Responsible Data Science

Anonymity and privacy (2)

April 14, 2025

Lucas Rosenblatt

Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

composition

Query sensitivity

The l_1 sensitivity of a query ***q***, denoted ***Δq***, is the maximum difference in the result of that query on a pair of neighboring databases

$$\Delta q = \max_{D,D'} \left| q(D) - q(D') \right|$$

query *q*	query sensitivity *Δq*
parallel composition	
select gender, count(*) from D group by gender	**1** (**disjoint groups**, presence or absence of one tuple impacts only one of the counts)
sequential composition	
an arbitrary list of *m* counting queries	***m*** (no assumptions about the queries, and so a single individual may change the answer of **every query** by 1)

r/ai

Proof?

Theorem 3.14. Let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1$ be an ε_1-differentially private algorithm, and let $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_2$ be an ε_2-differentially private algorithm. Then their combination, defined to be $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1 \times \mathcal{R}_2$ by the mapping: $\mathcal{M}_{1,2}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is $\varepsilon_1 + \varepsilon_2$-differentially private.

Proof. Let $x, y \in \mathbb{N}^{|\mathcal{X}|}$ be such that $\|x - y\|_1 \leq 1$. Fix any $(r_1, r_2) \in \mathcal{R}_1 \times \mathcal{R}_2$. Then:

$$\begin{aligned}
\frac{\Pr[\mathcal{M}_{1,2}(x) = (r_1, r_2)]}{\Pr[\mathcal{M}_{1,2}(y) = (r_1, r_2)]} &= \frac{\Pr[\mathcal{M}_1(x) = r_1] \Pr[\mathcal{M}_2(x) = r_2]}{\Pr[\mathcal{M}_1(y) = r_1] \Pr[\mathcal{M}_2(y) = r_2]} \\
&= \left(\frac{\Pr[\mathcal{M}_1(x) = r_1]}{\Pr[\mathcal{M}_1(y) = r_1]}\right) \left(\frac{\Pr[\mathcal{M}_2(x) = r_1]}{\Pr[\mathcal{M}_2(y) = r_1]}\right) \\
&\leq \exp(\varepsilon_1) \exp(\varepsilon_2) \\
&= \exp(\varepsilon_1 + \varepsilon_2)
\end{aligned}$$

By symmetry, $\frac{\Pr[\mathcal{M}_{1,2}(x)=(r_1,r_2)]}{\Pr[\mathcal{M}_{1,2}(y)=(r_1,r_2)]} \geq \exp(-(\varepsilon_1 + \varepsilon_2))$.

The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroth@cis.upenn.edu

https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Sequential composition

- Consider 4 queries executed in sequence
  - Q1: select count(*) from D under $\varepsilon_1 = 0.5$
  - Q2: select count(*) from D where sex = Male under $\varepsilon_2 = 0.2$
  - Q3: select count(*) from D where sex = Female under $\varepsilon_3 = 0.25$
  - Q4: select count(*) from D where age > 20 under $\varepsilon_4 = 0.25$

- $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = 1.2$ That is: all queries together are ε-differentially private for ε =1.2. **Can we make a stronger guarantee?**
- This works because **Laplace noise is additive**

More generally: set a **cumulative privacy budget**, and split it between all queries, pre-processing, other data manipulation steps of the pipeline

r/ai

Parallel composition

- If the inputs are disjoint, then the result is ε-differentially private for ε =max(***ε1, …, εk***)
  - Q1: select count(*) from D under ***ε1*** **= 0.5**
  - Q2: select count(*) from D where sex = Male under ***ε2*** **= 0.2**
  - Q3: select count(*) from D where sex = Female under ***ε3*** **= 0.25**
  - Q4: select count(*) from D where age > 20 under ***ε4*** **= 0.25**
- ***ε = ε1 + max(ε2, ε3)+ ε4*** **=1** That is: all queries together are ε-differentially private for ε =1.

Composition and consistency

- Consider again 4 queries executed in sequence
  - Q1: select count(*) from D under ε_1 = 0.5 returns **2005**
  - Q2: select count(*) from D where sex = Male under ε_2 = 0.2 returns **1001**
  - Q3: select count(*) from D where sex = Female under ε_3 = 0.25 returns **995**
  - Q4: select count(*) from D where age > 20 under ε_4 = 0.25 returns **1789**

Assuming that there are 2 genders in D, Male and Female, there is **no database consistent with these statistics**!

Also don't want any negative counts + may want to impose datatype checks, e.g., no working adults with age = 5 etc.

r/ai

DP synthetic data generation

r/ai

DP synthetic data

Lots of advantages

- Consistency is not an issue
- Analysts can treat synthetic data as a regular dataset, run existing tools
- No need to worry about the privacy budget
- Can answer as many queries as they want, and any kind of a query they want, including record-level queries

What's the catch?

Recall the Fundamental Law of Information Recovery. It tells us that we cannot answer all these queries accurately and still preserve privacy!

Therefore, when releasing synthetic data, we need to document it with which queries it supports well

r/ai

PrivBayes from scratch

r/ai

BayesNets: How do they work? How can we privatize?

Lucas Rosenblatt

For DataSynthesizer, the underlying construction is a Bayesian network with a differentially private guarantee i.e. PrivBayes

So, we're going to do a quick review of Bayesian networks.[1]

[1] Thanks to by Prof. Yi Mei, whose notes were helpful in making some of these slides https://github.com/meiyi1986/tutorials/

Why a Bayesian Network?

A Bayesian network let's us represent uncertainty about our data by parameterizing the probabilistic dependencies between variables.

What are benefits?

Handles missing data gracefully

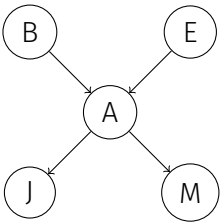
Learns "causal" relationships, so can be used to estimate consequences of intervention

Can easily incorporate ***causal*** *prior knowledge about data*

Can help avoid overfitting

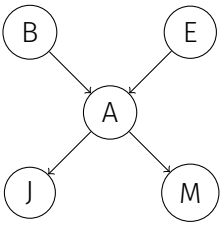
So, what actually is a BayesNet?

Here's a (relatively famous) one!



The edges represent *conditionality*...but how do we actually represent that?

We can write a *factorization* of the *joint probability distribution* of this DAG as follows:



$$P(B, E, A, J, M) = P(B) \times P(E) \times P(A|B, E) \times P(J|A) \times P(M|A).$$

And in general, from the product rule of probabilities over a joint distribution, we have,

$$P(X_1, \ldots, X_n) = P(X_1) \times P(X_2|X_1) \times \ldots \times P(X_n|X_1, \ldots, X_{n-1}) \tag{1}$$

which, for a Bayesian network, is just the transition probabilities of parents to children,

$$P(X_1, \ldots, X_n) = P(X_1|\text{parents}(X_1)) \times \ldots \times P(X_n|\text{parents}(X_n)) \tag{2}$$

because nodes in the DAG without directed pathways to other nodes are *conditionally independent*.

So, what are we *actually* doing when we are "constructing a Bayesian Network?"

Oversimplification, but we are:

1. Estimating *structure* i.e. *variable conditioning*, then
2. Estimating *free parameters* i.e. *probabilities for the conditionals from (1)*

1. Estimating *structure*

Challenging problem because the number of possible structures grows *super-exponentially* with the number of variables...

Common approach: let's be greedy!

Before giving you the structure algorithm, we need a tool that will help us with edge selection: Mutual Information (MI, or often formally $\mathbb{I}(X;Y)$).

MI is a fundamental concept - amount of information obtained about one random variable through observing another random variable

First[2], need standard idea of "*information* content" of an outcome x from a random variable X (denoted $\mathbb{I}(x)$):

$$\mathbb{I}(x) = -\log(p(x)) \tag{3}$$

Measure is *literally* in bits if log base 2 - quantifies how "surprising" or "informative" the outcome x is in our distribution

E.g. less probable an outcome, the more informative it is considered to be...

[2] All of this is Claude Shannon, who wrote a short paper "A Mathematical Theory of Communication" that fully invented the field of information theory. So cool!

How did Shannon come up with this? Wanted a functional idea of information that had properties...

