The National Academies of SCIENCES ENGINEERING MEDICINE

COMMITTEE ON NATIONAL STATISTICS
Division of Behavioral and Social Sciences and Education

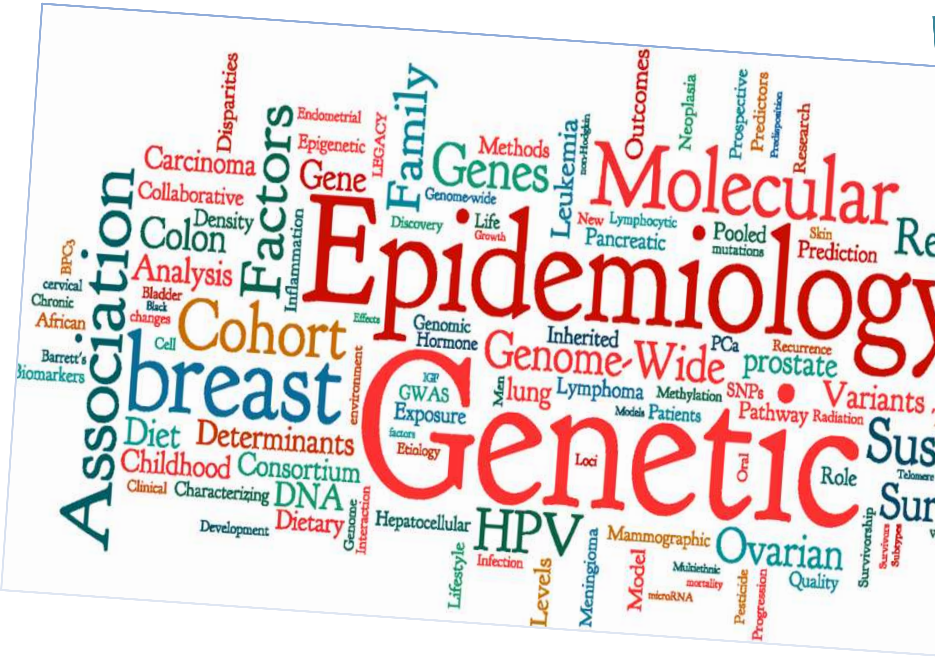
Workshop on 2020 Census Data Products:
Data Needs and Privacy Considerations
December 11 - 12, 2019

Differential “Privacy Guarantees” and Ethical Equipoise

Daniel C. Barth-Jones, M.P.H., Ph.D.
Assistant Professor of Clinical Epidemiology,
Mailman School of Public Health
Columbia University

db2431@columbia.edu

As an Epidemiologist and Statistical Disclosure Researcher, I see an Important Historic, Societal Debate underway…



