The (Real and Imagined) Bounds of Statistical Purpose

Alexandra Wood

Department of Political Science
Purdue University

Presentation for ASA Privacy Day Webinar

January 28, 2026

Source: US Census Bureau (1950), (1935)

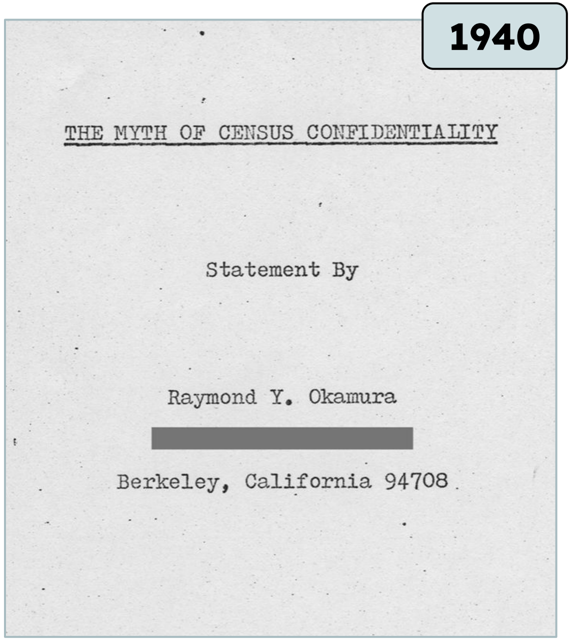
Motivation

Statistical purpose is a fundamental yet underexplored concept.

- Statistical purpose functions as a special case of purpose limitation
  - bounding the scope of permissible processing activities and
  - implicating regulatory requirements distinct from those applicable to processing for other purposes.
- In the absence of an established definition, organizations rely on different interpretations and assumptions about what counts as a statistical purpose.
- Clarifying statistical purpose is essential for
  - safeguarding statistical integrity,
  - maintaining public trust, and
  - ensuring legal frameworks protect personal information as intended.

Motivation

1940

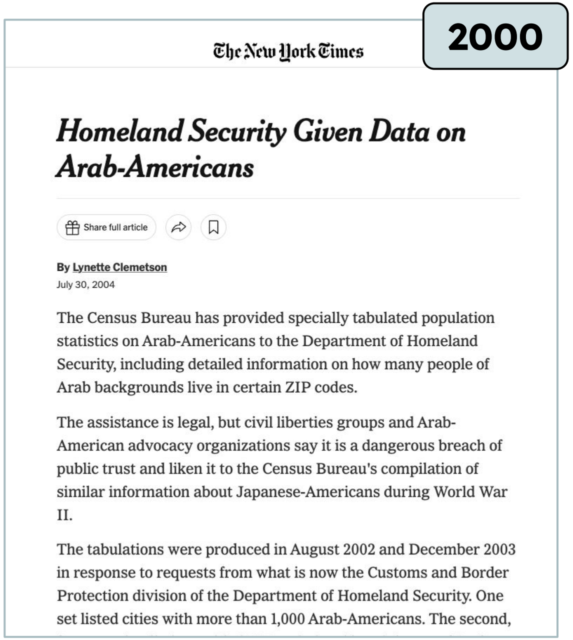


THE MYTH OF CENSUS CONFIDENTIALITY

Statement By

Raymond Y. Okamura
[redacted]
Berkeley, California 94708

2000



The New York Times

Homeland Security Given Data on Arab-Americans

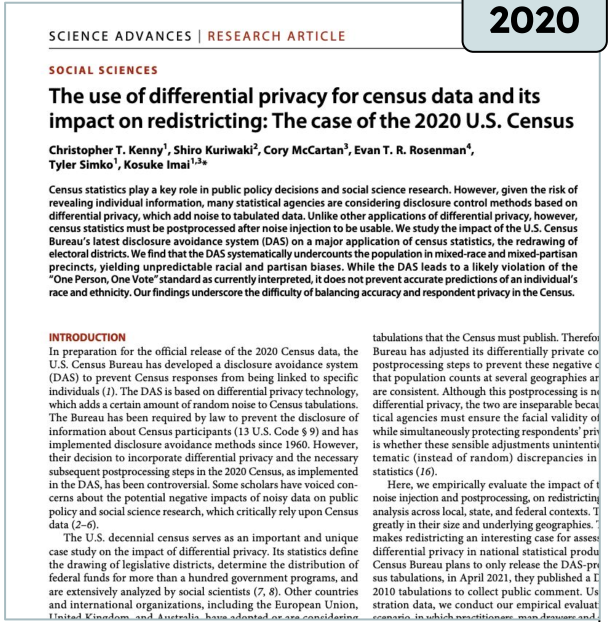
By Lynette Clemetson
July 30, 2004

The Census Bureau has provided specially tabulated population statistics on Arab-Americans to the Department of Homeland Security, including detailed information on how many people of Arab backgrounds live in certain ZIP codes.

The assistance is legal, but civil liberties groups and Arab-American advocacy organizations say it is a dangerous breach of public trust and liken it to the Census Bureau's compilation of similar information about Japanese-Americans during World War II.

The tabulations were produced in August 2002 and December 2003 in response to requests from what is now the Customs and Border Protection division of the Department of Homeland Security. One set listed cities with more than 1,000 Arab-Americans. The second,

2020



SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census

Christopher T. Kenny[1], Shiro Kuriwaki[2], Cory McCartan[3], Evan T. R. Rosenman[4], Tyler Simko[1], Kosuke Imai[1,3]*

Census statistics play a key role in public policy decisions and social science research. However, given the risk of revealing individual information, many statistical agencies are considering disclosure control methods based on differential privacy, which add noise to tabulated data. Unlike other applications of differential privacy, however, census statistics must be postprocessed after noise injection to be usable. We study the impact of the U.S. Census Bureau's latest disclosure avoidance system (DAS) on a major application of census statistics, the redrawing of electoral districts. We find that the DAS systematically undercounts the population in mixed-race and mixed-partisan precincts, yielding unpredictable racial and partisan biases. While the DAS leads to a likely violation of the "One Person, One Vote" standard as currently interpreted, it does not prevent accurate predictions of an individual's race and ethnicity. Our findings underscore the difficulty of balancing accuracy and respondent privacy in the Census.

INTRODUCTION

In preparation for the official release of the 2020 Census data, the U.S. Census Bureau has developed a disclosure avoidance system (DAS) to prevent Census responses from being linked to specific individuals (*1*). The DAS is based on differential privacy technology, which adds a certain amount of random noise to Census tabulations. The Bureau has been required by law to prevent the disclosure of information about Census participants (13 U.S. Code § 9) and has implemented disclosure avoidance methods since 1960. However, their decision to incorporate differential privacy and the necessary subsequent postprocessing steps in the 2020 Census, as implemented in the DAS, has been controversial. Some scholars have voiced concerns about the potential negative impacts of noisy data on public policy and social science research, which critically rely upon Census data (*2*–*6*).

The U.S. decennial census serves as an important and unique case study on the impact of differential privacy. Its statistics define the drawing of legislative districts, determine the distribution of federal funds for more than a hundred government programs, and are extensively analyzed by social scientists (*7*, *8*). Other countries and international organizations, including the European Union, United Kingdom, and Australia, have adopted or are considering

tabulations that the Census must publish. Therefor Bureau has adjusted its differentially private co postprocessing steps to prevent these negative c that population counts at several geographies ar are consistent. Although this postprocessing is nc differential privacy, the two are inseparable becau tical agencies must ensure the facial validity of while simultaneously protecting respondents' pri is whether these sensible adjustments unintenti tematic (instead of random) discrepancies in statistics (*16*).