(1. Continuous) Was continuous function of the probability of the event, p. Why? So only small changes in the probability produced small changes in the measure, or there are no abrupt jumps in "surprise."

(2. Monotonicity) If $p_1 < p_2$, then we must have $\mathbb{I}(p_1) > \mathbb{I}(p_2)$. This seems natural - information in rare events should be more informative than common events.

(3. Additivity for Independent Events) Recall that independent events in probability give you "no information about one another." Formally, we then want, for two independent events with probabilities p q; if the combined event has a probability $p \cdot q$, then the information content should satisfy

$$\mathbb{I}(p \cdot q) = \mathbb{I}(p) + \mathbb{I}(q). \tag{4}$$

$$\mathbb{I}(p \cdot q) = \mathbb{I}(p) + \mathbb{I}(q). \tag{5}$$

Think about it for a second: ...

...yup! The only family of functions that satisfies "$\mathbb{I}(p \cdot q) = \mathbb{I}(p) + \mathbb{I}(q)$" for $p, q \in [0, 1]$ is $\mathbb{I}(x) = -k \log(x)$.

We define *entropy* $H(X)$ as the expected information content of its outcomes:[3]

$$H(X) = -\sum_{x \in \Omega(X)} p(x) \log(p(x)) \tag{6}$$

"Average amount of information (or uncertainty) produced by a random variable"

We can then write *conditional* entropy, $H(Y|X)$, quantifying the amount of information (or uncertainty) that remains about Y after observing X e.g.

$$H(Y|X) = -\sum_{x \in \Omega(X)} \sum_{y \in \Omega(Y)} p(x, y) \log(p(y|x)) \tag{7}$$

[3] note here that $\Omega(X)$: the domain (set of possible values) of X

Finally, we can define Mutual Information!

MI ($\mathbb{I}(X;Y)$) is often viewed in terms of entropy, and is super natural:

$$\mathbb{I}(X;Y) = H(Y) - H(Y|X) \tag{8}$$

Formulating it this way gives mutual information as the *reduction* in *uncertainty* (entropy) about Y due to the knowledge of X.

In other words, its the reduction in entropy of Y when X is known compared to when X is not known. Super intuitive!

If we go ahead and substitute in the formulas and simplify, we get:

$$\mathbb{I}(X;Y) = \sum_{x \in \Omega(X)} \sum_{y \in \Omega(Y)} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{9}$$

1. Estimating *structure* Ok, back to our greedy algorithm. Maybe it'll seem simple...

- Initialize a Bayesian network with no edges.
- Calculate mutual information $\mathbb{I}(X_i; X_j)$ for every pair of nodes X_i and X_j.
- Sort these pairs in decreasing order of mutual information.
- For each pair (X_i, X_j) in the sorted list:
  - If adding the edge $X_i \to X_j$ does not introduce a cycle,
  - And if adding the edge increases the overall score of the network according to a scoring criteria (like likelihood - don't worry about this.),
  - Then add the edge $X_i \to X_j$ to the network.
- Repeat until no more edges can be added without violating a-cyclicity

2. Estimating free parameters i.e. probabilities Once we have a structure, all we need to do is estimate the conditional probabilities *based on our data*

A few ways to do this, but very commonly people use **MLE**.

For node X_i, parents $\text{Parents}(X_i)$, $P(X_i|\text{Parents}(X_i))$ can be estimated,

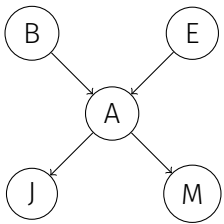
$$P(X_i = x|\text{Parents}(X_i) = \mathbf{pa}) \approx \frac{N(X_i = x, \text{Parents}(X_i) = \mathbf{pa})}{N(\text{Parents}(X_i) = \mathbf{pa})} \tag{10}$$

...where $N(X_i = x, \text{Parents}(X_i) = \mathbf{pa})$ is number of instances in the dataset where $X_i = x$ while parents are in configuration **pa**...

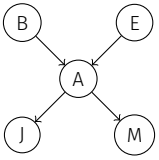
...and $N(\text{Parents}(X_i) = \mathbf{pa})$ is the number of instances where the parents are **pa** regardless of value of X_i.

I think its super clear with an example.

Remember the initial example structure I gave?



Let's update edge for $P(A|B,E)$ with MLE! (assume A, B, E all binary for simplicity).



Easy: we count the occurrences of A for each combination of B and E in our dataset (for b, e in $\{0, 1\}$):

$$P(A = 1|B = b, E = e) = \frac{N(A = 1, B = b, E = e)}{N(B = b, E = e)} \tag{11}$$

$$P(A = 0|B = b, E = e) = \frac{N(A = 0, B = b, E = e)}{N(B = b, E = e)} \tag{12}$$

Now we have a *data driven* estimate of $P(A|B, E)$. We repeat for all other edges, and then we have our joint *conditional* distribution!

Making BayesNets Private

In order to make BayesNets private (and thus create something like PrivBayes) we need to:

1. Think about the *sensitivity* of each step above that touches data.

2. Think about the type of *mechanism* we can use to ensure differential privacy at each step.

Question: can you think of the steps / mechanisms that need privatizing?

*Caveat: what follows is not *exactly* what PrivBayes does, they are more clever...but its close, and offers good intuition...*

1. Estimating *structure*

Non-private algorithm: selects edges to include in the network based on exact mutual information.

DP algorithm: selects edges to include in the network based on *a probability proportional* to the *exponential* with mutual information as the scoring function.

For candidate edge e between nodes X_i and X_j in the dataset D.

We set the *selection probability* for e in structure estimation as:

$$P(e|D) \propto \exp\left(\frac{\epsilon}{2\Delta\mathbb{I}} \times \mathbb{I}(X_i; X_j)\right) \tag{13}$$

where ϵ is the privacy budget and $\Delta\mathbb{I}$ is the sensitivity of the mutual information[4]

[4] The maximum amount by which $\mathbb{I}(X_i; X_j)$ can change with the addition or removal of a data point

2. Estimating free parameters

Bunch of ways to noise the probabilities - simple baseline based on counts...

Non-private algorithm: use simple *MLE* calculation based on data to estimate the probability of observing any given $X_i = x_i$, given values of parents of X_i.

DP algorithm: do the same calculation, but add an additive noise mechanism somewhere in the mix...

For a node X_i with parents Parents(X_i), compute noisy conditional probability $\tilde{P}(X_i|\text{Parents}(X_i))$,

$$\tilde{P}(X_i = x|\text{Parents}(X_i) = \mathbf{pa}) = \frac{N(X_i = x, \text{Parents}(X_i) = \mathbf{pa}) + \text{Lap}(\lambda)}{N(\text{Parents}(X_i) = \mathbf{pa}) + \sum_{x'} \text{Lap}(\lambda)} \tag{14}$$

...where Lap(λ) is a random variable drawn from the Laplace distribution with scale parameter $\lambda = \frac{\Delta f}{\epsilon}$

...and where Δf is the global sensitivity of the count function (i.e. 1!).

We made BayesNets private!

See PrivBayes paper for better way...

Quick note on KL-Divergence Used as a metric in the lab - a.k.a. "relative entropy."

Distance measure between two probability distributions over the *same variable*

Two discrete probability distributions P and Q, the KL divergence is:

$$D_{KL}(P\|Q) = \sum_{x\in\Omega} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{15}$$

"The amount of information lost when Q is used to approximate P."

Note: mutual information can be expressed as the KL divergence between the joint distribution $p(x, y)$ and the product of the marginal distributions $p(x)p(y)$:

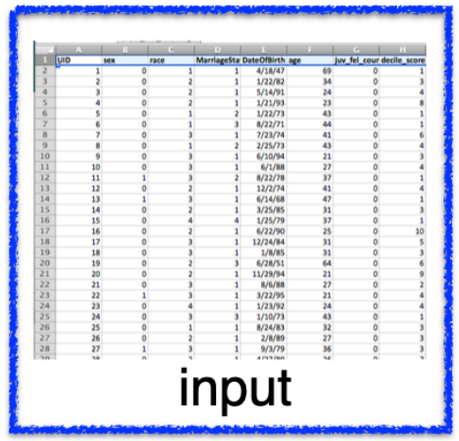
$$\mathbb{I}(X; Y) = D_{KL}(p(x, y)\|p(x)p(y)) \tag{16}$$

MI can thus be viewed as a *specific application* of KL divergence.

