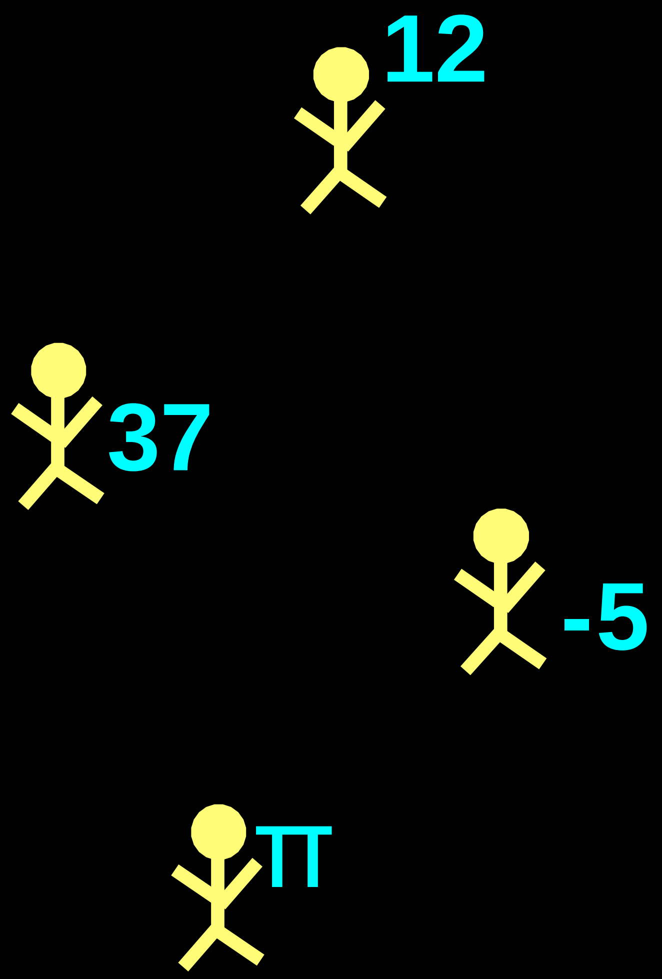
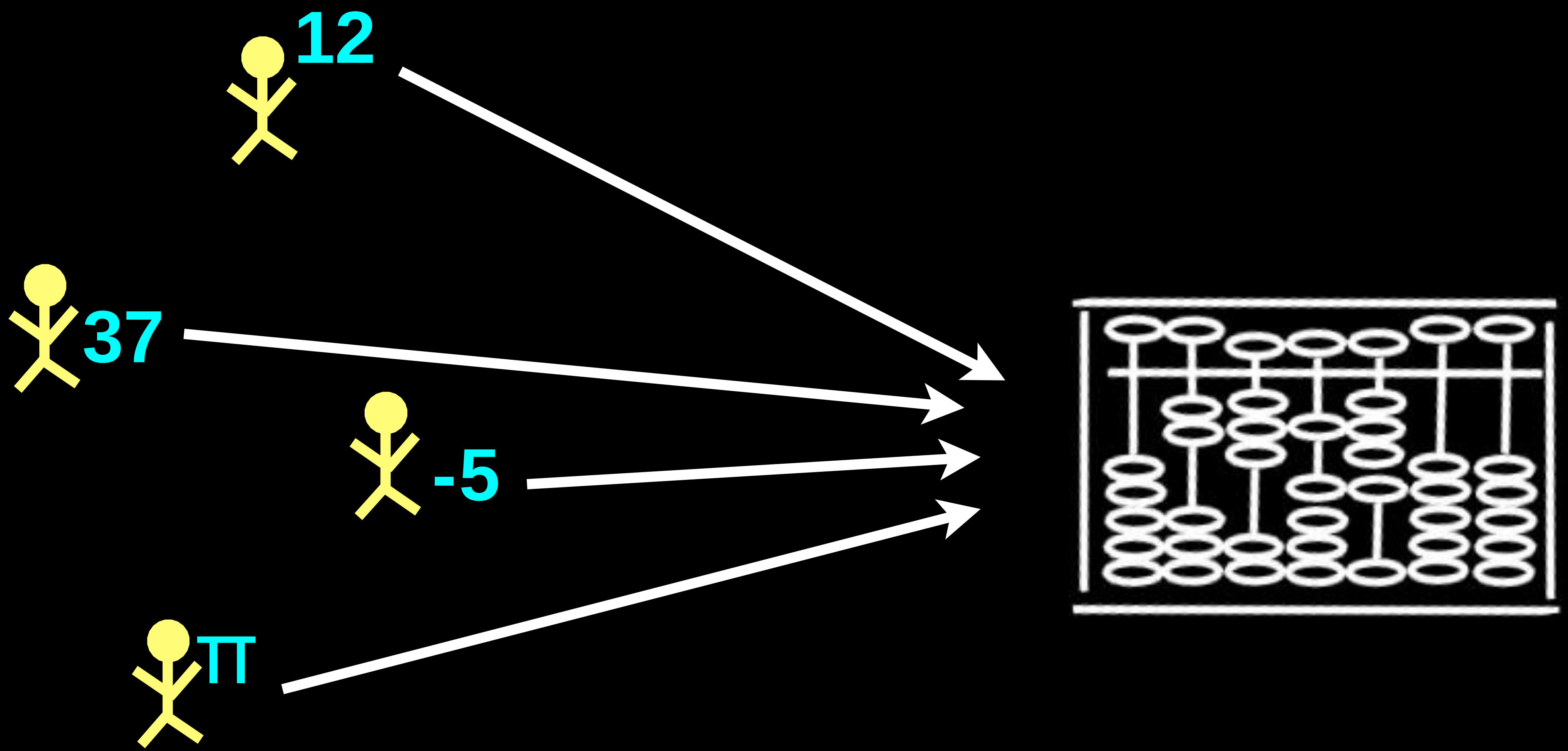
Introduction to Differential Privacy

Anudit Nagar

individuals have lots of interesting data...

