Staring Down the Database Reconstruction Theorem

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau

American Association for the Advancement of Science
Annual Meeting Saturday, February 16, 2019 3:30-5:00

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

The views expressed in this talk are my own and not those of the U.S. Census Bureau.

The challenges of a census:

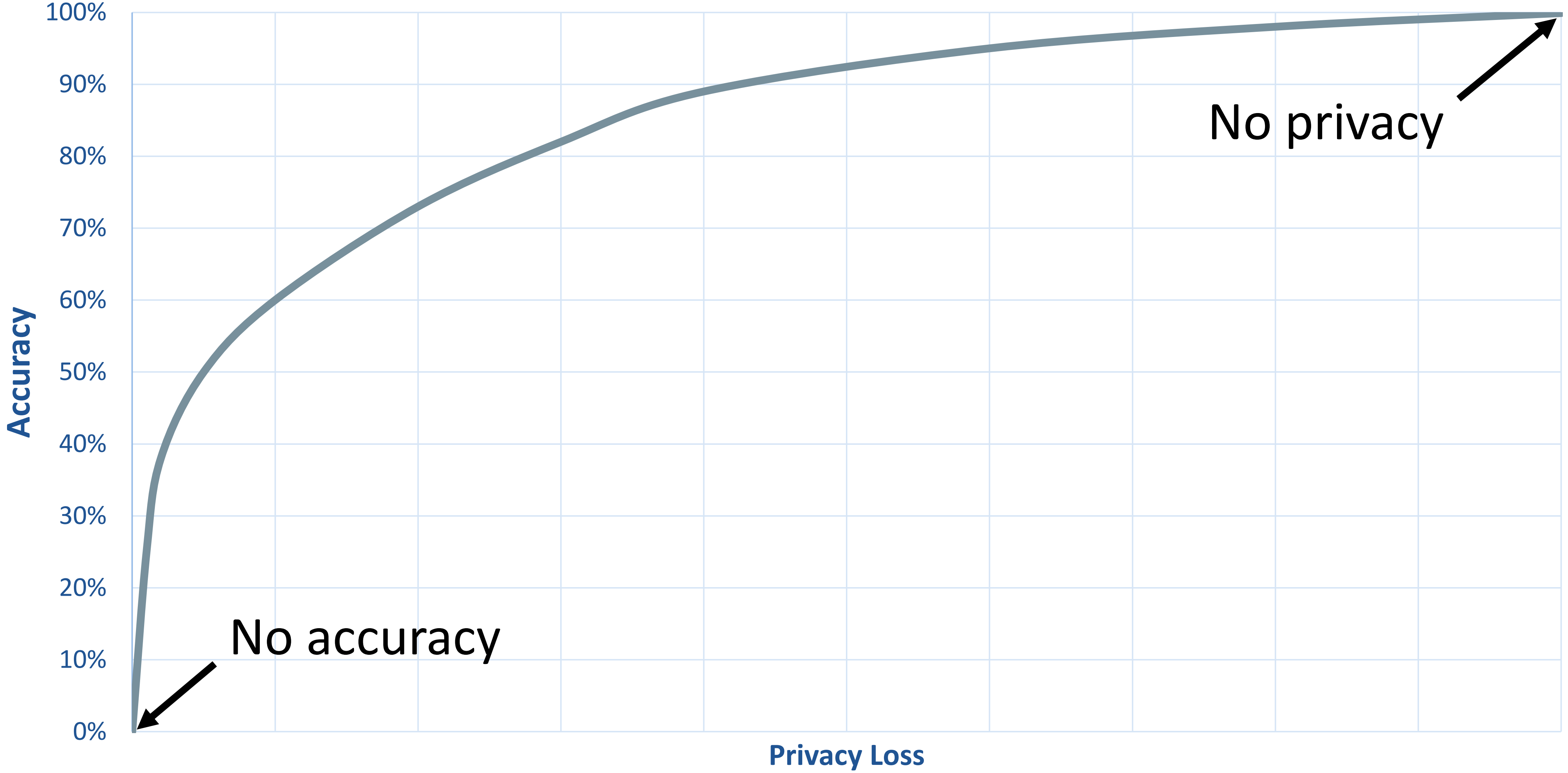
1. collect all of the data necessary to underpin our democracy;
2. protect the privacy of individual data to ensure trust and prevent abuse.