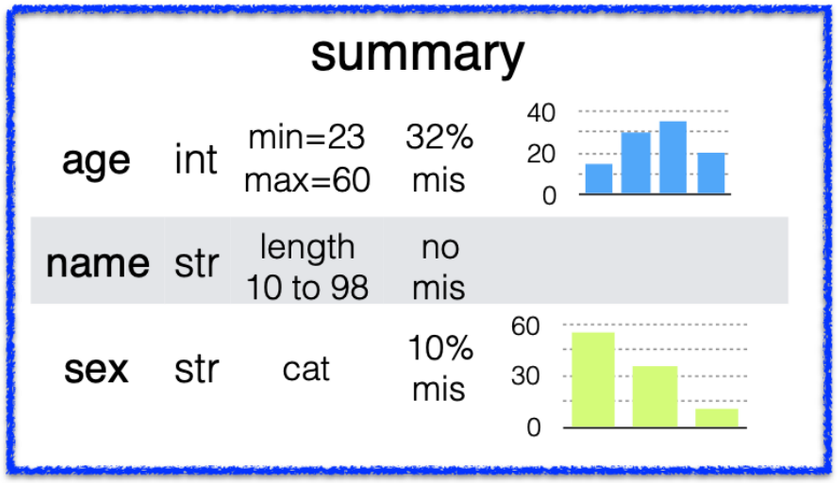
DataSynthesizer

r/ai

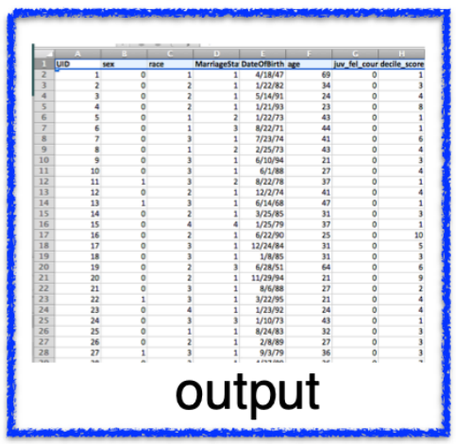
Data Synthesizer


input

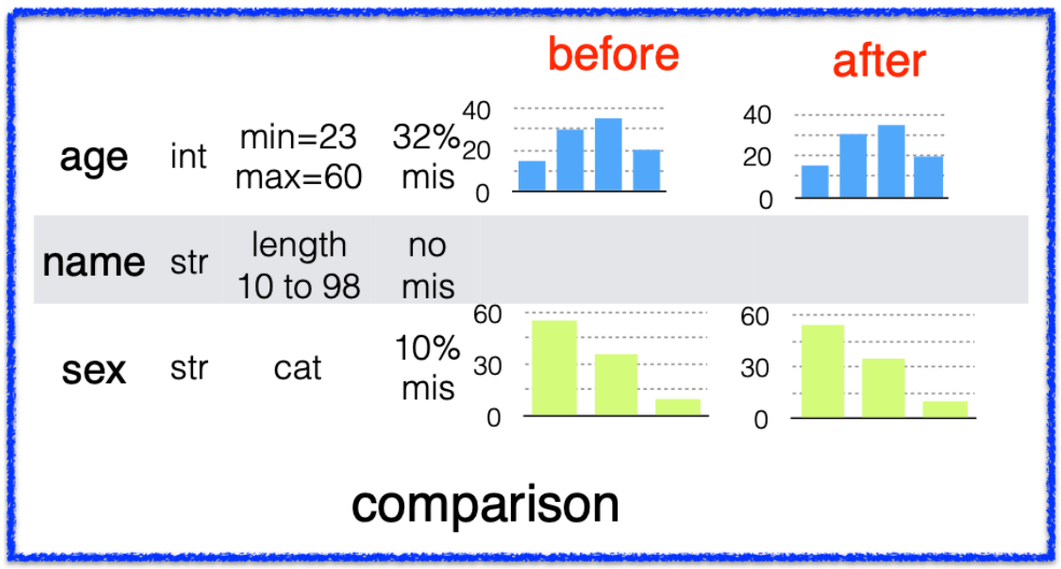
Data Describer



Data Generator


output

Model Inspector



[Ping, Stoyanovich, Howe *2017*]

http://demo.dataresponsibly.com/synthesizer/

Data Synthesizer

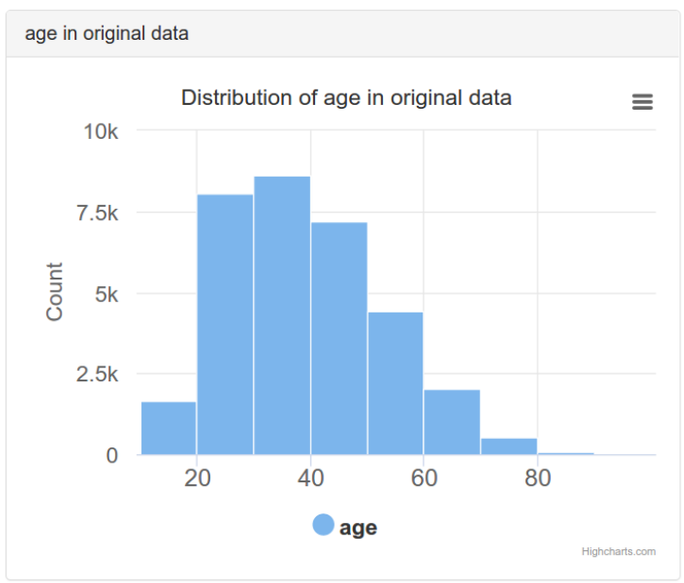
- The tool generates an output dataset of a specified size, in one of three modes
  - **random** - type-consistent random output
  - **independent attribute** - learn a noisy histogram for each attribute
  - **correlated attribute** - learn a noisy Bayesian network (BN)

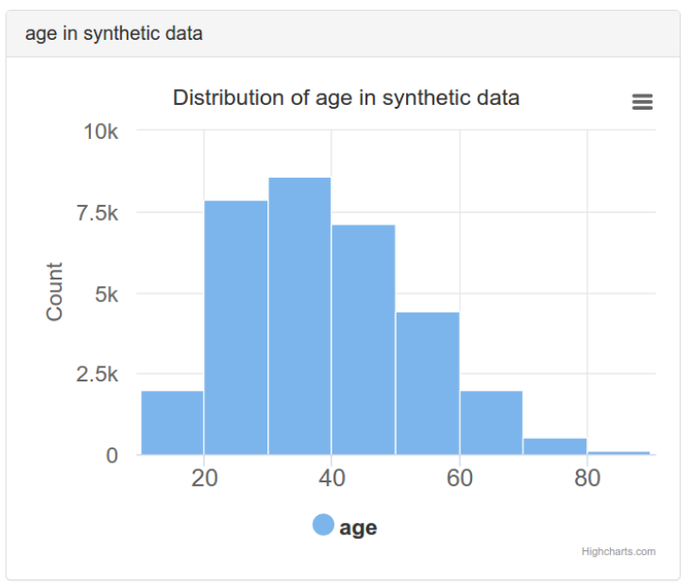
[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Data Synthesizer: Independent attributes

Given the over-all privacy budget **ε**, and an input dataset of size ***n***. Allocate ***ε/d*** of the budget to each attribute A_i in $\{A_1, .., A_d\}$. Then for each attribute:

- Compute the i^{th} histogram with ***t*** bins (***t***=20 by default), with query q_i
- The sensitivity Δq_i of this (or any other) histogram query is **2/n** **Why?**
- So, each bin's noisy probability is computed by adding $Lap\left(\frac{2d}{\varepsilon n}\right)$



