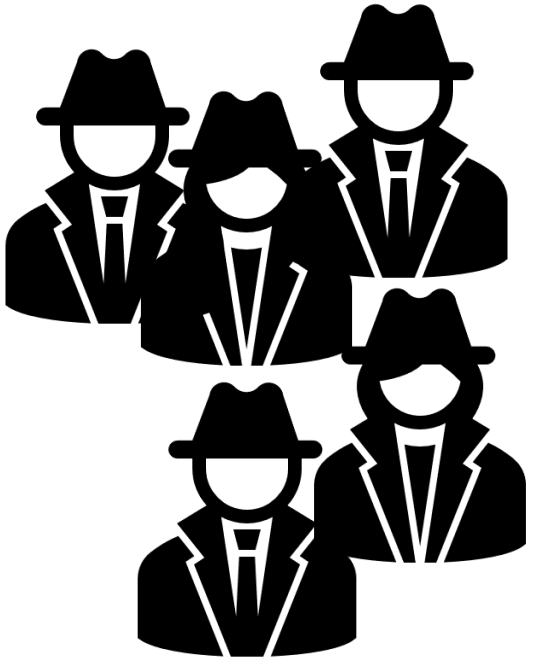
Disclosure Avoidance

What is disclosure?



What is disclosure avoidance?

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Types of Disclosure

Identity disclosure

- Data subject is identified from a released file

Attribute disclosure

- Sensitive information about a data subject is revealed through the released file

Inferential disclosure

- The released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods

Objectives:

- Prevent identification of individual respondents
- Prevent intentional and inadvertent disclosure of individuals' personal information

Method choice base on whether:

- Aggregate estimates (formatted as frequency counts or aggregate magnitude data)
- Micro-data (individual units)

United States Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Cell suppression

Using a threshold, individual cells get defined as sensitive and the values are suppressed.

Table 4: Example -- With Disclosure

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

Table 6: Example -- Without Disclosure, Protected by Suppression

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	D	D	D	20
Beta	20	10	10	15	55
Gamma	D	D	10	D	25
Delta	D	14	D	D	35
Total	50	35	30	20	135

NOTE: D indicates data withheld to limit disclosure.

Be aware of empty cells when producing tables. Releasing tables with large numbers of empty cells should be avoided. Tables with large numbers of empty cells can be aggregated to a higher geography if needed.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Random rounding

Cell values are rounded based on a random decision of whether to round up or down.

Table 4: Example -- With Disclosure

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

Table 7: Example -- Without Disclosure, Protected by Random Rounding

Number of Delinquent Children by County and Education Level of Household Head

Education Level of Household Head

County	Low	Medium	High	Very High	Total
Alpha	15	0	0	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	15	15	10	0	35
Total	50	35	30	20	135

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Controlled rounding

Cell values are rounded based on a random decision of whether to round up or down, but the rounding is controlled in some instances to ensure the sums of the rows and columns now work.

Table 4: Example -- With Disclosure

Number of Delinquent Children by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

Table 8: Example -- Without Disclosure, Protected by Controlled Rounding

Number of Delinquent Children by County and Education Level of Household Head

	Education Level of Household Head				
County	Low	Medium	High	Very High	Total
Alpha	15	0	5	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	10	15	5	5	35
Total	50	35	30	20	135

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Collapsing cells

Groups of cells (rows or columns) are combined into a single row.

Table 2: Example -- With Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A	2	4	18	20	7	1	52
B	--	-	7	9	-	-	16
C	--	6	30	15	4	-	55
D	-	-	2	--	-	-	2

Table 3: Example -- Without Disclosure

Number of Beneficiaries by Monthly Benefit Amount and County

	Monthly Benefit Amount						
County	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Total
A and B	2	4	25	29	7	1	68
C and D	--	6	32	15	4	-	57

Data Protection Methods: Controlled tabular adjustment

Sensitive cells are replaced with a value that is a "sufficient distance" from the true value.

Non-sensitive cells are minimally adjusted to ensure the totals match.

May be done in combination with rounding as an indication to the data user that a cell has been adjusted.

Data Protection Methods: Top-coding and Bottom-coding

Top-coding and Bottom-coding censor data high or lower than a certain value.

This data was top-coded above $250k in column 4 and bottom-coded below $10k in column 5.

