SURV 622/SURVMETH 622 Privacy & Confidentiality II

February 17, 2025

James Wagner

Disclosure, Risks, and Control (continued)

Statistical Disclosure Control

- Statistical disclosure control for household **microdata** (business microdata generally not releasable)
  - Removing obvious personal identifiers such as name, address, or SSN
  - Suppressing detail about variables such as geography, age, or date of birth
  - Top-coding variables related to income or wealth
  - Adding noise or swapping values across households
- Statistical disclosure control for **tabular** data
  - Cell suppression
  - Adding noise or swapping records in microdata before tabulating
  - Addition of noise to cell values
- **Privacy-utility trade-off**

Examples

Traditional approaches-tables

- Cell suppression
- Controlled tabular adjustment
- Rounding
- Cell perturbation

Traditional approaches-microdata

- Local suppression
- Global recoding
- Top coding
- Sampling
- Rounding
- Swapping
- Added noise
- Data shuffling
- …

Risk Assessment

- Determine *whether* **indirect** identifying variables put respondents at risk of disclosure
  - Typically use the “rule of 3”, where each combination of indirect identifying variables found in the data has has size 3
  - May be increased for more sensitive surveys
- Determine *which* **variables** put respondents at risk
  - Which categories of variables have very small sample sizes, either by themselves or in combination with other variables
- Both of these questions can be answered using extensive cross-tabulations
- Can target disclosure control solutions towards problematic variables

Specific Examples

- Top-coding
  - Upper limit on values of a given variable, all cases above a certain part of the distribution are placed into one single category
  - Mean-corrected topcoding: choose the value for topcoded cells such that the mean of the distribution is correct
- Noise addition
  - Multiplying or adding a stochastic or randomized number
  - Multiplicative noise: generating random numbers with mean=1
  - Differential Privacy

Specific Examples

- Grouping, aggregating
  - Geographic population thresholds
  - Sensitive variables (nationality)
- Rounding
- Data Swapping
  - Introduce uncertainty, does not change the marginal distribution
  - Can distort joint distributions of swapped and unswapped variables
  - Data intruder cannot be sure whether observations contain true responses or not

Specific Examples

- Synthetic Data
    - Using the private survey data, estimate parameters for the joint distribution between each variable
    - Generate 1 or more data sets based on the joint distribution
    - Can be difficult for more complicated surveys
    - Can result in incorrect conclusions if the joint distribution not specified correctly
- Partially synthetic data
    - Randomly generate specific variables or observations
    - Randomly generate specific variables or observations with a given (non-constant) probability

