Privacy and Confidentiality

How was privacy invented?

Why do we need confidentiality?

What are the top research priorities?

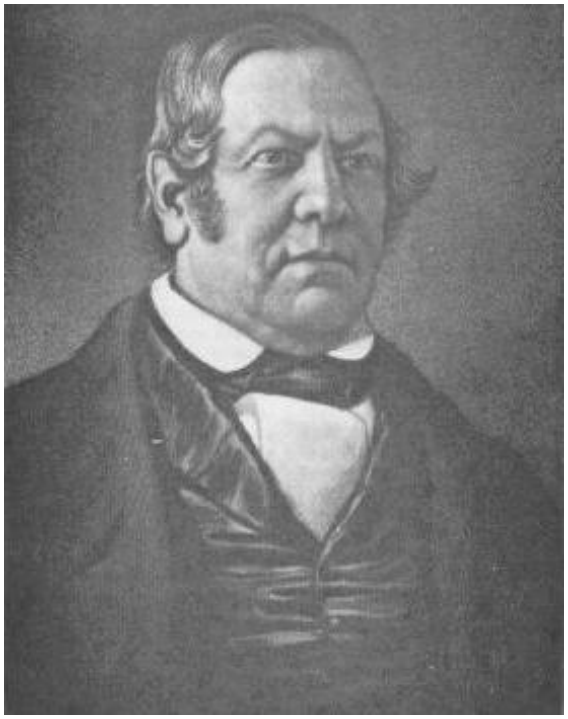
100

SCHEDULE of the whole number of Persons

Name of County, City, Ward, Town, Township, Parish, Precinct, Hundred, or District.	NAMES of HEADS OF FAMILIES.	FREE WHITE PERSONS — MALES												
		Under five years of age.	Of five and under ten.	Of ten and under fifteen.	Of fifteen and under twenty.	Of twenty and under thirty.	Of thirty and under forty.	Of forty and under fifty.	Of fifty and under sixty.	Of sixty and under seventy.	Of seventy and under eighty.	Of eighty and under ninety.	Of ninety and under one hundred.	Of
		under 5	5 to 10	10 to 15	15 to 20	20 to 30	30 to 40	40 to 50	50 to 60	60 to 70	70 to 80	80 to 90	90 to 100	10
	Jacob Stevenson						1							
	Stephen Faulk													
	Lewis Johnson													
	John Anderson													
	Benjamin James													
	William Hemphill													
	Abraham Miller													
	Mable Thatcher													
	Thomas Snyder	1	1				1							
	John Zebley	1				1			1					
						1								

1790-1840:
The census marshals instructed to post the returns in "two of the most important places" of their enumeration district.

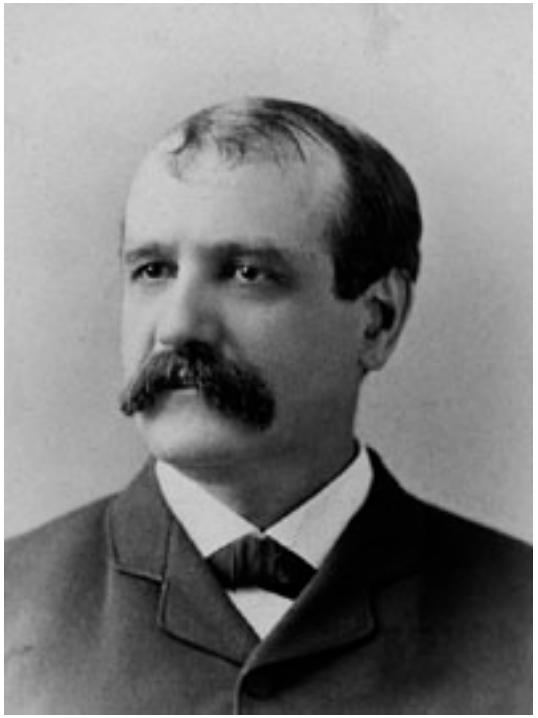
1850: Thomas McKennan



“[Census information] is not to be used in any way to the gratification of curiosity, the exposure of any man's business or pursuits, or for the private emolument of the marshals or assistants.”