Id	Age	Actual Wealth	Top-coded	Bottom-coded
1	26	$25988	$25988	$25988
2	34	$32458	$32458	$32458
3	22	$75483	$75483	$75483
4	45	$18574	$18574	$18574
5	64	$15302956	$250000	$15302956
6	41	$192834	$192834	$192834
7	19	$33859	$33859	$33859
8	42	$196	$196	$10000
9	37	$274858	$250000	$274858
10	42	$6492	$6492	$10000

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Recoding

Recoding changes every data value in a set of microdata. Usually, this means the data is recoded into an interval, either directly or through rounding.

This data was recoded into intervals of $25k.

Id	Age	Actual Wealth	Recoded
1	26	$25988	$25k-$50k
2	34	$32458	$25k-$50k
3	22	$75483	$50k-$75k
4	45	$18574	$0k-$25k
5	64	$15302956	$250k+
6	41	$192834	$150k-$200k
7	19	$33859	$25k-$50k
8	42	$196	$0k-$25k
9	37	$274858	$250k+
10	42	$6492	$0k-$25k

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Data swapping

Id	Age	Actual Wealth	Postal code
1	26	$25988	20942
2	34	$32458	47892
3	22	$75483	91003
4	45	$18574	47743
5	64	$15302956	10293
6	41	$192834	88391
7	19	$33859	20341
8	42	$196	33061
9	37	$274858	09281
10	42	$6492	77801

Id	Age	Actual Wealth	Postal code
1	26	$25,988	47892
2	34	$32,458	20942
3	22	$33,859	91003
4	45	$15,302,956	47743
5	64	$18,574	10293
6	41	$274,858	88391
7	19	$75,483	20341
8	42	$196	77801
9	37	$192,834	9281
10	42	$6,492	33061

Data Protection Methods: Synthetic data

Synthetic data are modeled statistical outputs released in a format that closely resembles confidential data.

Synthetic data are useful when other methods produce results that do not make sense.

To create synthetic data, a model is built on existing data, unique records in the data are identified, and replaced with a value generated from the model.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Noise infusion

Data from each respondent is perturbed by a small amount in either direction.

Noise can be added mainly to sensitive cells while leaving other cells basically unchanged.

Id	Firm	Material Costs	Noisy Costs
1	26	$25988	$26994
2	34	$32458	$32078
3	22	$75483	$76464
4	45	$18574	$18368
5	64	$15302956	$13798675
6	41	$192834	$173570
7	19	$33859	$35572
8	42	$196	$215
9	37	$274858	$250616
10	42	$6492	$6851

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Data Protection Methods: Microdata

When releasing microdata:

1. Eliminate information that directly identifies individuals, geolocation data in particular.
2. Suppress data that may indirectly identify individuals.
3. Introducing uncertainty into the reported data. Swapping data.

To reduce the potential for disclosure, most public-use microdata files:

- Include data from only a sample of the population.
- Do not include obvious identifiers.
- Limit geographic detail.
- Limit the number and detailed breakdown of categories within variables on the file.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Disclosure Avoidance Discussion

What are your organization's requirements for limiting disclosure of PII?

Which of these techniques do you use? And, in what context?

Differential Privacy

“Giving overly accurate answers to too many questions will inevitably destroy privacy.”

- Dwork and Roth, The Algorithmic Foundations of Differential Privacy

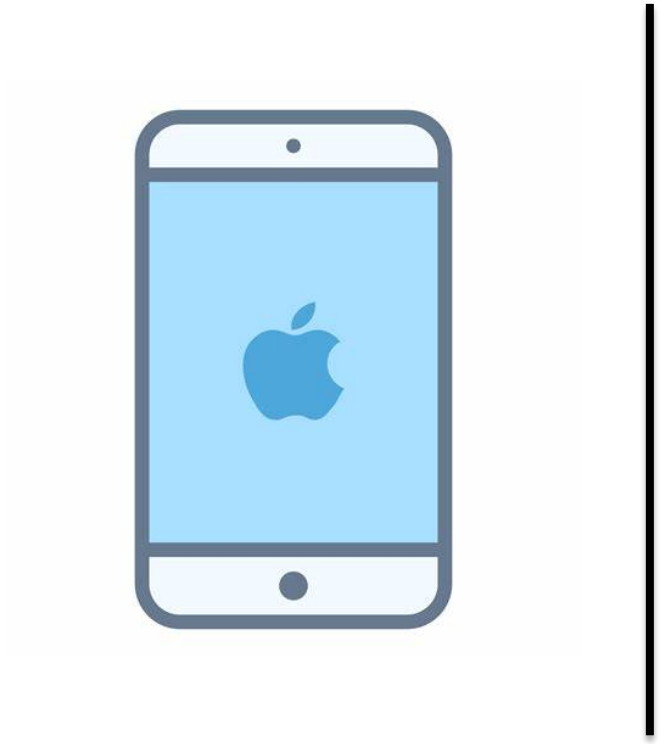
“Too many statistics published too accurately from a confidential database exposes the entire database with near certainty”