Statistical Disclosure Risks: Tabular Data

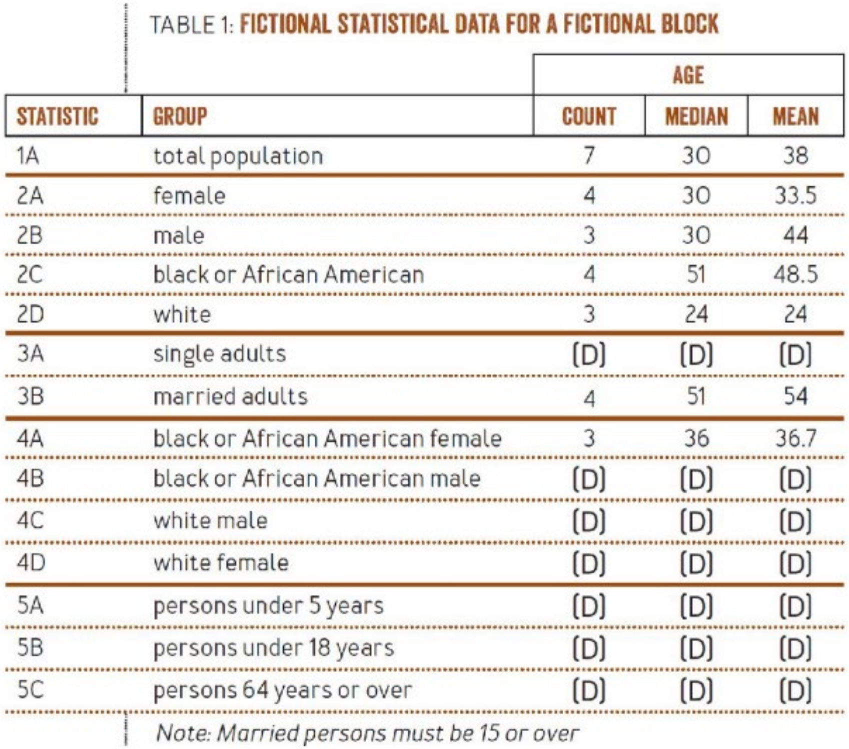


TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

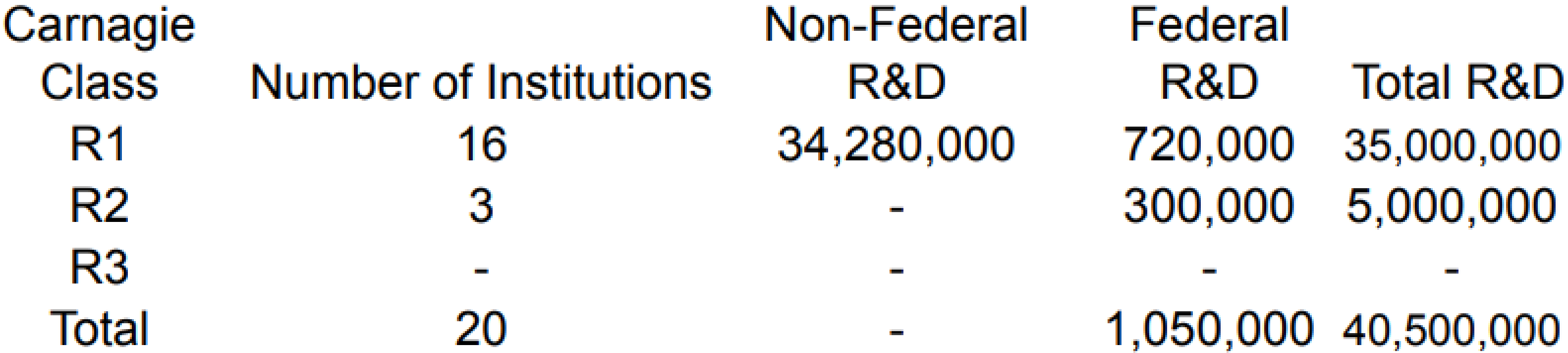
		AGE		
STATISTIC	**GROUP**	**COUNT**	**MEDIAN**	**MEAN**
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American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

- Table at left contains fictional data for a block with 7 residents
- Table implies a set of constraints on the values in individual records
    - For example, in the row for males, there are only 30 possible combinations of three ages between 1 and 125 that have a median of 30 and a mean of 44
- Combination of all implied constraints can be solved for the possible values of age, sex, race, and marital status that are consistent with the tabular data

Statistical Disclosure Risks: Secondary Disclosure

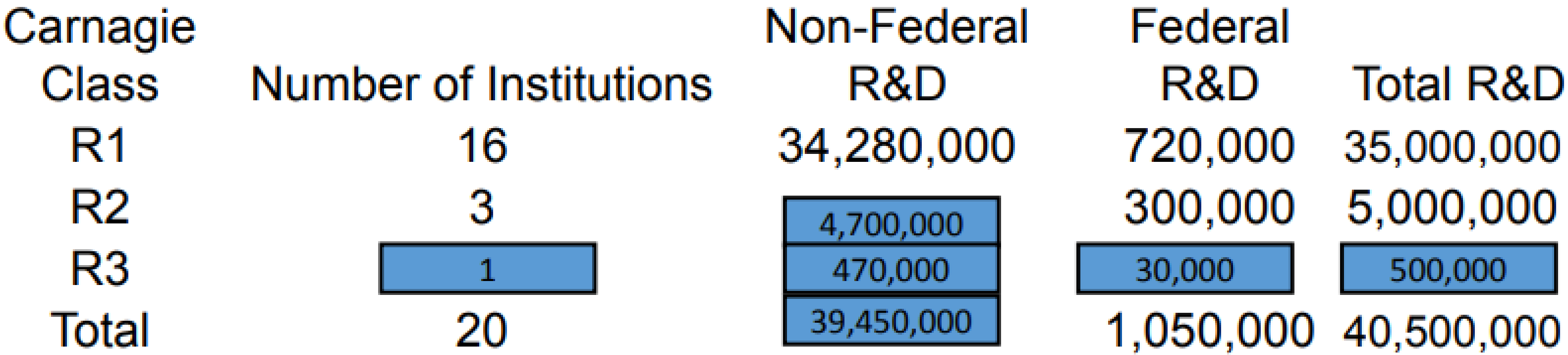
Hiding all non-zero values below the threshold doesn’t necessarily make a table safe…



Carnagie		Non-Federal	Federal	
Class	Number of Institutions	R&D	R&D	Total R&D
R1	16	34,280,000	720,000	35,000,000
R2	3	-	300,000	5,000,000
R3	-	-	-	-
Total	20	-	1,050,000	40,500,000

