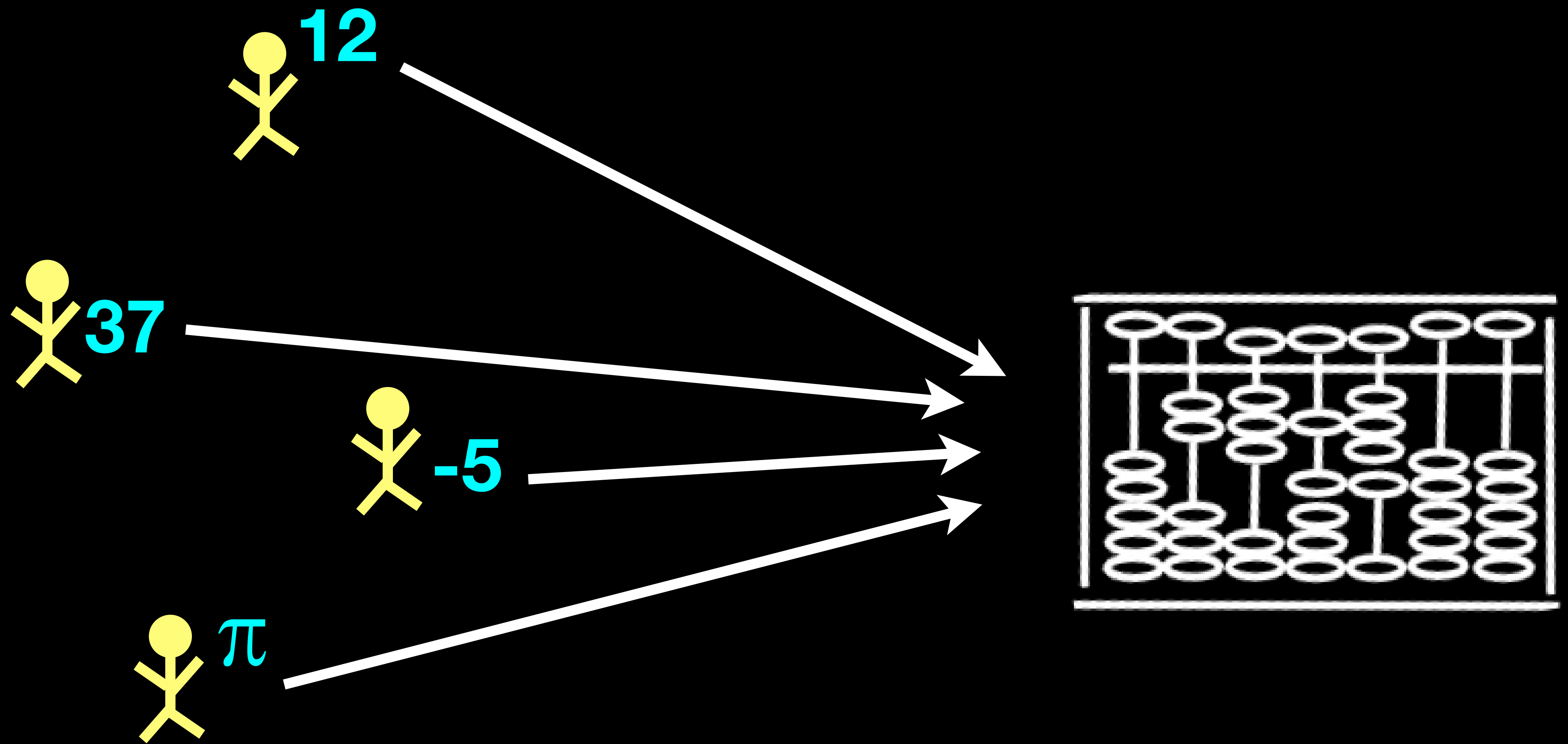
Data Privacy: Tensions and Opportunities

Katrina Ligett
Assistant Professor of Computer Science and Economics
Caltech

individuals have lots of interesting data...



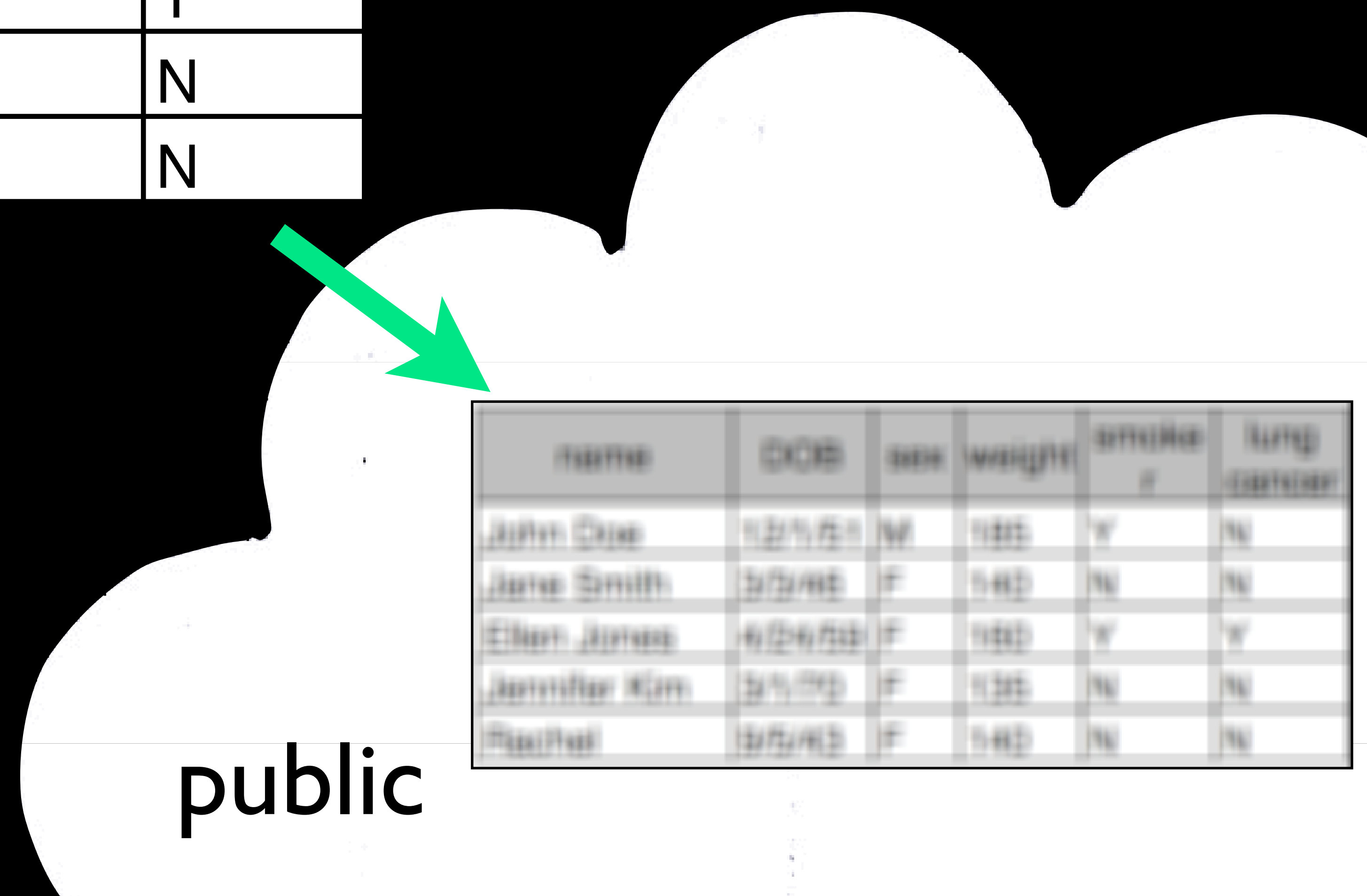
what's the problem?

want to compute on it

individuals hold data...