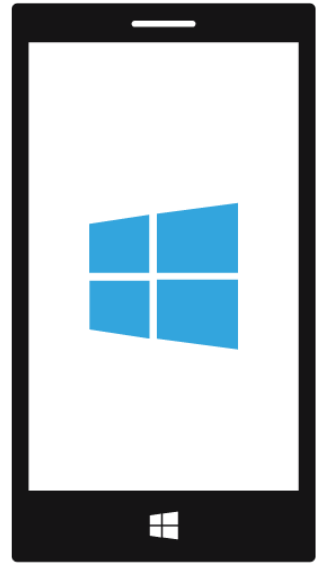
– Dinur and Nissim 2003

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Have you interacted with a differentially private system?



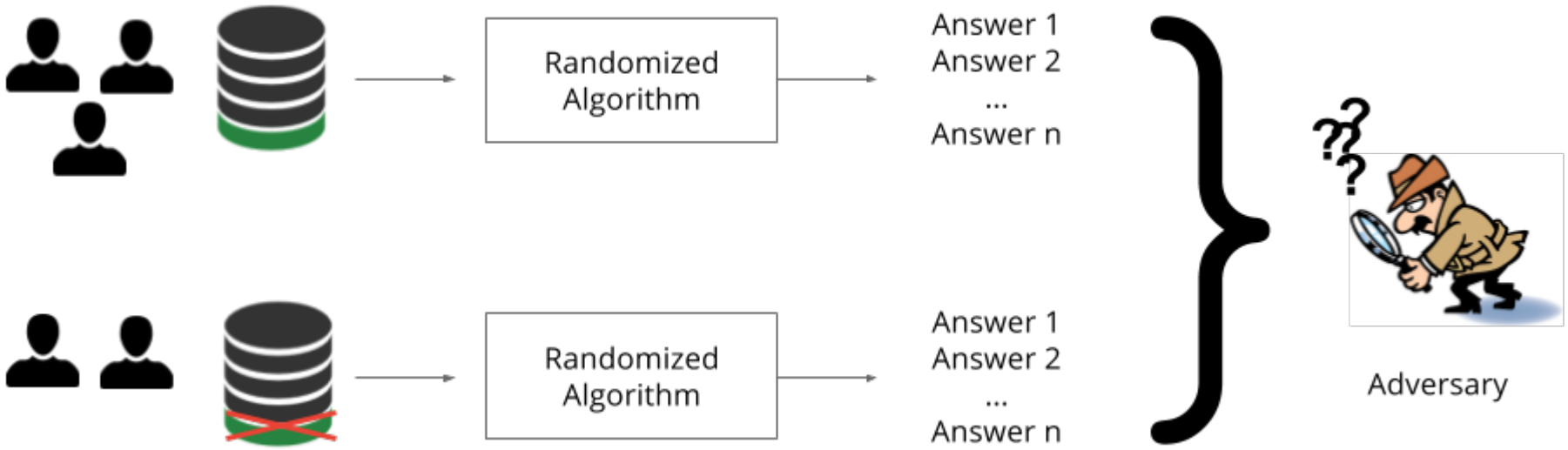


United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

What is Differential Privacy?

Differential Privacy is a mathematical framework for measuring the precise disclosure risk associated with each release of confidential data.