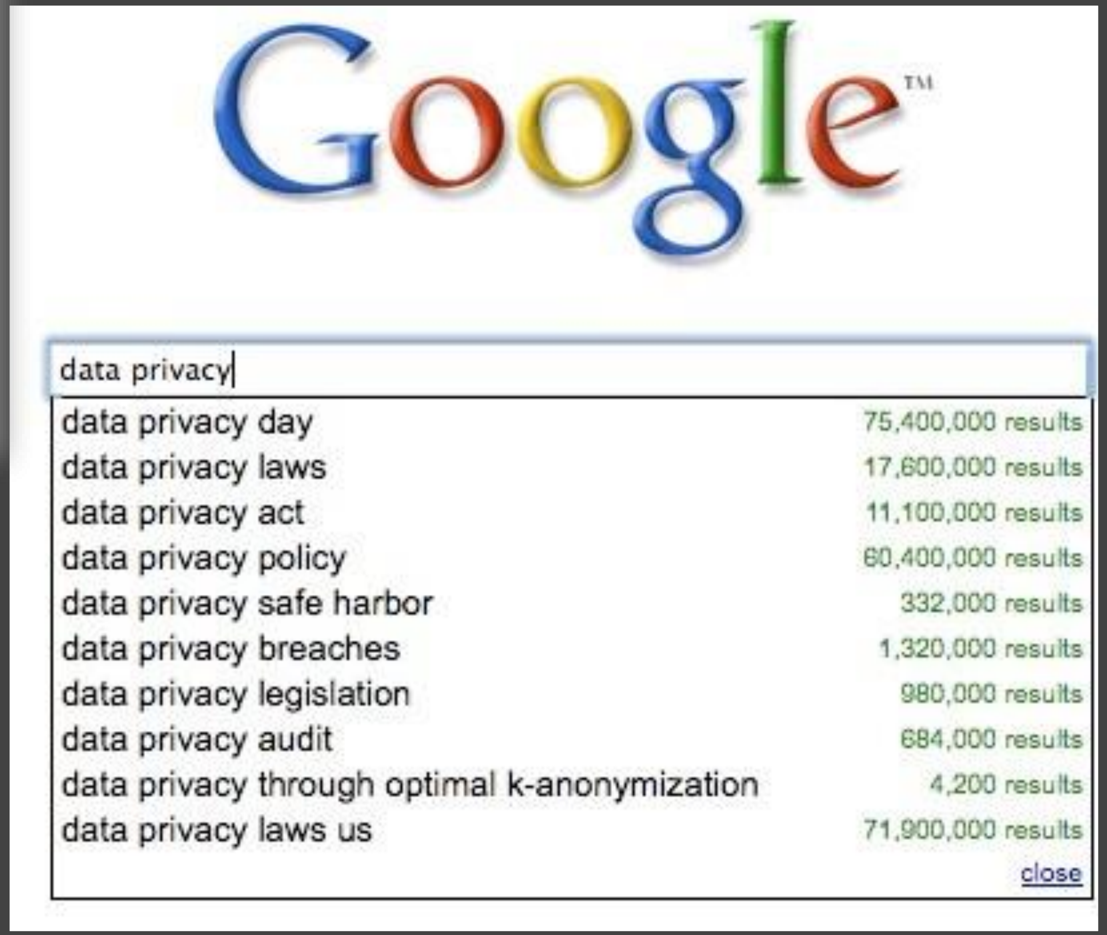


individuals have lots of interesting data...



...we want to compute on it

Gmail
by Google BETA



Google

data privacy

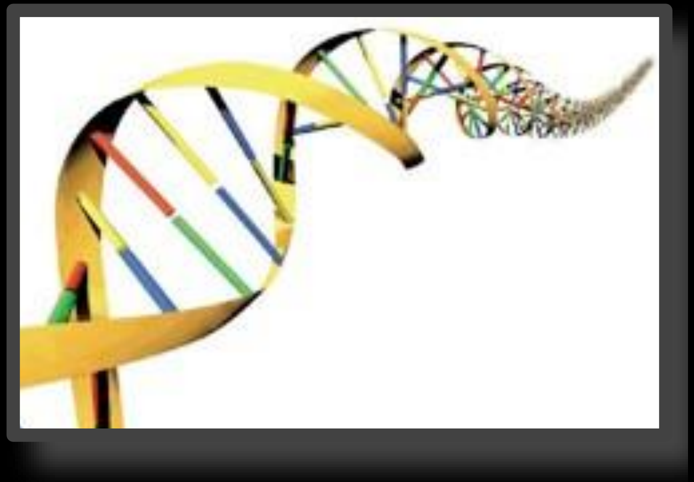
data privacy day	75,400,000 results
data privacy laws	17,600,000 results
data privacy act	11,100,000 results
data privacy policy	60,400,000 results
data privacy safe harbor	332,000 results
data privacy breaches	1,320,000 results
data privacy legislation	980,000 results
data privacy audit	684,000 results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

close

facebook

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

cicada PatientCare Portal

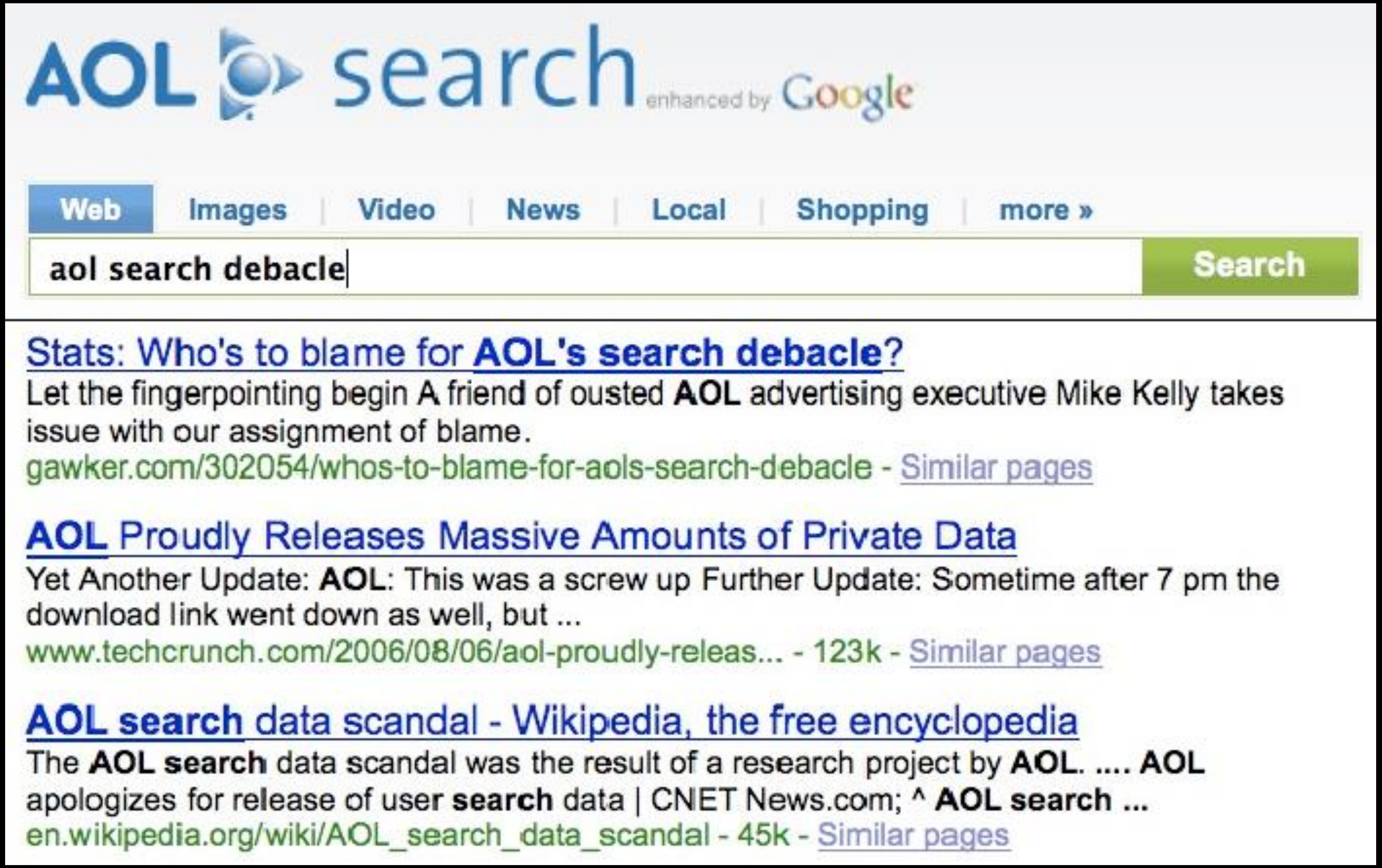


- finding statistical correlations
  - genotype/phenotype associations
  - correlating medical outcomes with risk factors or events