Here, we empirically evaluate the impact of t noise injection and postprocessing, on redistricting analysis across local, state, and federal contexts. T greatly in their size and underlying geographies. makes redistricting an interesting case for asses differential privacy in national statistical produ Census Bureau plans to only release the DAS-pr sus tabulations, in April 2021, they published a D 2010 tabulations to collect public comment. Us stration data, we conduct our empirical evaluat scenario in which practitioners map drawers and

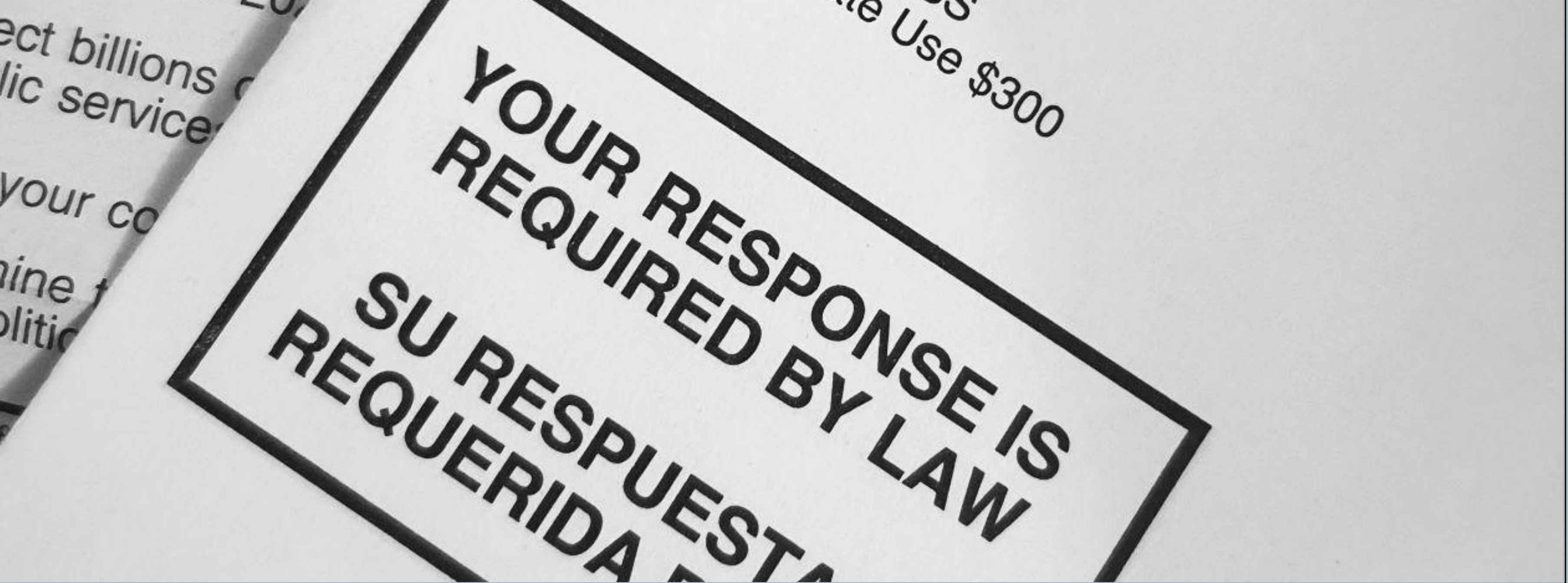
Outline

1 The **Origins and Functions** of Statistical Purpose

2 The **Many Interpretations** of Statistical Purpose

3 **External Authorities** as Guides to (Re-)Interpretation

4 Practical **Interventions & Implications** for Practice



1 The Origins and Functions of Statistical Purpose

The Origins of Statistical Purpose

Sources: margoanderson.org, C-SPAN (2001)

Anderson and Seltzer (2009) trace the roots of statistical confidentiality to the origins of the term "statistics."[1]

- i.e., the analysis of data about the state, conducted by early practitioners known as statists.

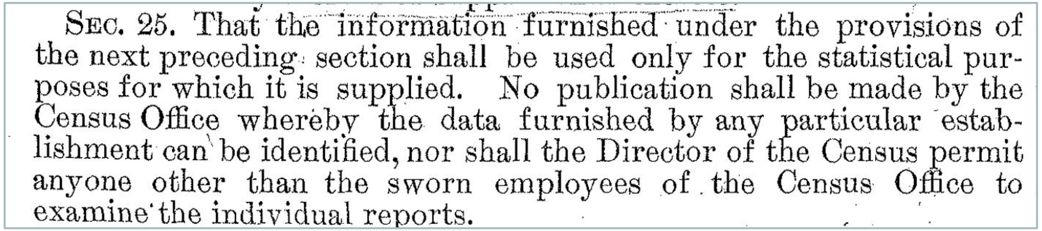
Statists recognized the potential for generating new knowledge from administrative data.[1]

- They developed the first techniques for statistical analysis, which were devised to draw collective insights from aggregates, stripped of identifiers.
- Over time, they pushed to establish practices for collecting or using data for statistical purposes.

Early Tensions, Then Codification

Anderson and Seltzer (2009) argue that, while statisticians recognized that maintaining a separation between administrative and statistical data was essential, this was not understood outside of the statistical community.[1]

- Repeated administrative requests for statistical data were resisted by statistical agencies, producing ongoing tension.
- The statistical community advanced ethical and confidentiality principles to ensure public trust, independence, and data quality.
- In 1909, the Thirteenth Census Act introduced confidentiality protection for establishments.



SEC. 25. That the information furnished under the provisions of the next preceding section shall be used only for the statistical purposes for which it is supplied. No publication shall be made by the Census Office whereby the data furnished by any particular establishment can be identified, nor shall the Director of the Census permit anyone other than the sworn employees of the Census Office to examine the individual reports.

Codification of Individual Protection

Administrative requests for individual-level census information persisted, and sometimes met by resistance from census officials. For example, requests came from:

- The Justice Department, for citizenship status, to be used in deportation cases,[2]
- Draft enforcement officials, which was ultimately successful and inspired a failed proposal to redesign the 1920 Census as a process for registering for the Selective Service,[2] and
- The Internal Revenue Bureau, for records pertaining to the ages of children, for the purposes collecting taxes from businesses employing child labor, which was also successful.[2]

In 1929, the Fifteenth Census Act was amended to extend confidentiality protections to individuals. (Pub. L. 71-12 § 11).

- It also introduced the provision that "in no case shall information furnished under the authority of this Act be used to the detriment of the person or persons to whom such information relates." (§ 18).

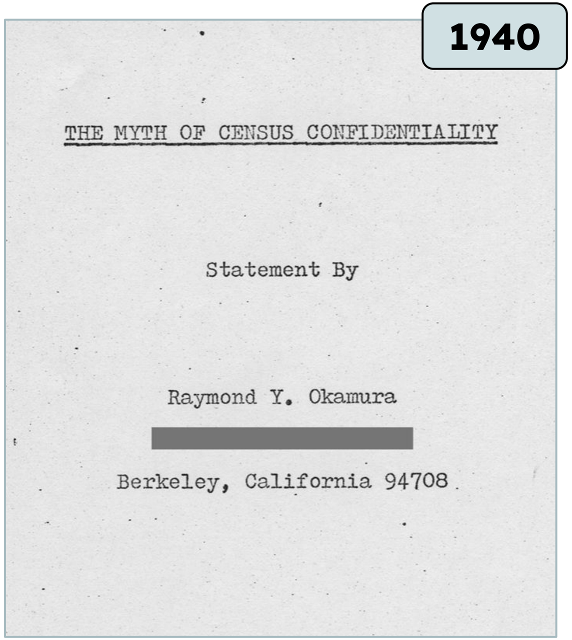
Second War Powers Act

Administrative requests expanded with the perception of compelling public purposes at times of crisis.