1870: Francis Amasa Walker

“No graver offense can be committed by assistant marshals than to divulge information acquired in the discharge of their duty.”



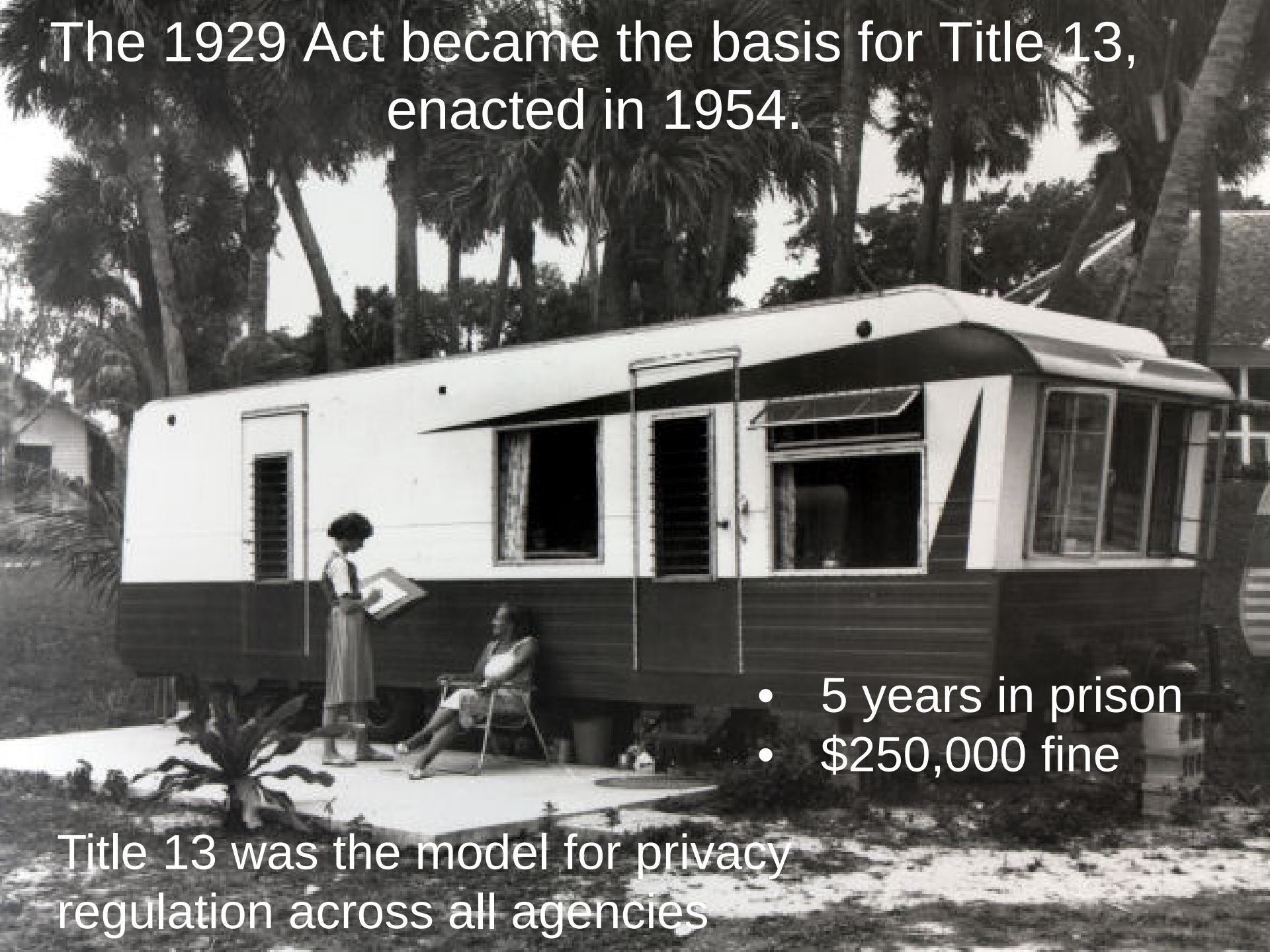
1910: William Howard Taft

“Every employee of the Census Bureau is prohibited, under heavy penalty, from disclosing any information which may thus come to his knowledge.”