...what if it's sensitive?

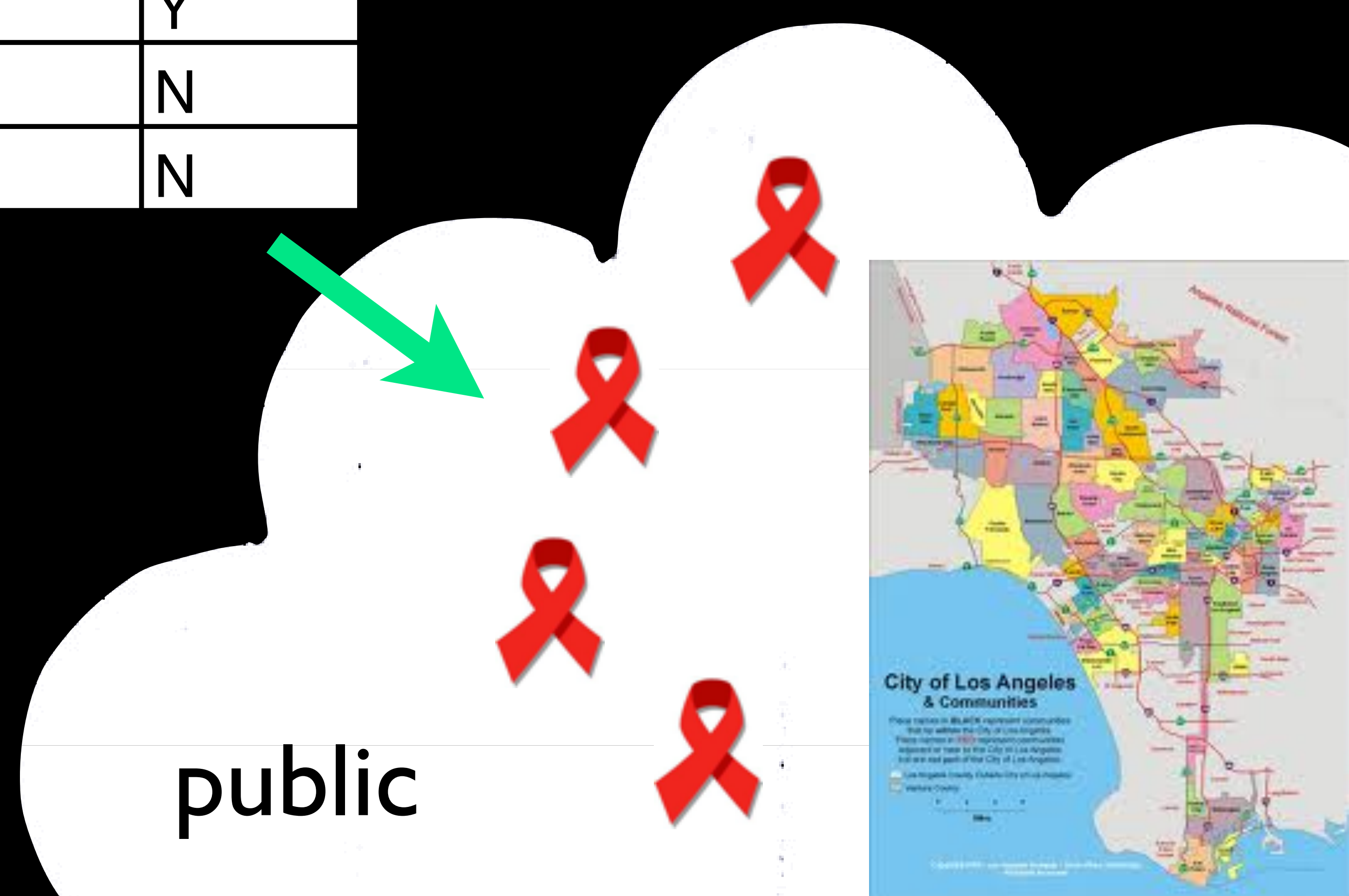
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



individuals hold data...

...what if it's sensitive?

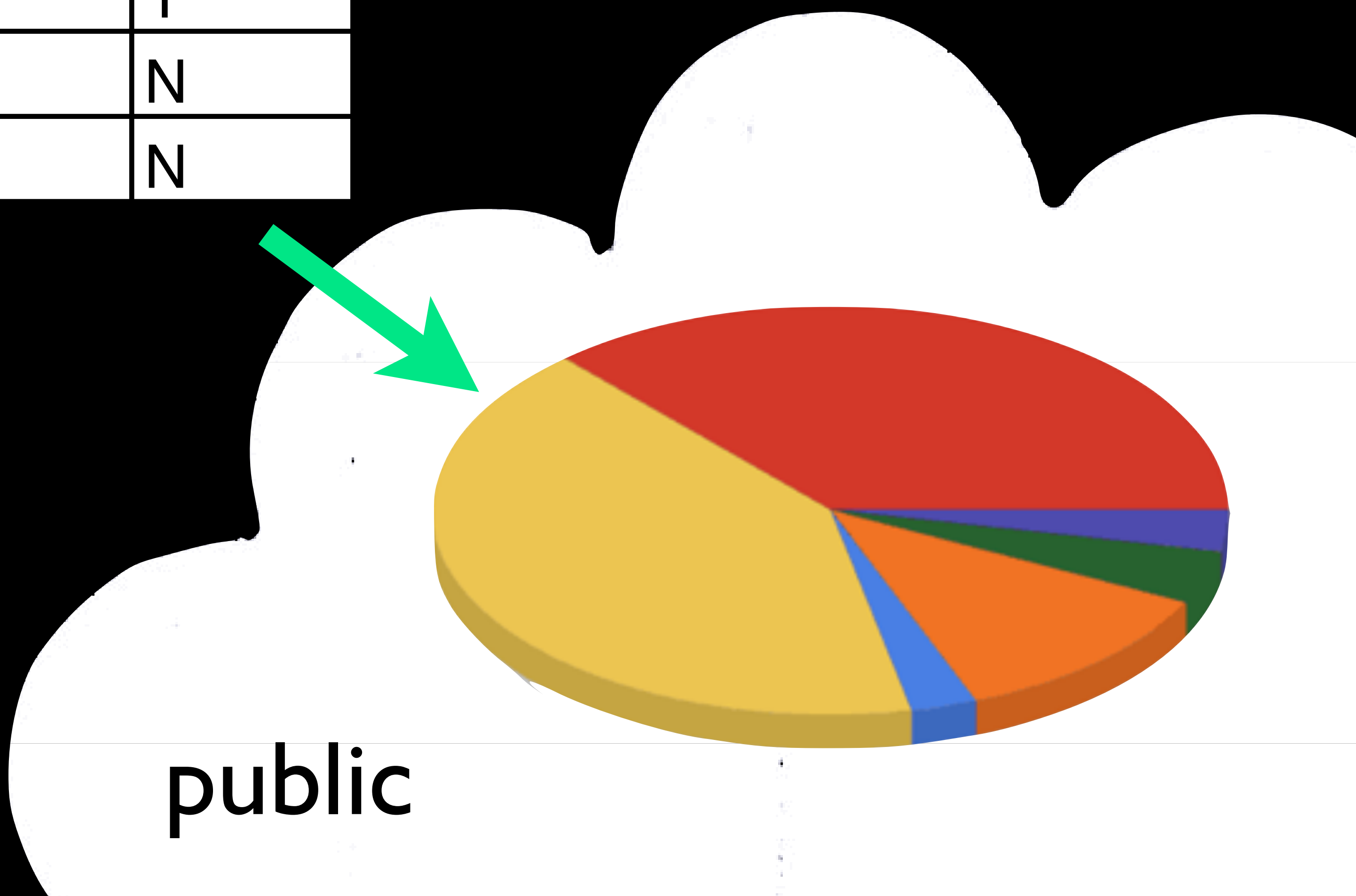
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