- During WWII, the Bureau initially denied requests for census data on the grounds of the new statutory confidentiality protections and norms and practices of official statistics.[2]
    - Extensive legislative and administrative efforts to curtail the Census Bureau's confidentiality protections were blocked by the Census Bureau and the Attorney General.[2]
- Enactment of the Second War Powers Act suspended the census confidentiality protections and authorized access to census records "for use in connection with the conduct of the war" until the law's repeal in 1947 (56 Stat. 186, § 1402)..
    - The contemporaneously stated purpose was to enable the Army to obtain counts as well as the names and addresses of people of Japanese ancestry on the West Coast in support of the forced evacuation and internment program.[2]

Use of 1940 Census Data

1940



THE MYTH OF CENSUS CONFIDENTIALITY

Statement By

Raymond Y. Okamura
[redacted]
Berkeley, California 94708

"While it was technically correct to state that names and addresses were not revealed, such a statement was extremely misleading because sufficient other census information was provided to the War Department for the purpose of locating and imprisoning Japanese Americans. . . .

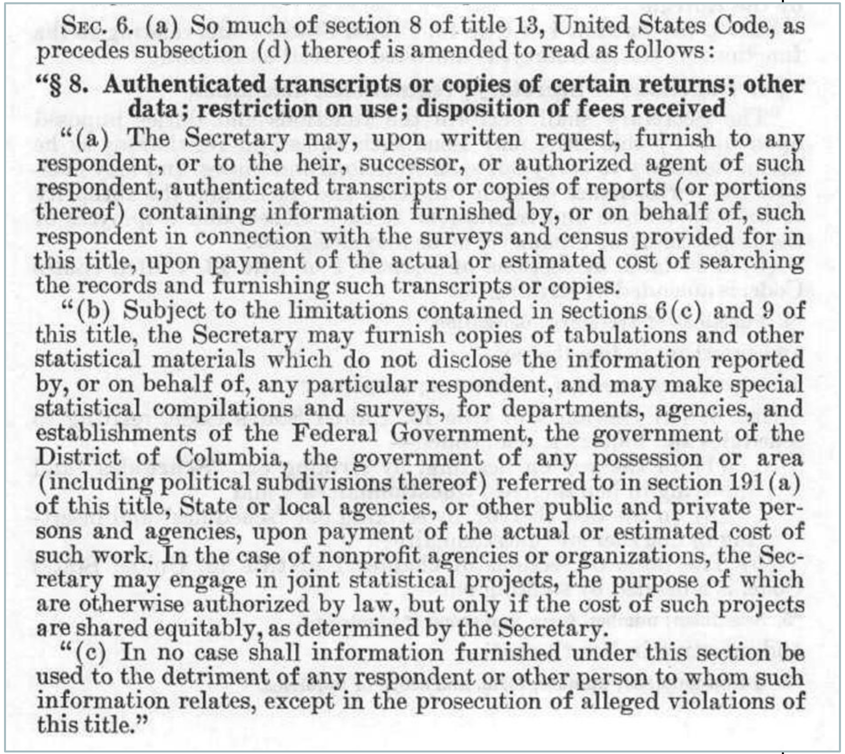
Such incarceration could not have been one of the 'statistical purposes' for which census respondents supplied information regarding their race or ancestry . . ."[3]

Expanded Individual Protection

Congress amended Title 13 in 1976 “to add further protection of privacy” by

- prohibiting the disclosure of information “reported by, or on behalf of, any respondent,”
- restricting the scope of permissible recipients of raw census data to “respondents, their heirs, or authorized agents” and not “Governors of States, territories, and courts of record,” and
- prohibiting “the furnishing of statistical information which would disclose information reported by or on behalf of any respondent.”

Public Law 94-521 § 6.



SEC. 6. (a) So much of section 8 of title 13, United States Code, as precedes subsection (d) thereof is amended to read as follows:

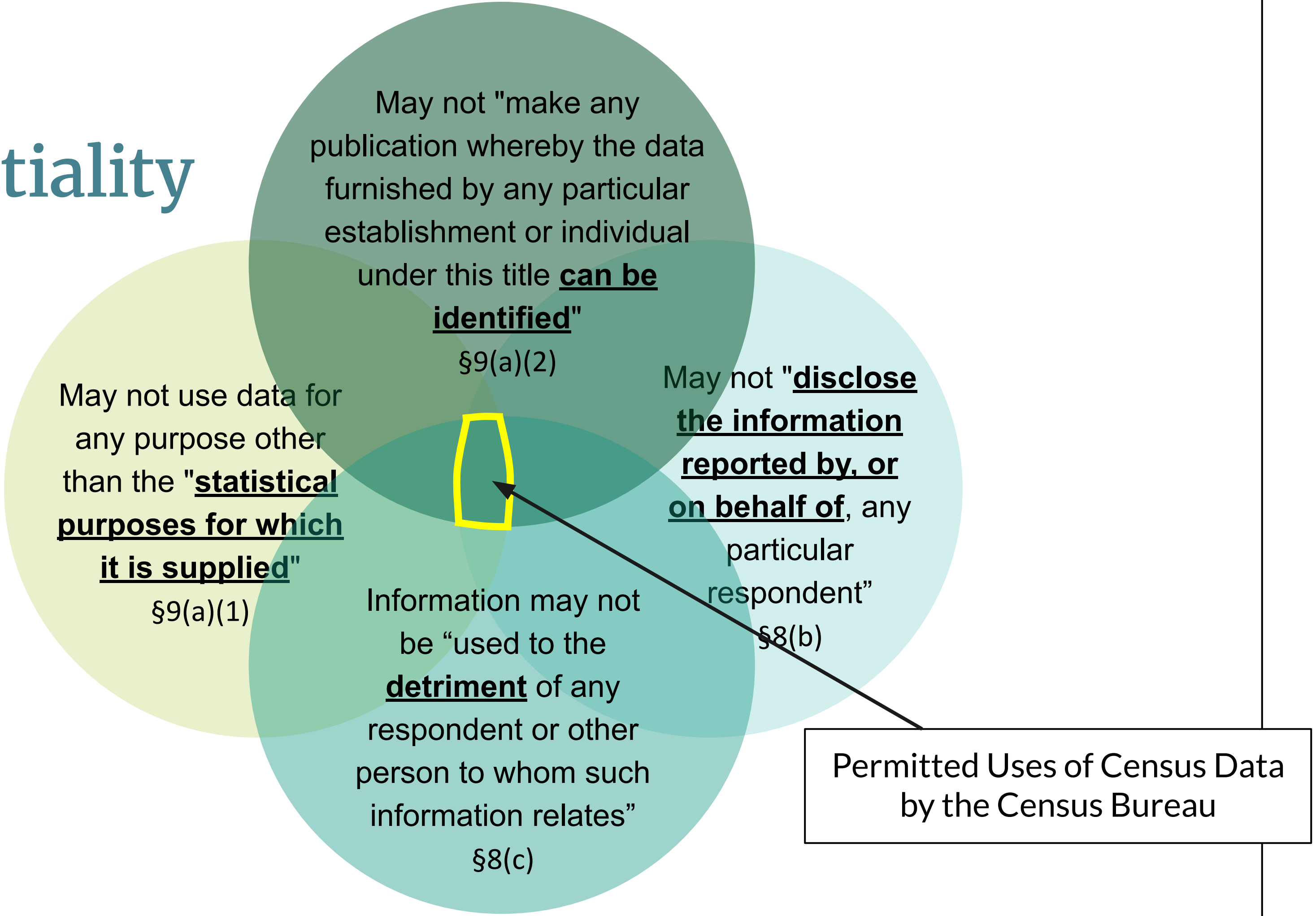
“§ 8. Authenticated transcripts or copies of certain returns; other data; restriction on use; disposition of fees received