- publishing aggregate statistics

- noticing events/outliers
  - intrusion detection
  - disease outbreaks

- datamining/learning tasks
  - use customer data to update strategies



[Stats: Who's to blame for **AOL's search debacle**?](gawker.com/302054/whos-to-blame-for-aols-search-debacle)

Let the fingerpointing begin A friend of ousted **AOL** advertising executive Mike Kelly takes issue with our assignment of blame.

gawker.com/302054/whos-to-blame-for-aols-search-debacle - Similar pages

AOL Proudly Releases Massive Amounts of Private Data

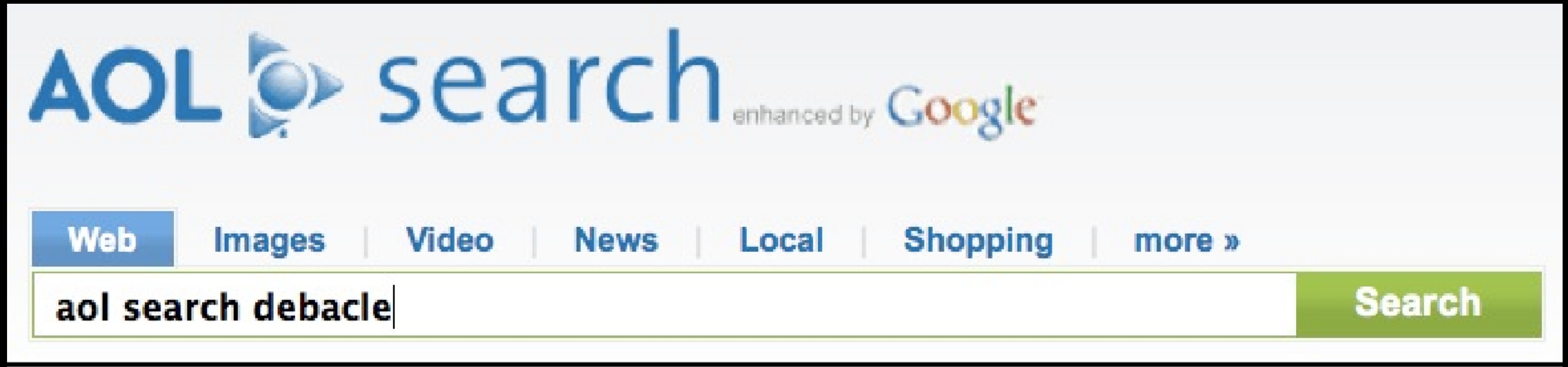
Yet Another Update: **AOL**: This was a screw up Further Update: Sometime after 7 pm the download link went down as well, but ...

www.techcrunch.com/2006/08/06/aol-proudly-releas... - 123k - Similar pages

AOL search data scandal - Wikipedia, the free encyclopedia

The **AOL search** data scandal was the result of a research project by **AOL**. .... **AOL** apologizes for release of user **search** data | CNET News.com; ^ **AOL search** ...

en.wikipedia.org/wiki/AOL_search_data_scandal - 45k - Similar pages



aol search debacle

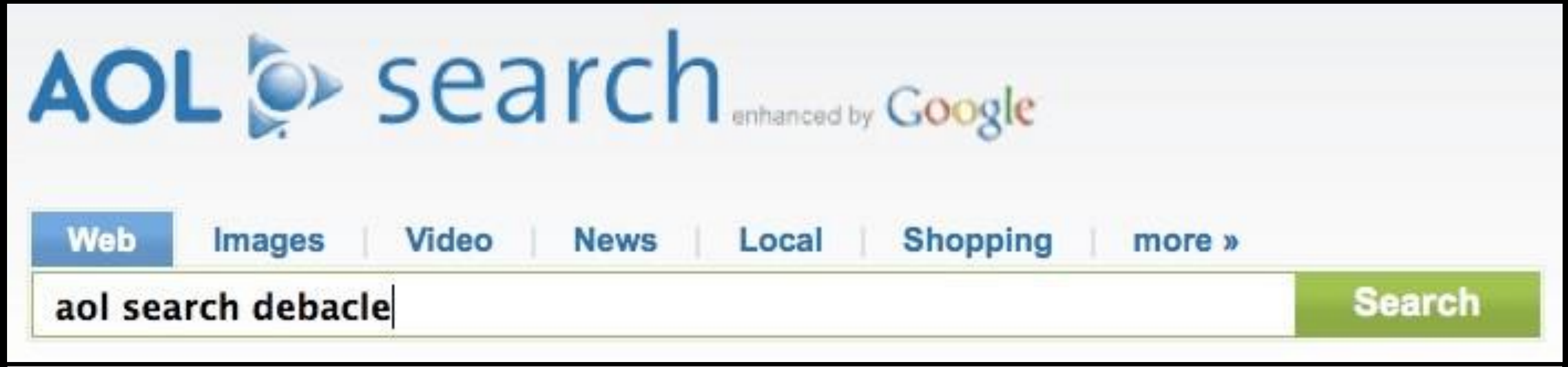
No“personally identifiable information” was released

John Doe	tax forms
John Doe	error in form 1099
Katrina Ligett	data privacy
Katrina Ligett	aol search debacle
Katrina Ligett	Ligett DBLP
Katrina Ligett	computer science news
Katrina Ligett	Caltech rankings
Katrina Ligett	weather Pasadena
Jane Smith	youtube
Jane Smith	free tv download
Jane Smith	streaming tv
Chris Jones	childrens books
Chris Jones	dr seuss
Chris Jones	“the cat and the hat”
Chris Jones	gifts for children

aol search debacle

No“personally identifiable information” was released

John Doe	tax forms
John Doe	error in form 1099
user195023	data privacy
user195023	aol search debacle
user195023	Ligett DBLP
user195023	computer science news
user195023	Caltech rankings
user195023	weather Pasadena
Jane Smith	youtube
Jane Smith	free tv download
Jane Smith	streaming tv
Chris Jones	childrens books
Chris Jones	dr seuss
Chris Jones	“the cat and the hat”
Chris Jones	gifts for children



aol search debacle

No "personally identifiable information" was released

John Doe	tax forms
John Doe	error in form 1099
user195023	data privacy
user195023	aol search debacle
user195023	Ligett DBLP
user195023	computer science news
user195023	Caltech rankings
user195023	weather Pasadena
Jane Smith	youtube
Jane Smith	free tv download
Jane Smith	streaming tv
Chris Jones	childrens books
Chris Jones	dr seuss
Chris Jones	"the cat and the hat"
Chris Jones	gifts for children

This doesn’t apply to me! I don’t want to publish the whole dataset!