individuals hold data...

...what if it's sensitive?

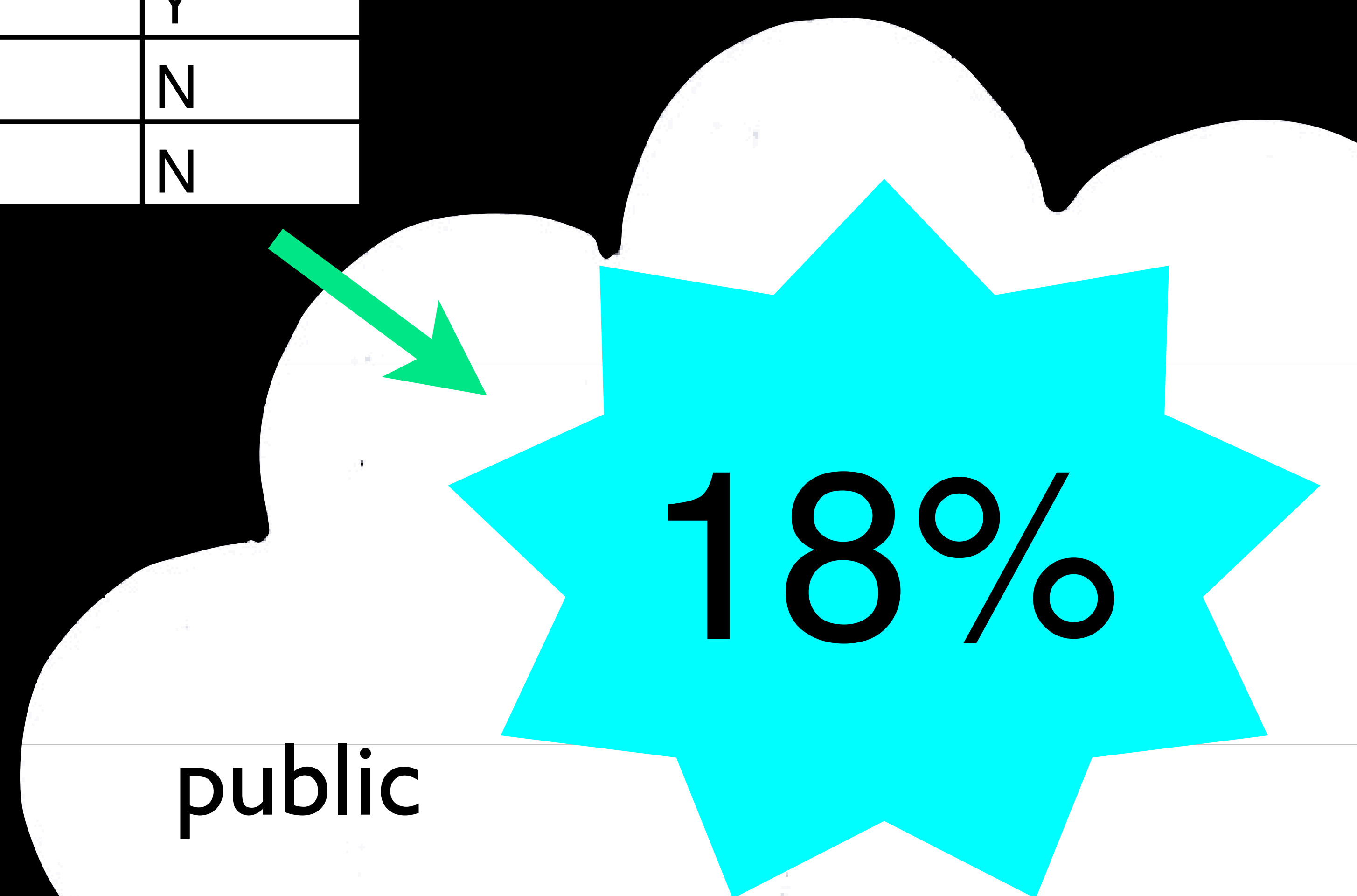
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



individuals hold data...

...what if it's sensitive?

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



Gmail by Google BETA

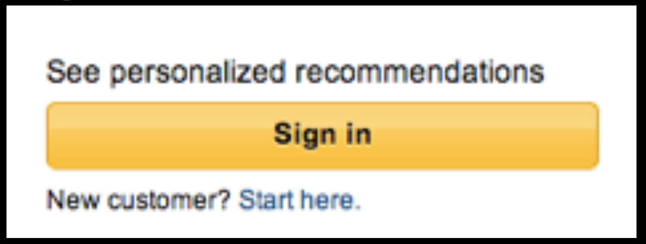
facebook

Form 1040 Department of the Treasury—Internal Revenue Service U.S. Individual Income Tax Return 2008

cicada PatientCare Portal

- 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



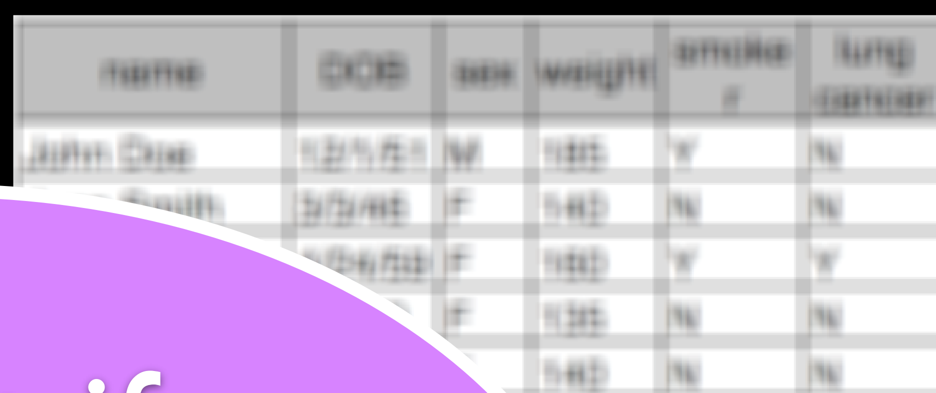
what to promise about output?