“(a) The Secretary may, upon written request, furnish to any respondent, or to the heir, successor, or authorized agent of such respondent, authenticated transcripts or copies of reports (or portions thereof) containing information furnished by, or on behalf of, such respondent in connection with the surveys and census provided for in this title, upon payment of the actual or estimated cost of searching the records and furnishing such transcripts or copies.

“(b) Subject to the limitations contained in sections 6(c) and 9 of this title, the Secretary may furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent, and may make special statistical compilations and surveys, for departments, agencies, and establishments of the Federal Government, the government of the District of Columbia, the government of any possession or area (including political subdivisions thereof) referred to in section 191(a) of this title, State or local agencies, or other public and private persons and agencies, upon payment of the actual or estimated cost of such work. In the case of nonprofit agencies or organizations, the Secretary may engage in joint statistical projects, the purpose of which are otherwise authorized by law, but only if the cost of such projects are shared equitably, as determined by the Secretary.

“(c) In no case shall information furnished under this section be used to the detriment of any respondent or other person to whom such information relates, except in the prosecution of alleged violations of this title.”

Title 13's Confidentiality Mandate



May not "make any publication whereby the data furnished by any particular establishment or individual under this title **can be identified**"
§9(a)(2)

May not use data for any purpose other than the "**statistical purposes for which it is supplied**"
§9(a)(1)

May not "**disclose the information reported by, or on behalf of**, any particular respondent"
§8(b)

Information may not be "used to the **detriment** of any respondent or other person to whom such information relates"
§8(c)

Permitted Uses of Census Data by the Census Bureau

Baldrige v. Shapiro (1982)

Title 13, § 9(a)(1) played a key role in the 1982 Supreme Court case, *Baldrige v. Shapiro.*

- Denver officials sought to challenge the 1980 count and compel the Census Bureau to disclose vacancy information, “for precisely those statistical purposes for which that information is intended, i.e., to ensure an accurate count of the population of the City and County of Denver in the 1980 Decennial Census.”
- The Court rejected the Denver officials’ argument: “Subsection 9(a)(1) permits use of the data only for ‘the statistical purposes for which it is supplied.’ There is no indication in the Census Act that the hundreds of municipal governments in the 50 states were intended by Congress to be the ‘monitors’ of the Census Bureau.’”

Source: US Census Bureau (1980)

Evolution and Expansion

Statistical confidentiality protections have been refined over time in response to challenges.

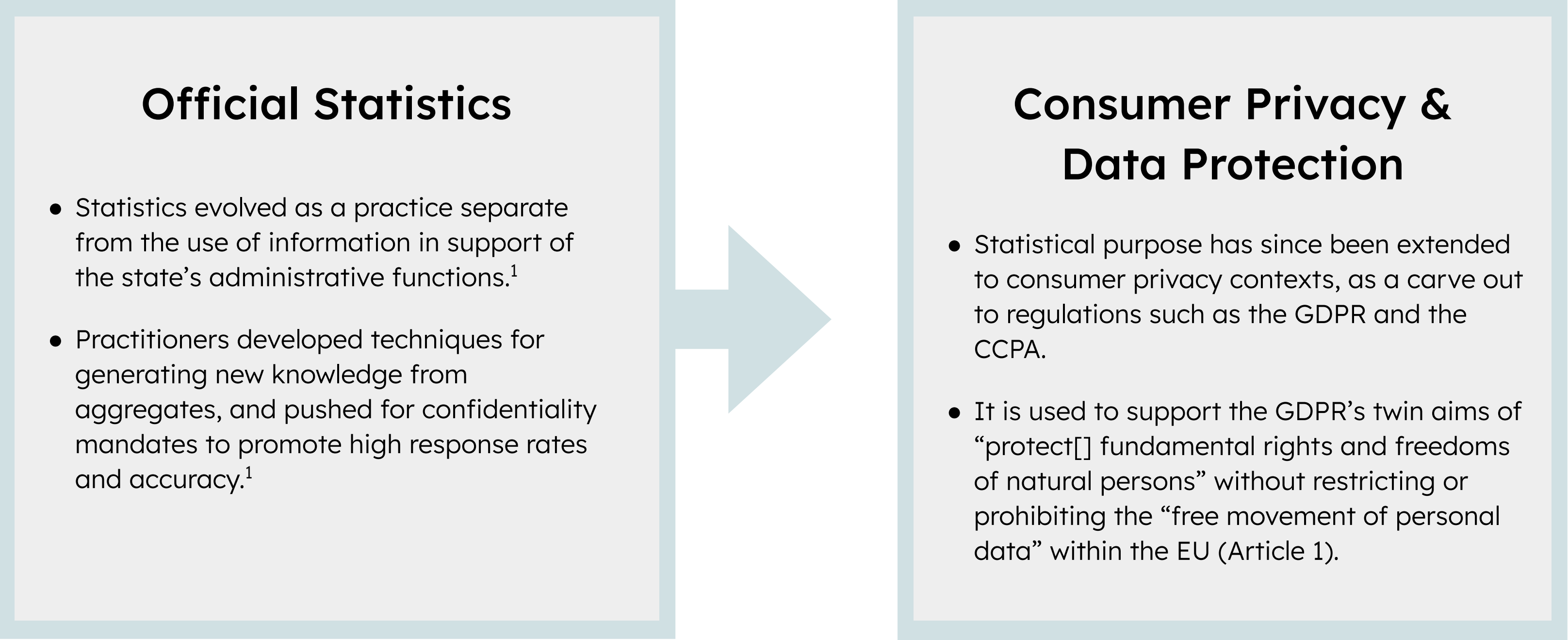
- Reflected in a wide body of statistical confidentiality principles, ethical codes, policies, and statutory mandates.

They have also inspired exemptions to consumer privacy and data protection regulations.

- E.g., under Article 5 of the GDPR, further processing for statistical purposes is not considered to be incompatible with the original purpose, thereby exempting it from the purpose limitation principle where appropriate safeguards are in place.
- Similar exemptions appear in the California Consumer Privacy Act and Canada’s proposed Consumer Privacy Protection Act, among others.

Constraining and Enabling Roles

- Statistical purpose exceptions play a critical role in balancing the dual purposes of data access and data protection under many different regulatory frameworks.