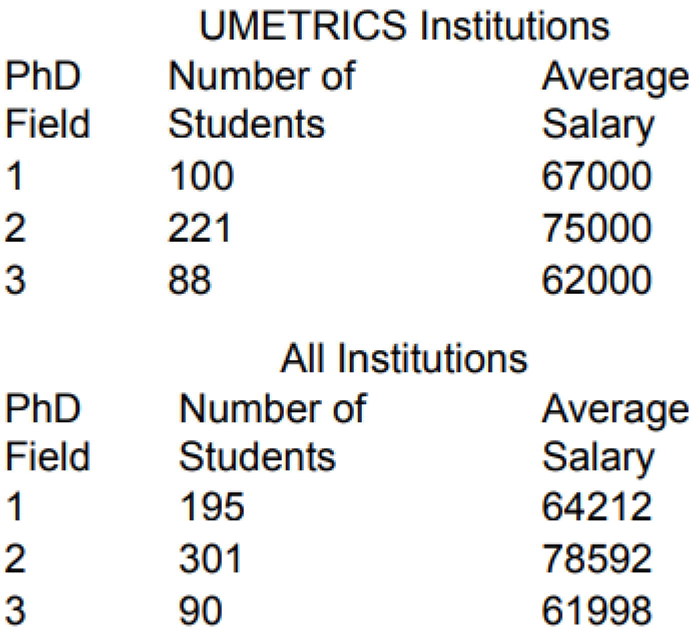
Statistical Disclosure Risks: Secondary Disclosure

Hiding all non-zero values below the threshold doesn’t necessarily make a table safe…



Carnagie Class	Number of Institutions	Non-Federal R&D	Federal R&D	Total R&D
R1	16	34,280,000	720,000	35,000,000
R2	3	4,700,000	300,000	5,000,000
R3	1	470,000	30,000	500,000
Total	20	39,450,000	1,050,000	40,500,000

Statistical Disclosure Risks:
osure



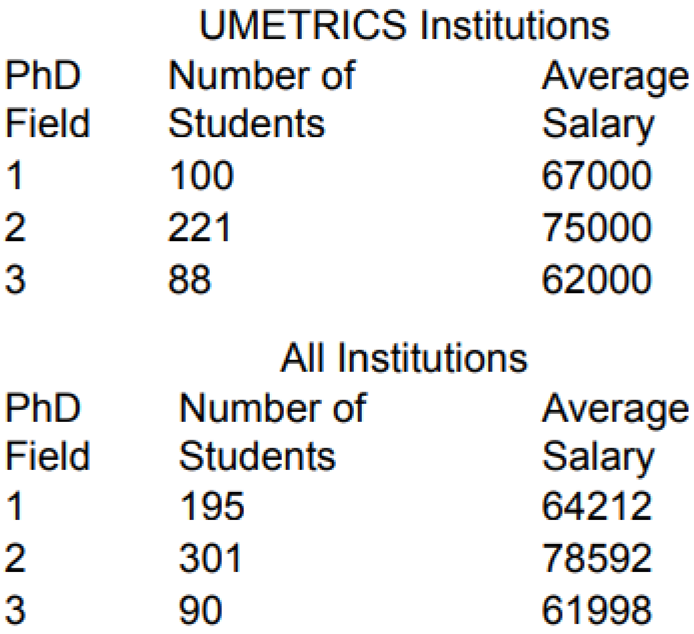
UMETRICS Institutions

PhD Field	Number of Students	Average Salary
1	100	67000
2	221	75000
3	88	62000

All Institutions

PhD Field	Number of Students	Average Salary
1	195	64212
2	301	78592
3	90	61998

Statistical Disclosure Risks: Secondary Disclosure



UMETRICS Institutions

PhD Field	Number of Students	Average Salary
1	100	67000
2	221	75000
3	88	62000

All Institutions

PhD Field	Number of Students	Average Salary
1	195	64212
2	301	78592
3	90	61998

2 students are in PhD field 3 and are not attending UMETRICS institutions

Limitations of Older Statistical Disclosure Control Methodologies

- Because statistical agencies cannot reveal exactly what they have done to the data to limit statistical disclosure, there is a risk that analysts will draw erroneous conclusions
    - Example: In microdata, adding noise to a variable (e.g., age) can distort the relationships between age and other variables in the data sets, but if noise mechanism is not known analyst cannot take distortion into account
    - Example: In tabular data, if rules call for cells that are too heavily influenced by a few values to be suppressed, may lose meaningful variation
- Methods are not provably private
    - Cannot quantify degree to which confidentiality of individuals information has been protected