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?



what if someone knew Alice is a smoker?

public

there is a correlation of xxx

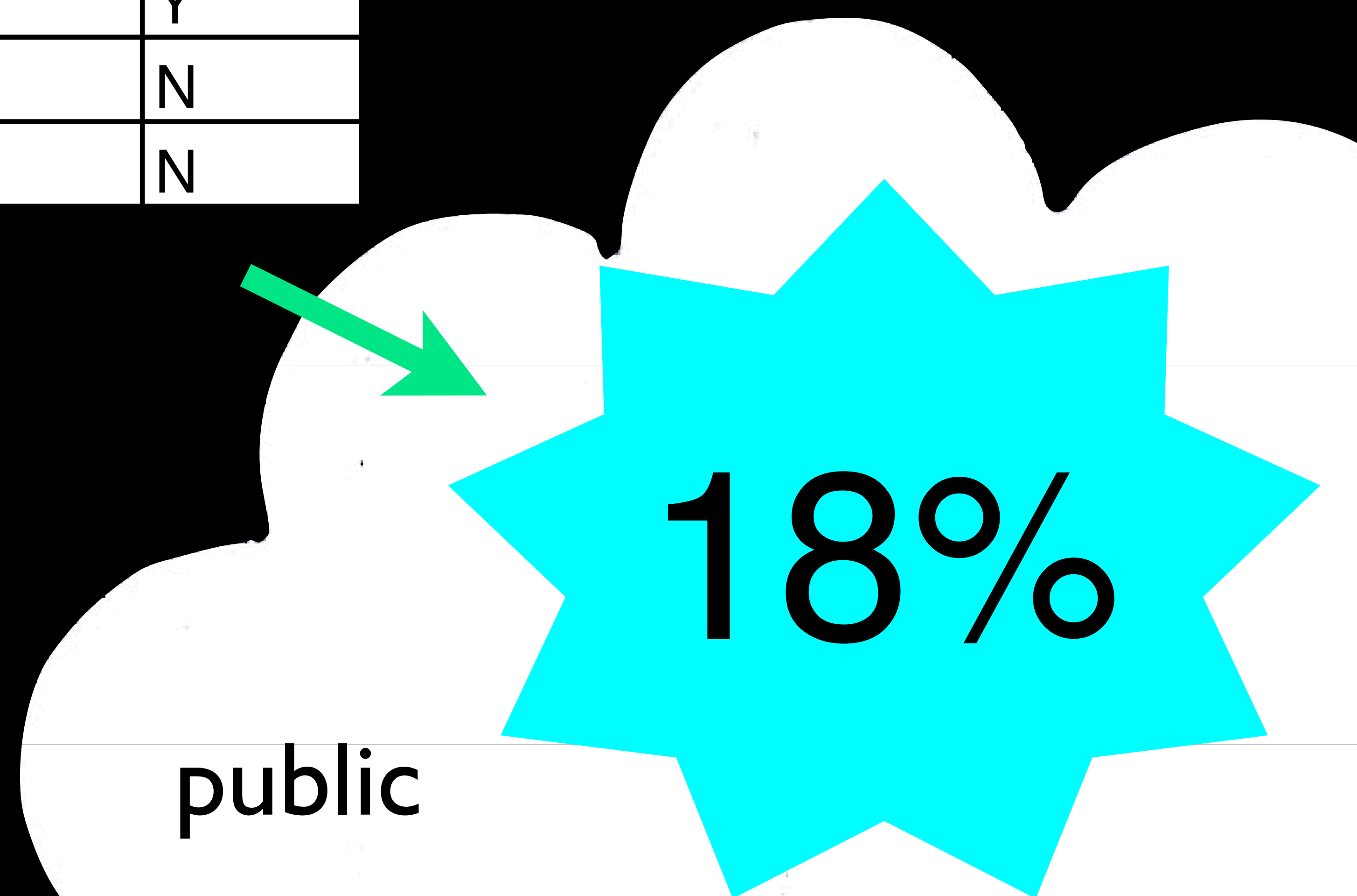
what to promise about output?

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

not possible!

what to promise about output?

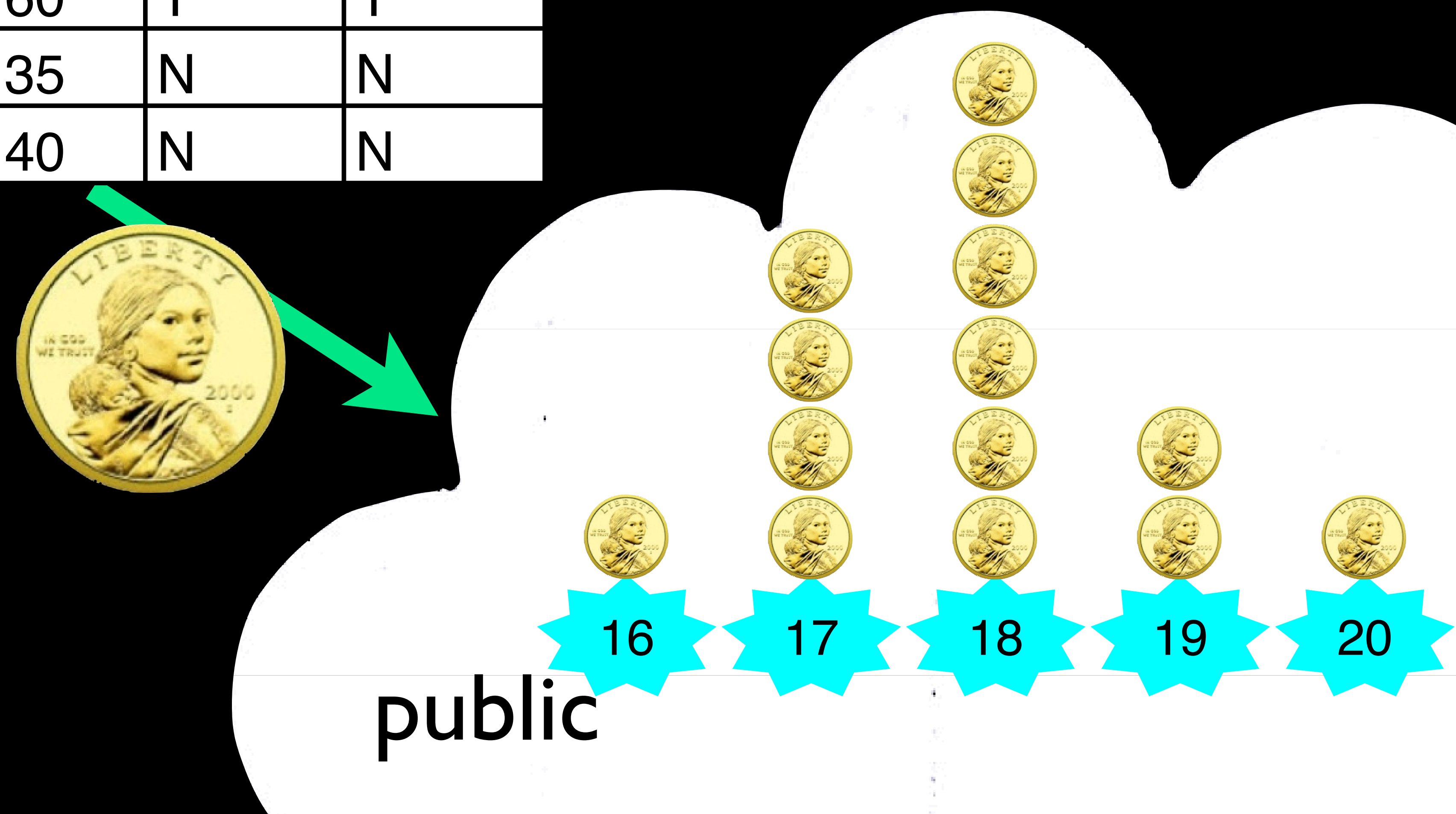
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



