Privacy and Confidentiality

Introduction to Big Data for Social Science

Frauke Kreuter[1] ...

June 3–4, 2019

[1]fkreuter@umd.edu

Table of contents

1. Intro to Privacy and Confidentiality

2. Why access is important

3. Current access modalities

3.1 Legal framework

4. Big data challenges

4.1 Legal framework

4.2 One possible statistical framework

5. The future

Intro to Privacy and Confidentiality

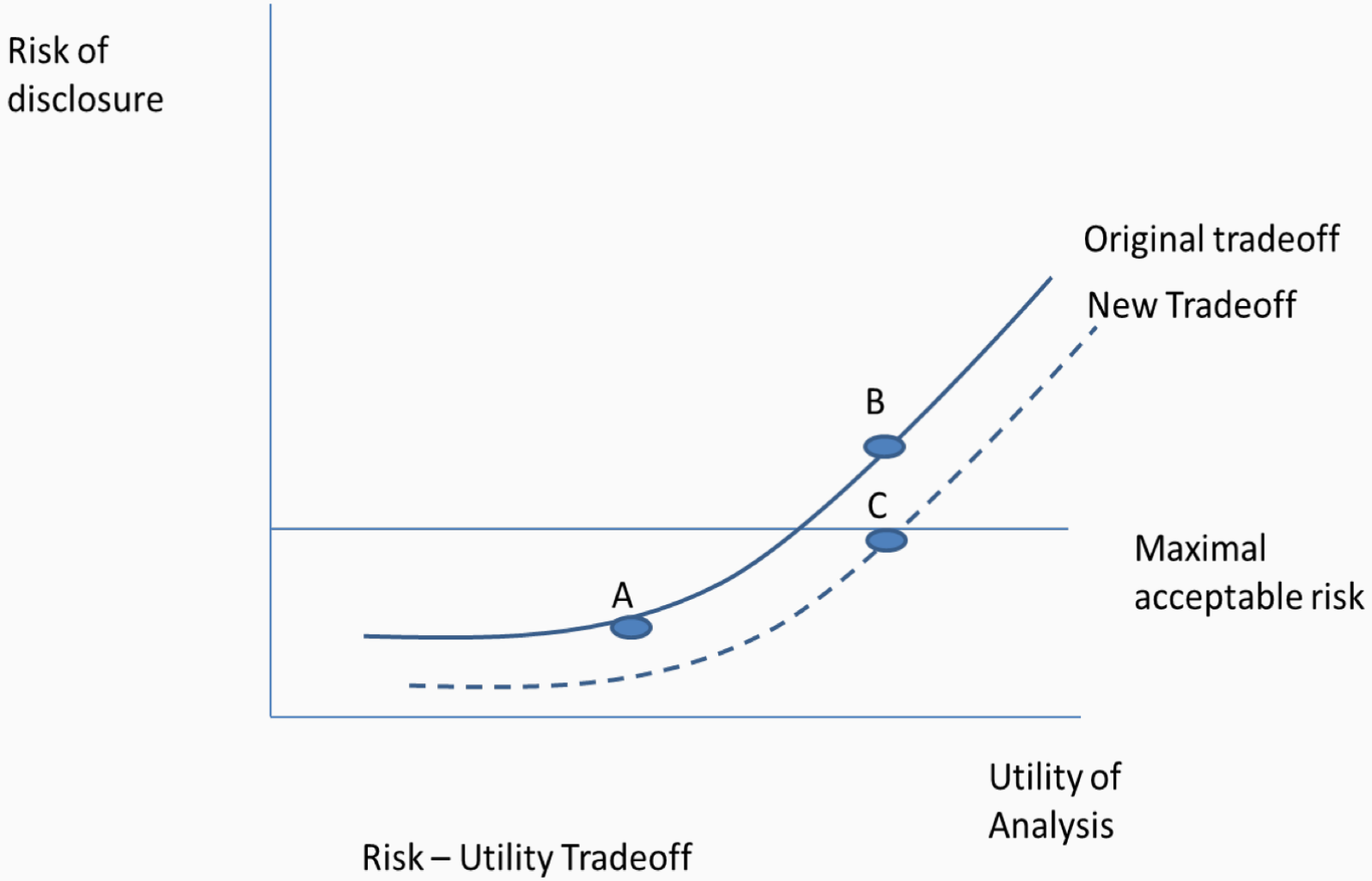
What is ...

Privacy
Includes the famous “right to be left alone,” and the ability to share information selectively but not publicly (White House 2014)

Confidentiality
means “preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information” (McCallister, Grance, and Scarfone 2010).

Challenge

How to balance the ***risk*** of providing access with the associated utility?



Risk – Utility Tradeoff

Why access is important

Why is Access Important?

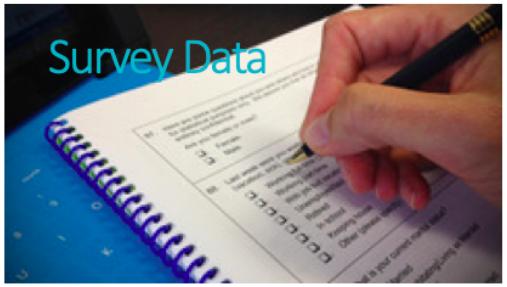
- Linkage validation
- Replication
- Building knowledge infrastructure

Current access modalities

Legal framework

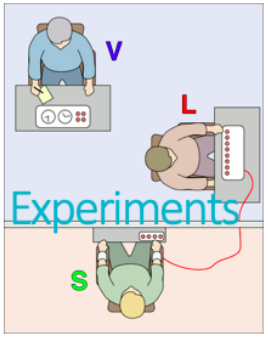
- Data controlled by statistical agencies
- Title 26
- Title 13
- CIPSEA
- Twin pillars of anonymization and consent

Informed Consent...But



Real value is in combined data products.

Problems:
Access
Consent



Some Linkage Consent Research Results

- Opt-in vs. opt-out wording
- Gain vs. loss framing
- Front vs. back placement

Background: Opt-in vs. opt-out wording

- Evidence from behavioral economics and psychology: default option has strong effect (Thaler and Sunstein 2008; Schwartz 2014)
- Strong effect in linkage requests (Bates 2005; Pascale 2011)
- However: Not always clear if behavior matches intentions (Ellikson and Hawes 1989, cited in Singer 1993)
- Concerns about bias (Das and Couper 2013)