1929 Census Act

Before the 1929 Census Act the Director had the discretion to provide access to census information, and did so often, “for genealogy or other proper purposes.”

Under the 1929 Census Act, all disclosure—even to other government agencies—became illegal.

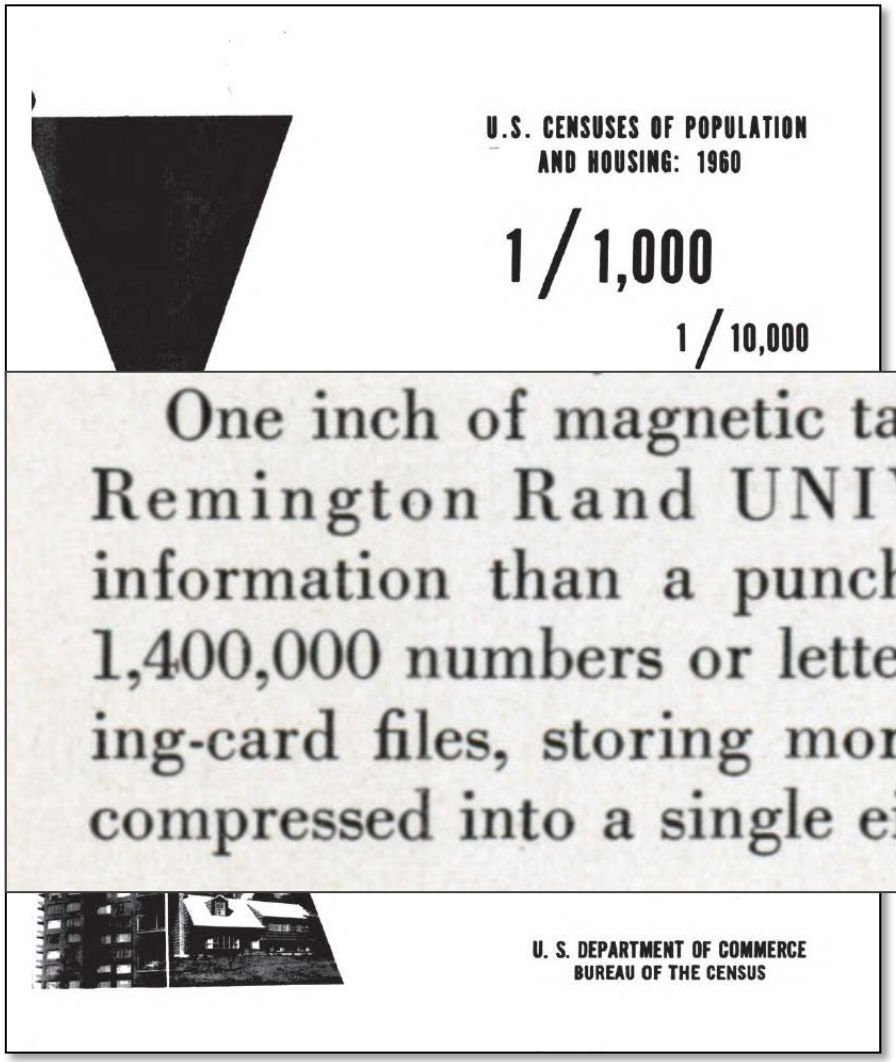


The 1929 Act became the basis for Title 13, enacted in 1954.