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?

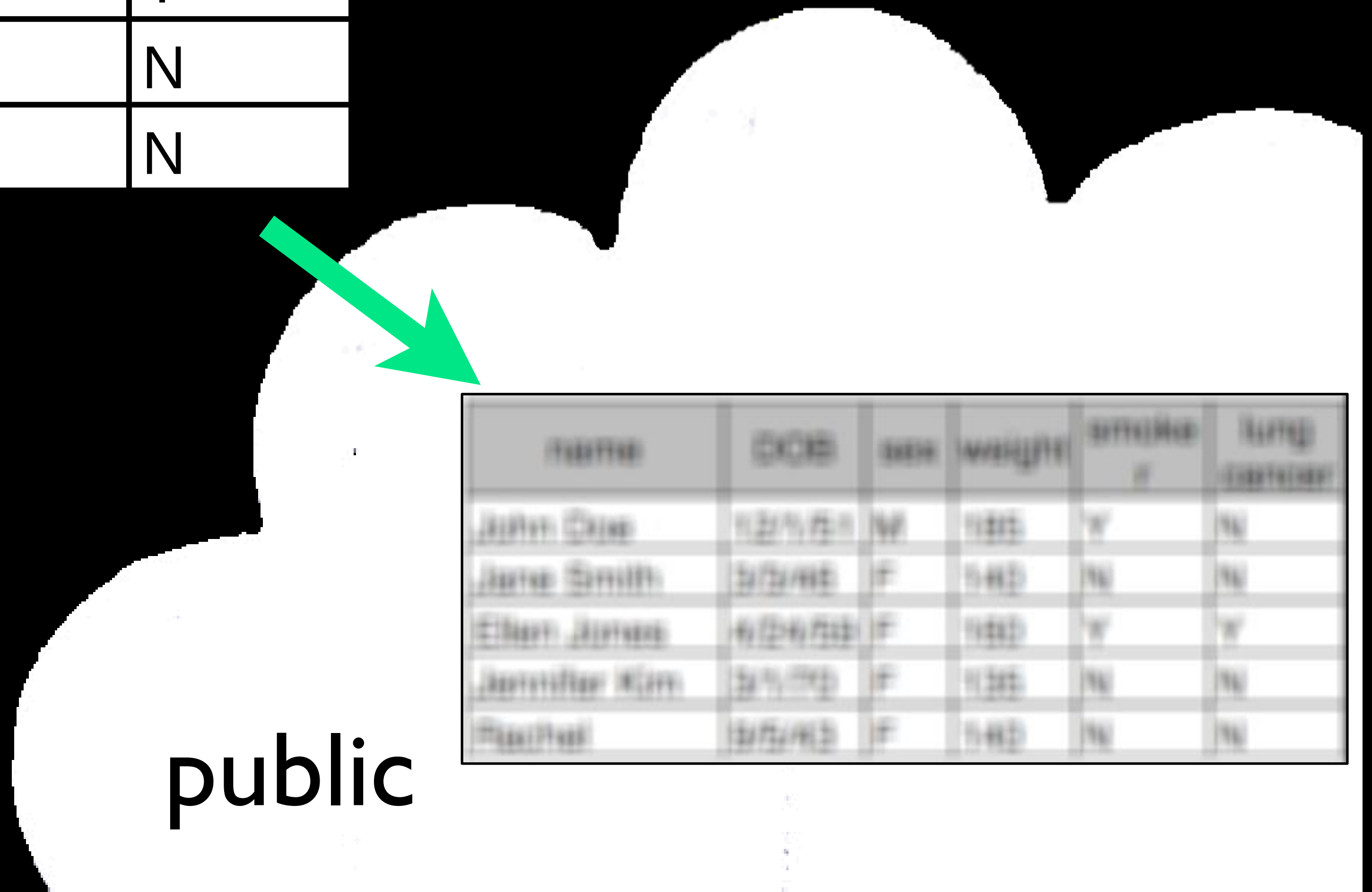
name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

public

individuals hold data...

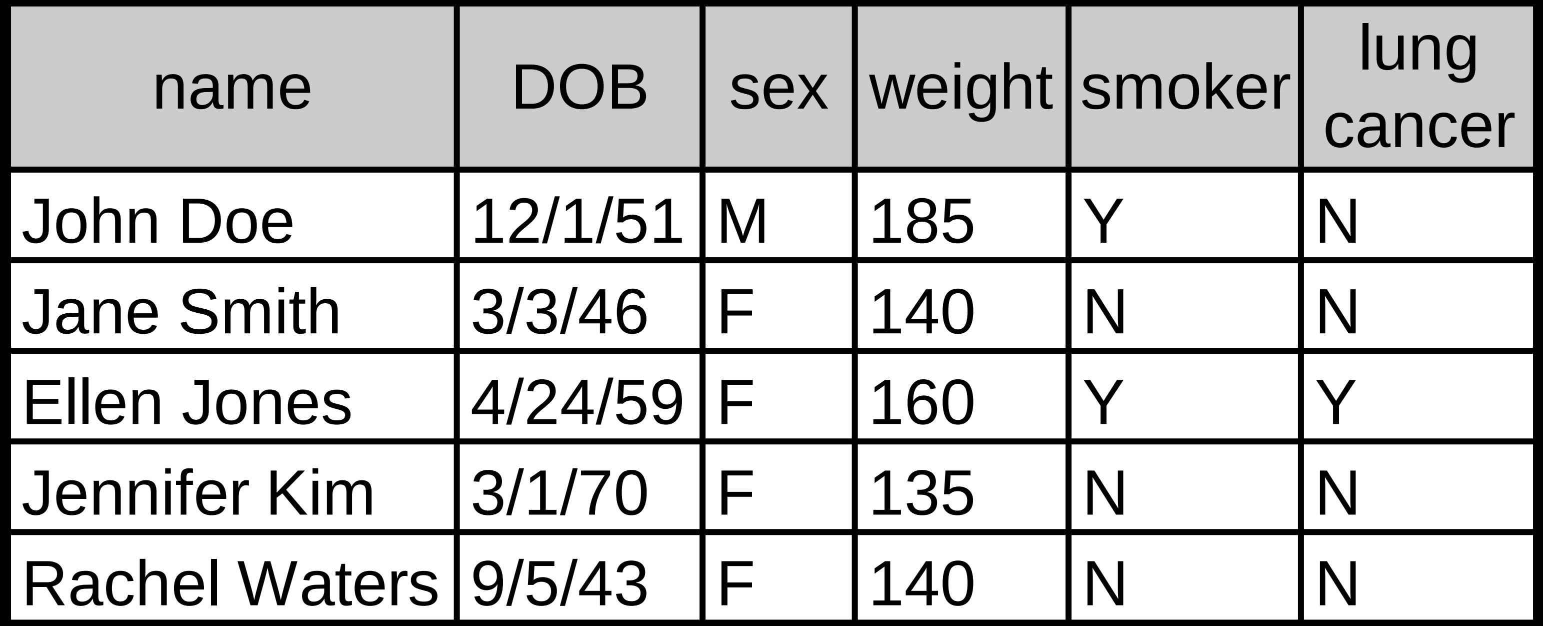
...what if it’s sensitive?