Official Statistics

- Statistics evolved as a practice separate from the use of information in support of the state's administrative functions.[1]
- Practitioners developed techniques for generating new knowledge from aggregates, and pushed for confidentiality mandates to promote high response rates and accuracy.[1]

Consumer Privacy & Data Protection

- Statistical purpose has since been extended to consumer privacy contexts, as a carve out to regulations such as the GDPR and the CCPA.
- It is used to support the GDPR's twin aims of "protect[] fundamental rights and freedoms of natural persons" without restricting or prohibiting the "free movement of personal data" within the EU (Article 1).

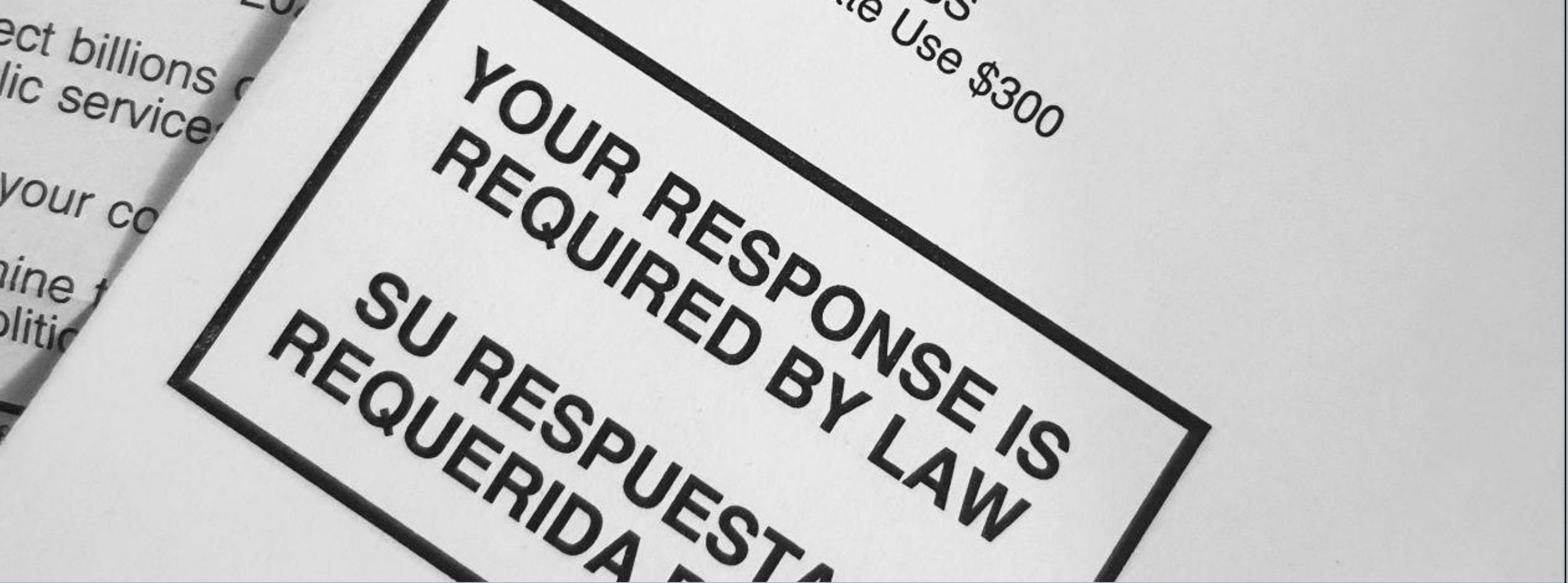
General Data Protection Regulation

- **Exemption from the purpose limitation principle:** A compatibility presumption that further processing for statistical purposes is not incompatible with the initial purposes (Article 5).
- **Basis for lawful processing under Article 6**: “Further processing for . . . statistical purposes should be considered to be compatible lawful processing operations.” (Recital 50).
- **Expanded retention period:** permits the retention of personal data for longer periods for statistical purposes provided appropriate technical and organizational measures are implemented (Article 5).
- **Very broad scope:** “any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results” (Recital 162)
  - Potentially encompasses a very wide class of statistical and machine learning analyses, including analyses carried out within commercial contexts.[4]
  - Because Recital 162 provides that the result of processing of personal data for statistical purposes is non-personal data, the output from a very wide class of analysis may fall outside of the scope of data protection rules.

The Problem

Despite its significant role in determining the boundaries of permissible data use, the term *statistical purpose* is often either underspecified or left undefined altogether.

- In the absence of a clear definition, a multitude of, often conflicting, interpretations have emerged.
- Nonetheless, regulators and regulated entities often treat statistical purpose as having a commonly accepted meaning.
- Historical challenges within official statistics suggest that interpretations of statistical purpose exemptions are likely to be contested, especially as demand for statistical analyses grows.



2 The Many Interpretations of Statistical Purpose

The Many Interpretations of Statistical Purpose

Statistical purpose is a multidimensional concept, shaped by many interacting elements, each of which has be interpreted in different ways.

- Although legislation is often left underspecified due to “the limits of human foresight, the ambiguities of language, and the high cost of legislative deliberation” thereby leaving “many areas of uncertainty . . . to be resolved by the courts” (Landes & Posner 1975), statutory meanings of the term statistical purpose seem especially indeterminate.
- As an aid in understanding the many meanings of statistical purpose, the following typology characterizes various elements which combine in different ways to make up interpretations of statistical purpose in current regulatory frameworks.
- Even among examples in the same category, there is wide variation, reflecting substantial uncertainty regarding questions such as how to process personal data for statistical purposes without revealing information specific to individual data subjects.

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 Aggregation

5 Non-administrative

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

The Many Interpretations of Statistical Purpose

1 **Statistical Production**

2 Population Analysis

3 Non-identification

4 Aggregation

5 Non-administrative

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

Representing the essential role of statistical purpose as the production of statistical results, as in the meaning of production used in EU Regulation 223/2009 on European statistics ("all the activities related to the collection, storage, processing, and analysis necessary for compiling statistics").

- Modernized Convention 108: "the elaboration of statistical surveys or the production of statistical, aggregated results."
- GDPR, Recital 162: "any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results."

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 **Population Analysis**

3 Non-identification

4 Aggregation

5 Non-administrative

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

Emphasizing the purpose of characterizing groups at a collective level but not as individuals.

- Recommendation No. R (97) 18, explanatory memorandum: the purpose of statistical activity “is merely to describe the characteristics of that population as a whole.” (and information from many different individuals is used by a statistician to “elaborate[] results designed to ‘characterise a collective phenomenon.’”)
- CIPSEA: statistical purpose is “the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups.”

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 **Non-identification**

4 Aggregation

5 Non-administrative

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

Complicated by the lack of a consistently understood and meaningful definition of non-identifying generally.

Some instances are intended to refer to non-identifying uses.

- CIPSEA: “the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups.”

Others refer to non-identifying results.

- GDPR, Recital 162: “[t]he statistical purpose implies that the result of processing for statistical purposes is not personal data” (i.e., the result is not “information relating to an identified or identifiable natural person”)

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 **Aggregation**

5 Non-administrative

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

Multiple meanings, overlapping with previous:

1. *the goal of statistical production*

   (Convention 108, explanatory report: statistical purpose means “the elaboration of statistical surveys or the production of statistical, aggregated results.”)

2. *for the analysis of collective phenomena*

   (Regulation (EC) No 223/2009: statistics are “quantitative and qualitative, aggregated and representative information characterizing a collective phenomenon in a considered population.”)

3. *without identifying individuals*

   (GDPR, Recital 162: results of processing for statistical purpose are “not personal data, but aggregate data.”)