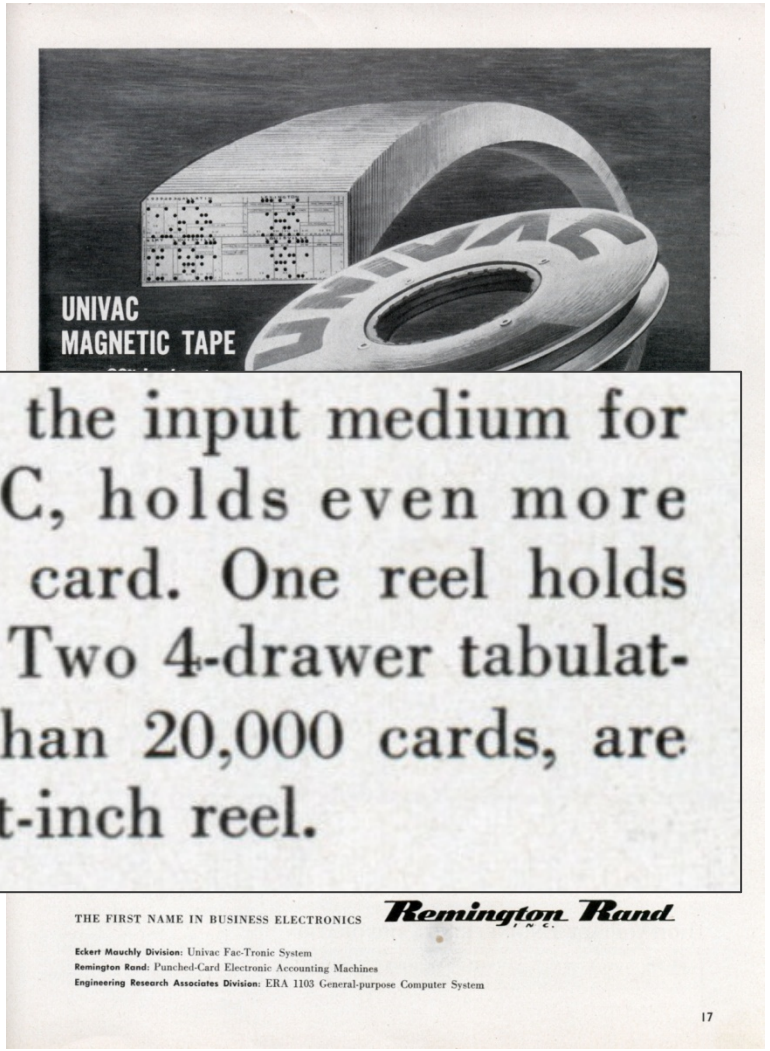
- 5 years in prison
- $250,000 fine

Title 13 was the model for privacy regulation across all agencies

1962: The Invention of Public-Use Microdata



Cover, 1960 Census Microdata Codebook



One inch of magnetic tape, the input medium for Remington Rand UNIVAC, holds even more information than a punched card. One reel holds 1,400,000 numbers or letters. Two 4-drawer tabulating-card files, storing more than 20,000 cards, are compressed into a single eight-inch reel.

Distributed on 13 Univac Tapes
(or 18,000 punchcards)

Strategies of Disclosure Control: Public Use Microdata Files

- Reduced detail variables describing small and visible populations (e.g. high incomes).
- Introduction of "noise" (small amounts of variation) into selected data items.
- Data swapping (i.e., swapping pairs of households across geographic areas to add uncertainty)
- No documented cases of harm through re-identification have occurred in 52 years of intensive use and open distribution.

Public Opinion Quarterly, Vol. 74, No. 3, Fall 2010, pp. 551–569

THE POLLS–REVIEW
INACCURATE AGE AND SEX DATA IN THE CENSUS PUMS FILES: EVIDENCE AND IMPLICATIONS

J. TRENT ALEXANDER
MICHAEL DAVERN
BETSEY STEVENSON*

Abstract We discover and document errors in public-use microdata samples ("PUMS files") of the 2000 Census, the 2003–2006 American Community Survey, and the 2004–2009 Current Population Survey. For women and men age 65 and older, age- and sex-specific population estimates generated from the PUMS files differ by as much as 15 percent from counts in published data tables. Moreover, an analysis of labor-force participation and marriage rates suggests the PUMS samples are not representative of the population at individual ages for those age 65 and over. PUMS files substantially underestimate labor-force participation of those near retirement age and overestimate labor-force participation rates of those at older ages. These problems were an unintentional byproduct of the misapplication of a newer generation of disclosure-avoidance proce-