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?

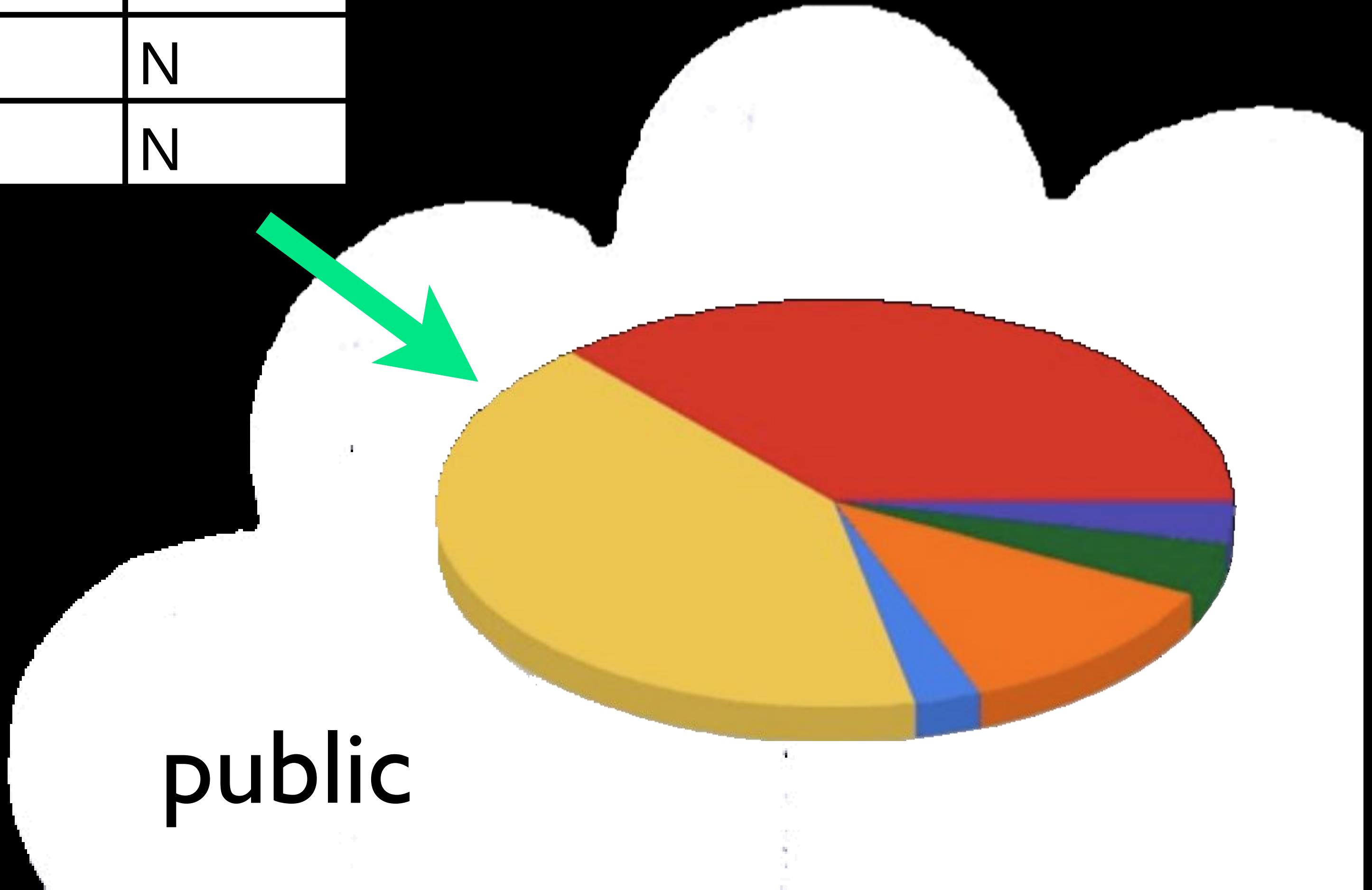


name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

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...

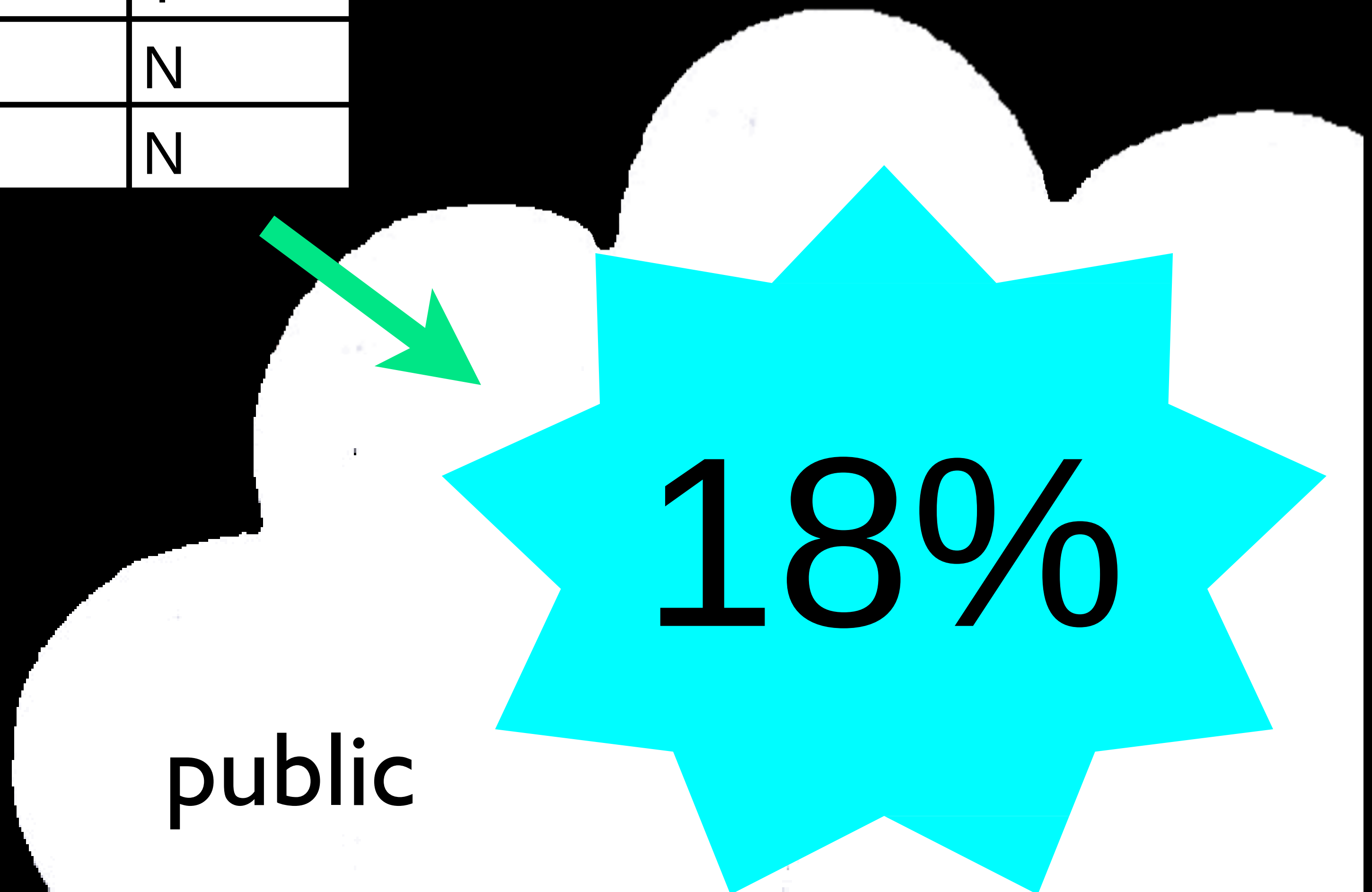
...what if it’s sensitive?