Differential Privacy

Differential Privacy

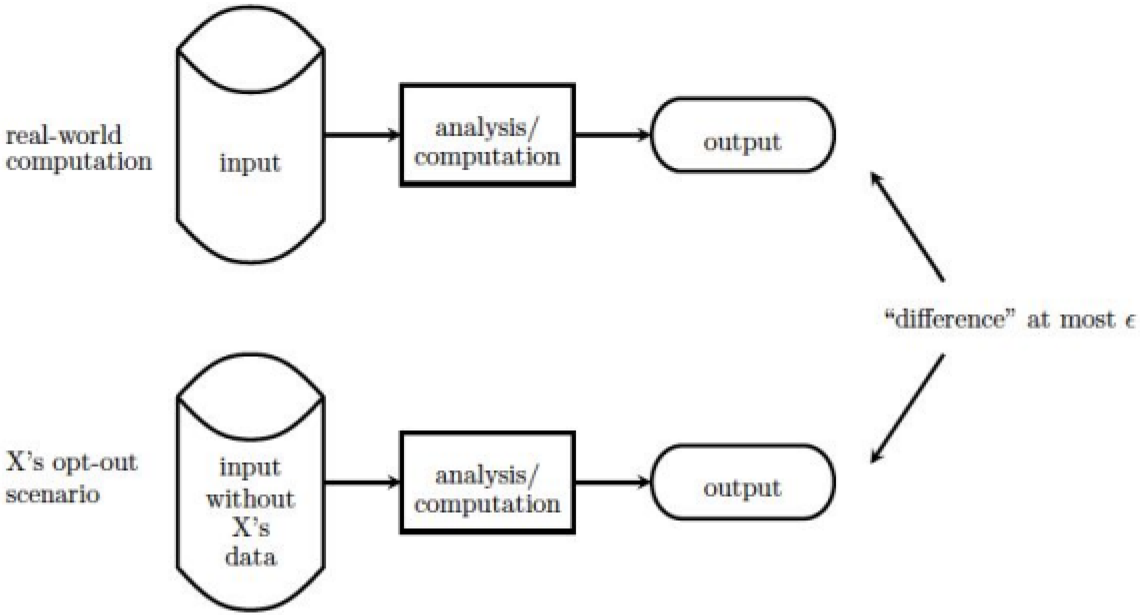
- *Differential privacy* is a system that tries to give a rigorous mathematical definition of privacy
- A computation is differentially private if, for any two neighboring datasets X, X′ that differ with respect to just one element x_i, the difference between the output of a computation based on the two data sets never differs by "too much"
  - A mechanism is said to be ε-differentially private if, for all computations in allowable set

$$e^{-\varepsilon} \le \Pr(A(X)=v) / \Pr(A(X^{'})=v) \le e^{\varepsilon}$$

$$e^{-\varepsilon} \approx 1-\varepsilon, \quad e^{\varepsilon} \approx 1+\varepsilon$$

  - Differential privacy rests on added risk of disclosure for most vulnerable case in the data
  - Mechanism for differential privacy is addition of noise to output data
  - Amount of noise that is needed to create required uncertainty will depend on the number of people in the daat set as well as on distribution of those people's characteristics
- If two independent releases based on the same dataset provide ε-differential privacy, the two together provide 2ε-differential privacy
  - Successive releases make additive claims on the "privacy budget"

How Do We Add Randomness?



Epsilon: privacy loss parameter

Captures deviation between opt-out and real world scenario

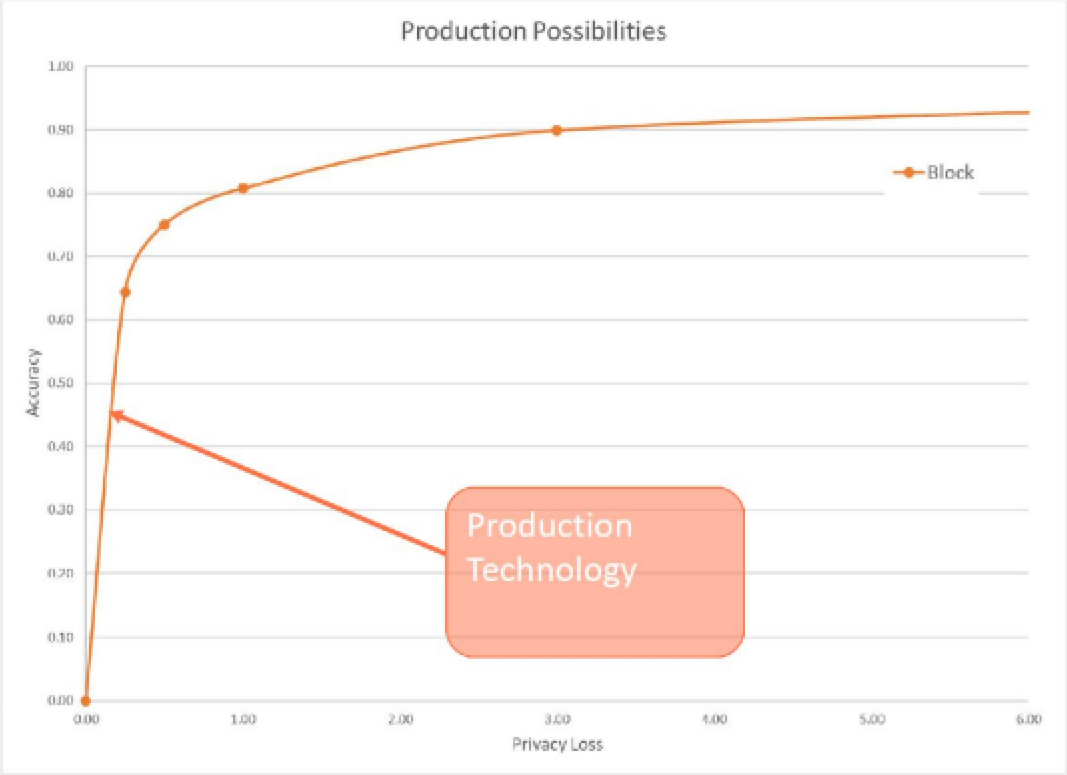
The effect of each individual's information on the output of the analysis