what to promise about output?

think of output as randomized

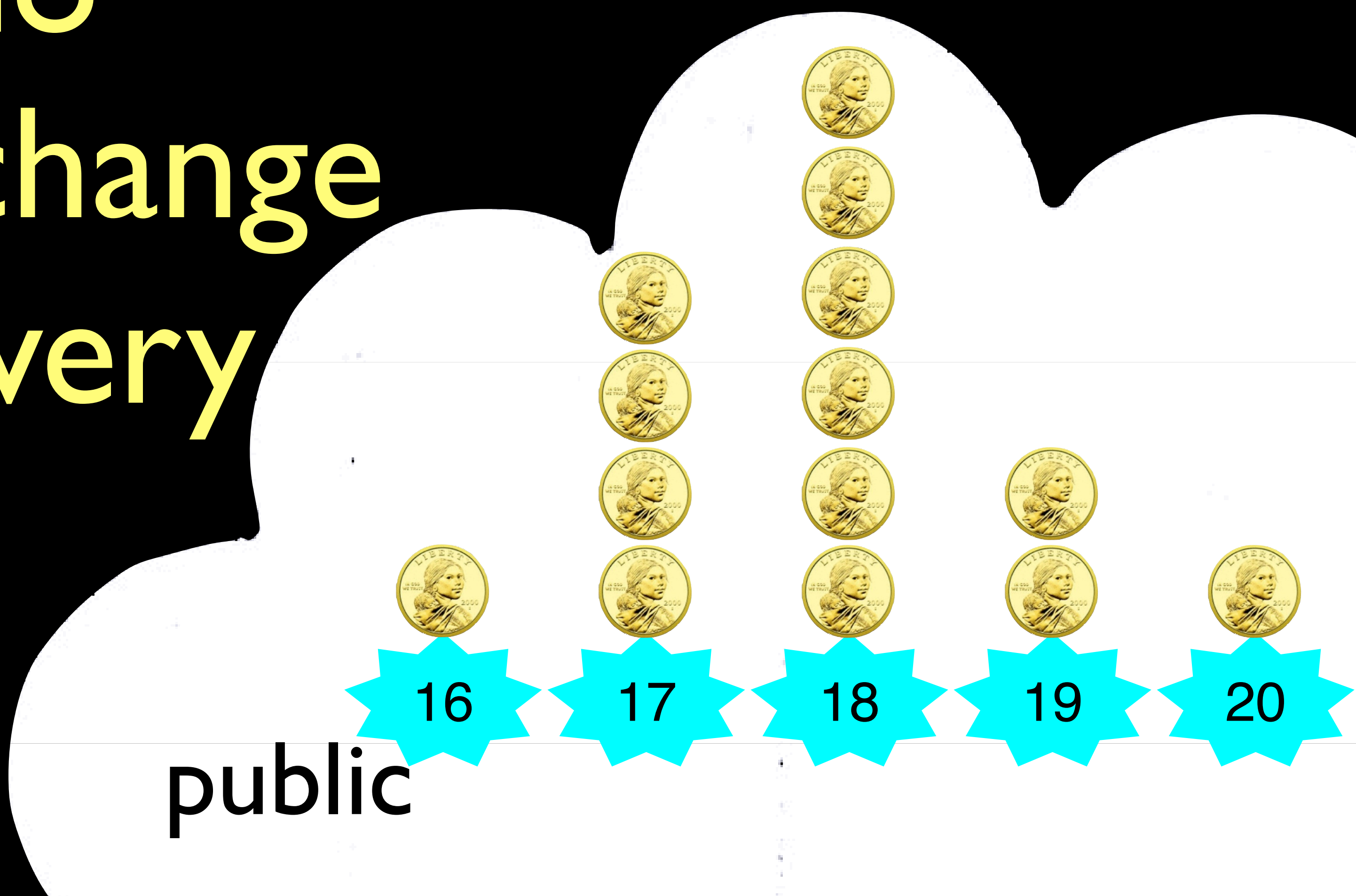
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



what to promise about output?