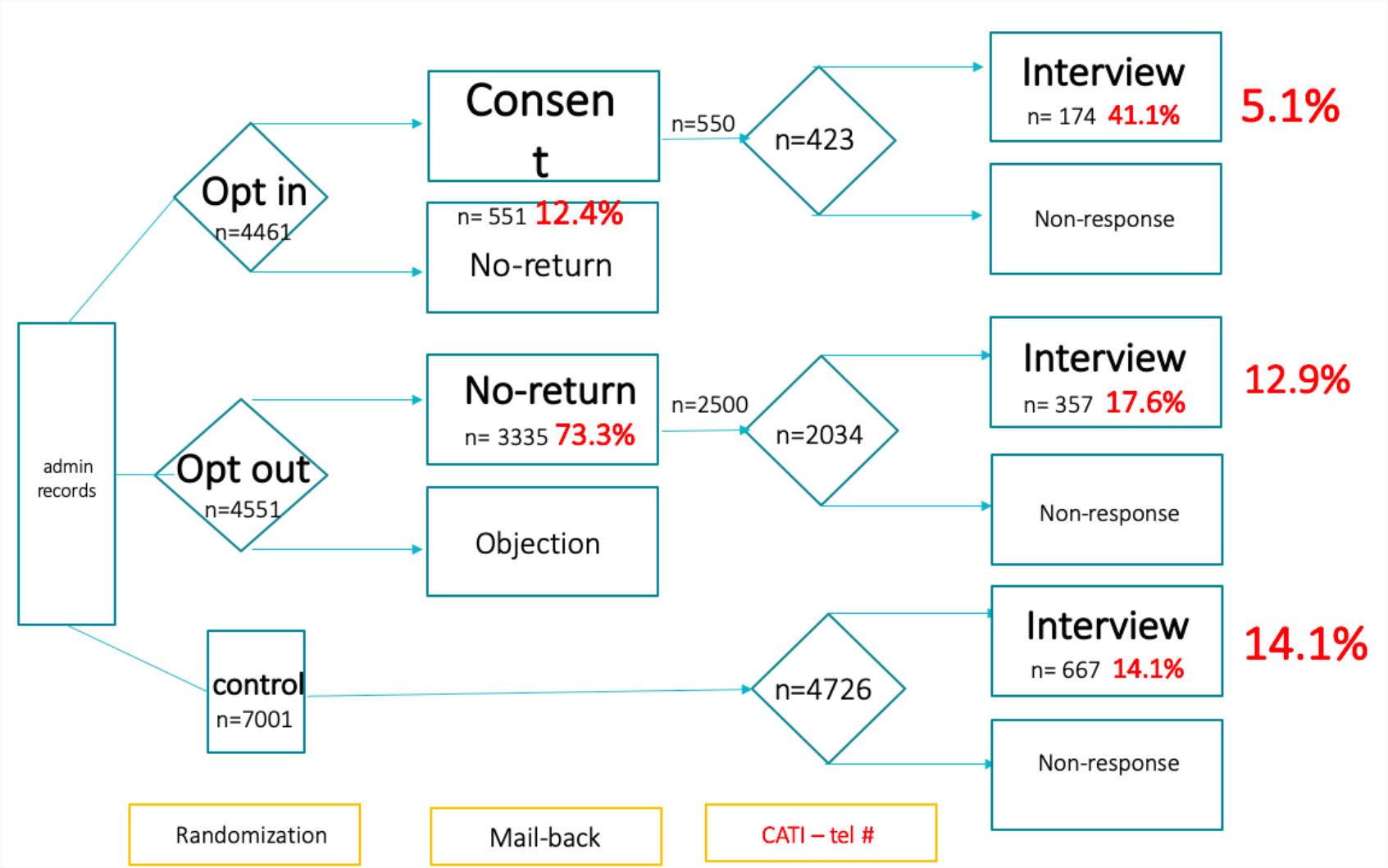
Background: Gain vs. loss framing

- Most common: Emphasis on benefit works hypothetically (Bates, Wroblewski, and Pascale, 2012), but not better in practice (Pascale, 2011; Sakshaug et al. 2013)
- Why? Faced with risky choices, decision-making is influenced by framing in terms of gains or losses (Kahneman and Tversky, 1979)
- Survey context: Panel consent (Tourangeau and Ye 2009) Linkage to voting records (Kreuter et al. 2015)

Background: Front vs. back placement

- Most common: Back placement (Sakshaug, Tutz, Kreuter, 2013)
- Anecdotal rational: Rapport
- Experiments suggest higher rates when asked in front or in the context of related survey items (Sala, Knies, and Burton, 2013; Sakshaug, Tutz, and Kreuter, 2013)
- Open Question: Loss-frame vs. front placement

Flow Charts for Consenters

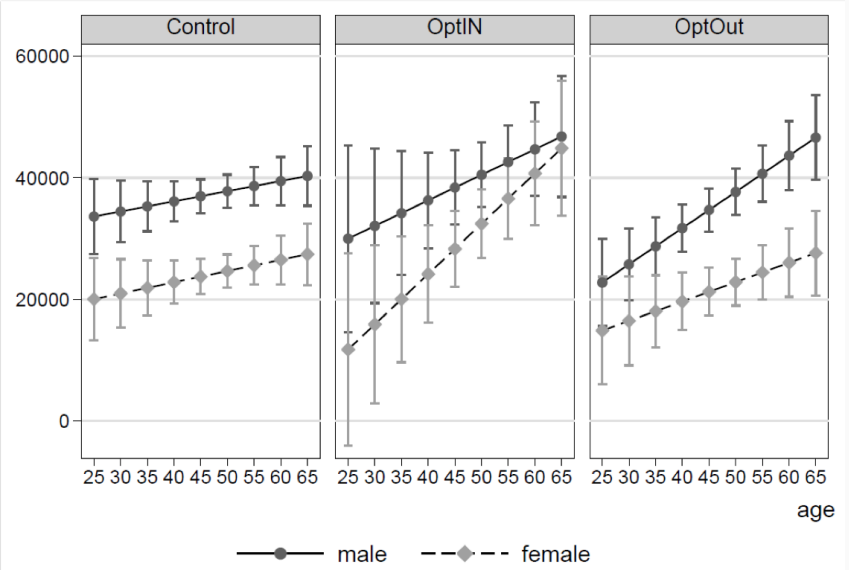


Research Questions & Results

1. Is there indication for bias in opt-in and opt-out groups compared to control?
2. How does gain vs. loss framing interfere with front vs. back?
3. Do people understand what they consent to? Does gain vs. loss framing affect understanding of linkage consent?

Research Questions & Results

1. Is there indication for bias in opt-in and opt-out groups compared to control?
2. How does gain vs. loss framing interfere with front vs. back?
3. Do people understand what they consent to? Does gain vs. loss framing affect understanding of linkage consent?



Regression Income (Euros) on age and gender

Research Questions & Results

1. Is there indication for bias in opt-in and opt-out groups compared to control?
2. How does gain vs. loss framing interfere with front vs. back?
3. Do people understand what they consent to? Does gain vs. loss framing affect understanding of linkage consent?

Research Questions & Results

1. Is there indication for bias in opt-in and opt-out groups compared to control?
2. How does gain vs. loss framing interfere with front vs. back?
3. Do people understand what they consent to? Does gain vs. loss framing affect understanding of linkage consent?

Placement and Framing (in German)

- Gain frame:
  - Front: *Die Informationen, die Sie **uns im Laufe** des Interviews **geben werden**, sind **nützlicher**, wenn Sie dem Zusammenspielen mit den Daten der Bundesagentur zustimmen.Sind Sie mit der Zuspielung der Informationen einverstanden?*
  - Back: *Die Informationen, die Sie **uns im Laufe** des Interviews **gegeben haben**, sind **nützlicher**, wenn Sie dem Zusammenspielen mit den Daten der Bundesagentur zustimmen. Sind Sie mit der Zuspielung der Informationen einverstanden?*
- Loss frame:
  - Front: *Leider sind die Informationen, die Sie uns im Laufe des Interviews **geben werden, weniger nützlich,** wenn Sie dem Zusammenspielen mit den Daten der Bundesagentur nicht zustimmen*
  - Back: *Leider sind die Informationen, die Sie uns im Laufe des Interviews **gegeben haben, weniger nützlich**, wenn Sie dem Zusammenspielen mit den Daten der Bundesagentur nicht zustimmen*

Consent to Linkage by Framing and Mode in %

Phone	Front	Back	Total n
Gain	90.8	78.7	598
Loss	90.5	81.2	610
Total n	613	595	1208

Phone	Front	Back	Total n
Gain	90.8	78.7	598
Loss	90.5	81.2	610
Total n	613	595	1208

Research Questions & Results

1. Is there indication for bias in opt-in and opt-out groups compared to control?
2. How does gain vs. loss framing interfere with front vs. back?
3. Do people understand what they consent to? Does gain vs. loss framing affect understanding of linkage consent?