Smaller value is more privacy

Differential Privacy

- Advantages of using a differentially private mechanism:
  - Can tell data users the *form* and *magnitude* of the noise that has been added to the data, making it possible for them to draw statistically valid conclusions from the published output
  - Can make defensible statements about the degree of privacy protection provided
- Private companies including Amazon, Apple, and Google are using differential privacy in their analyses of customer data
- Census Bureau adopted differential privacy as the basis for decisions about releases of data from the 2020 Census; expect other US statistical agencies to follow suit
  - Potential loss of public use microdata files has generated considerable controversy

Differential Privacy



Choice of ε (across all releases based on a given data set) is a choice about the appropriate tradeoff between accuracy and privacy

Source: Abowd (2018)

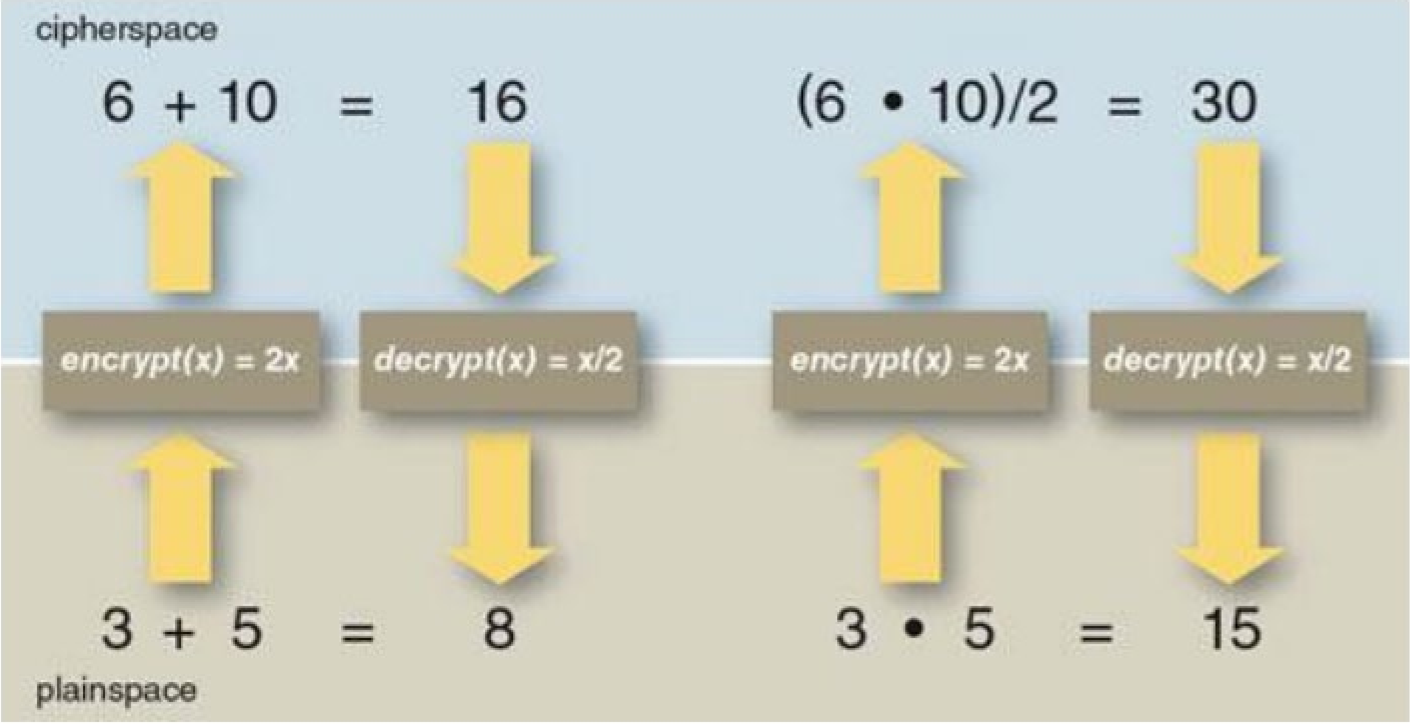
Differential Privacy

- Unresolved issues with respect to the application of differential privacy in the federal statistical context (Abraham 2019)
    - How to decide on the right tradeoff between accuracy and privacy
    - How to decide on the best way to allocate a limited “privacy budget”
    - How to marshal the resources within the federal statistical system to develop and apply differential privacy to its numerous data products

Homomorphic Encryption

Homomorphic Encryption

- **Idea:** Encrypt data using a function that preserves mathematical operations, allows researchers to do analysis on encrypted data, then decrypt to view results.
- **Homomorphism:** A map such that f(x*y)=f(x)*f(y)
- **Workflow**
  - Take data X and Y. Encrypt with a function so that you have f(X) and f(Y).
  - Analysis is done on f(X) and f(Y) (e.g., hypothesis test, regression models).
  - Results are returned to data owner.



Source: https://www.americanscientist.org/article/alice-and-bob-in-cipherspace

Encryption

- Encryption is not noise, so it is theoretically stronger protection…as long as it’s not decrypted
  - Example: RSA Encryption with a public/private key. Anyone who gets access to your private key could get the unaltered micro-data