- Too many statistics

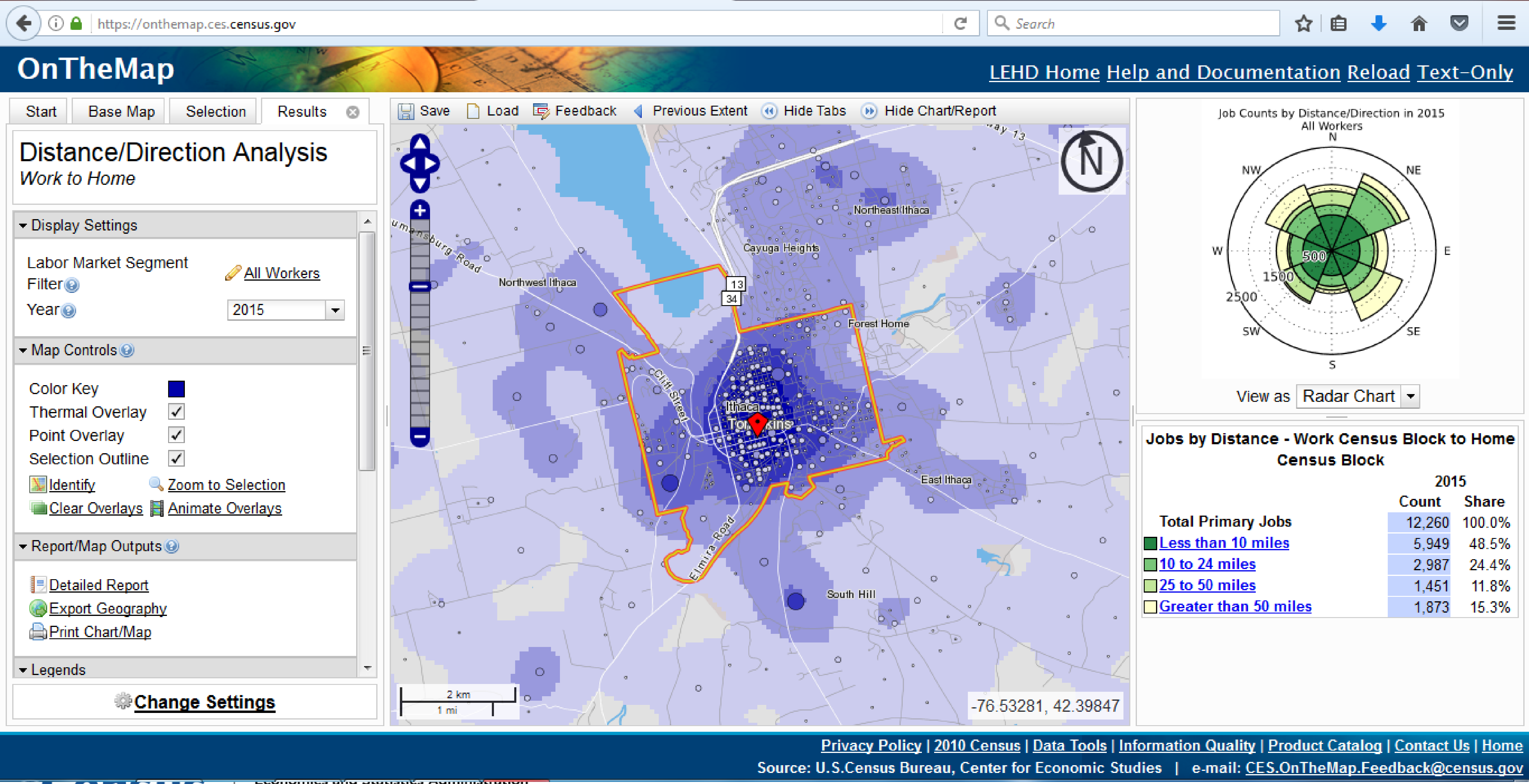
- Noise infusion is necessary

- Transparency about methods helps rather than harms

Fundamental Tradeoff betweeen Accuracy and Privacy Loss



Good science and privacy protection are partners



https://onthemap.ces.census.gov

OnTheMap

LEHD Home Help and Documentation Reload Text-Only

Start | Base Map | Selection | Results

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report

Distance/Direction Analysis

Work to Home

Display Settings

Labor Market Segment Filter: All Workers

Year: 2015

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

Change Settings

Job Counts by Distance/Direction in 2015
All Workers

View as: Radar Chart

Jobs by Distance - Work Census Block to Home Census Block

	2015	
	Count	Share
Total Primary Jobs	12,260	100.0%
🟩 Less than 10 miles	5,949	48.5%
🟩 10 to 24 miles	2,987	24.4%
🟩 25 to 50 miles	1,451	11.8%
🟨 Greater than 50 miles	1,873	15.3%

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

What we did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

What we found

- Census block correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age, race, ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically

We fixed this for the 2020 Census by implementing differential privacy

Acknowledgments

- The Census Bureau’s 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Rob Sienkiewicz (ACC Disclosure Avoidance, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Craig Corl, Aref Dajani, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, Cecil Washington, and Pavel Zhuralev

Thank you.

John.Maron.Abowd@census.gov

More Background on the 2020 Census Disclosure Avoidance System

- September 14, 2017 CSAC (overall design)
  https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#
- August, 2018 KDD’18 (top-down v. block-by-block)
  https://digitalcommons.ilr.cornell.edu/ldi/49/
- October, 2018 WPES (implementation issues)
  https://arxiv.org/abs/1809.02201
- October, 2018 *ACMQueue* (understanding database reconstruction)
  https://digitalcommons.ilr.cornell.edu/ldi/50/ or
  https://queue.acm.org/detail.cfm?id=3295691
- December 6, 2010 CSAC (detailed discussion of algorithms and choices)
  https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

Selected References

- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. In Halevi, S. & Rabin, T. (Eds.) Calibrating Noise to Sensitivity in Private Data Analysis Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg, 265-284, DOI: 10.1007/11681878_14.
- Fellegi, Ivan P. 1972. On the Question of Statistical Confidentiality. Journal of the American Statistical Association, Vol. 67, No. 337 (March):7-18, stable URL http://www.jstor.org/stable/2284695.
- Ganda, Srivatsava, Shiva Kasiviswanathan and Adam Smith. 2008. Composition Attacks and Auxiliary Information in Data Privacy. In Knowledge, Discovery and Datamining, Las Vegas, NV, doi:10.1145/1401890.1401926.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- McKenna, Laura. 2018. Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47.
- Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. 2012. Exploring Re-Identification Risks in Public Domains, Tenth Annual International Conference on Privacy, Security and Trust, IEEE, doi:10.1109/PST.2012.6297917.
- U.S. Census Bureau. 2019. LEHD Origin-Destination Employment Statistics (2002-2015) [computer file]. Washington, DC: U.S. Census Bureau, Longitudinal-Employer Household Dynamics Program [distributor], accessed on February 15, 2019 at https://onthemap.ces.census.gov.