Phone	Consenters %correct	Non-consenters %correct
Answers sent to IAB	88.3	57.8
Merged with IAB	93.3	36.7
Name/Adress saved	68.3	38.8
Result lead to you	63.4	--
IAB only access	85.6	--
Public access to identifiable data	87.5	--

Statistical Framework for Anonymization

1. Aggregated tabular data
2. Public use files
3. Licensing
4. Synthetic Data
5. Research Data Centers

Types of Access

1. Aggregated tabular data

Marital Status	**Male**	**Female**	**Total**
Married	38	17	55
Divorced	7	4	11
Single	3	**1**	4
Total	48	22	70

Types of Access

1. Aggregated tabular data

Marital Status	Male	Female	Total
Married	38	17	55
Divorced	7	4	11
Single	3	**1**	4
Total	48	22	70

Marital status/ Hours worked	Male			Female			Total
	More than 30	16-30	15 or less	More than 30	16-30	15 or less	
Married	30	6	2	14	3	**0**	55
Divorced	3	4	0	2	2	**0**	11
Single	2	0	1	0	0	**1**	4
Total	35	10	3	16	5	**1**	70

But

- These answer preposed questions
- Don’t allow analysis of marginal effects

Types of Access

2. Public use files
3. Licensing

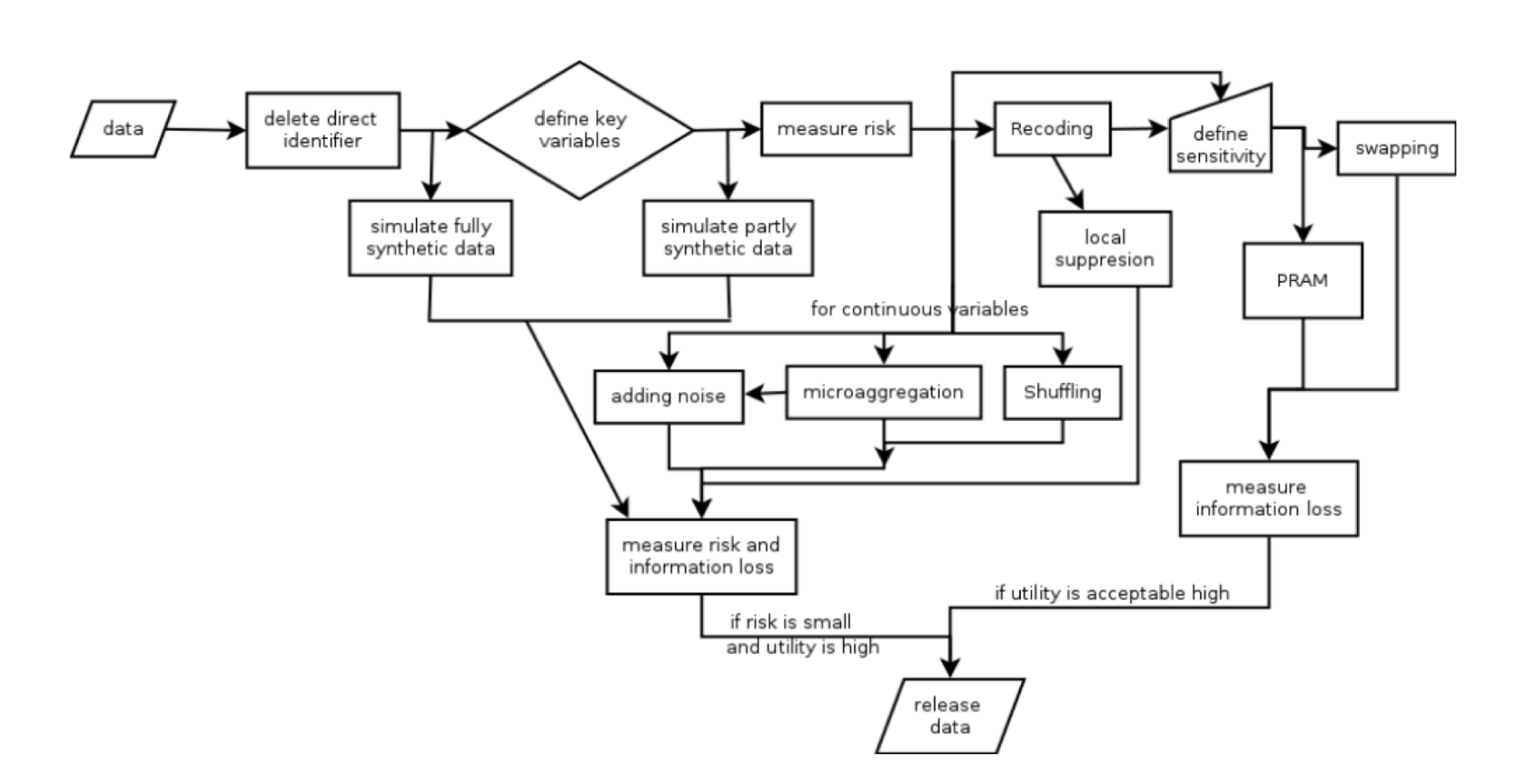
Past Ideas for Protection

Traditional approaches - tables

- cell suppression
- controlled tabular adjustment
- rounding
- cell pertubation

Traditional approaches - microdata

- local suppression
- global recoding
- top coding
- sampling
- rounding
- swapping
- added noise
- data shuffling



Introduction to Statistical Disclosure Control (SDC) Matthias Templ, Bernhard Meindl and Alexander Kowarik

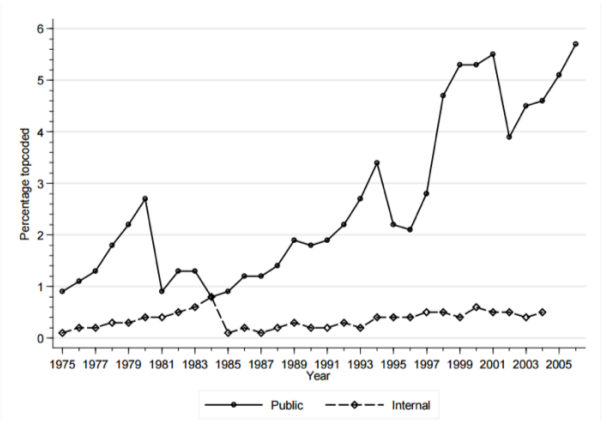
`http://www.data-analysis.at`

But

- Public use files destroy utility
- Licensing has versioning and cost challenges

Consequences

Figure 1: Percentage of Individuals with Censored Household Income in March CPS, by Year

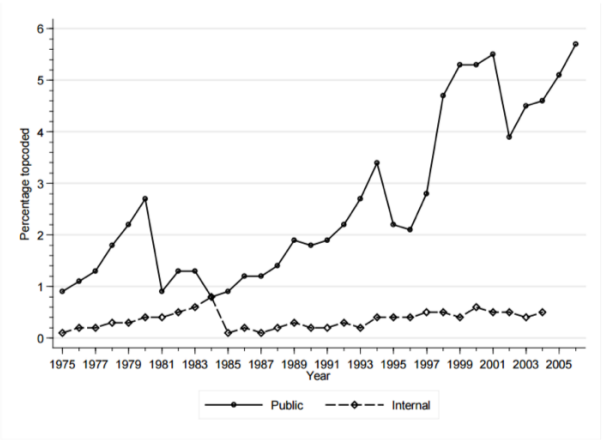


Source: authors' calculations from internal and public use data files of March CPS. Internal data were not available for years after 2004.

Burkhauser, Richard V., Feng, Shuaizhang, Jenkins, Stephen P. and Larrimore, Jeff (2010) Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring. The Journal of Economic Inequality, 9 (3). pp. 393-415 ISSN 1569- 1721 DOI: 10.1007/s10888-010-9131-6

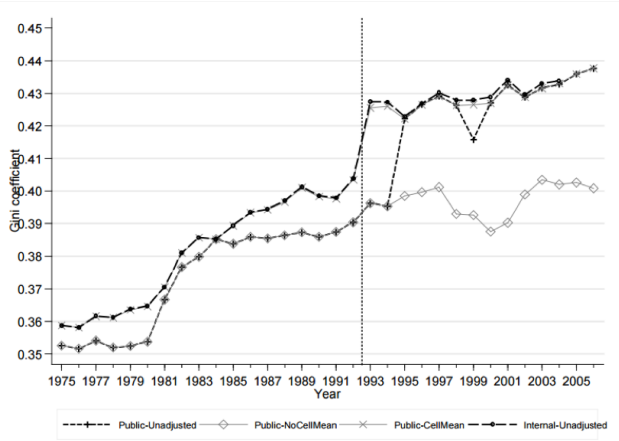
Consequences

Figure 1: Percentage of Individuals with Censored Household Income in March CPS, by Year



Source: authors' calculations from internal and public use data files of March CPS. Internal data were not available for years after 2004.