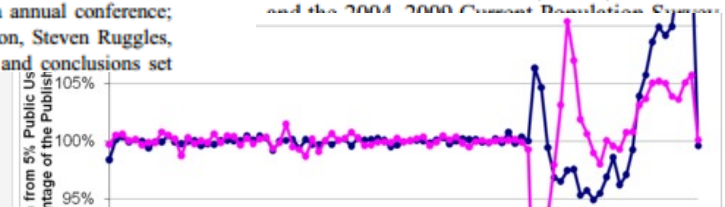
J. TRENT ALEXANDER is a Research Scientist at the Minnesota Population Center at the University of Minnesota, Minneapolis, MN, USA. MICHAEL DAVERN is the Vice President and Director of Public Health Research at NORC at the University of Chicago, Chicago, IL, USA. BETSEY STEVENSON is an Assistant Professor of Business and Public Policy at the Wharton School at the University of Pennsylvania, Philadelphia, PA, USA. The authors would like to thank seminar participants at Wharton, the Minnesota State Demographers' Brownbag Series, and the University of Pennsylvania Population Studies Center; participants at the 2009 Joint Statistical Meetings and the 2010 Population Association of America annual conference; as well as Orley Ashenfelter, Carolyn Liebler, Peter Graven, David Johnson, Steven Ruggles, Stephen Tordella, and Justin Wolfers for useful discussions. The analysis and conclusions set

erturbation Goes Wrong and Why
onal Relies Instead on Sampling,
ing, and Other Minimally Harmful
ct Privacy of Census Microdata

t McCaa, Steven Ruggles, and Matthew Sobek

a Population Center, 50 Willey Hall
nneapolis MN 55455 USA
ontact: rmccaa@umn.edu

l disseminates population census microdata at no cost
a series of 212 samples totaling almost a half billion
researchers. Registration is required for researchers
. Statistics from Google Analytics show that IPUMS-
ing registration form is an effective deterrent for
otect data privacy, we rely principally on sampling,
il, swapping of records across geographic boundaries,
methods such as top and bottom coding. We do *not*
ethods. A recent case of perturbation gone wrong—
2000 census of the USA (PUMS), the 2003–2006



More photos and interactive graphics

how tricky it car
accuracy.

The Census Bureau, which needs Americans to cooperate wit

« PREVIOUS POST
The Economics of Superstardom

SUPPORT FREAK

FREAKON

POPULAR | LATES

How Can Tiny Norw
Teslas? A New Freal
Until recently, tiny N
million) has been the
Teslas (after the U.S.)

Tell Me Something I
Freakonomics Radio
A few years ago, I dev
sitting next to me on
wanted to have a

Varieties of Confidential Data

- 13 Federal statistical agencies
  - Most have both public use and restricted files
- Federal administrative records
  - Wide variation in access (CMS to IRS)
- Administrative records from schools and health providers (FERPA and HIPAA)
- Academic studies funded by Federal agencies, e.g. NIH and NSF (Common Rule)
- Private-sector data

What are we afraid of?

- Statistical agencies fear increase in non-response rates to surveys
- Some members of the public fear that the government will use data for social control
- Many people fear (more justifiably) abuse of data by the private sector
  - Identity theft
  - Public dissemination of private information

Much NSF Privacy Research focuses on Private-sector data

- Privacy for smart meter data
- Smartphone user privacy
- Privacy in a camera-rich world

CyberInfrastructure Research

- Critical research areas
  - Cybersecurity/Trustworthy systems
  - Controlled remote access through virtualization

- Less promising research areas
  - Synthetic data: disappointing so far, except for the simplest data collections
  - Perturbation
  - Data shuffling

Conclusions

- Most scientific research on human populations is based on data from federal statistical agencies, administrative records, or federally funded academic studies.
  - All of these are well-regulated
  - The track record of confidentiality protection is strong
- Most of the risk to confidentiality comes from the private sector, not from scientific research

Conclusions

- We need to improve cybersecurity
- We need new regulatory controls for private-sector data.
  - This is as much an institutional problem as a technical one.