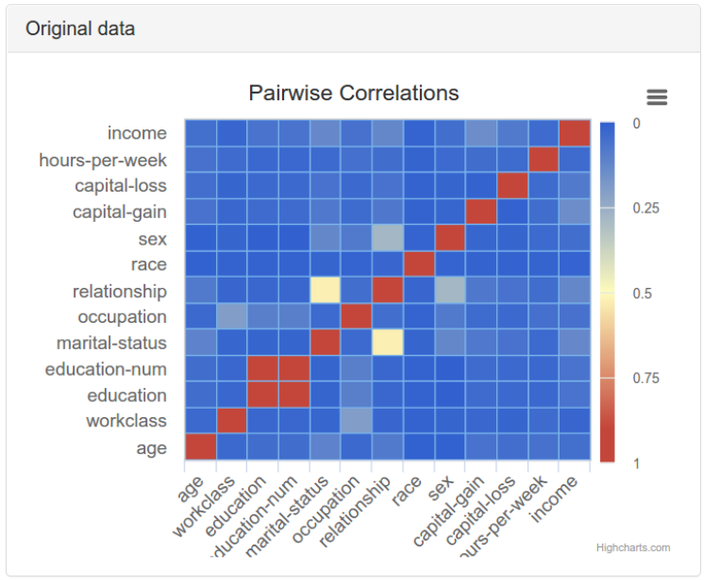


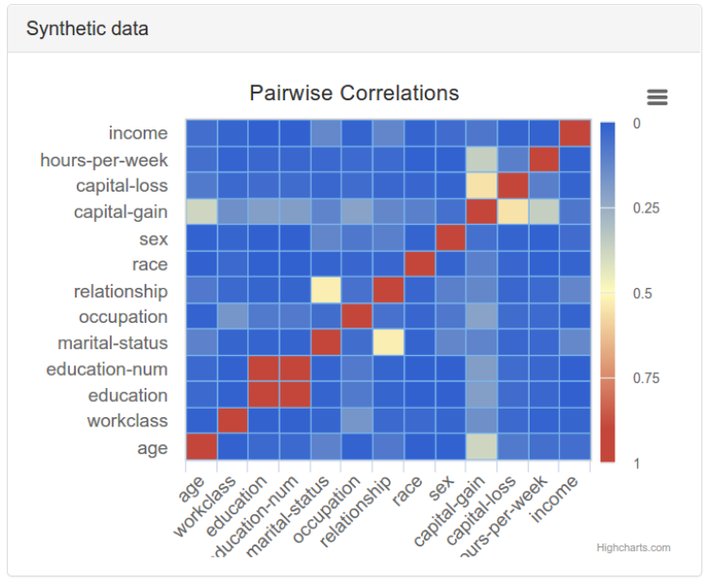
[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

r/ai

Data Synthesizer: Correlated attributes

- Learn a differentially private Bayesian network (BN)
- Use the method called **PrivBayes** [Zhang, Cormode, Procopiuc, Srivastava, Xiao, *2016*]
- Privacy budget is split equally between (a) network structure computation and (b) populating the conditional probability tables of each BN node
- User inputs privacy budget **ε** and the maximum number of parents for a BN node ***k*** - you'll play with these settings as part of HW2
- The tool treats a missing attribute value as one of the values in the attribute's domain (not shown in the examples in the next two slides)

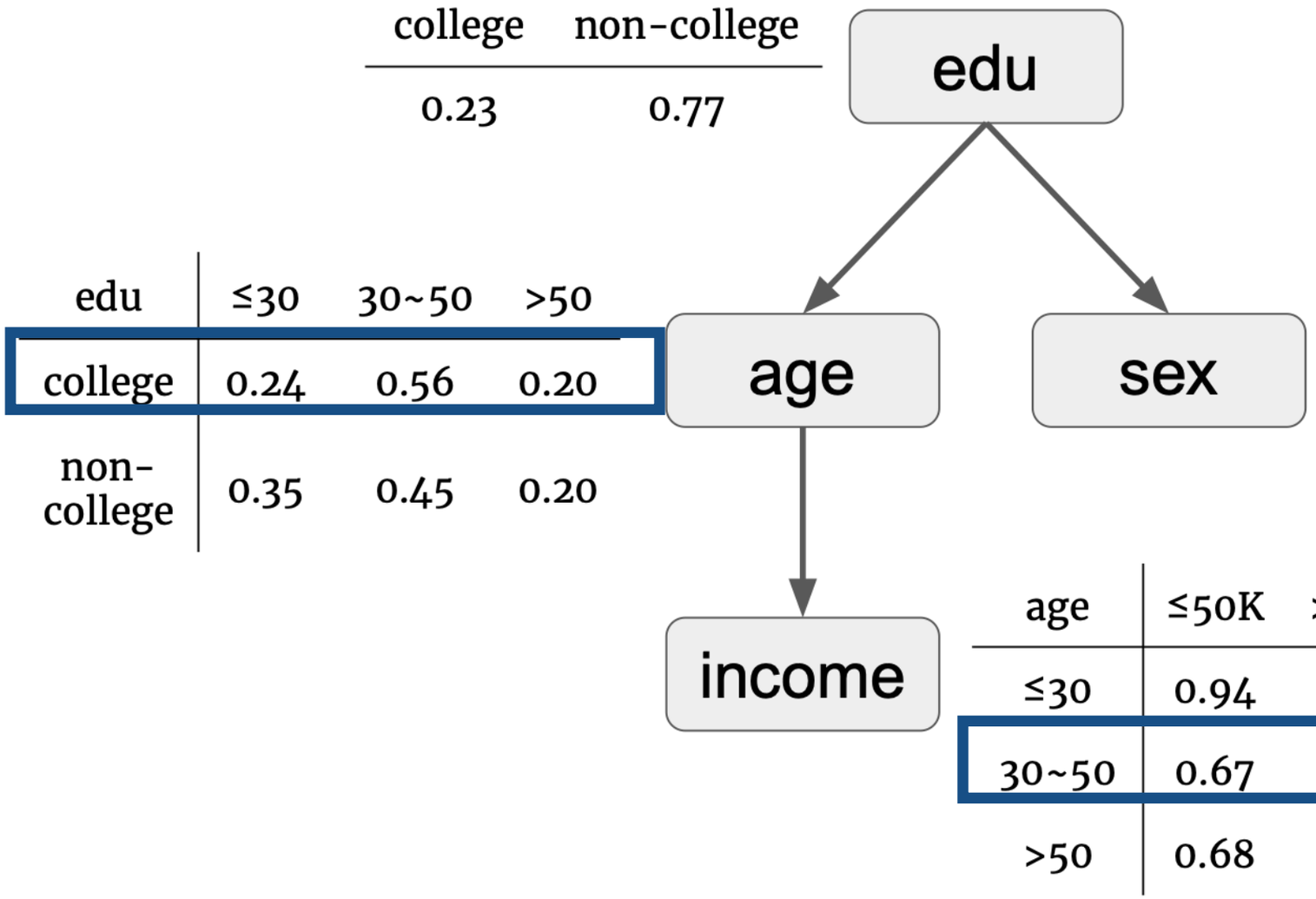




[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Data Synthesizer: Correlated attributes

K=1 not a causal DAG, a regular Bayesian network!