Figure 2: Gini Coefficient Estimates Derived Using Four Censoring Adjustment Methods

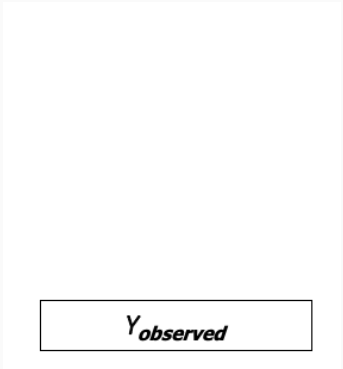


Source: authors' calculations from internal and public use data files of the March CPS. There was a major change in CPS data collection methods between 1992 and 1993. Internal data were not available for years after 2004. See text for definitions of the series.

Burkhauser, Richard V., Feng, Shuaizhang, Jenkins, Stephen P. and Larrimore, Jeff (2010) Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring. The Journal of Economic Inequality, 9 (3). pp. 393-415 ISSN 1569- 1721 DOI: 10.1007/s10888-010-9131-6

Types of Access

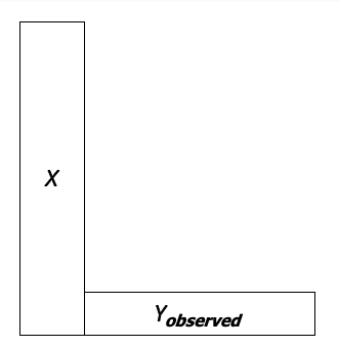
4. Synthetic Data



Drechsler

Types of Access

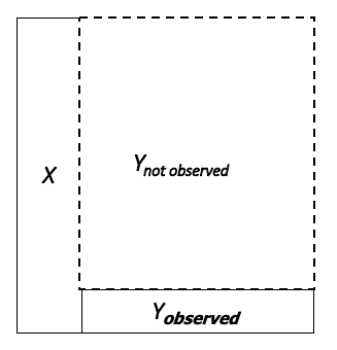
4. Synthetic Data



Drechsler

Types of Access

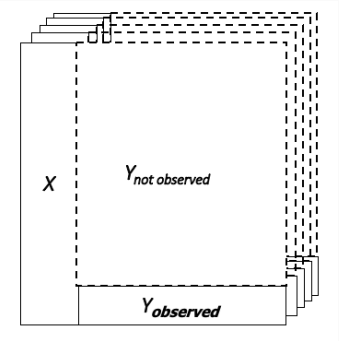
4. Synthetic Data



Drechsler

Types of Access

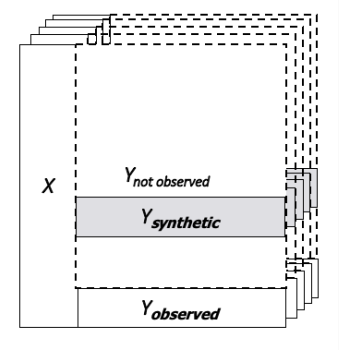
4. Synthetic Data



Drechsler

Types of Access

4. Synthetic Data



Drechsler

Synthetic Data cont'd

- Closely related to multiple imputation for nonresponse
- Generate synthetic datasets by drawing from a model fitted to the original data
- Not the missing values but the sensitive values are replaced with a set of plausible values given the original data
- Generate multiple draws to be able to obtain valid variance estimates from the synthetic data

Drechsler

Synthetic Data cont'd

- three steps necessary for data release:
  - fit model to the original data
  - repeatedly draw from that model to generate multiple synthetic datasets
  - release these datasets to the public
- over the years different designs for generating synthetic data evolved
- two main approaches: **fully synthetic** datasets and **partially synthetic** datasets
- only partially synthetic datasets have been released so far

Drechsler

Fully synthetic

- goes back to Rubin (1993)
- a useful SDC method should fulfil three goals
  - preserve confidentiality
  - maintain valid inferences
  - allow the user to rely on standard statistical software
- masking techniques very popular at that time
- can fulfil the first two goals in certain settings
- Rubin criticises masking as an approach to protect confidentiality

Drechsler

Fully Synthetic Data cont'd

- requires special software to obtain valid inferences
- requires complicated error-in-variables models
- no special software will be developed for each *analysis method x masking method x database type*
- users have their own science to worry about
- shouldn't be expected to become experts in demasking programs

Fully Synthetic Data cont'd

- Rubin suggests an alternative approach for releasing confidential microdata
- instead of applying masking procedures, completely synthetic data should be released
- approach is based on the ideas of multiple imputation
- all units that did not participate in the survey are treated as missing data
- missing data are multiply imputed
- samples from the generated synthetic populations are released to the public

Implications

- implications for four groups
  - survey units
  - data snoopers
  - legitimate date users
  - data producers
- implications for the survey units
  - no confidential data would be released
  - hopefully survey units will be more likely to respond and to respond truthfully
- implications for the data snooper
  - the fact that data are synthetic will be well documented
  - knowing this should destroy intruder's interest in snooping at the individual level

Implications cont'd

- implications for the legitimate data user
  - is not interested in any individual's data
  - microdata only to allow analyses for population estimands using straightforward statistical tools
  - data should be analyzable using the full range of standard complete-data statistical tools
  - valid inferences should be easy to obtain
  - the proposed synthetic data fulfils these requirements, but maybe not for all inferences
- implications for the data producer
  - proposal requires a heavy investment to generate synthetic data
  - data producer has the knowledge and resources to do that
  - unrealistic to expect a similar investment from the user
  - resources should be allocated where the problem can be solved

Pros and Cons

advantages of the approach

- data are fully synthetic
- re-identification of single units almost impossible
- all variables are still fully available
- valid inferences can be obtained using simple combining rules

But

disadvantages of the approach

- strong dependence on the imputation model
- setting up a model might be difficult/impossible