However, all uses and releases of aggregate statistics carry some risks to individuals.

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 Aggregation

5 **Non-administrative**

6 Ancillary Activities

7 Prescribed Purpose

8 Harm Prevention

Emphasizing that a statistical purpose is one that excludes administrative, regulatory, and law enforcement purposes as well as any other use in support of individual decisions.

- CIPSEA: distinguishes a statistical purpose from a nonstatistical purpose, which is defined as "the use of data in identifiable form for any purpose that is not a statistical purpose, including any administrative, regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileged, or benefits of a particular identifiable respondent."
- GDPR, Recital 162: prohibits the use of the "result [of processing for statistical purposes] or the personal data . . . in support of measures or decisions regarding any particular natural person."

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 Aggregation

5 Non-administrative

6 **Ancillary Activities**

7 Prescribed Purpose

8 Harm Prevention

Some definitions include other processing activities in support of processing for statistical purposes.

- CIPSEA: statistical purpose "includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support [such statistical] purposes."
- Federal Statistical Confidentiality Order: statistical purpose includes "the development, implementation, or maintenance of methods, procedures, or information resources that support such [statistical] purposes," and "specifically includes records management and archival functions" authorized by statute and "conducted under information security and confidentiality restrictions consistent with" the order.

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 Aggregation

5 Non-administrative

6 Ancillary Activities

7 **Prescribed Purpose**

8 Harm Prevention

Subsection 9(a)(1) of Title 13 restricts the Census Bureau's use of census data to "the statistical purposes for which it was supplied."

- Baldrige v. Shapiro (1982): Officials from the City of Denver sought to compel the Census Bureau to disclose vacancy information on its master address list "for precisely those statistical purposes for which that information is intended, i.e., to ensure an accurate count of the population of the City and County of Denver in the 1980 Decennial Census."
- However, the Supreme Court rejected this interpretation, holding that "[t]here is no indication in the Census Act that the hundreds of municipal governments in the 50 states were intended by Congress to be the 'monitors' of the Census Bureau."

The Many Interpretations of Statistical Purpose

1 Statistical Production

2 Population Analysis

3 Non-identification

4 Aggregation

5 Non-administrative

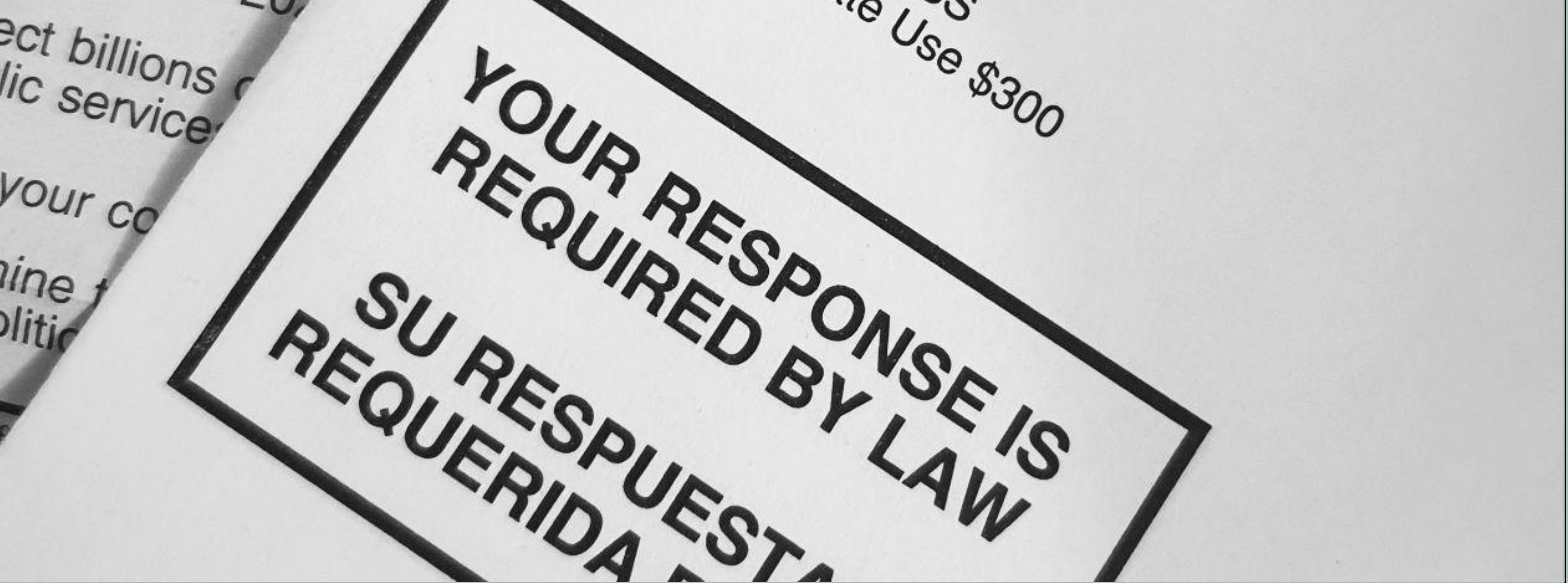
6 Ancillary Activities

7 Prescribed Purpose

8 **Harm Prevention**

Derives from principles of statistical confidentiality and statistical ethics and reflected, in part, in, e.g., 13 U.S.C. 8(c) (prohibiting information furnished to the Census Bureau to "be used to the detriment of any respondent or other person to whom such information relates"), but deeply contested:

- Raymond Okamura: statistical information provided by the Census Bureau to the War Department in support of Japanese internment "was not for the purpose for which it was supplied, and was clearly to the detriment of the persons to whom such information related (people were imprisoned as a result)."
- Hermann Habermann: Title 13's confidentiality provisions "are not intended to prohibit the use of aggregate statistical information," including "uses that may result in harm to particular groups within society."



3 External Authorities as Guides to (Re-)Interpretation

Statistical Policy and Standards

Example standards

- OMB standards and policy directives, e.g., Statistical Policy Directive No. 1 (2014),
- European Statistics Code of Practice,
- UN Fundamental Principles of Official Statistics, and implementation guidelines,
- National Academy of Sciences' principles to guide the practices of federal statistical agencies.

Example requirements[5]

- Confidentiality protections implementations across the full statistical lifecycle
- Penalties for confidentiality breaches
- Publicly available confidentiality policy
- Clear respondent notice on use and protections
- Controlled research access to transformed microdata
- Data minimization and secure data handling
- Disclosure risk controls for aggregate releases
- Formal review and anonymization for microdata access
- Strict, contract-based researcher access conditions
- Ongoing monitoring and enforcement of confidentiality
- Staff confidentiality obligations and security safeguards

Statistical Ethics

Professional codes of statistical ethics are foundational to statistical governance.