college	non-college
0.23	0.77

edu	≤30	30~50	>50
college	0.24	0.56	0.20
non-college	0.35	0.45	0.20

edu	female	male
college	0.30	0.70
non-college	0.34	0.66

age	≤50K	>50K
≤30	0.94	0.07
30~50	0.67	0.33
>50	0.68	0.32

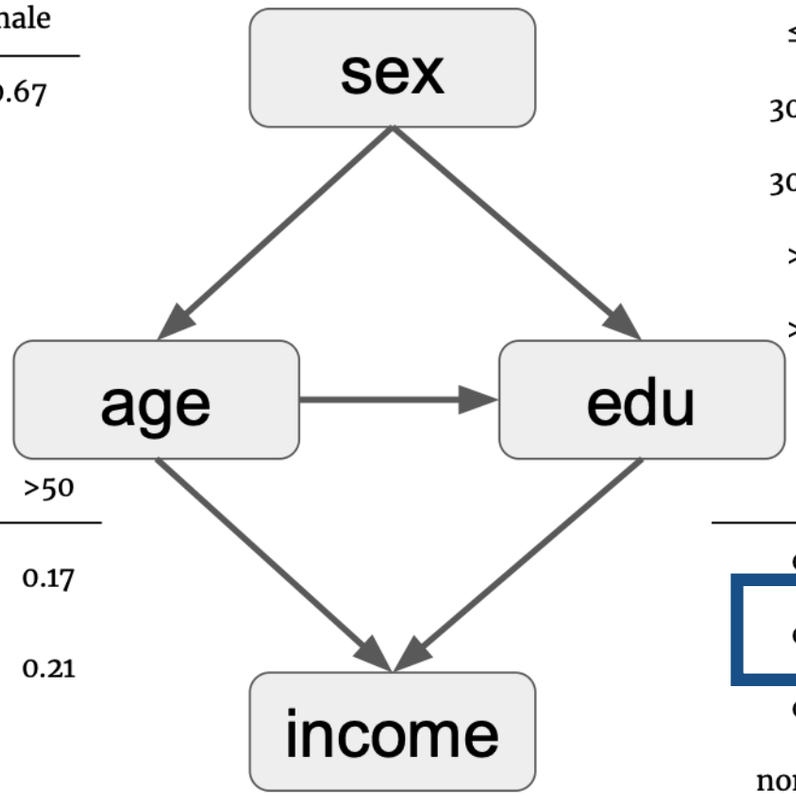
[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Data Synthesizer: Correlated attributes

not a causal DAG, a regular Bayesian network!

K=2

female	male
0.33	0.67



age	sex	college	non-college
≤30	female	0.18	0.82
≤30	male	0.16	0.84
30~50	female	0.25	0.75
30~50	male	0.28	0.72
>50	female	0.17	0.83
>50	male	0.25	0.75

sex	≤30	30~50	>50
female	0.40	0.43	0.17
male	0.29	0.59	0.21

edu	age	≤50K	>50K
college	≤30	0.83	0.17
college	30~50	0.45	0.55
college	>50	0.41	0.59
non-college	≤30	0.96	0.04
non-college	30~50	0.76	0.24
non-college	>50	0.75	0.25

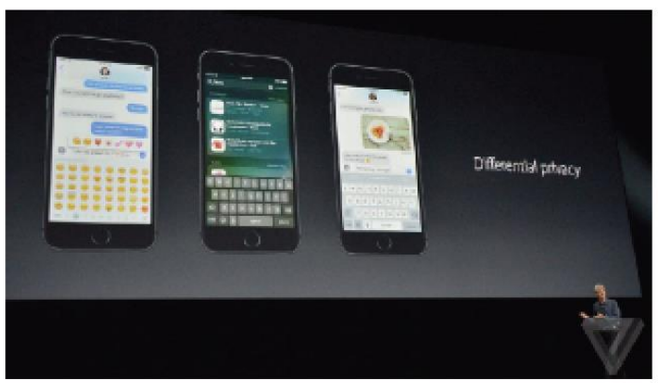
[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

DP in the field

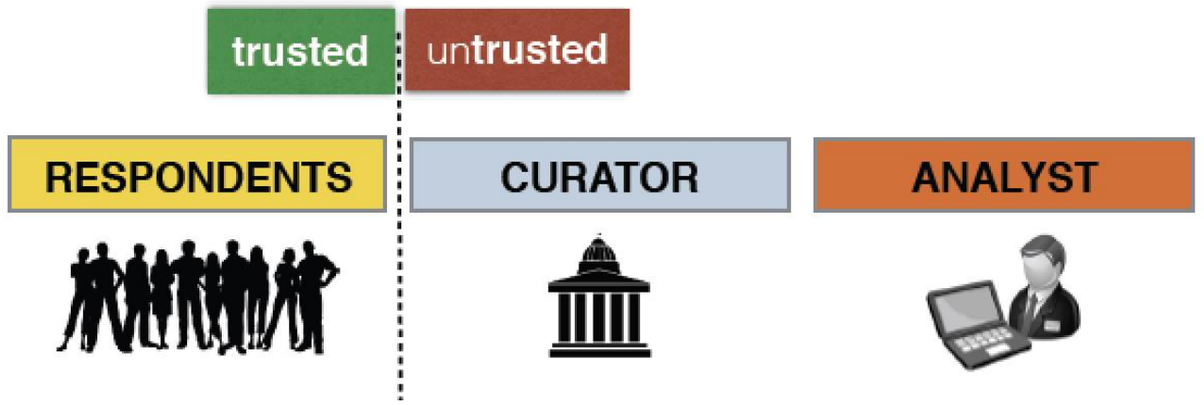
r/ai

Differential privacy in the field

Apple



Google



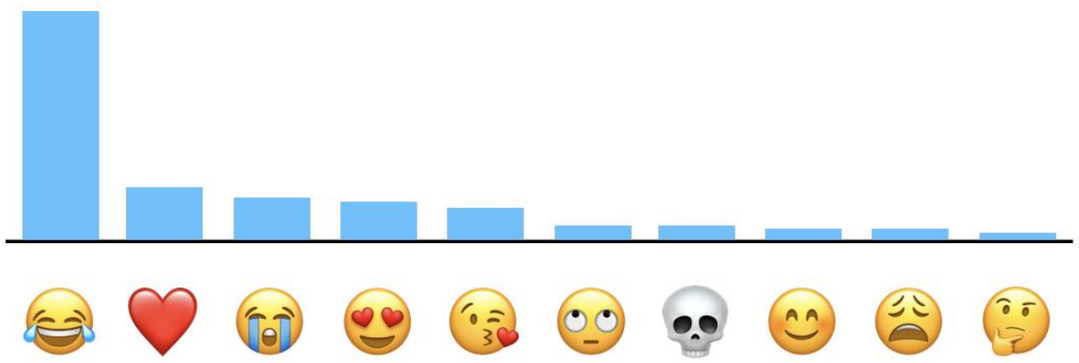
slide by Gerome Miklau

r/ai

Example: What's your favorite emoji?

A privacy-preserving system

Apple has adopted and further developed a technique known in the academic world as *local differential privacy* to do something really exciting: gain insight into what many Apple users are doing, while helping to preserve the privacy of individual users. It is a technique that enables Apple to learn about the user community without learning about individuals in the community. Differential privacy transforms the information shared with Apple before it ever leaves the user's device such that Apple can never reproduce the true data.

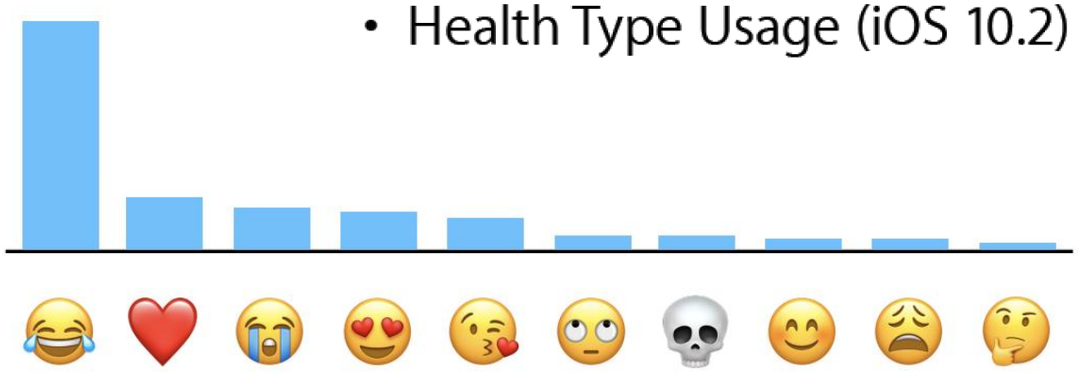


https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Example: What's your favorite emoji?