Public Policy Collision Course

Epidemiologic Triad: Characteristics of Person, Place and Time =

Re-identification Quasi-identifiers: Characteristics of Person, Place and Time

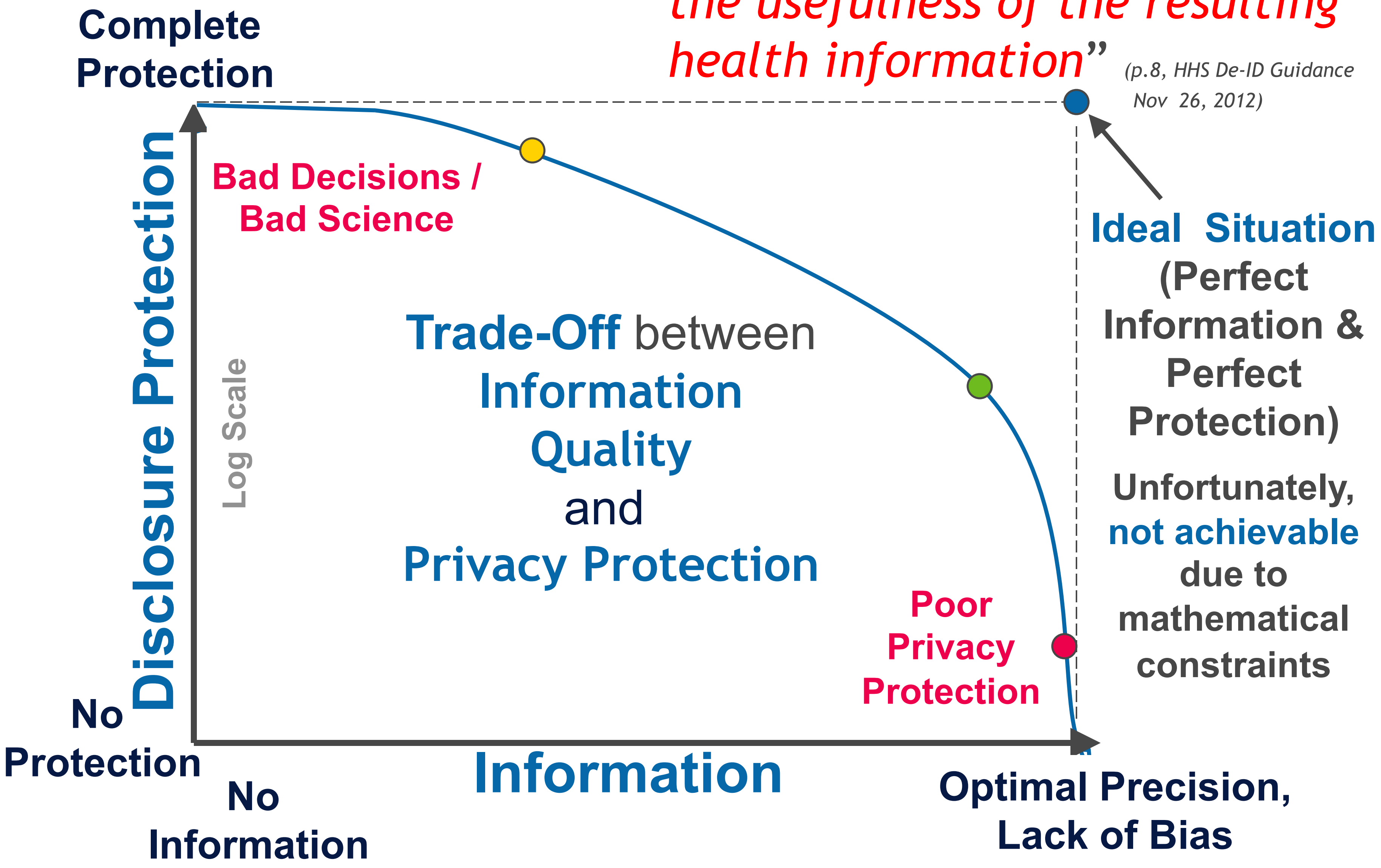
Census Data is an Invaluable "Public Good"

It serves as an essential foundation for a variety of political, social and scientific purposes. In particular, it is essential for supporting a whole host of vital social and health research activities.

- We can gain a useful "big picture perspective" of the nuanced balancing that is required in approaching the data privacy and data utility trade-offs to get this task as "close to right" as possible to adopt a lens from the area of research ethics.
- Beneficence: "**Maximize possible benefits, minimize possible harms**"
- Justice/Injustice: Injustice occurs **when benefits** of research for individuals **are unequally denied** and/or **when risks are inequitably distributed**.
- **Rarity - being unusual -** has profound ramifications operating at both **individual and group levels** for both **re-identification risks and benefits.** These **trade-offs (and** the associated **autonomous choices** that individuals and groups will want to make **about** where the right **balance points** are) should be expected to **vary dramatically.**

The Inconvenient Truth:

"*De-identification leads to information loss which may limit the usefulness of the resulting health information*" *(p.8, HHS De-ID Guidance Nov 26, 2012)*



Complete Protection

No Protection

Disclosure Protection

Log Scale

Information

No Information

Optimal Precision, Lack of Bias

Bad Decisions / Bad Science

Trade-Off between **Information Quality** and **Privacy Protection**

Poor Privacy Protection

Ideal Situation

(Perfect Information & Perfect Protection)

Unfortunately, **not achievable** due to mathematical constraints

My Original Question from Harvard Petrie-Flom Re-identification Symposium:

Ethical Equipose:

"Is it an ethically compromised position, particularly in the coming age of personalized medicine, if we end up purposefully masking the racial, ethnic or other group membership status information (e.g. American Indians or LDS Church members, etc.) for certain individuals, or for those with certain rare genetic diseases/disorders, in order to protect them against supposed re-identifications? In making this ethical determination, we must, of course, recognize that by doing so, we would also deny them the benefits of research conducted with de-identified data that could help address their health disparities, find cures for their rare diseases, or facilitate "orphan drug" research that would otherwise not be economically viable."