not always necessary to synthesize all variables

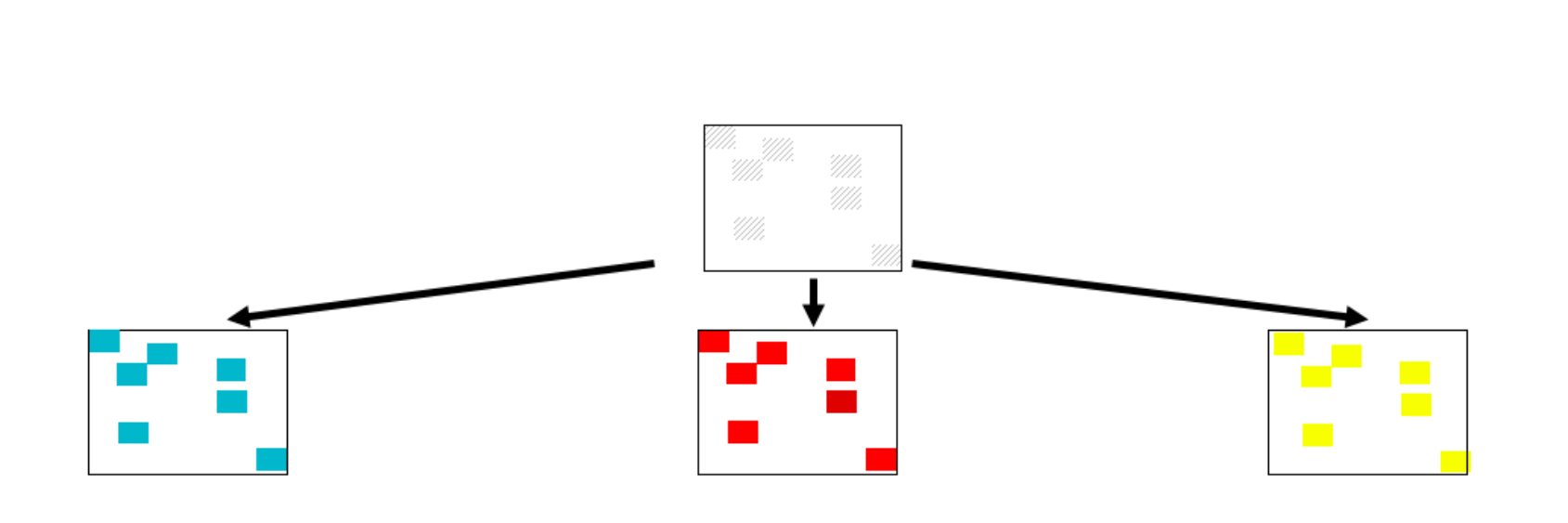
alternative: partially synthetic data

Drechsler

Multiple Imputation for Nonresponse and Confidentiality

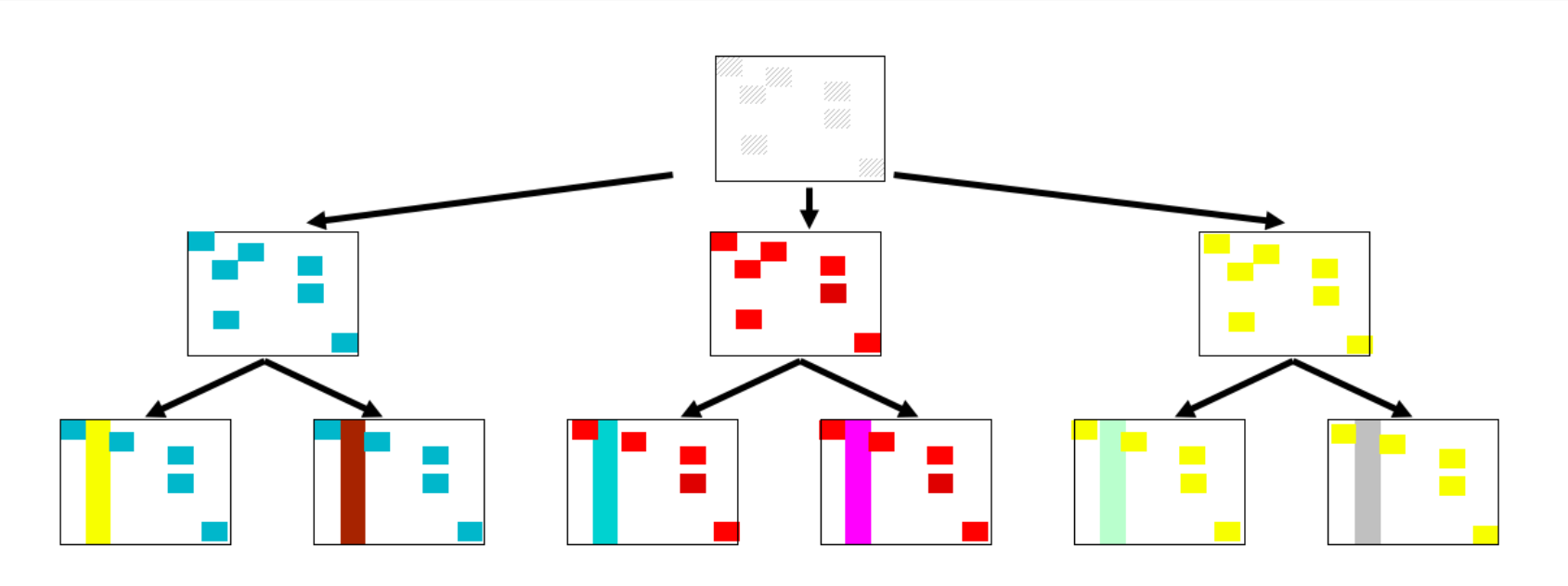
Drechsler

Multiple Imputation for Nonresponse and Confidentiality



Drechsler

Multiple Imputation for Nonresponse and Confidentiality



Drechsler

Synthetic Data Compared to Other SDC Techniques

advantages

- tries to preserve the multivariate relationship between the variables and not only specific statistics
- suitable for any variable type
- most SDC methods cannot address some of the problems typically encountered in practice
  - item nonresponse
  - skip patterns
  - logical constraints

disadvantage

- lot of work
- depends heavily on the quality of the imputation models
- User understanding?

Big data challenges

In Big Data Era

- Most data **no longer collected** by the government (internet search logs, Twitter, supermarket scanners...)
- Question **how to share collected information** without violating privacy guarantees becomes more relevant
- three famous privacy breaches stimulate the discussions on data confidentiality
    - identification of a city mayor in “anonymised” medical records in the USA
    - A Face Is Exposed for AOL Searcher No. 4417749
    - Netflix Spilled Your Brokeback Mountain Secret

What is disclosure

- Identity disclosure linkage with external available data
- Attribute disclosure
- Inferential disclosure

The Massachusetts Group Insurance Commission had a bright idea back in the mid-1990s—it decided to release "anonymized" data on state employees that showed every single hospital visit. The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number. But a graduate student in computer science saw a chance to make a point about the limits of anonymization.

Latanya Sweeney requested a copy of the data and went to work on her "reidentification" quest. It didn't prove difficult. Law professor Paul Ohm describes Sweeney's work:

> At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.

Boom! But it was only an early mile marker in Sweeney's career; in 2000, she showed that 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex.

What is disclosure

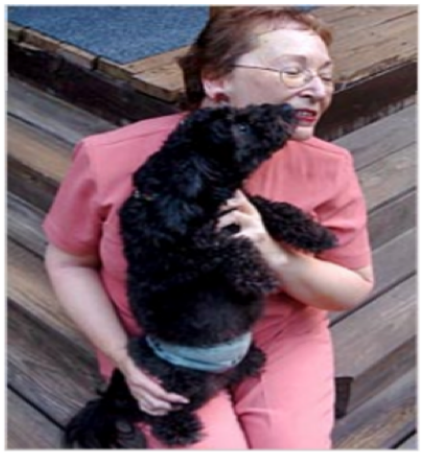
- Identity disclosure linkage with external available data
- Attribute disclosure
- Inferential disclosure

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

E-MAIL
PRINT
REPRINTS

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

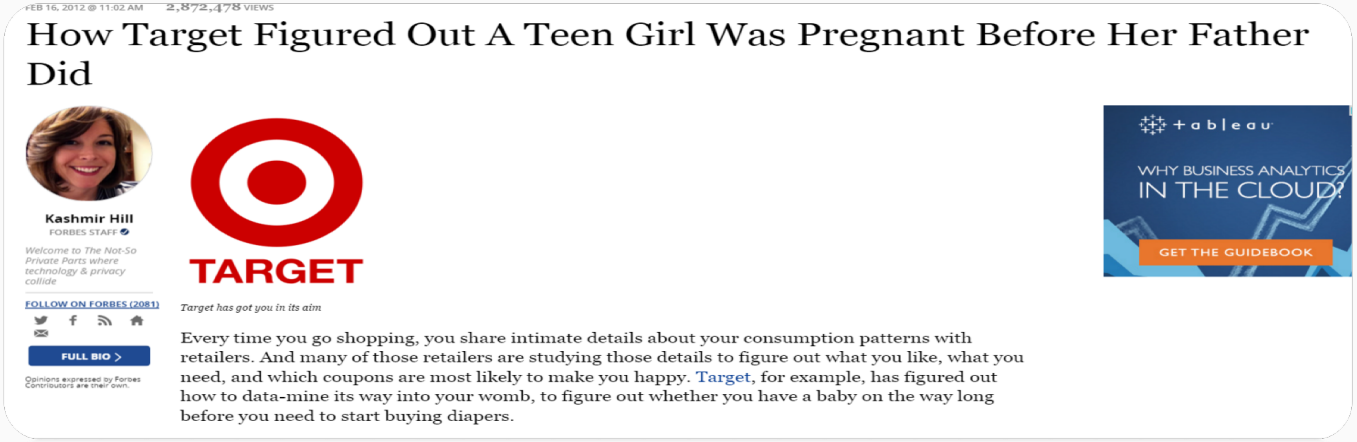
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

What is disclosure

- Identity disclosure linkage with external available data
- Attribute disclosure
- Inferential disclosure



FEB 16, 2012 @ 11:02 AM 2,872,478 VIEWS

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill
FORBES STAFF

Welcome to The Not-So Private Parts where technology & privacy collide

FOLLOW ON FORBES (2081)

FULL BIO >

Opinions expressed by Forbes Contributors are their own.