Analysis

- Analysis is done with limited amount of noise added to data.
- Encryption can be quite strong depending on sophistication of models/methods being used.
- Can be practical to outsource
    - With a cloud computing environment that can take in data and output predictions, data owner sends encrypted data to the cloud environment, a model is built within the cloud environment, encrypted results are returned. Data owner then decrypts results.

Disadvantages

- Some amount of noise still needs to be added to avoid chosen plain text attacks.
  - Deterministic encryption means attacker might be able to figure out key.
- Can be quite slow.
  - Faster computing does mean it’s not necessarily prohibitively slow.
- Limited in types of methods that can be used.
  - If the model is too complicated, homomorphic encryption might take too long and be computationally infeasible
  - Practically, many analyses will be ok

Multi-Party Computation

Multi-Party Computing

- *Secure multi-party computing (MPC)* refers to methods developed by computer scientists since early 1980s for carrying out computations using data from multiple sources without the source data ever having to be shared.
- At least in principle, MPC offers an alternative for statistical calculations to creating large linked data sets, but some caveats:
    - Methods well developed for some types of calculations but not others
    - Proof-of-concept testing has used records that contain a clean linkage variable
    - Given current computing technology, *much* more time consuming than computations made directly on linked data

Multi-Party Computation within the Cloud

Example process:

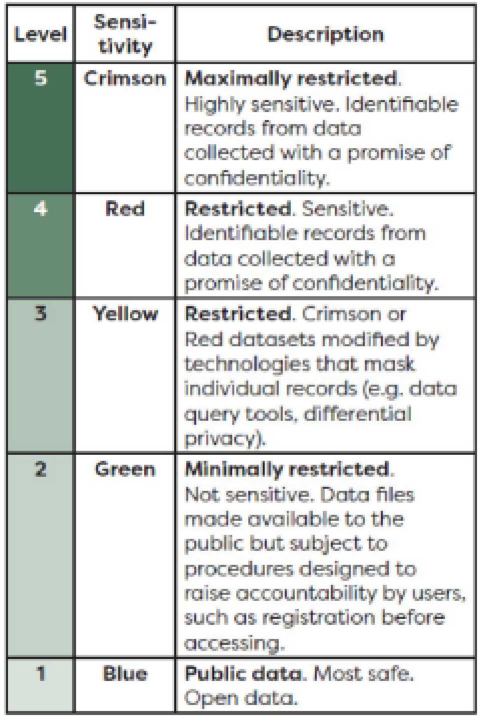
- Four sites run a trial with a different treatment
- Each site uploads data to the cloud environment
- Once everyone has submitted their data, an ANOVA is run using the data
- Output: there is a difference between treatments

Possibilities with Cloud Computing

- Could be very useful in the medical world:
    - High need for privacy
    - Useful to be able to input your own data and get some sort of results back