Reference provided in slide

Language matters for establishing and maintaining Trust and Transparency

- Title 13 Section 9 – Speaks of "Confidentiality"
- ***"Differential vs Formal Privacy"***
  - ***Will the public understand the "differential" aspect?***
- ***"Privacy Guarantee"?***
  - *The use of the term "Guarantee" is questionable when re-identifiability depends on the selected value of Epsilon" - and the relationship between ε and re-identifiability is not easily explicated. (Ref: Work of Chris Clifton, see reference slides)*
  - *We saw yesterday that within the ranges of epsilon being contemplated, risks of re-identification that are beyond "de minimis" levels will still remain.*
- ***Disclosure Avoidance /(Disclosure Reduction?)***

Statistical Disclosure Limitation versus Differential Privacy

- *Quasi-identifiers vs. "Everything is Personally Identifiable Information"*
- *Assumptions of Differential Privacy*
  - *All data elements are potentially knowable by data intruders and equally as useful for re-identification or attribute inference.*
  - *All data elements are equally sensitive or able to invoke privacy harms (e.g., Vacancy, where's the privacy harm?)*
- *Assumptions of SDL -- Re-identification risks depend importantly on:*
  - *Replicability*
  - *Accessibility*
  - *Distinguishablity*
  - *Ability to build a comprehensive population register*

What’s to Love about Differential Privacy?

- *Privacy Guarantees*
- *Mathematical Elegance*
- *Broad assumptions about data intruder knowledge and capabilities (nearly omniscience, omnipotence and constantly co-conspiring)*
- *Broad assumptions about what might be harmful in terms of data privacy attacks, both re-identification risk and attribute inference.*
- *Composability*
- *Consistency*

For implementation by the U.S. Census, there is no organization I would trust to do this as best it can be done. And for a complete population Decennial Census, it’s hard to think of a case where it would be needed as much as in this Use Case.

What’s Not to Love about Differential Privacy?

- *Privacy “Guarantees”*
- *The complexity of communicating what it does and how it does it to the public*
  - *Trust and Transparency Issues*
- *The “accuracy costs” that are incurred by its very broad assumptions*
- *The accuracy costs incurred because of the Census need for certain “Invariants” -- and the ethical dilemmas posed by the transfer of these accuracy costs to data for other purposes and individuals*
- *Differential Privacy strictly enforces the “privacy”, but only optionally enforces the accuracy issues through a wise, reasoned and empirically analyzed and justifiable selection of epsilon.*
- *It is not without completely free of potential avenues of attack*
  - *Repeated instantiations can be revelatory*
  - *Correlated observations don’t receive the same guarantees*

Some Final Not-so-Random Concerns

- *Off-Spline Geographies (e.g. ZCTAs)*
- *Subtraction Geographies - simultaneous reporting of overlayed geographies with differing boarders can be used by attackers to target small areas.*
- *The competition between individuals, groups, researchers and politicians for “Privacy Loss” Budgets*

Reserve Slides for Questions

Bill of Health

Examining the intersection of law and health care, biotech & bioethics
A blog by the Petrie-Flom Center and friends

Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations

- http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/

- https://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/

- http://blogs.law.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/

References for Differential Privacy Concerns

1. Clifton, C.; Tassa, T. On Syntactic Anonymity and Differential Privacy. Transactions On Data Privacy 6 (2013) 161–183
2. Lee J., Clifton C. How Much Is Enough? Choosing ε for Differential Privacy. In: Lai X., Zhou J., Li H. (eds) Information Security. ISC 2011. Lecture Notes in Computer Science, vol 7001. Springer, Berlin, Heidelberg
3. Lee, J.; Clifton C. Differential Identifiability. KDD ’12, August 12–16, 2012, Beijing, China.