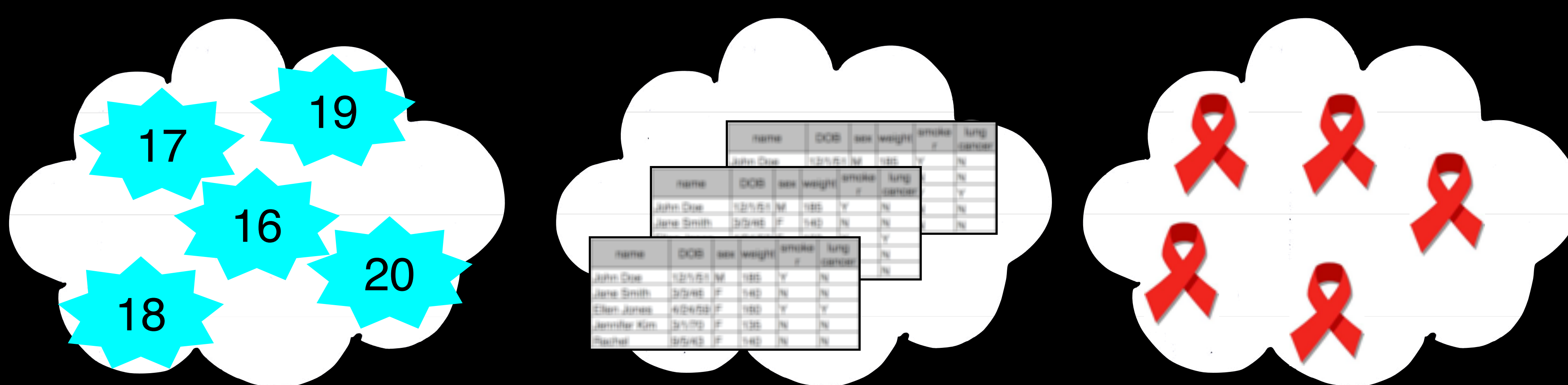
think of output as randomized

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



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

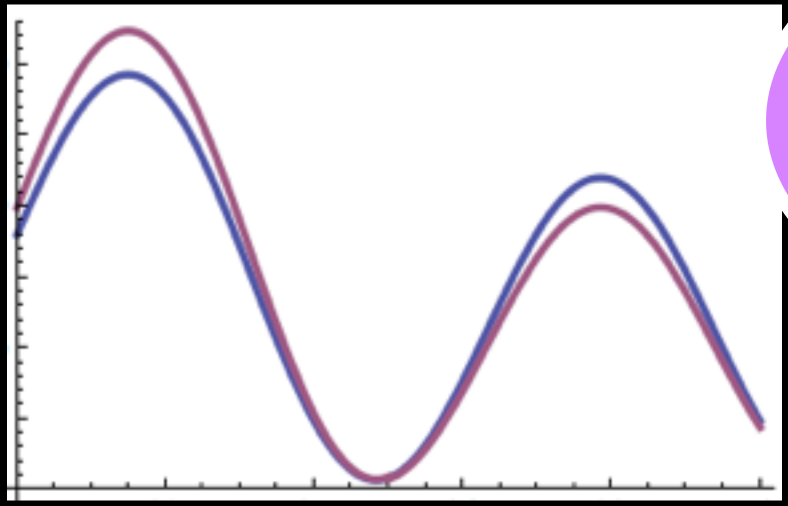
[DinurNissim03, DworkNissimMcSherrySmith06]

ε-Differential Privacy for mechanism M:

For any two neighboring data sets D_1, D_2,

any $C \in \text{range}(M)$,

$$\Pr[M(D_1) \in C] \leq e^{\varepsilon} \Pr[M(D_2) \in C]$$

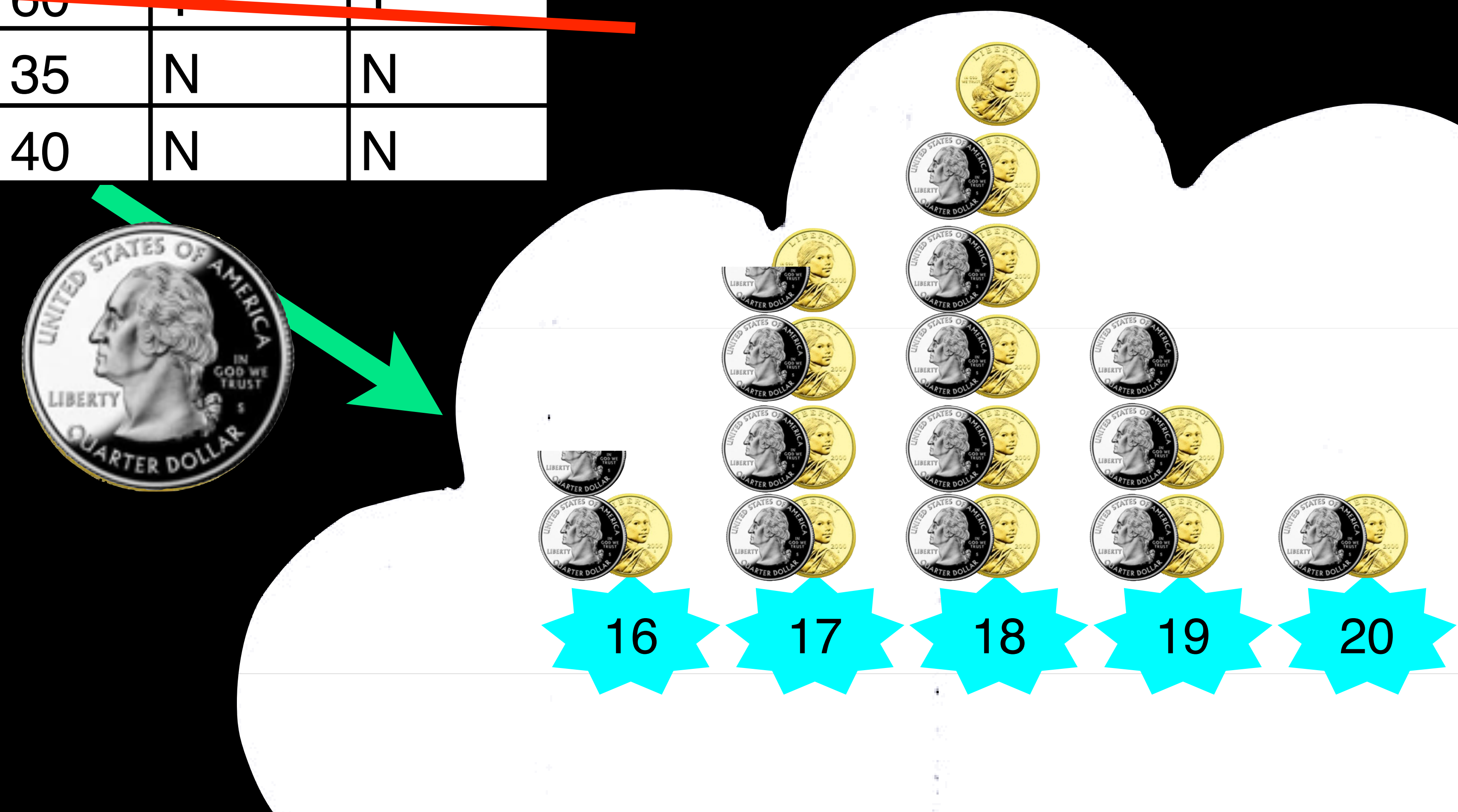


$e^{\varepsilon} \sim (1 + \varepsilon)$

differential privacy

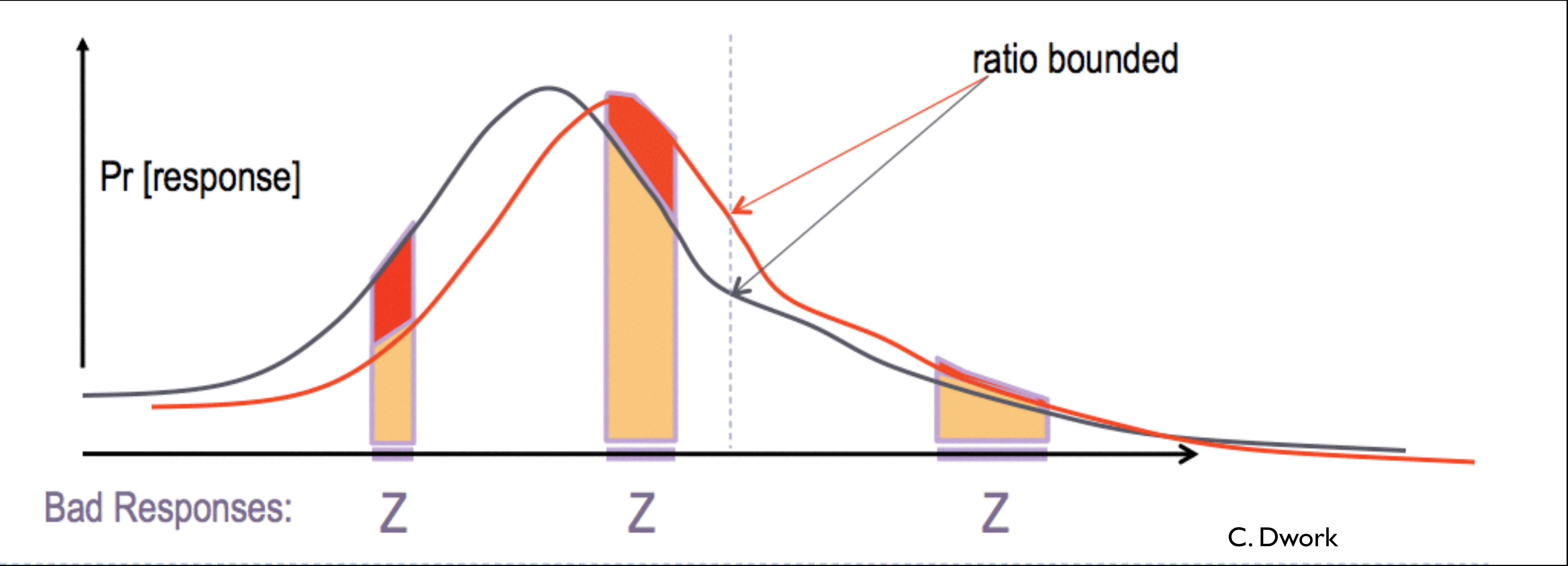
$$\Pr[M(D_1) \in C] \le e^{\varepsilon} \Pr[M(D_2) \in C]$$

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
~~Ellen Jones~~	~~4/24/55~~	~~F~~	~~160~~	~~Y~~	~~Y~~
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