Source Environment

Tiered Access and Statistical



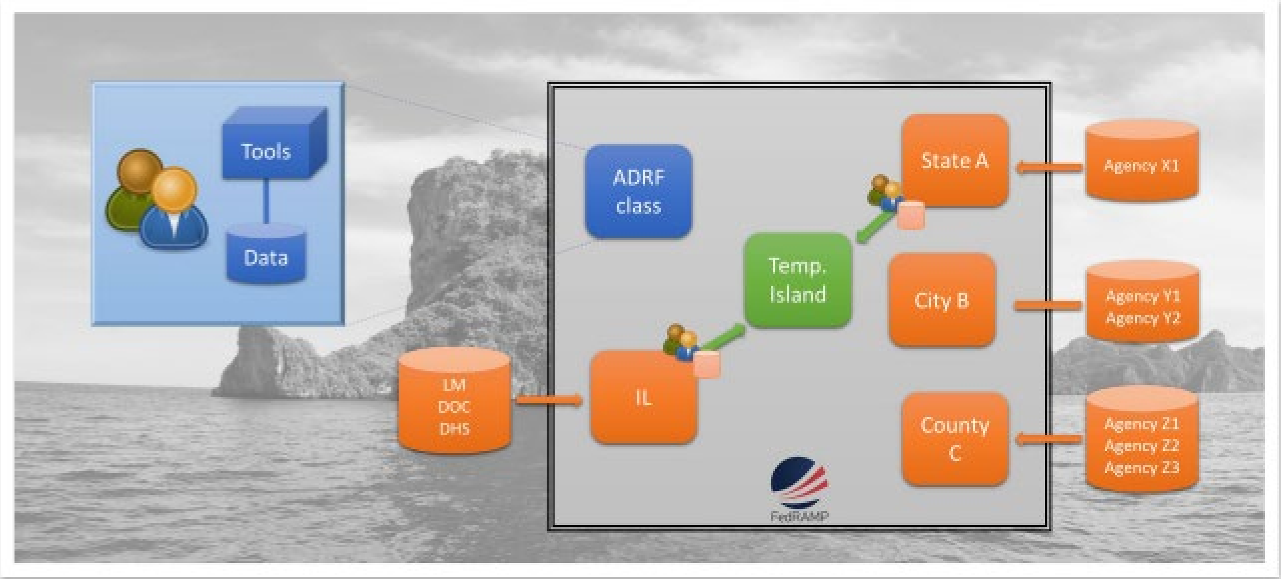
Level	Sensi-tivity	Description
5	Crimson	**Maximally restricted.** Highly sensitive. Identifiable records from data collected with a promise of confidentiality.
4	Red	**Restricted.** Sensitive. Identifiable records from data collected with a promise of confidentiality.
3	Yellow	**Restricted.** Crimson or Red datasets modified by technologies that mask individual records (e.g. data query tools, differential privacy).
2	Green	**Minimally restricted.** Not sensitive. Data files made available to the public but subject to procedures designed to raise accountability by users, such as registration before accessing.
1	Blue	**Public data.** Most safe. Open data.