Apple uses local differential privacy to help protect the privacy of user activity in a given time period, while still gaining insight that improves the intelligence and usability of such features as:

- QuickType suggestions
- Emoji suggestions
- Lookup Hints
- Safari Energy Draining Domains
- Safari Autoplay Intent Detection (macOS High Sierra)
- Safari Crashing Domains (iOS 11)
- Health Type Usage (iOS 10.2)



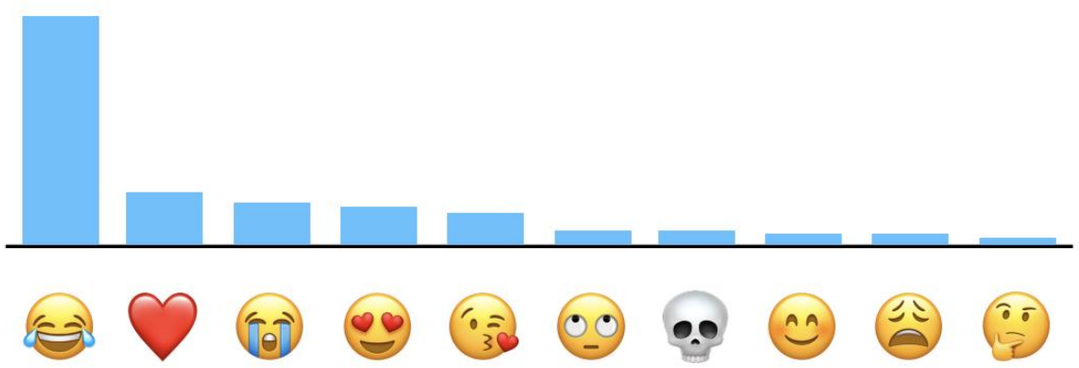
https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

r/ai

Example: What's your favorite emoji?

Privacy budget

The Apple differential privacy implementation incorporates the concept of a per-donation *privacy budget* (quantified by the parameter epsilon), and sets a strict limit on the number of contributions from a user in order to preserve their privacy. The reason is that the slightly-biased noise used in differential privacy tends to average out over a large numbers of contributions, making it theoretically possible to determine information about a user's activity over a large number of observations from a single user (though it's important to note that Apple doesn't associate any identifiers with information collected using differential privacy).



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

r/ai

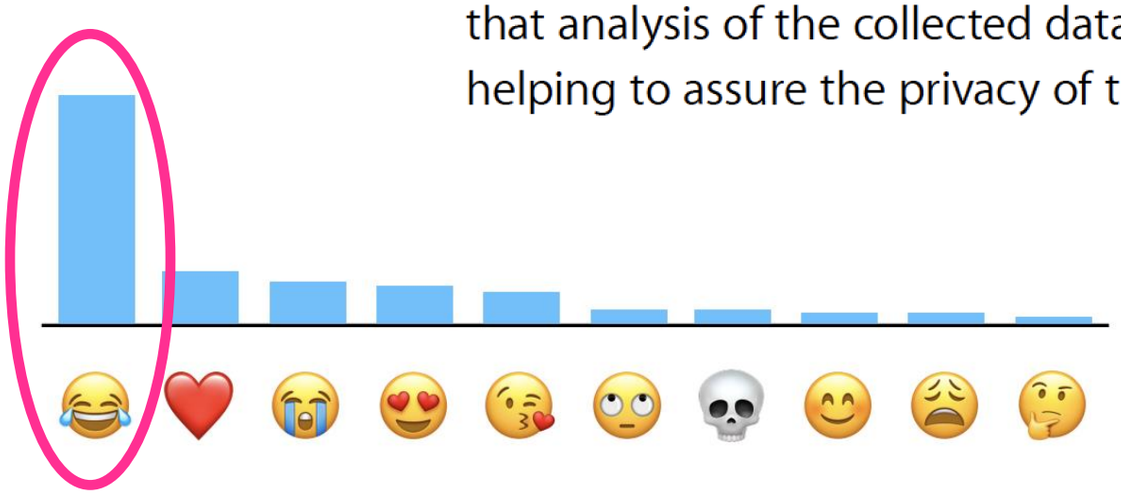
Example: What's your favorite emoji?

Count Mean Sketch

In our use of the Count Mean Sketch technique for differential privacy, the original information being processed for sharing with Apple is encoded using a series of mathematical functions known as *hash functions*, making it easy to represent data of varying sizes in a matrix of fixed size.

The data is encoded using variations of a SHA-256 hash followed by a privatization step and then written into the sketch matrix with its values initialized to zero.

The noise injection step works as follows: After encoding the input as a vector using a hash function, each coordinate of the vector is then flipped (written as an incorrect value) with a probability of $1/(1 + e^{\varepsilon/2})$, where ε is the privacy parameter. This assures that analysis of the collected data cannot distinguish actual values from flipped values, helping to assure the privacy of the shared information.



https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

r/ai

Transparency is important!

ANDY GREENBERG SECURITY 09.15.2017 09:28 AM

How One of Apple's Key Privacy Safeguards Falls Short

Apple has boasted of its use of a cutting-edge data science known as "differential privacy." Researchers say they're doing it wrong.

WIRED

Epsilon, Epsilon

"…[Researchers] examined how Apple's software injects random noise into personal information—ranging from emoji usage to your browsing history to HealthKit data to search queries—before your iPhone or MacBook upload that data to Apple's servers.

Ideally, that obfuscation helps protect your private data from any hacker or government agency that accesses Apple's databases, advertisers Apple might someday sell it to, or even Apple's own staff. But **differential privacy's effectiveness depends on a variable known as the "privacy loss parameter," or "epsilon,"** which determines just how much specificity a data collector is willing to sacrifice for the sake of protecting its users' secrets. By taking apart Apple's software to determine the epsilon the company chose, the researchers found that **MacOS uploads significantly more specific data than the typical differential privacy researcher might consider private.** iOS 10 uploads even more. And perhaps most troubling, according to the study's authors, is that **Apple keeps both its code and epsilon values secret**, allowing the company to potentially change those critical variables and erode their privacy protections with little oversight…."