TARGET

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

What is disclosure

- Identity disclosure linkage with external available data
- Attribute disclosure
- Inferential disclosure

> "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"
>
> The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

(Nice customer service, Target.)

> On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

Target's Andrew Pole (from LinkedIn)

What Target discovered fairly quickly is that it creeped people out that the company knew about their pregnancies in advance.

> "If we send someone a catalog and say, 'Congratulations on your first child!' and they've never told us they're pregnant, that's going to make some people uncomfortable," Pole told me. "We are very conservative about compliance with all privacy laws. **But even if you're following the law, you can do things where people get queasy.**"

Bold is mine. That's a quote for our times.

Additional Problems

- What is the legal framework when the ownership of data is unclear?

  Collection and analysis often no longer within same entity. Ownership of data less clear.

- Who has the legal authority to make decisions about permission, access and dissemination and under what circumstances?

  The challenge in the case of big data is that data sources are often combined, collected for one purpose and used for another and users often have no good understanding of it or how their data will be used.

=> Concepts Out of Date

- notification is either comprehensive or comprehensible, but not both.

  (Nissenbaum 2011)
- Understanding of the nature of harm has diffused over time..
- Consumers value their own privacy in variously flawed ways.

  (Acquisti 2014)

Summary: Tensions

- Privacy vs. Public Good
- Convenience vs. Accusation

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Tensions cont'd

- Privacy vs. Public Good
- Convenience vs. Accusation
- Identifiable vs. Reachable
- "Tyranny of Minority"

Tensions cont'd

- Privacy vs. Public Good
- Convenience vs. Accusation

- Identifiable vs. Reachable
- "Tyranny of Minority"

- Consent vs. Confusion
- Privacy vs. Data quality

- Europe vs. USA
- "Big Brother" vs. "Big Mother"

Differential Privacy

- proposed by Dwork (2006)
- received a lot of attention in recent years
- only mechanism that ensures formal privacy guarantees

- designed as a query response system
- no microdata should be released
- direct relationship to remote access

Formal Privacy Guarantees

- difficult to ensure formal guarantees
- with background knowledge information about individual could be revealed even if she is not in the database

Example queries:

1. the number of members of the U.S. House of Representatives with the sickle cell trait
2. the number of members of the House of Representatives, other than the Speaker of the House, with the sickle cell trait.

“How many people with the following identifying characteristics [description of Aunt Wilma and only Aunt Wilma] have had at least three pregnancies?”

Question

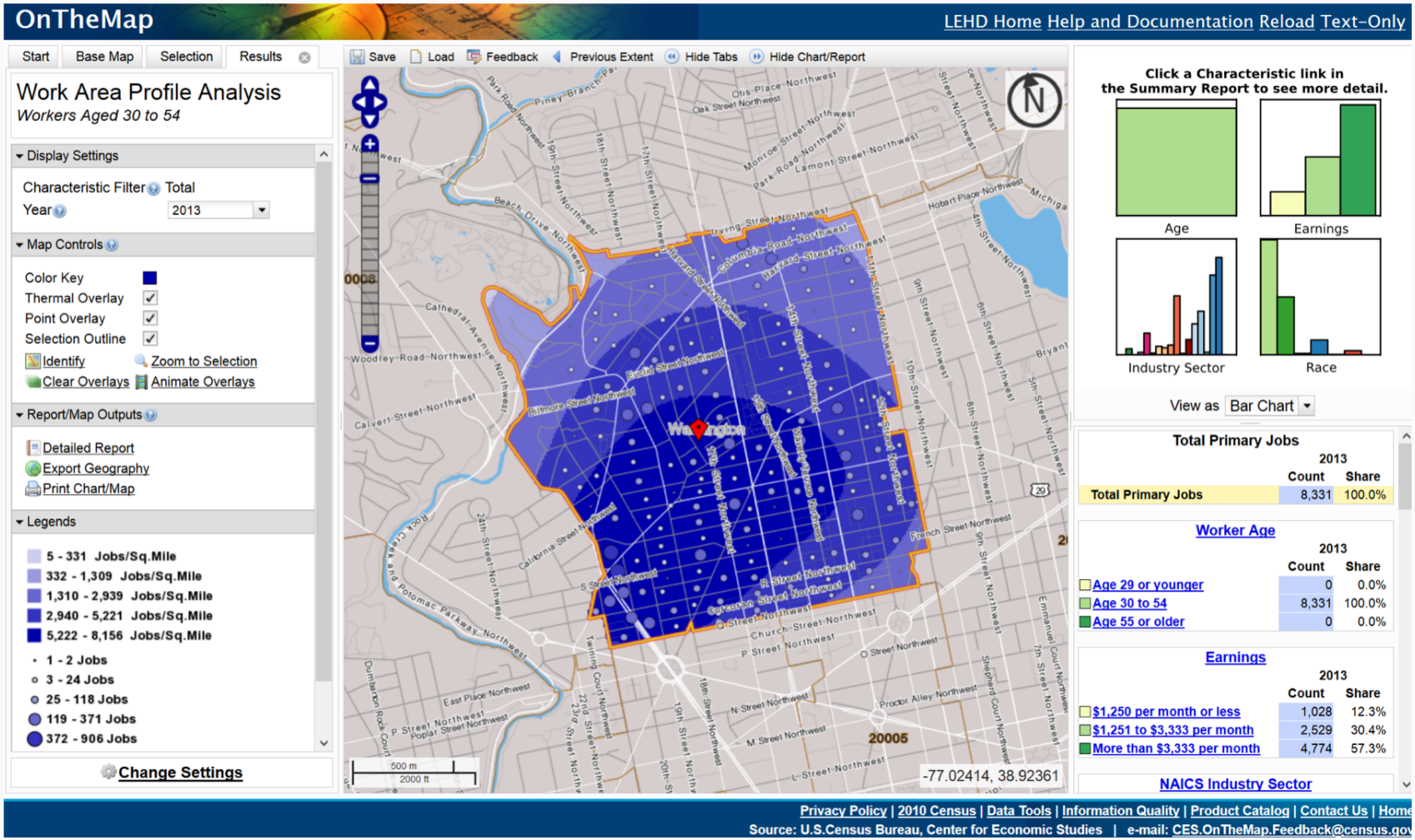
Can introducing small inaccuracies into the query responses protect the family's privacy?

- The degree to which small distortions can protect against arbitrary counting query sequences depends on the size of 'small' compared to the number of queries.
- The general form of the bound is: if the magnitudes of the errors are all bounded by E, then at least k - 4E2 bits can be correctly reconstructed (Dwork and Yekhanin 2008).

Fundamental Problem with Learning Useful Things

"If the database teaches that smoking causes cancer, the bad (pays higher insurance premiums) and good (joins smoking cessation program) consequences for an individual smoker will be incurred independent of whether or not the particular smoker is in the database." (Dwork 2014)

- **differential privacy**: privacy is ensured through randomization (think randomized response in the survey context "Did you ingest a controlled substance in the past week?" "Is your mother born in December?")
- ensures that the addition or removal of a single row in the database has (almost) no impact on the results of interest (think sickle cell query above)
- implies that individuals can participate without risk


OnTheMap
LEHD Home Help and Documentation Reload Text-Only
Start
Base Map
Selection
Results
Save
Load
Feedback
Previous Extent
Hide Tabs
Hide Chart/Report
Work Area Profile Analysis
Workers Aged 30 to 54
Display Settings
Characteristic Filter Total
Year 2013
Map Controls
Color Key
Thermal Overlay
Point Overlay
Selection Outline
Identify
Zoom to Selection
Clear Overlays
Animate Overlays
Report/Map Outputs
Detailed Report
Export Geography
Print Chart/Map
Legends
5 - 331 Jobs/Sq.Mile
332 - 1,309 Jobs/Sq.Mile
1,310 - 2,939 Jobs/Sq.Mile
2,940 - 5,221 Jobs/Sq.Mile
5,222 - 8,156 Jobs/Sq.Mile
1 - 2 Jobs
3 - 24 Jobs
25 - 118 Jobs
119 - 371 Jobs
372 - 906 Jobs
Change Settings
Click a Characteristic link in the Summary Report to see more detail.
Age
Earnings
Industry Sector
Race
View as Bar Chart
Total Primary Jobs
2013
Count
Share
Total Primary Jobs 8,331 100.0%
Worker Age
Age 29 or younger 0 0.0%
Age 30 to 54 8,331 100.0%
Age 55 or older 0 0.0%
Earnings
$1,250 per month or less 1,028 12.3%
$1,251 to $3,333 per month 2,529 30.4%
More than $3,333 per month 4,774 57.3%
NAICS Industry Sector
-77.02414, 38.92361
Privacy Policy | 2010 Census | Data Tools | Information Quality | Product Catalog | Contact Us | Home
Source: U.S.Census Bureau, Center for Economic Studies | e-mail: CES.OnTheMap.Feedback@census.gov