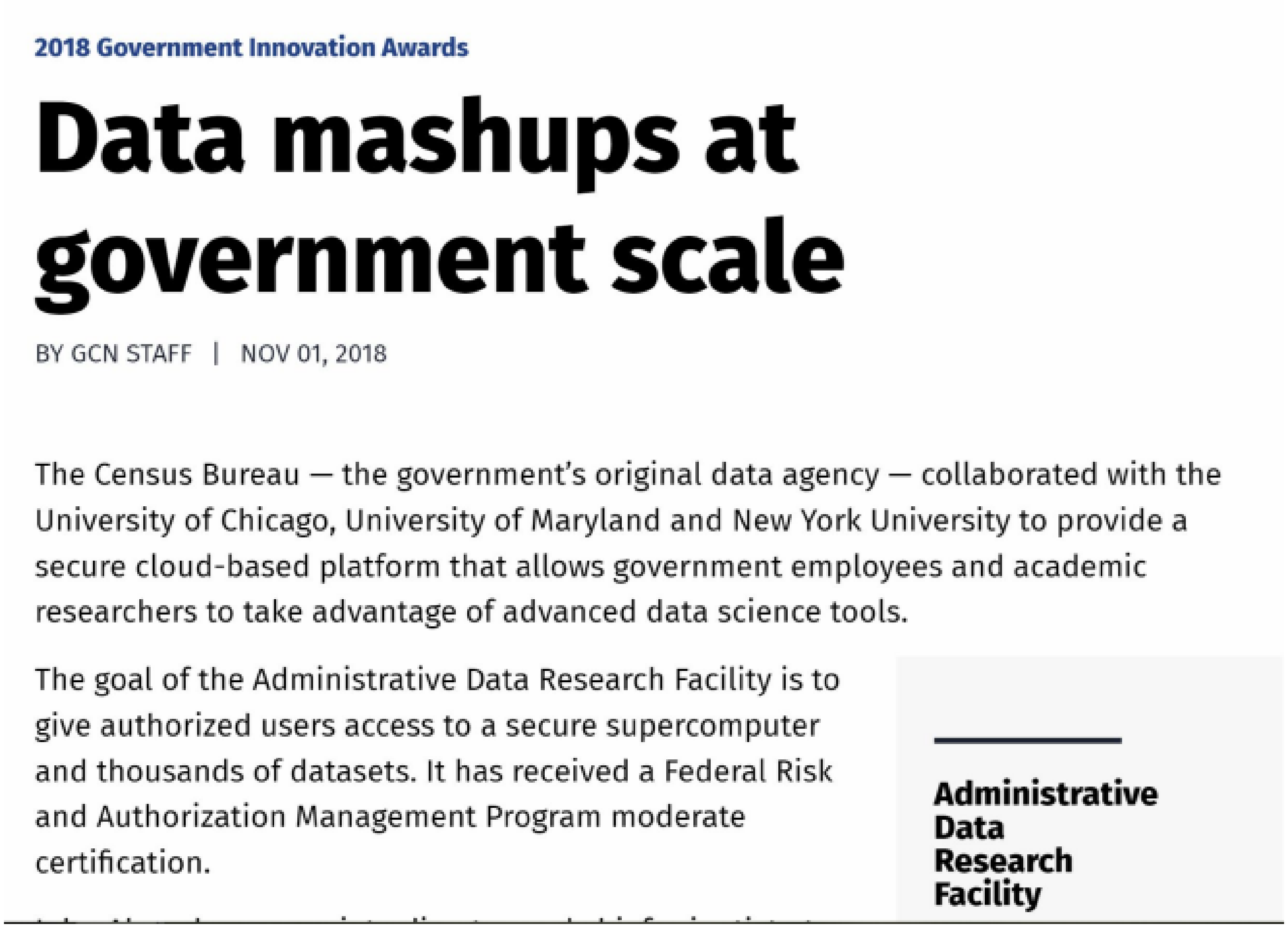
- Access depends on data sensitivity
  - Most sensitive data available only in a secure enclave (physical or virtual)
  - Modified or less sensitive data may e.g. be accessible using data query tool or released as restricted use data file under a data use agreement
  - Only least sensitive data released as unrestricted public use data file
- Access to restricted data matched to user needs
  - Many data users do not need access to unadulterated microdata
  - Role for synthetic data combined with a validation server

Source: Commission on Evidence-Based Policymaking, *Promise of Evidence-based Policymaking,* 2017

Data Security

- Approaches that are community led and that build value
- Frameworks that establish adoptable approaches for the secure handling of essential data
- Processes that are built on an overarching st





2018 Government Innovation Awards

Data mashups at government scale

BY GCN STAFF | NOV 01, 2018

The Census Bureau — the government's original data agency — collaborated with the University of Chicago, University of Maryland and New York University to provide a secure cloud-based platform that allows government employees and academic researchers to take advantage of advanced data science tools.

The goal of the Administrative Data Research Facility is to give authorized users access to a secure supercomputer and thousands of datasets. It has received a Federal Risk and Authorization Management Program moderate certification.

Administrative Data Research Facility

https://gcn.com/Articles/2018/11/01/PSI_Administrative-Data-Research-Facility.aspx?p=1

Summary

- In the modern world, views and concepts of privacy are rapidly changing.
- Lots of potential for interesting, innovative, and beneficial research, but need to be careful with the data that is available now.
- Tradeoff between utility and safety.
- Many methods to protect data privacy, but no “gold standard”

Also

- Homomorphic Encryption https://www.youtube.com/watch?v=vUtyuw7YLVM
- Privacy preservation techniques in big data analytics: a survey
  - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0141-8

- Testimony to Commission of Evidence Based Policy https://www.cep.gov/
- Recent Gates Foundation funded workshop (in The ANNALS of the American Academy of Political and Social Science)
- http://policydatainfrastructure.com/author-contributions.html

- The modernization of statistical disclosure limitation at the U.S. Census Bureau
- Aref N. Dajani1, Amy D. Lauger1, Phyllis E. Singer1, Daniel Kifer2, Jerome P. Reiter3, Ashwin Machanavajjhala4, Simson L. Garfinkel1, Scot A. Dahl6, Matthew Graham7, Vishesh Karwa8, Hang Kim9, Philip Leclerc1, Ian M. Schmutte10, William N. Sexton11, Lars Vilhuber7, 11, and John M. Abowd5

- An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices; John M. Abowd and Ian M. Schmutte