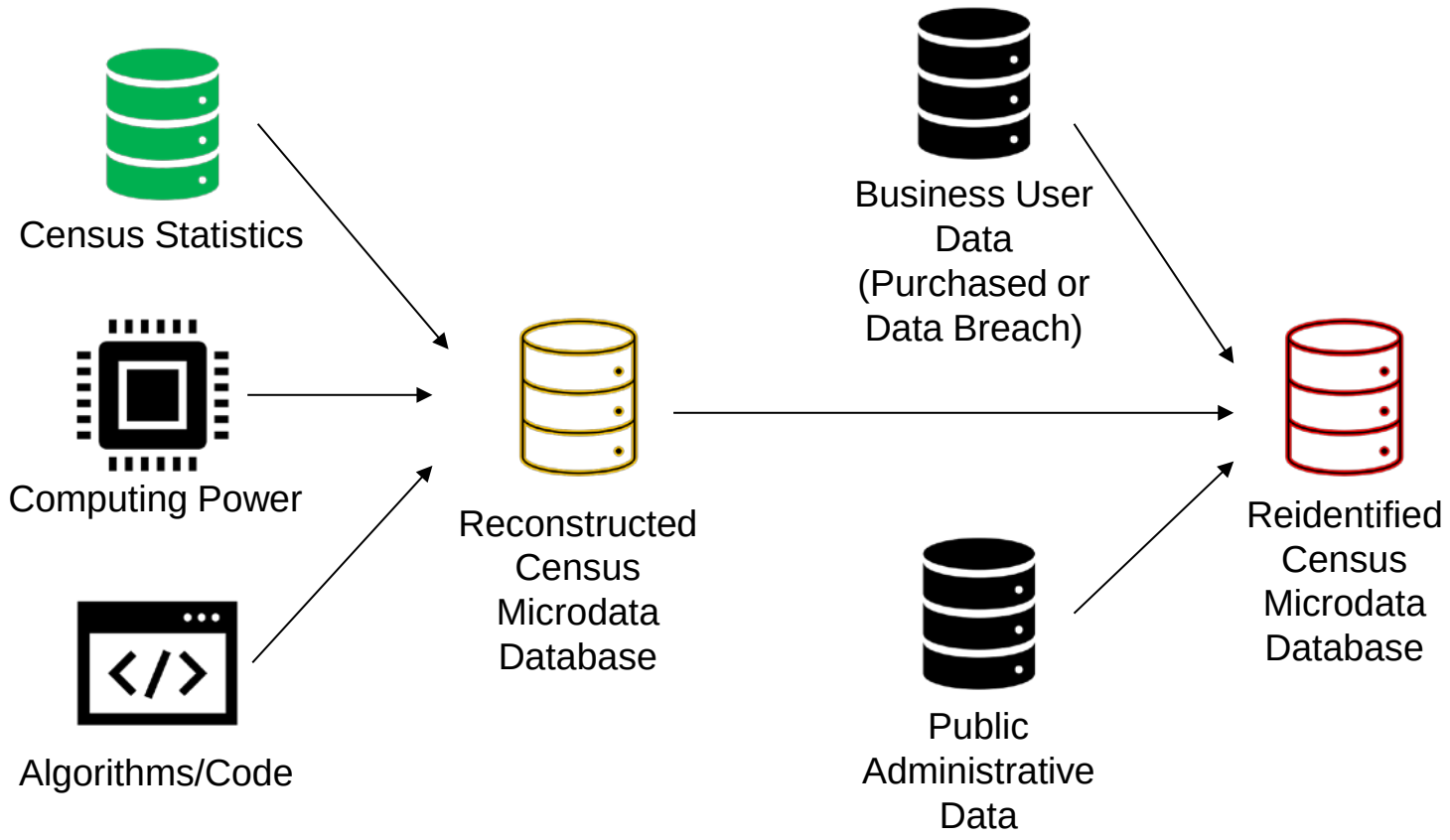
[Privacy and machine learning: two unexpected allies?]