Why Differential Privacy is so Attractive

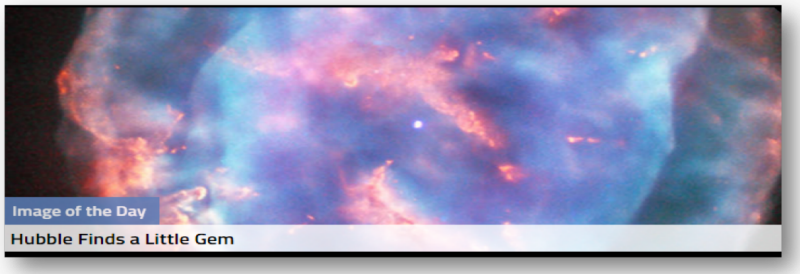
- only concept that offers formal privacy guarantees
- independent of the data
- doesn't make any assumption about background knowledge
- no ex-post risk assessment required
- privacy guarantees still hold even if other data sources are published later

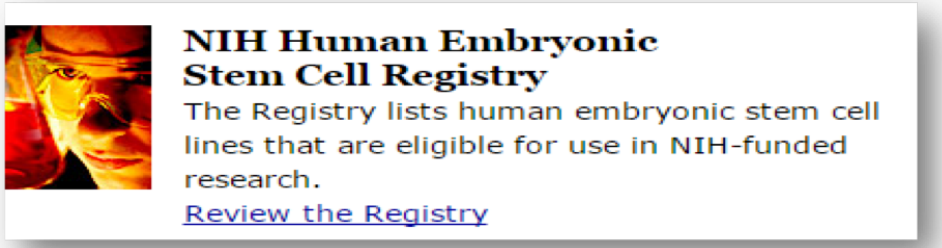
But

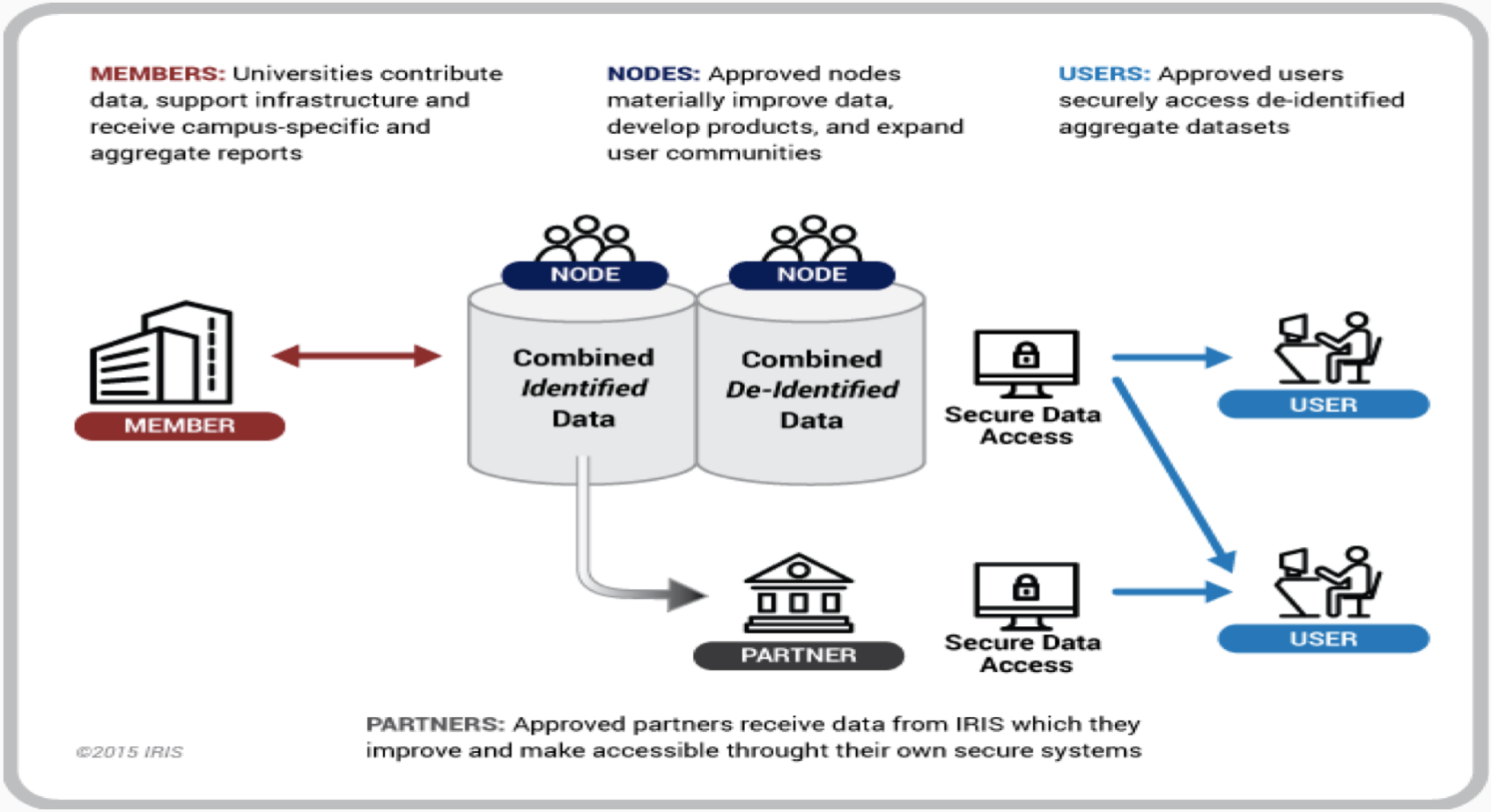
- no interpretation for ε
- which level of ε is acceptable
- very conservative assumption about intruder knowledge
- intruder knows every record except one
- strong requirements for the protection mechanism
- large amount of noise needs to be added
- results that fulfill DP often not useful in practice
- especially true of differentially private microdata should be generated