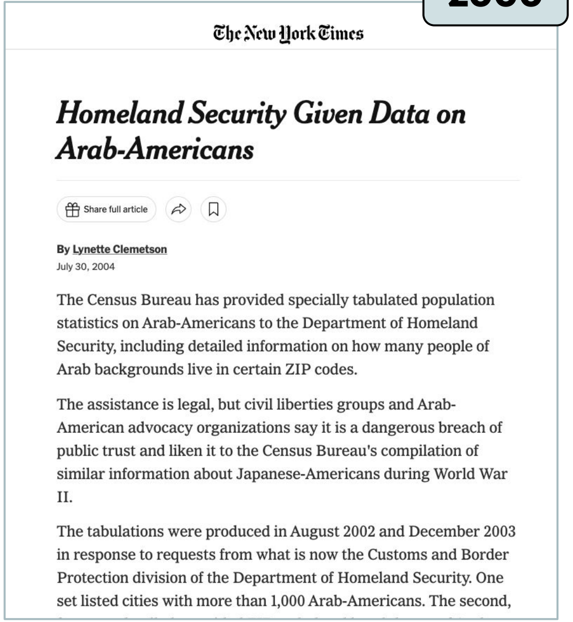
- e.g., American Statistical Association's Ethical Guidelines for Statistical Practice require:[6]
    - Respect for the rights, interests, and welfare of human subjects, including
        - Using data only as permitted by consent, or considering their interests and welfare when consent is not required,
        - Refraining from collecting or using more data than necessary,
        - Minimizing data re-identification risk,
        - Explaining any impact of de-identification on accuracy of results.
    - Addressing risks of harm to individuals and groups, especially vulnerable populations, including
        - Assessing impacts on society, groups, and individuals "the impact of statistical practice on society, groups, and individuals,
        - Recognizing potential adverse effects or effects on public perception of marginalized groups,
        - Avoiding "statistical practices that exploit vulnerable populations or create or perpetuate unfair outcomes."

Post-9/11 Special Tabulations

"The Census Bureau has provided specially tabulated population statistics on Arab-Americans to the Department of Homeland Security, including detailed information on how many people of Arab backgrounds live in certain ZIP codes. . . .

The categories provided were Egyptian, Iraqi, Jordanian, Lebanese, Moroccan, Palestinian, Syrian and two general categories, 'Arab/Arabic' and 'Other Arab.'"

2000



The New York Times

Homeland Security Given Data on Arab-Americans

Share full article

By Lynette Clemetson
July 30, 2004

The Census Bureau has provided specially tabulated population statistics on Arab-Americans to the Department of Homeland Security, including detailed information on how many people of Arab backgrounds live in certain ZIP codes.

The assistance is legal, but civil liberties groups and Arab-American advocacy organizations say it is a dangerous breach of public trust and liken it to the Census Bureau's compilation of similar information about Japanese-Americans during World War II.

The tabulations were produced in August 2002 and December 2003 in response to requests from what is now the Customs and Border Protection division of the Department of Homeland Security. One set listed cities with more than 1,000 Arab-Americans. The second,

Lynette Clemetson, *Homeland Security Given Data on Arab-Americans*, New York Times, July 30, 2004.

"[A] spokeswoman for Customs and Border Protection, said the requests were made to help the agency identify in which airports to post signs and pamphlets in Arabic. 'The information is not in any way being used for law enforcement purposes,' . . .

But critics . . . said general demographic snapshots could be derived without such detailed information, and that the ZIP-code-level data with its breakdowns of ancestral origin seemed particularly excessive, since for all of the groups only English or Arabic need be used."

Privacy Research

Key findings from privacy attacks research:

- Lack of rigor leads to unanticipated privacy failures.
  - New attack modes emerge as research progresses.
  - Redaction of identifiers, release of aggregates, etc. is insufficient.
- Auxiliary information must be taken into consideration.
- Any useful analysis of personal data must leak some information about individuals.
- Information leakages accumulate with multiple analyses/releases.

New line of privacy work in theoretical computer science (beginning ~2003) yields a new concept: Differential privacy (2006)[7]

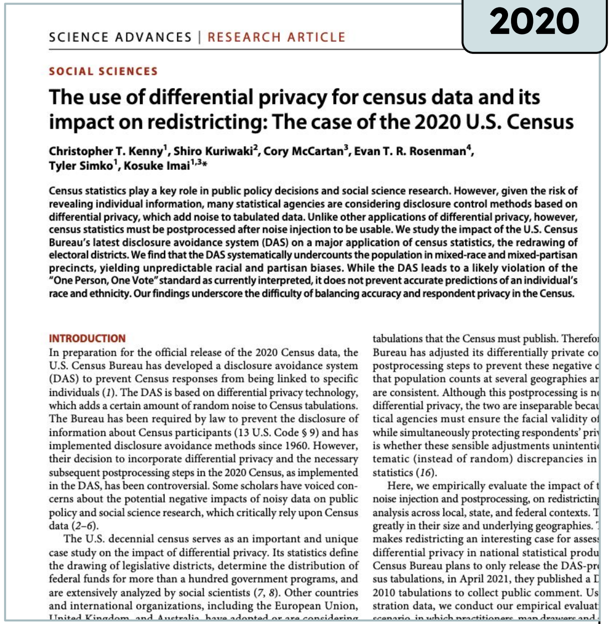
- Mathematically provable guarantee of privacy
- Supported by rich theory
- Provides a definition (i.e., a standard) of privacy for “statistical releases”

Statistical Inference and Confidentiality in the 2020 Census

“Researchers have developed methods to predict the race and ethnicity of individual voters using Census data. Since *Gingles*, voting rights cases have required evidence that an individual’s race is highly correlated with candidate choice. Statistical methods must therefore estimate this individual quantity from aggregate election results and aggregate demographic statistics.

. . .

Our analysis shows that across three main racial and ethnic groups, the predictions based on the DAS data appear to be as accurate as those based on the 2010 Census data. The finding suggests that, although the new DAS methodology may protect differential privacy, it may not prevent accurate prediction of sensitive attributes any more than the swapping methodology used in the 2010 Census.”



2020

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census

Christopher T. Kenny[1], Shiro Kuriwaki[2], Cory McCartan[3], Evan T. R. Rosenman[4], Tyler Simko[1], Kosuke Imai[1,3]*

Census statistics play a key role in public policy decisions and social science research. However, given the risk of revealing individual information, many statistical agencies are considering disclosure control methods based on differential privacy, which add noise to tabulated data. Unlike other applications of differential privacy, however, census statistics must be postprocessed after noise injection to be usable. We study the impact of the U.S. Census Bureau’s latest disclosure avoidance system (DAS) on a major application of census statistics, the redrawing of electoral districts. We find that the DAS systematically undercounts the population in mixed-race and mixed-partisan precincts, yielding unpredictable racial and partisan biases. While the DAS leads to a likely violation of the “One Person, One Vote” standard as currently interpreted, it does not prevent accurate predictions of an individual’s race and ethnicity. Our findings underscore the difficulty of balancing accuracy and respondent privacy in the Census.