https://www.wired.com/story/apple-differential-privacy-shortcomings/

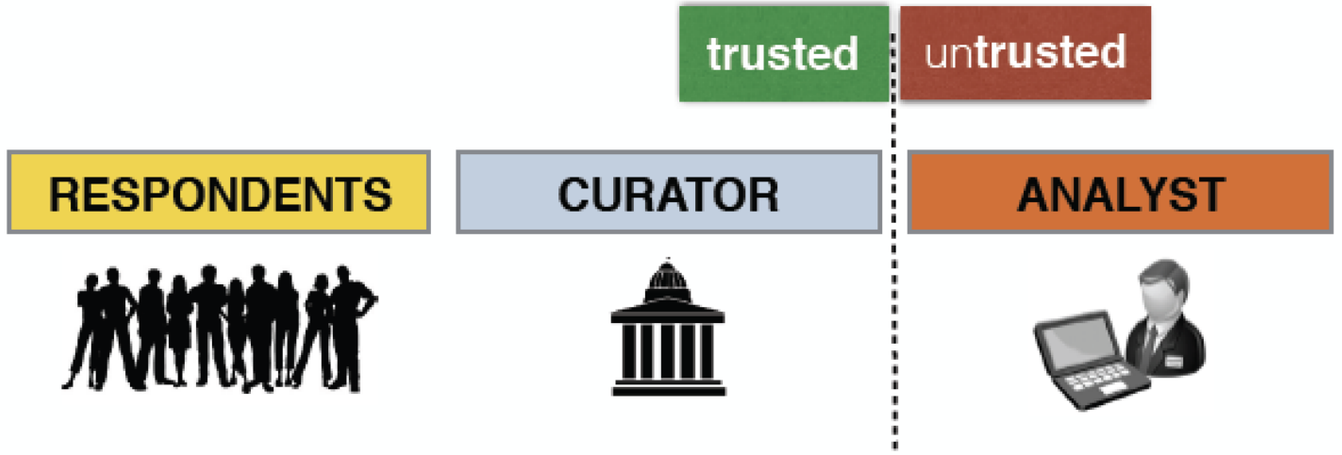
r/ai

DP & the US Census

r/ai

Differential privacy in the field

United States™ Census Bureau



Decennial Census 2020

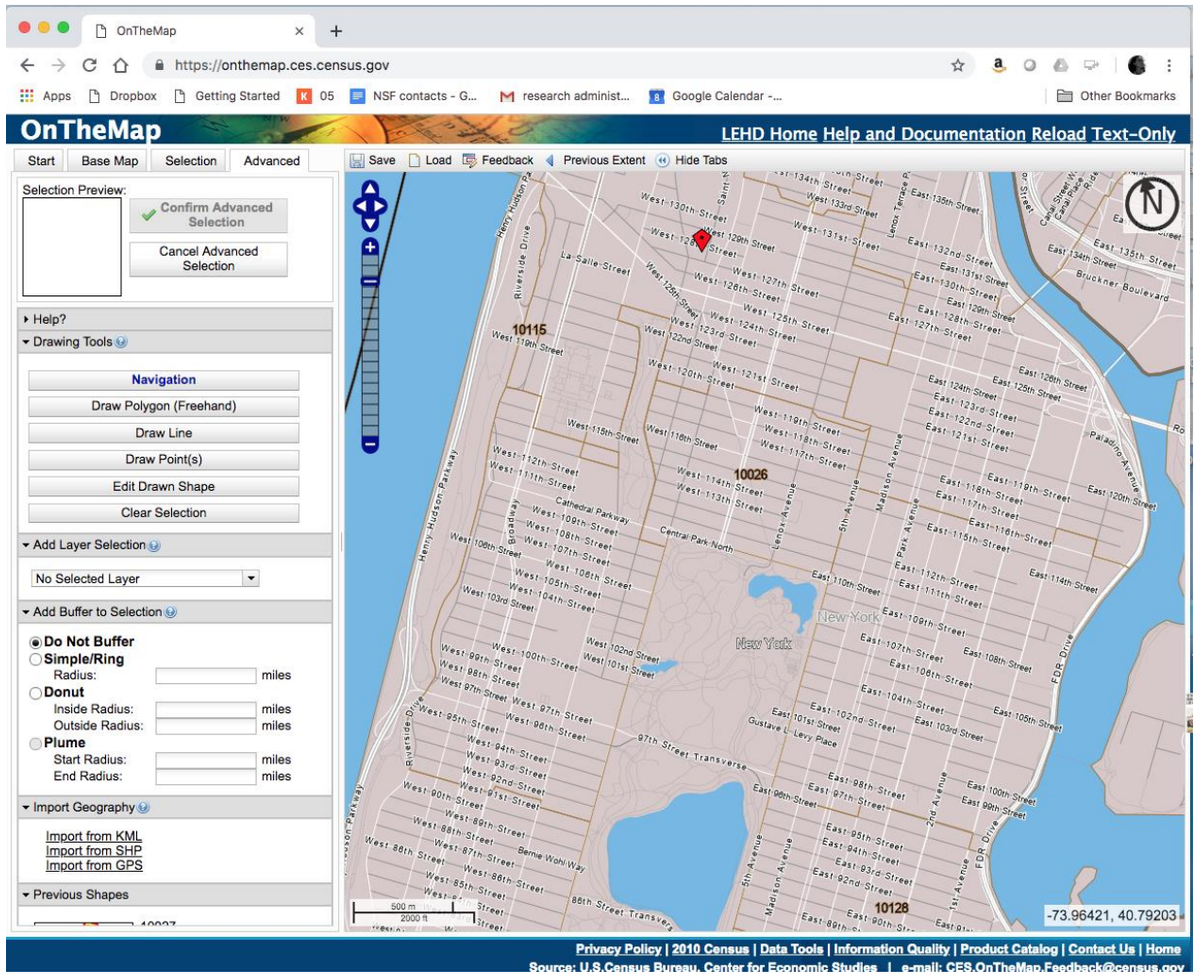
slide by Gerome Miklau

r/ai

Differential privacy in the field

First adoption by the US Census Bureau:

OnTheMap (2008), synthetic data about where people in the US live and work



Differential privacy in the field

TheUpshot

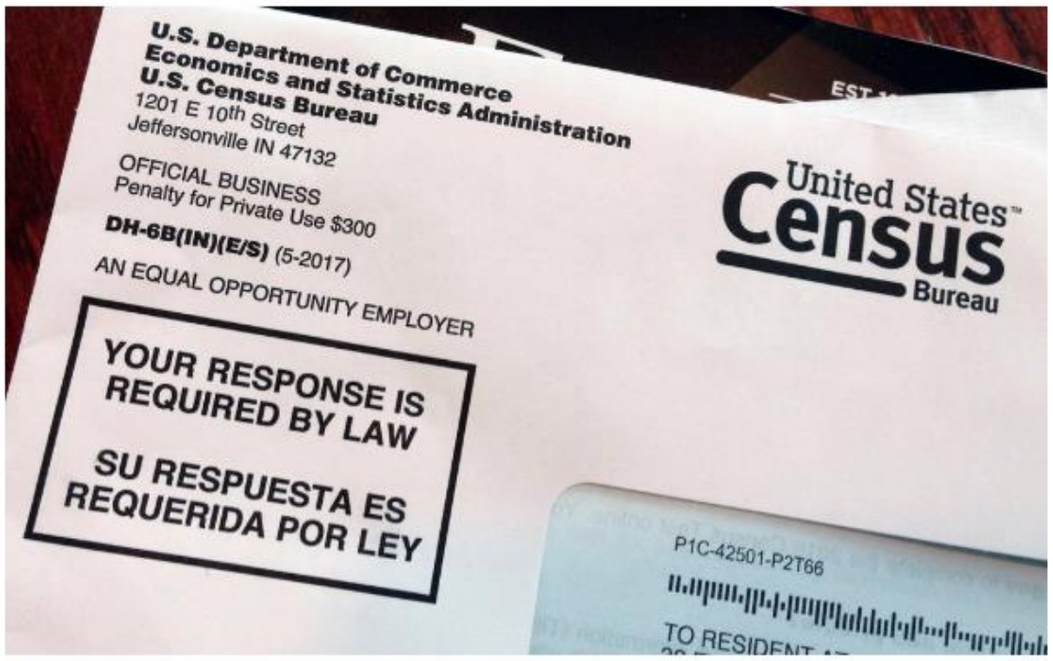
To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.

The New York Times

By Mark Hansen

Dec. 5, 2018



A 2018 census test letter mailed to a resident in Providence, R.I. The nation's test run of the 2020 Census is in Rhode Island. Michelle R. Smith/Associated Press

At the root of the problem are the tables of aggregate statistics that the bureau publishes. There are hundreds of tables — sex by age, say, or ethnicity by race — summarizing the population at several levels of geography, from areas the size of a city block all the way up to the level of a state or the nation. In 2010, the bureau released tables with nearly eight billion numbers in all. That was about 25 numbers for each person living in the United States, even though Americans were asked only 10 questions about themselves. In other words, the tables were generated in so many ways that the Census Bureau ended up releasing more data in aggregate then it had collected in the first place.

r/ai

Reconstruction attack: an example

TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

		AGE		
STATISTIC	**GROUP**	**COUNT**	**MEDIAN**	**MEAN**
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American male	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

Note: Married persons must be 15 or over

[Garfinkel, Abowd and Martindale, *ACM Queue 2018*]

r/ai

Reconstruction attack: an example

Let's assume that the oldest person is 125 years old, and that everyone's age is different. How many possible age combinations are there?

$$\binom{125}{3} = 317{,}750$$

But only 40 combinations have median = 30 and mean = 44

Idea: extract all such constraints, represent them as a mathematical model, have an automated solver find a solution.

TABLE 2: **POSSIBLE AGES FOR A MEDIAN OF 30 AND MEAN OF 44**

A	B	C	A	B	C	A	B	C
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

[Garfinkel, Abowd and Martindale, *ACM Queue 2018*]

r/ai

What does the law say?

Title 13 of U.S. Code authorizes data collection and publication of statistics by the Census Bureau.

Section 9 of Title 13 requires privacy protections: “Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, ... may ... make **any publication whereby the data furnished by any particular establishment or individual under this title can be identified**” (Title 13 U.S.C. § 9(a)(2), Public Law 87-813).

In 2002, Congress further clarified the concept of identifiable data: it is prohibited to publish “**any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means**” (Pub. L. 107–347, Title V, §502 (4), Dec. 17, 2002, 116 Stat. 2969).