Reconstruction and Reidentification Attack

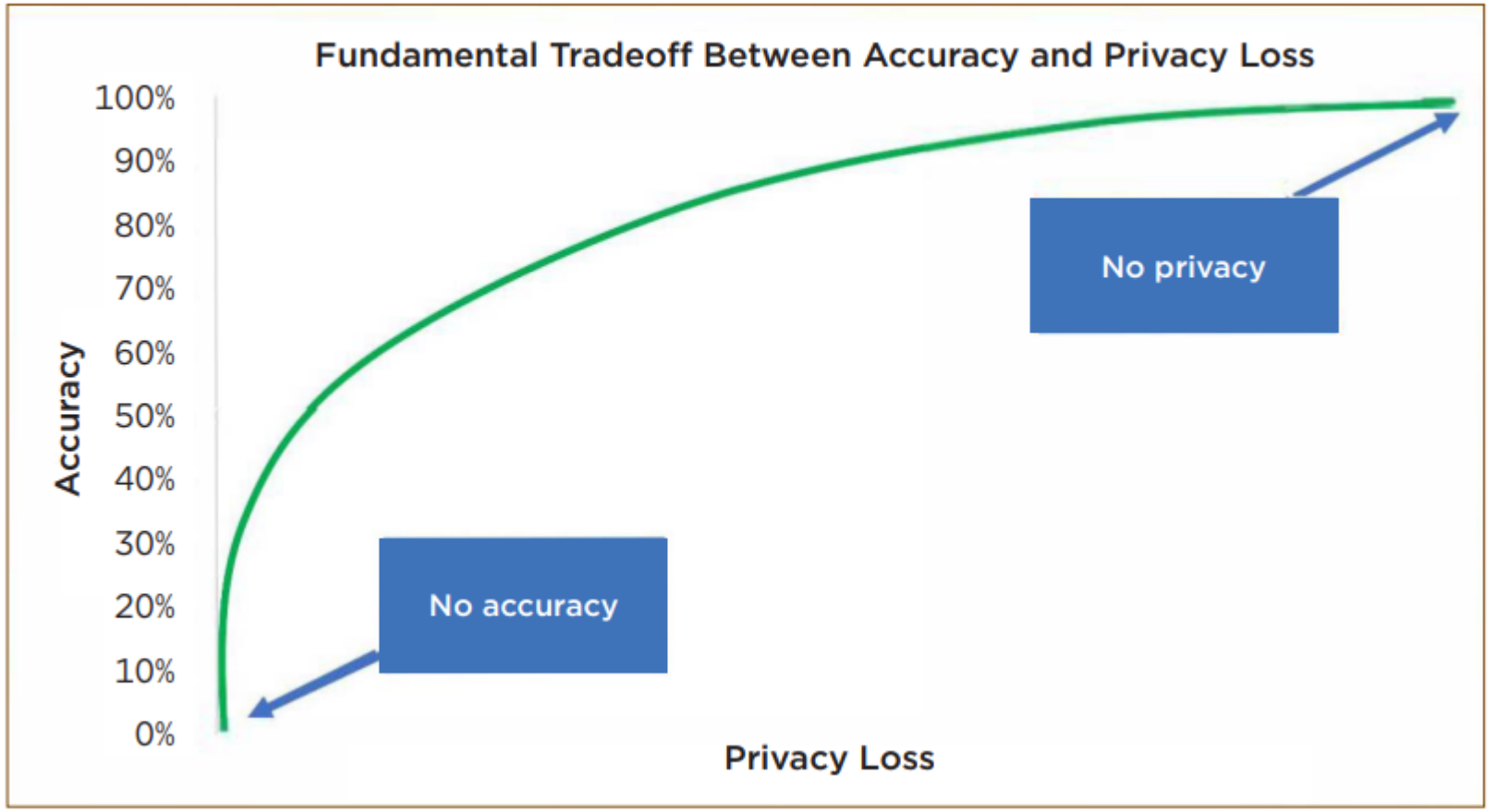
For the 2010 census, the Census Bureau published 150 billion statistics about a population of 310 million people. This is almost **500** statistics for every person from a 10-question questionnaire!

Reconstruction Attack – using publicly available data or statistics to (partially) create a database that is identical to a private database. Statistics published about a database allow this to happen.

Reidentification Attack (linkage attack) – using publicly available data or auxiliary information to identify individuals in a private database.