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

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...

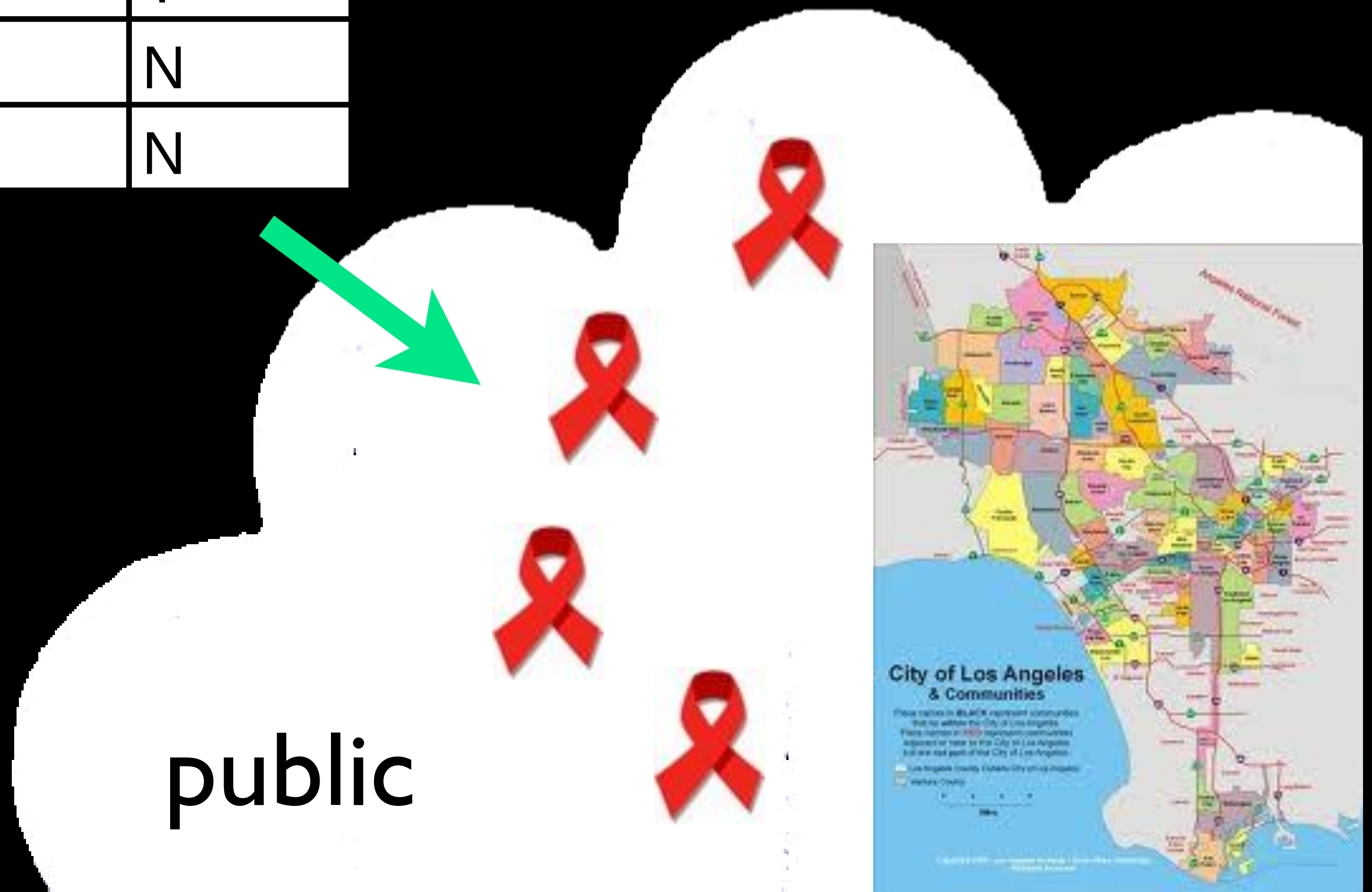
...what if it’s sensitive?

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



This doesn’t apply to me! I don’t want to publish the whole dataset!

This doesn’t apply to me! I don’t want to publish the whole dataset!

not so fast...

This doesn’t apply to me! I don’t want to publish the whole dataset!

not so fast...

see, e.g., Korolova 2011’s Facebook microtargeting attack

This doesn’t apply to me! I don’t want to publish the whole dataset!

not so fast...

see, e.g., Korolova 2011’s Facebook microtargeting attack

...must pay attention to *all* uses of sensitive data

facebook

Summing up.

an individual should not enable one to learn anything about another individual that could not be learned without access is this possible?