Section 214 of Title 13 outlines penalties: fines up to $5,000 or imprisonment up to 5 years or both per incident (data item), up to $250,000 in total.

r/ai

DP in the 2020 Census: pushback

UNIVERSITY OF MINNESOTA

Implications of Differential Privacy for Census Bureau Data and Research

Task Force on Differential Privacy for Census Data †
Institute for Social Research and Data Innovation (ISRDI)
University of Minnesota

November 2018
Version 2
Working Paper No. 2018-6

- noisy data - **impact on critical decisions**
- difficult to explain differential privacy / privacy budget to the public - **how do we set epsilon**?
- disagreement about whether using differential privacy is legally required
- messaging is difficult to get right "**the result doesn't change whether or not you participate**" - might discourage participation!

Revealing **characteristics** of individuals vs. their **identity**, is there a distinction?

But the Census collects "generic" **harmless data**, is this really a big deal?

What sorts of trade-offs should we be aware of? Who should decide?

r/ai

beyond differential privacy

r/ai

The Strava Heat Map

WIRED

JEREMY HSU SECURITY JAN 29, 2018 7:14 PM

The Strava Heat Map and the End of Secrets

The US military is reexamining security policies after fitness tracker data shared on social media revealed bases and patrol routes

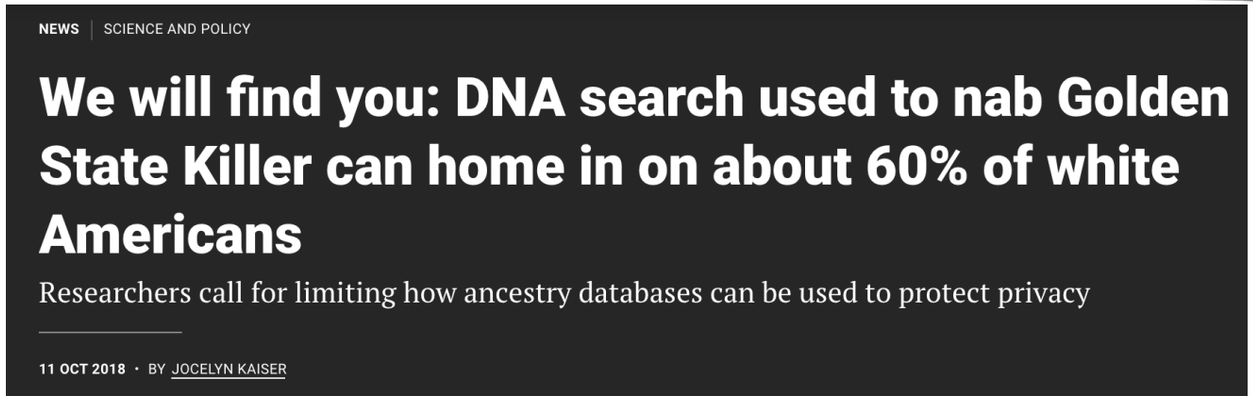
”Over the weekend, researchers and journalists raised the alarm about how **anyone can identify secretive military bases and patrol routes** based on public data shared by a “social network for athletes” called Strava.

This past November, the San Francisco-based Strava announced a huge update to its global heat map of user activity that displays 1 billion activities—including running and cycling routes—undertaken by exercise enthusiasts wearing Fitbits or other wearable fitness trackers. […]

But the biggest danger may come from potential adversaries figuring out “patterns of life,” by tracking and even identifying military or intelligence agency personnel as they go about their duties or head home after deployment. These digital footprints that echo the real-life steps of individuals underscore a greater challenge to governments and ordinary citizens alike: each person’s connection to online services and personal devices makes it increasingly difficult to keep secrets.”

r/ai

Is genetic data your own?



NEWS | SCIENCE AND POLICY

We will find you: DNA search used to nab Golden State Killer can home in on about 60% of white Americans

Researchers call for limiting how ancestry databases can be used to protect privacy

11 OCT 2018 · BY JOCELYN KAISER

Science

If you're white, live in the United States, and a distant relative has uploaded their DNA to a public ancestry database, there's a good chance **an internet sleuth can identify you from a DNA sample you left somewhere**. That's the conclusion of a new study, which finds that by combining an anonymous DNA sample with some basic information such as someone's rough age, researchers **could narrow that person's identity to fewer than 20 people by starting with a DNA database of 1.3 million individuals**. […]

The study was sparked by **the April arrest of the alleged "Golden State Killer," a California man accused of a series of decades-old rapes and murders**. To find him—and more than a dozen other criminal suspects since then—law enforcement agencies first test a crime scene DNA sample, which could be old blood, hair, or semen, for hundreds of thousands of DNA markers—signposts along the genome that vary among people, but whose identity in many cases are shared with blood relatives. They then upload the DNA data to **GEDmatch, a free online database where anyone can share their data from consumer DNA testing companies such as 23andMe and Ancestry.com to search for relatives who have submitted their DNA**. Searching GEDMatch's nearly 1 million profiles revealed several relatives who were the equivalent to third cousins to the crime scene DNA linked to the Golden State Killer. Other information such as genealogical records, approximate age, and crime locations then allowed the sleuths to home in on a single person.

https://www.science.org/content/article/we-will-find-you-dna-search-used-nab-golden-state-killer-can-home-about-60-white

r/ai

the origins of data protection

r/ai

Barrow, Alaska, 1979

Native leaders and city officials, worried about drinking and associated violence in their community, **invited a group of sociology researchers** to assess the problem and work with them to devise solutions.

Methodology:

- 10% representative sample (N=88) of everyone over the age of 15 using a 1972 demographic survey
- Interviewed on attitudes and values about use of alcohol
- Obtained psychological histories & drinking behavior
- Given the Michigan Alcoholism Screening Test
- Asked to draw a picture of a person (to determine cultural identity)

based on a slide by Bill Howe

r/ai

Study “results”

Alcohol Plagues Eskimos; Alcoholism Plagues Eskimo Village

DAVA SOBEL ();
January 22, 1980,
, Section Science Times, Page C1, Column , words

PERMISSIONS

[DISPLAYING ABSTRACT]

THE Inupiat Eskimos of Alaska's North Slope, whose culture has been overwhelmed by energy development activities, are "practically committing suicide" by mass alcoholism, University of Pennsylvania researchers said here yesterday. The alcoholism rate is 72 percent among the 2,000 Eskimo men and women in the village of Barrow, where violence is becoming the ...

At the conclusion of the study researchers formulated a report entitled **“The Inupiat, Economics and Alcohol on the Alaskan North Slope”**, released **simultaneously** at a press release and to the Barrow community.

The press release was picked up by the New York Times, who ran a front page story entitled **“Alcohol Plagues Eskimos”**

based on a slide by Bill Howe

r/ai

Harms and backlash

Article Preview

Eskimos Irate Over Alcoholism Study

[DISPLAYING ABSTRACT]

PERMISSIONS

BARROW, ALASKA HOT tempers and tension arising from a scientific report that found a high rate of alcoholism in this predominantly Eskimo community have abated somewhat after two days of meetings here at the northernmost point of Alaska.

Study **results were revealed** in the context of a press conference that was held far from the Native village, and **without the presence, much less the knowledge or consent**, of any community member who might have been able to present any context concerning the socioeconomic conditions of the village.

Study results suggested that nearly all adults in the community were alcoholics. In addition to the shame felt by community members, the town's Standard and Poor bond rating suffered as a result, which in turn decreased the tribe's ability to secure funding for much needed projects.

based on a slide by Bill Howe

r/ai

Problems

Edward F. Foulks, M.D., “Misalliances In The Barrow Alcohol Study”

Methodological

- “The authors once again met with the Barrow Technical Advisory Group, who stated their concern that only Natives were studied, and that outsiders in town had not been included.”

- “The **estimates of the frequency of intoxication based on association with the probability of being detained** were termed "ludicrous, both logically and statistically.”

Ethical

- Participants not in control of how their data is used
- Significant harm: social (stigmatization) and financial (bond rating)

can differential privacy help with this?

based on a slide by Bill Howe

r/ai

need an ethical framework!

r/ai

Responsible Data Science

Anonymity and privacy

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai center for responsible ai