Privacy vs. Accuracy



Disclosure Avoidance for the 2020 Census: An Introduction

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Advantages and Disadvantages

Advantages

- Closed under composition
- Robust to post-processing
- Future-proof
- Provable and tunable
- Public and explainable
- Protects against database reconstruction attacks

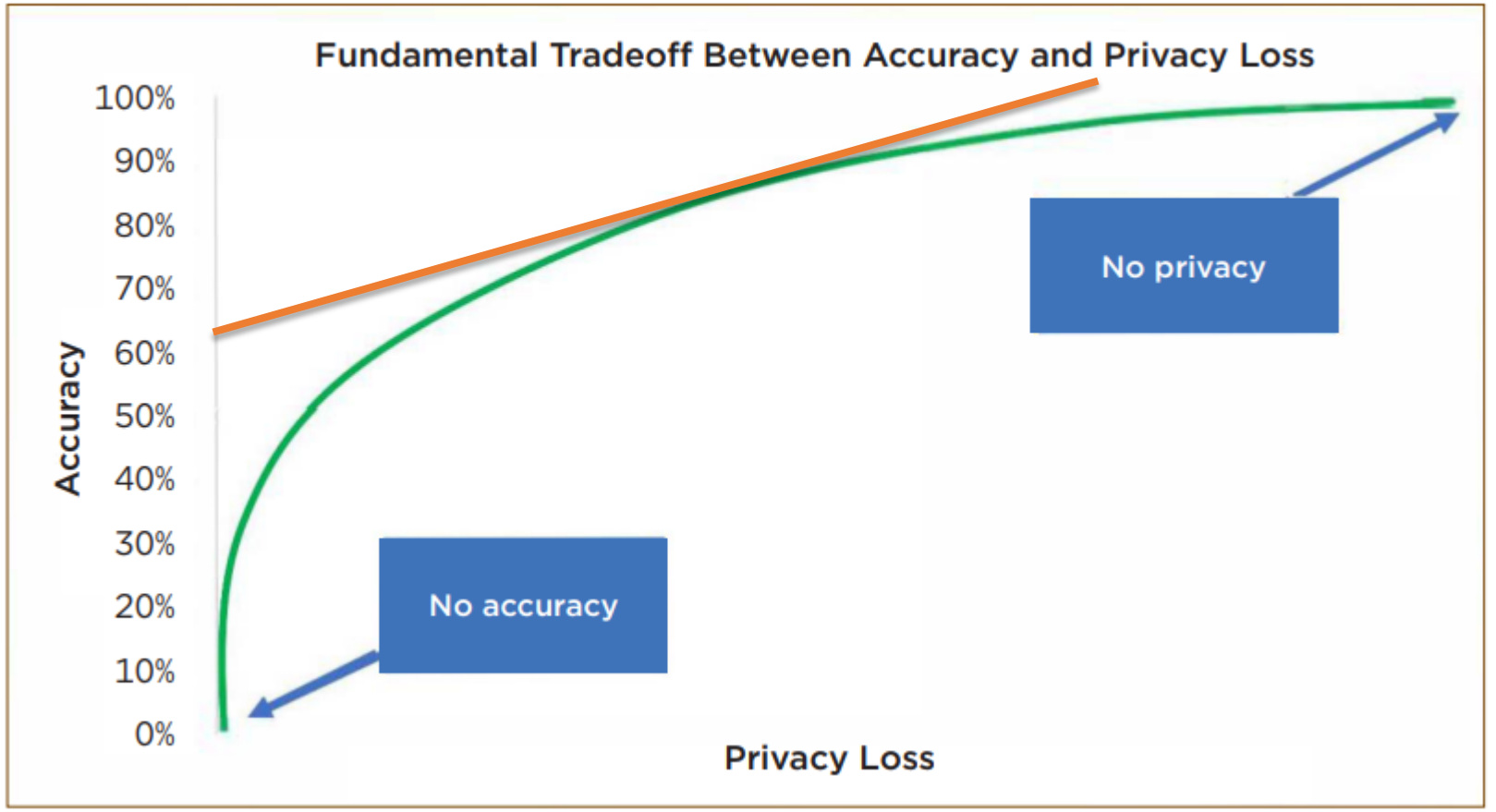
Disadvantages

- Entire country must be processed at once for best accuracy.
- Every use of the private data must be tallied in the privacy-loss budget.

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Privacy-Loss vs. Accuracy as Social Choices



Disclosure Avoidance for the 2020 Census: An Introduction

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE

Privacy-Loss vs. Accuracy Discussion

What sorts of factors influence how your organizations approach the privacy-loss vs. accuracy tradeoff?

How do you think your stakeholders view the privacy-loss vs. accuracy tradeoff? Does their view align with yours?

Are there any constraints or obstacles that you might face when trying to protect privacy or improve accuracy?

United States® Census Bureau

USAID FROM THE AMERICAN PEOPLE