INTRODUCTION

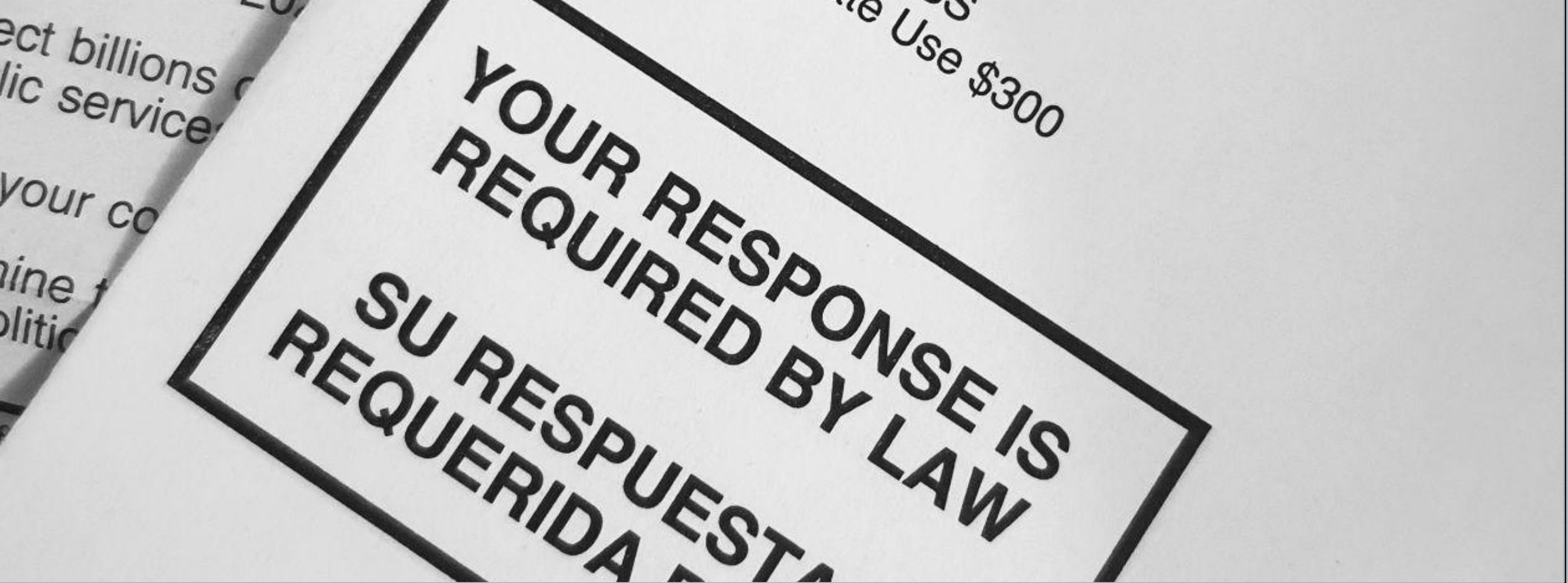
In preparation for the official release of the 2020 Census data, the U.S. Census Bureau has developed a disclosure avoidance system (DAS) to prevent Census responses from being linked to specific individuals (*1*). The DAS is based on differential privacy technology, which adds a certain amount of random noise to Census tabulations. The Bureau has been required by law to prevent the disclosure of information about Census participants (13 U.S. Code § 9) and has implemented disclosure avoidance methods since 1960. However, their decision to incorporate differential privacy and the necessary subsequent postprocessing steps in the 2020 Census, as implemented in the DAS, has been controversial. Some scholars have voiced concerns about the potential negative impacts of noisy data on public policy and social science research, which critically rely upon Census data (*2*–*6*).

The U.S. decennial census serves as an important and unique case study on the impact of differential privacy. Its statistics define the drawing of legislative districts, determine the distribution of federal funds for more than a hundred government programs, and are extensively analyzed by social scientists (*7*, *8*). Other countries and international organizations, including the European Union, United Kingdom, and Australia, have adopted or are considering

tabulations that the Census must publish. Therefor
Bureau has adjusted its differentially private co
postprocessing steps to prevent these negative c
that population counts at several geographies ar
are consistent. Although this postprocessing is n
differential privacy, the two are inseparable becau
tical agencies must ensure the facial validity of
while simultaneously protecting respondents’ priv
is whether these sensible adjustments unintentio
tematic (instead of random) discrepancies in
statistics (*16*).

Here, we empirically evaluate the impact of t
noise injection and postprocessing, on redistricting
analysis across local, state, and federal contexts. T
greatly in their size and underlying geographies. T
makes redistricting an interesting case for assess
differential privacy in national statistical produ
Census Bureau plans to only release the DAS-pr
sus tabulations, in April 2021, they published a D
2010 tabulations to collect public comment. Us
stration data, we conduct our empirical evaluat
scenario, in which practitioners, map drawers and

Christopher T. Kenny et al., *The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census*, 7 SCIENCE ADVANCES eabk3283 (2021).



4 Practical Interventions & Implications for Practice

Clarifying Definitions and Related Guidance

Statistical purposes refer to activities related to the collection, storage, processing, and analysis necessary for producing statistics that characterize a collective phenomenon. Statistical purposes exclude activities for administrative, regulatory, and law enforcement purposes as well as any other use in support of individual decisions. Further, such activities must be carried out using reasonable and appropriate safeguards that ensure that no individual incurs more than a minimal risk of harm from the use of their information, including with respect to future processing of or releases of data resulting from such activities. (Such activities are limited to the statistical purposes for which the information was supplied, though they may include activities to develop, implement, or maintain methods, procedures, or resources that support such statistical purposes.)

- Statistical purpose is inherently a multi-dimensional concept, subject to different interpretations and assumptions.
- There is an opportunity to clarify interpretation by developing a comprehensive definition informed by guidance from statistical policy, professional ethics, and the theoretical computer science literature on privacy.

Layering Legal, Technical, and Organizational Measures for Data Protection

Addressing statistical confidentiality requires expanding the range of safeguards to include "substantive, technical, operational, legal, policy, and ethical safeguards designed to deter the most likely and persistent 'intruders,' that is, other agencies of government with investigative, intelligence, or prosecutorial agendas."[1]

Examples include:

- Anonymization and privacy enhancing technologies to ensure functional separation[8]
- Interactive mechanisms as part of a tiered access system[9]
- Data minimization, particularly limiting data collection with respect to sensitive information,[10] and considering non-release or releases with less geographic detail in high-risk contexts[11]
- Commitments to and training in statistical ethics[12]

Conclusion

- Statistical purpose limitation acts as a constraint and enabler of data processing, aiming to balance access and protection.
- Often underspecified, especially in consumer privacy and data protection regulation, and the resulting uncertainty will likely lead to suboptimal trade-offs between data access and privacy.
- Statistical policy and standards, statistical ethics, and computer science research on privacy in statistical computation provide potential foundations for defining and guiding core concepts and practices with respect to statistical purpose provisions.
- Establishing clarifying definitions and practical guidance on layering legal, technical, and organizational controls will be critical to ensuring greater certainty for practitioners and adequate protection for data subjects.

References

1. Margo Anderson & William Seltzer, *Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues*, 1 J. Privacy & Confidentiality 7, 10 (2009).
2. Margo Anderson & William Seltzer, *Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965*, 23 J. Official Statistics 1, 6 (2007).
3. Raymond Y. Okamura, *The Myth of Census Confidentiality*, 8 Amerasia 111, 113 (1981).
4. Viktor Mayer-Schönberger & Yann Padova, *Regime Change? Enabling Big Data Through Europe’s New Data Protection Regulation*, 17 Columbia Science & Technology Law Review 315, 326 (2016).
5. United Nations, United Nations Fundamental Principles of Official Statistics: Implementation Guidelines (January 2015).
6. American Statistical Association, Ethical Guidelines for Statistical Practice (2022).
7. Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam D. Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, J. Privacy & Confidentiality 7 (3): 17–51 (2016).
8. Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation 27 (2013).
9. Micah Altman, Alexandra Wood, David R. O’Brien, Salil Vadhan & Urs Gasser, *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 Berkeley Tech. L. J. 1967 (2015).
10. Hermann Habermann, *Ethics, Confidentiality, and Data Dissemination*, 22 J. Official Statistics 599, 609 (2006).
11. Dennis Trewin, *Discussion*, 22 J. Official Statistics 631, 632 (2006).
12. Stephen E. Fienberg, *Discussion*, 22 J. Official Statistics 615, 617 (2006).