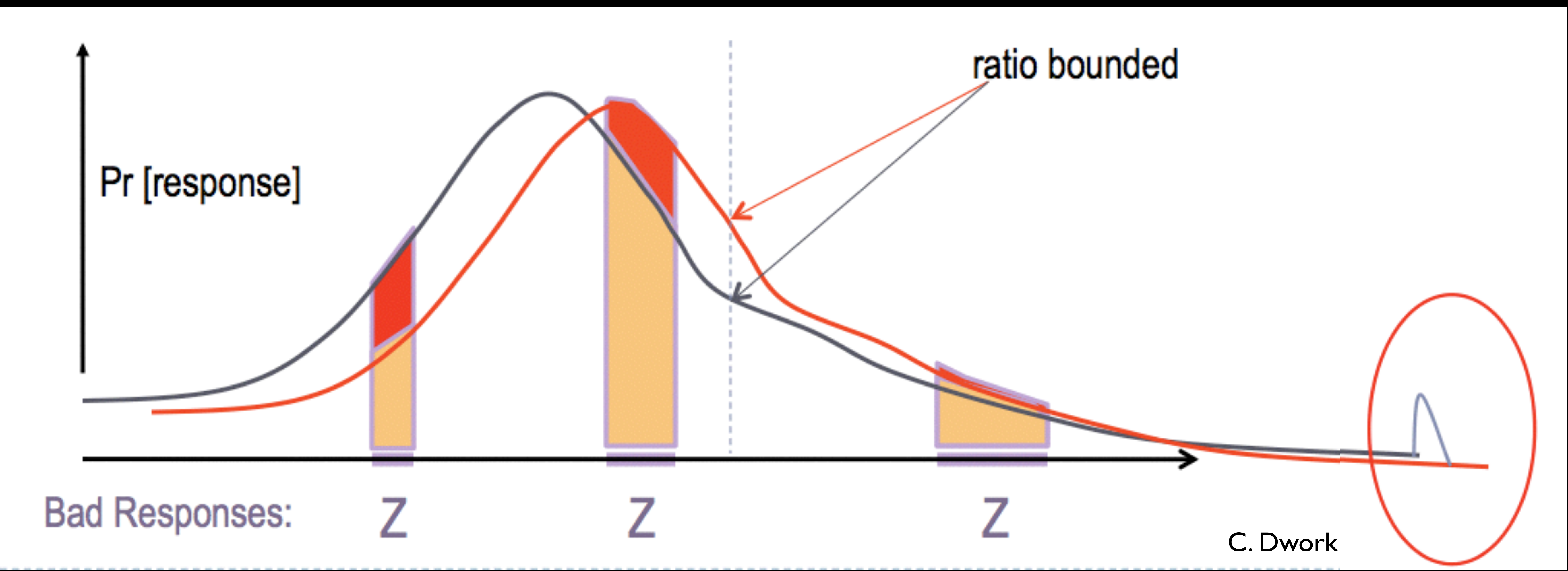
differential privacy

$$\Pr[M(D_1) \in C] \leq e^{\varepsilon} \Pr[M(D_2) \in C]$$



(ε,δ)-differential privacy

$$\Pr[M(D_1) \in C] \leq e^{\varepsilon} \Pr[M(D_2) \in C] + \delta$$



differential privacy

$$\Pr[M(D_1) \in C] \leq e^{\varepsilon} \Pr[M(D_2) \in C]$$

Is a statistical property of mechanism behavior

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

differential privacy

$$\Pr[M(D_1) \in C] \leq e^{\varepsilon} \Pr[M(D_2) \in C]$$

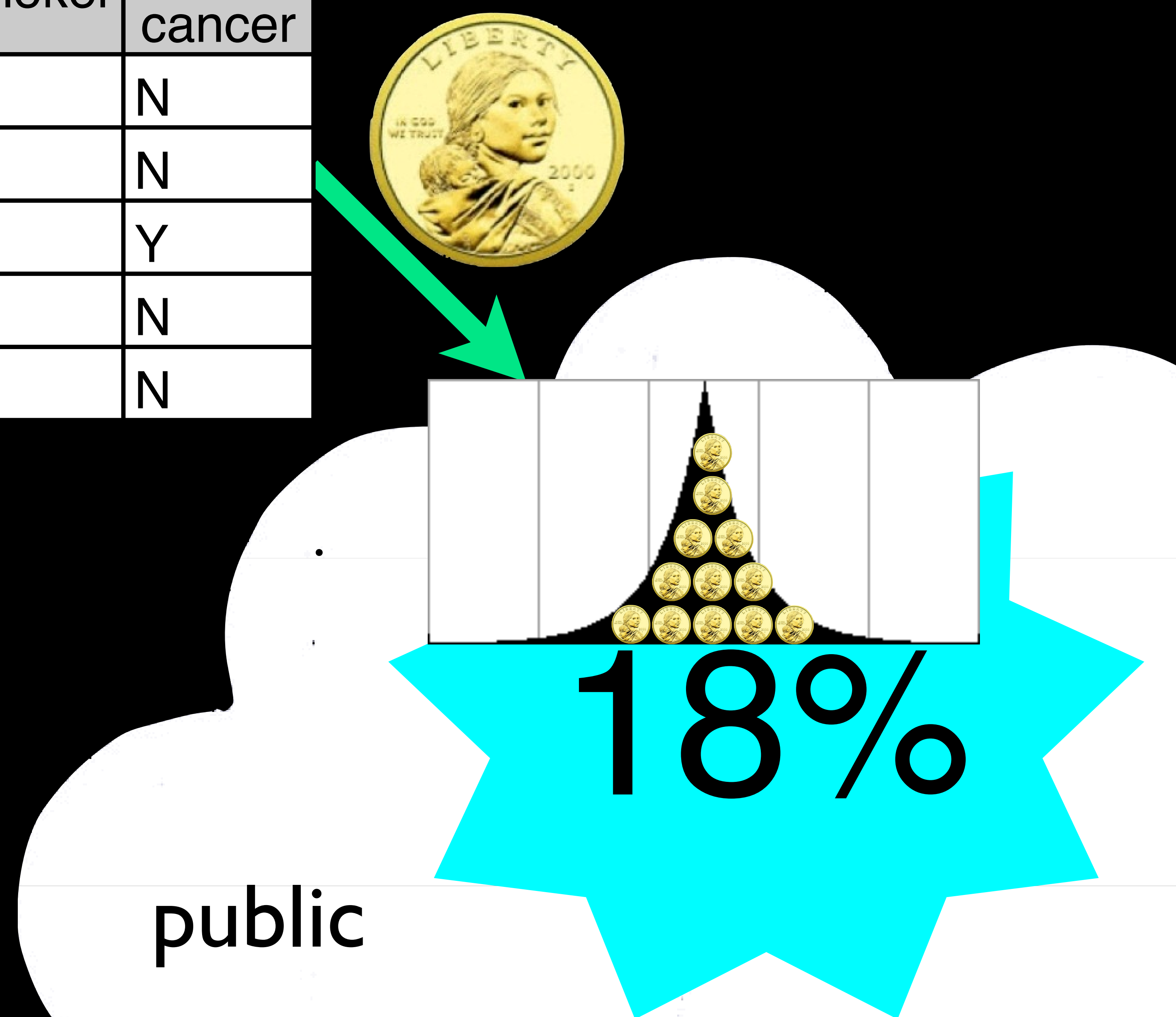
promise: if you leave the database, no outcome will change probability by very much

is this achievable?

yes!

if your output is a number...

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



add noise with particular shape

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

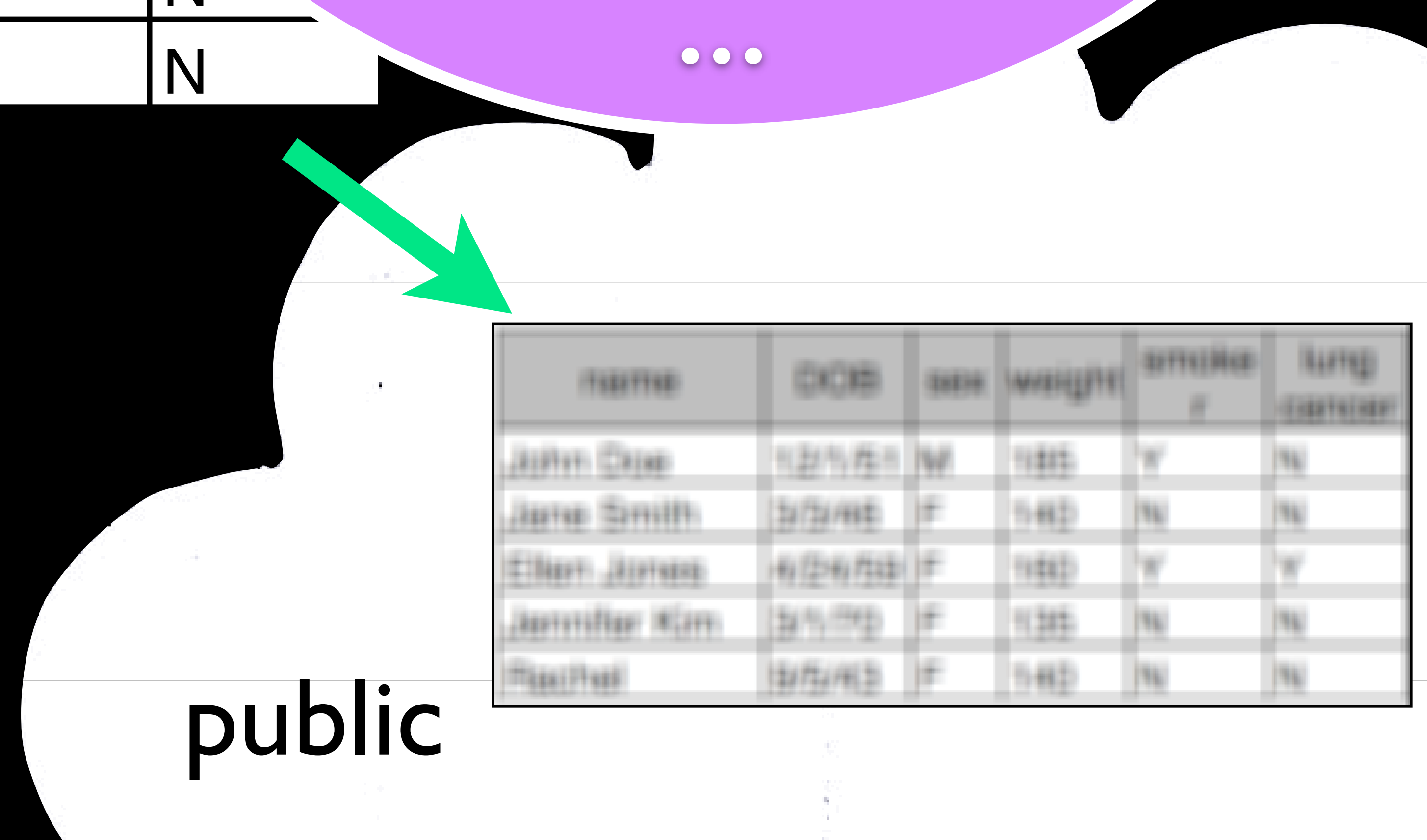
- measures how much one person can affect output
- sensitivity is 1 for counting queries that count number of rows satisfying a predicate

more concretel

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

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

...



Hardt-Ligett-McSherry algorithm

repeat:

1. 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 Multiplicative Weights 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
katrina@caltech.edu