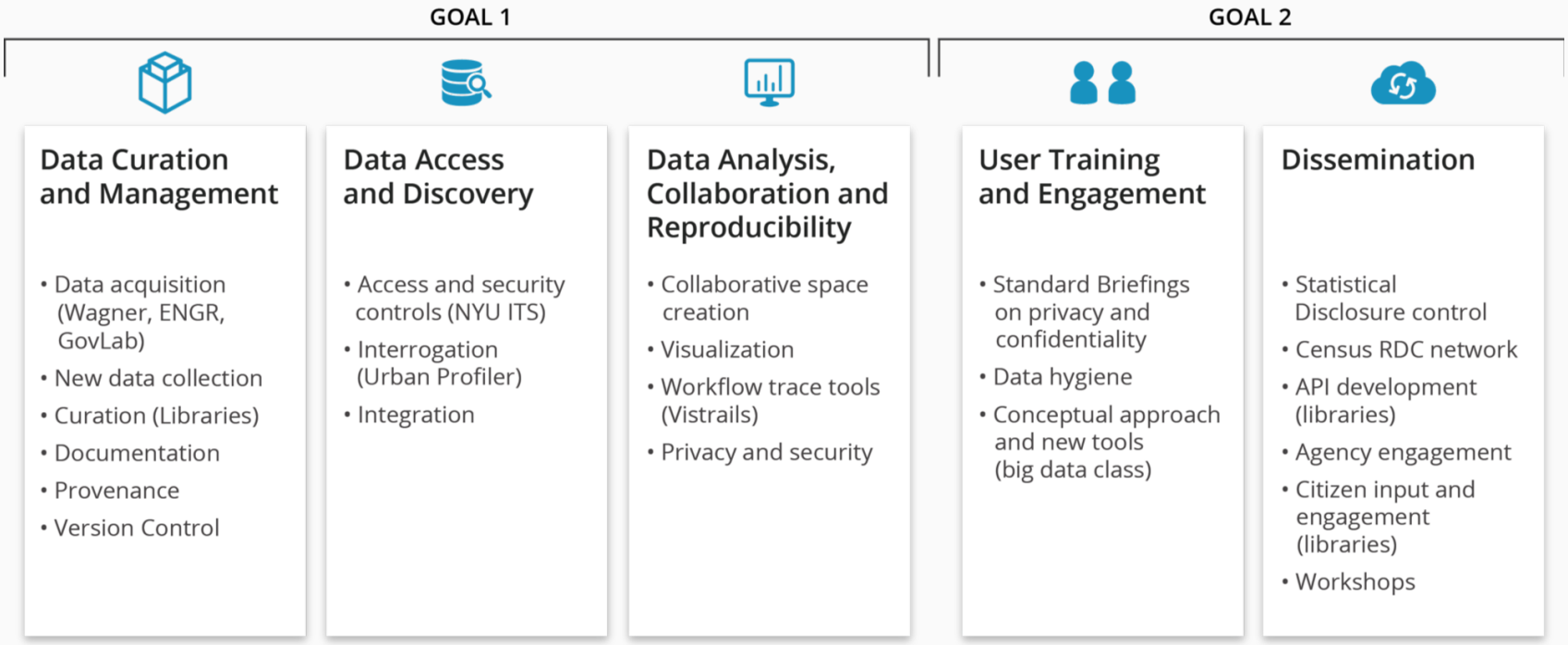
The future

New Approach

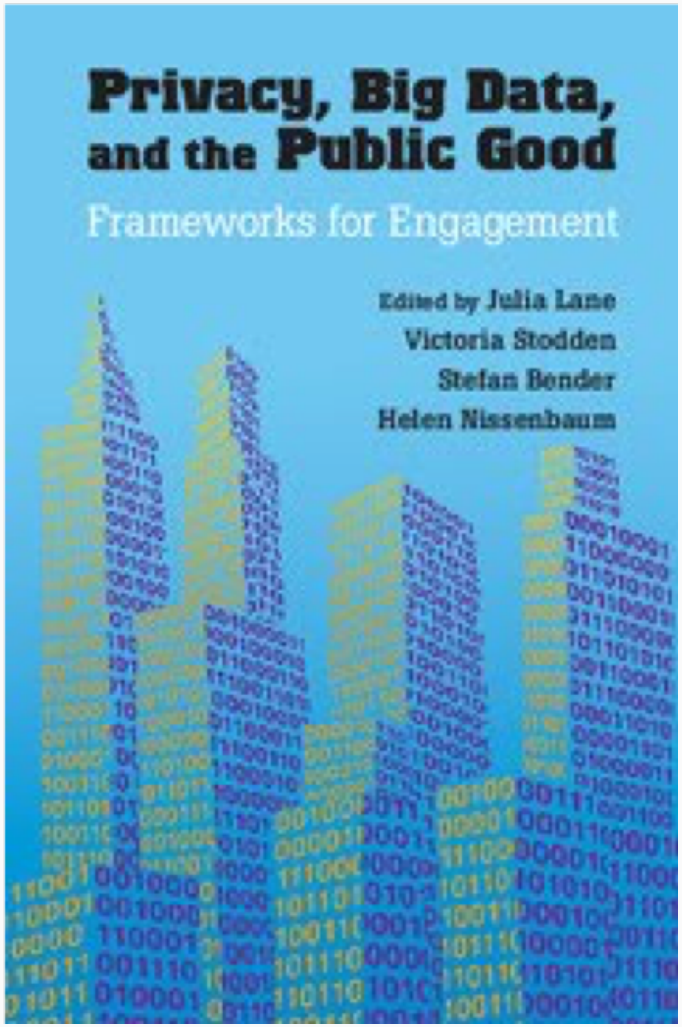





MEMBERS: Universities contribute data, support infrastructure and receive campus-specific and aggregate reports
NODES: Approved nodes materially improve data, develop products, and expand user communities
USERS: Approved users securely access de-identified aggregate datasets
NODE
NODE
MEMBER
Combined Identified Data
Combined De-Identified Data
Secure Data Access
USER
PARTNER
Secure Data Access
USER
©2015 IRIS
PARTNERS: Approved partners receive data from IRIS which they improve and make accessible throught their own secure systems


GOAL 1
GOAL 2
Data Curation and Management
• Data acquisition (Wagner, ENGR, GovLab)
• New data collection
• Curation (Libraries)
• Documentation
• Provenance
• Version Control
Data Access and Discovery
• Access and security controls (NYU ITS)
• Interrogation (Urban Profiler)
• Integration
Data Analysis, Collaboration and Reproducibility
• Collaborative space creation
• Visualization
• Workflow trace tools (Vistrails)
• Privacy and security
User Training and Engagement
• Standard Briefings on privacy and confidentiality
• Data hygiene
• Conceptual approach and new tools (big data class)
Dissemination
• Statistical Disclosure control
• Census RDC network
• API development (libraries)
• Agency engagement
• Citizen input and engagement (libraries)
• Workshops

Must Read



- What are the legal requirements?
- What are the rules of engagement?
- What are the best ways to provide access while also protecting confidentiality?
- Are there reasonable mechanisms to compensate citizens for privacy loss?
- How can we built trustworthy curators?
- What do we (need to) know about the data generating process?
- How can we increase linkage without increasing risks?

Thank you

References

Acquisti, A. (2014). The economics and behavioral economics of privacy. *Privacy, big data, and the public good: Frameworks for engagement*, *1*, 76-95.

Bates, A. T. (2005). Technology, e-learning and distance education. Routledge.

Bates, N., Wroblewski, M. J., & Pascale, J. (2012). Public Attitudes toward the Use of Administrative Records in the US Census: Does Question Frame Matter?. *Survey Methodology*, 04.

Burkhauser, R. V., Feng, S., Jenkins, S. P. & Larrimore, J. (2010). Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring. *The Journal of Economic Inequality, 9* (3). 393-415

Drechsler, J. (2011). Synthetic datasets for statistical disclosure control: theory and implementation (Vol. 201). Springer Science & Business Media.

References

Dwork C. (2006). Differential Privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) *Automata, Languages and Programming*. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Berlin: Springer.

Dwork, C., & Yekhanin, S. (2008). New efficient attacks on statistical disclosure control mechanisms. *Annual International Cryptology Conference*. 469-480

Dwork, C. (2014). 14 Differential Privacy: A Cryptographic Approach to Private Data Analysis. *Privacy, big data, and the public good: Frameworks for engagement*

Ellickson, P. L., & Hawes, J. A. (1989). An assessment of active versus passive methods for obtaining parental consent. *Evaluation Review, 13*(1), 45-55.

Kreuter, F., Sakshaug, J. W., & Tourangeau, R. (2015). The framing of the record linkage consent question. *International Journal of Public Opinion Research, 28*(1), 142-152.

McCallister, E., Grance, T. & Scarfone, K. A. (2010). Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*NIST SP 800-122*

References

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4), 32-48.

Sakshaug, J., Tutz, V., & Kreuter, F. (2013). Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data. *Survey Research Methods, 7* (2), 133-144.

Templ, M., Meindl, B. & Kowarik, A. (2017). Introduction to Statistical Disclosure Control (SDC). http://www.data-analysis.at

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: improving decisions about health. Wealth, and Happiness, 6.

Tourangeau, R., & Ye, C. (2009). The framing of the survey request and panel attrition. *Public Opinion Quarterly, 73*(2), 338-348.