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

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

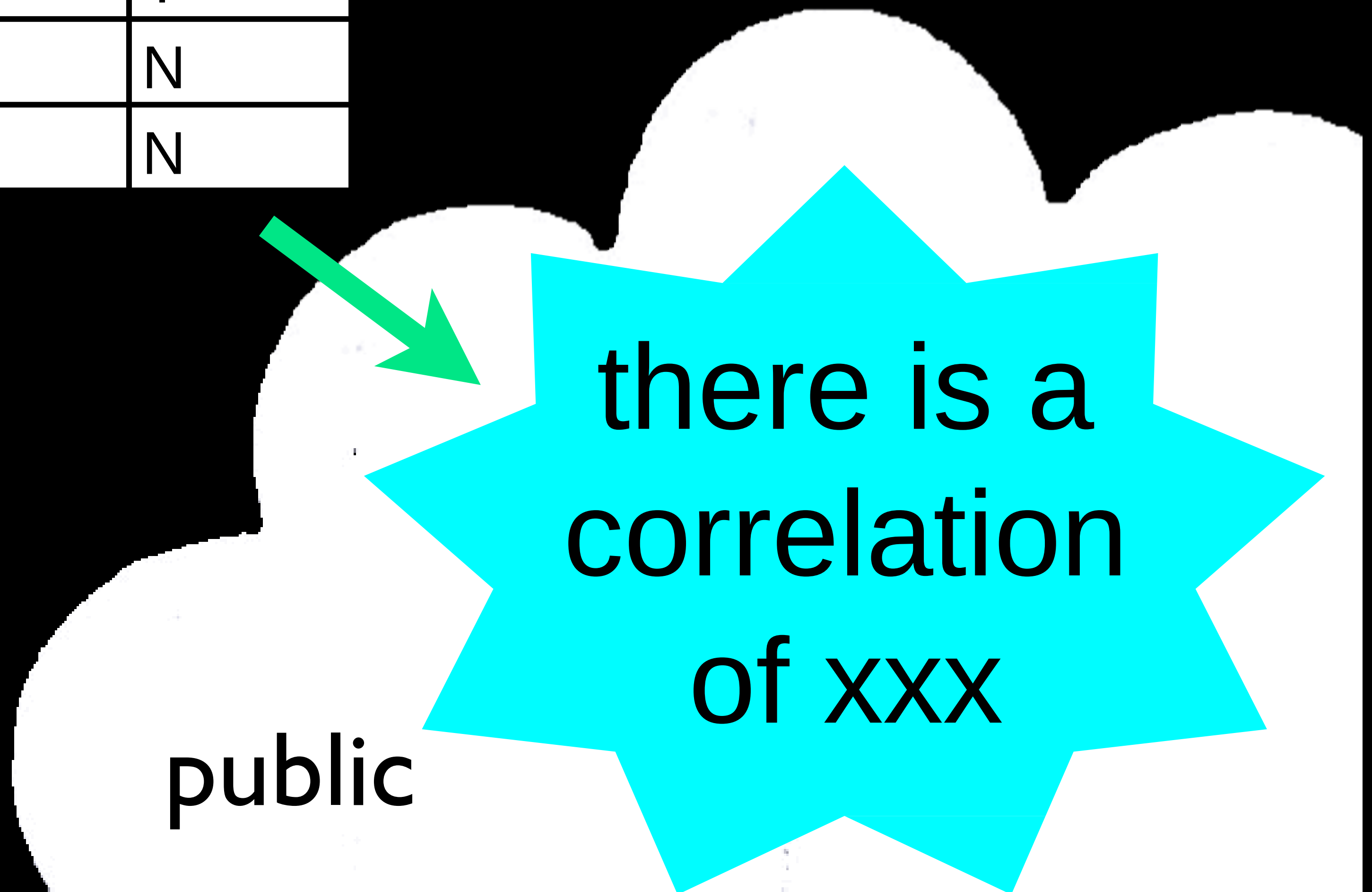
what if wanted to do a study about smoking and cancer?

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

public

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

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



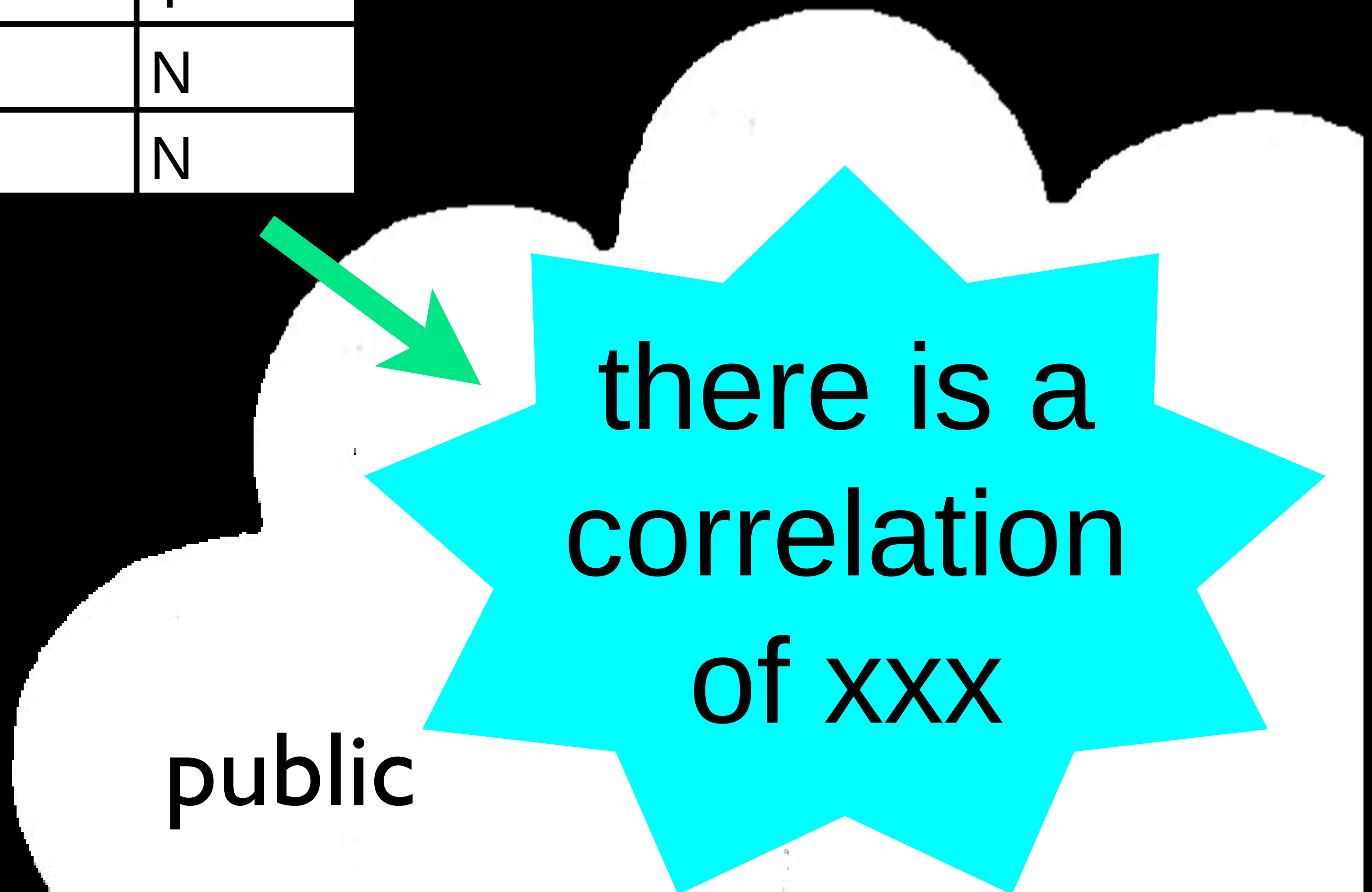
there is a correlation of xxx

public

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

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	2/4/56	F	122	N	N

what if someone knew Alice was a smoker?



there is a correlation of xxx

public

differential privacy

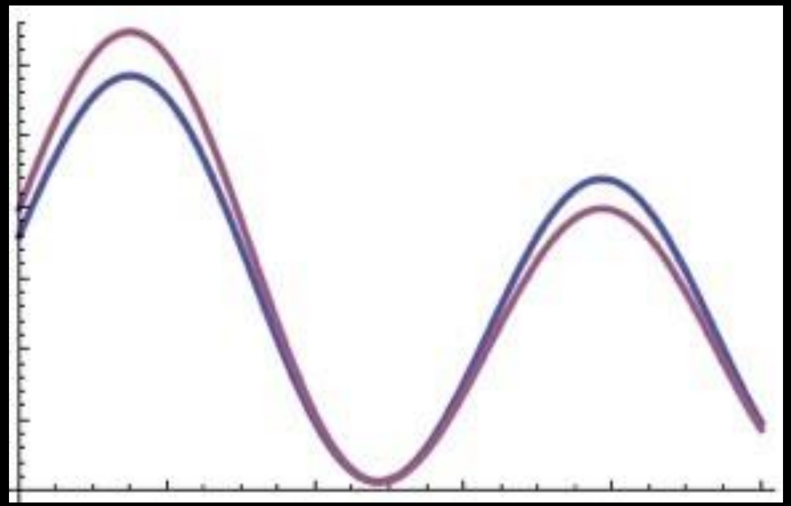
[DinurNissim03, DworkNissimMcSherrySmith06]

ε-Differential Privacy for mechanism M:

for any two neighboring data sets D_1, D_2,

any C **<** range(M),

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



differential privacy

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

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!

sensitivity of a function f

$$\Delta f = \max_{D1,D2} |f(D_1) - f(D_2)|$$

for neighboring data sets D_1, D_2

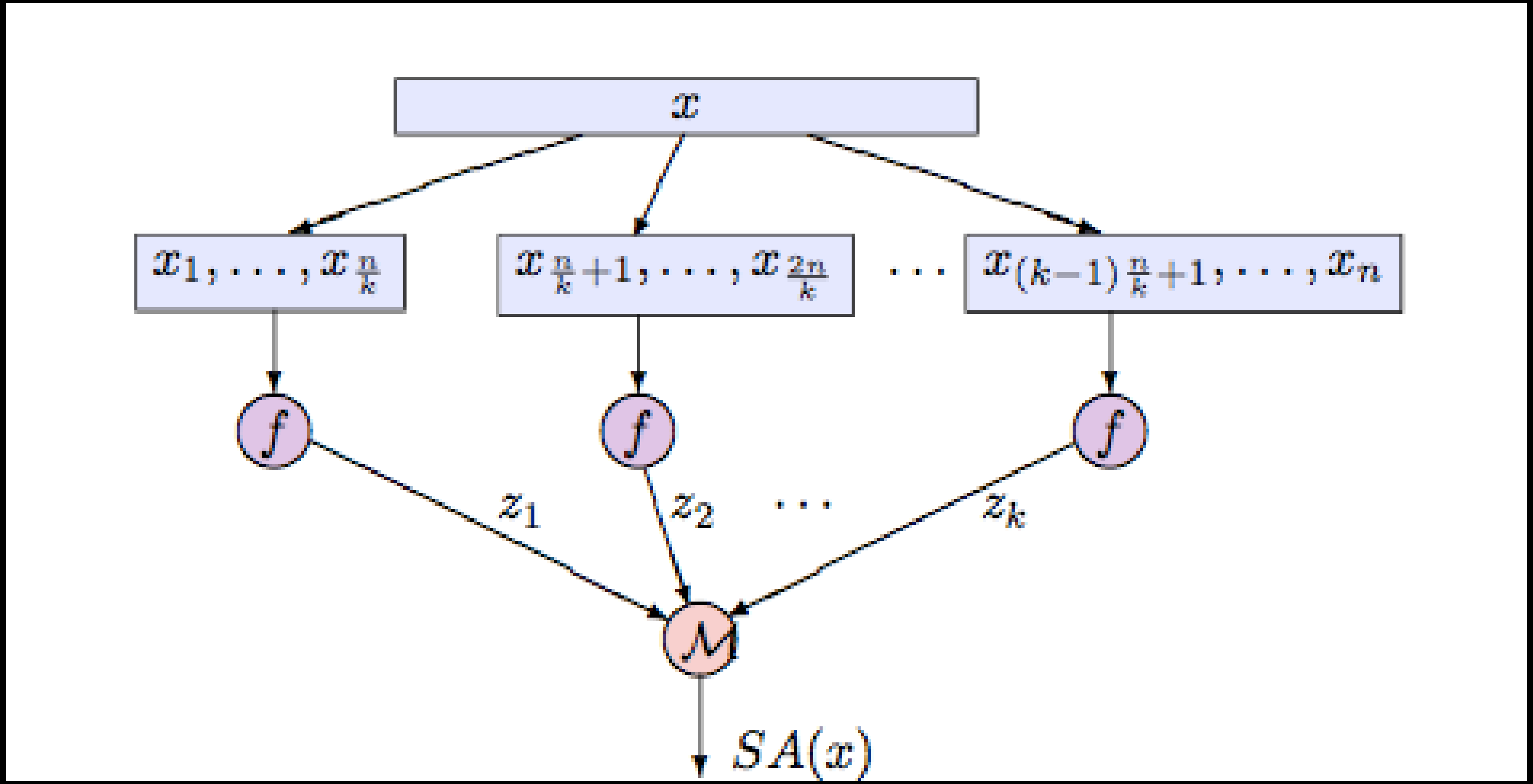
- measures how much one person can affect output
- sensitivity is 1 for counting queries that count number of rows satisfying a predicate

scale noise with sensitivity

$$\Delta f = \max_{D1,D2} |f(D_1) - f(D_2)|$$

on query f, can add scaled symmetric noise Lap(b) with b = $\Delta f/\varepsilon$, to achieve ε-differential privacy.

bootstrap for privacy = subsample and aggregate



x
$x_1, \ldots, x_{\frac{n}{k}}$
$x_{\frac{n}{k}+1}, \ldots, x_{\frac{2n}{k}}$
$\cdots$
$x_{(k-1)\frac{n}{k}+1}, \ldots, x_n$
f
f
f
z_1
z_2
$\cdots$
z_k
$\mathcal{M}$
$SA(x)$

summary

- privacy easy to get wrong; DP provides compelling definition and useful dose of paranoia
- powerful tools exist (some with no cost of privacy, and some with no noise!)
- powerful intuition from notions of robustness
- many nearly ready (and quite relevant) to common big data applications
- no ready-to-use, commercial- grade applications: need demand!

Fin.