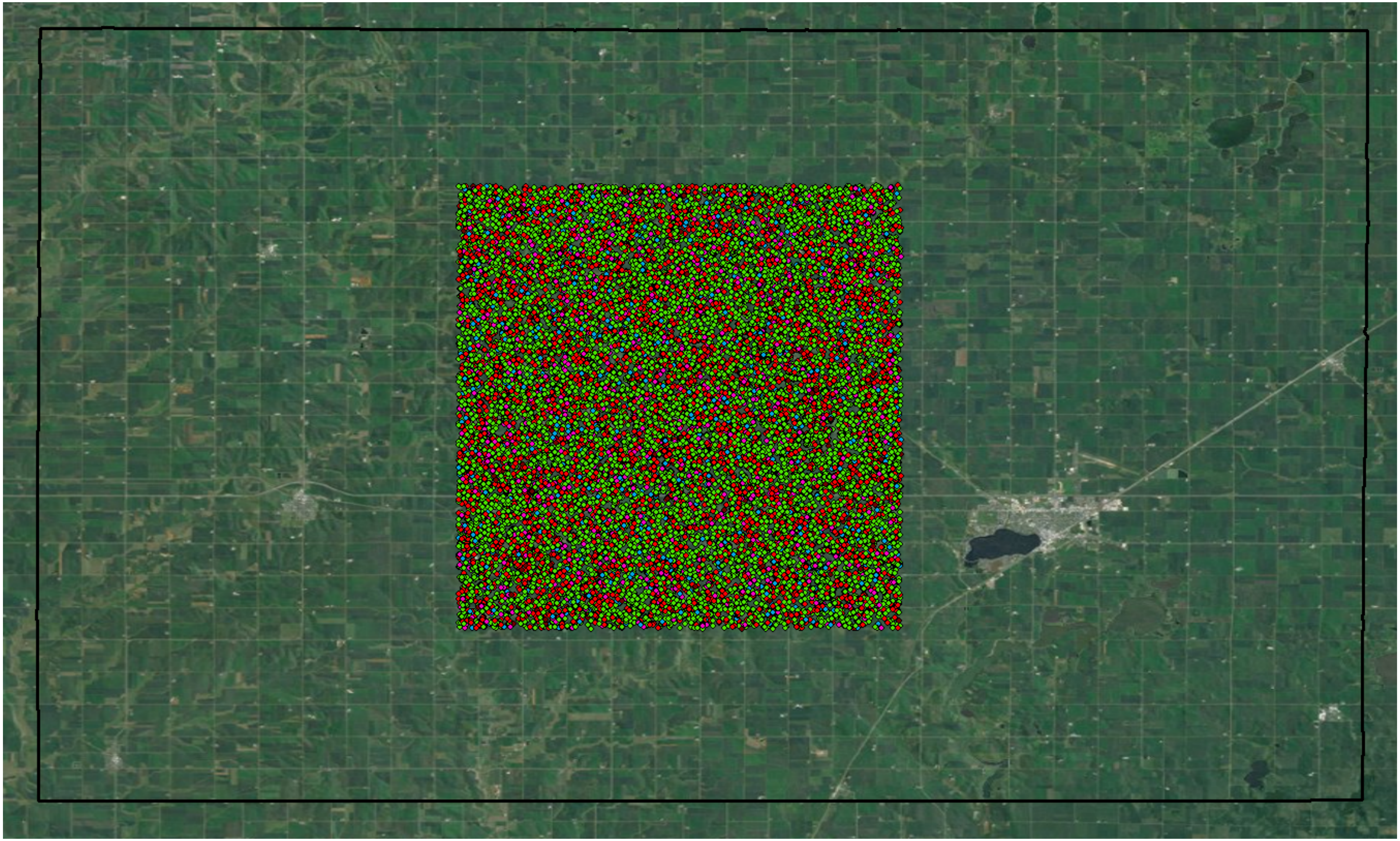
My Equivalent Question for the CNStat Workshop:

Ethical Equipose:

“Is it an ethically compromised position, particularly in the coming age of “big data”, if we end up purposefully masking the racial, ethnic or other group membership status information (e.g. American Indians or LDS Church members, etc.) for certain individuals, in order to protect them against potential re-identification threats? In making this ethical determination, we must, of course, recognize that by doing so, we will likely seriously distort important epidemiologic measurements and deny them the full benefits of research conducted with de-identified data that could help address their health disparities, find cures for their rare diseases, or facilitate “orphan drug” research that would otherwise not be economically viable.”

- ● Asian
- ● Black
- ● Hispanic
- ● Other
- ● White

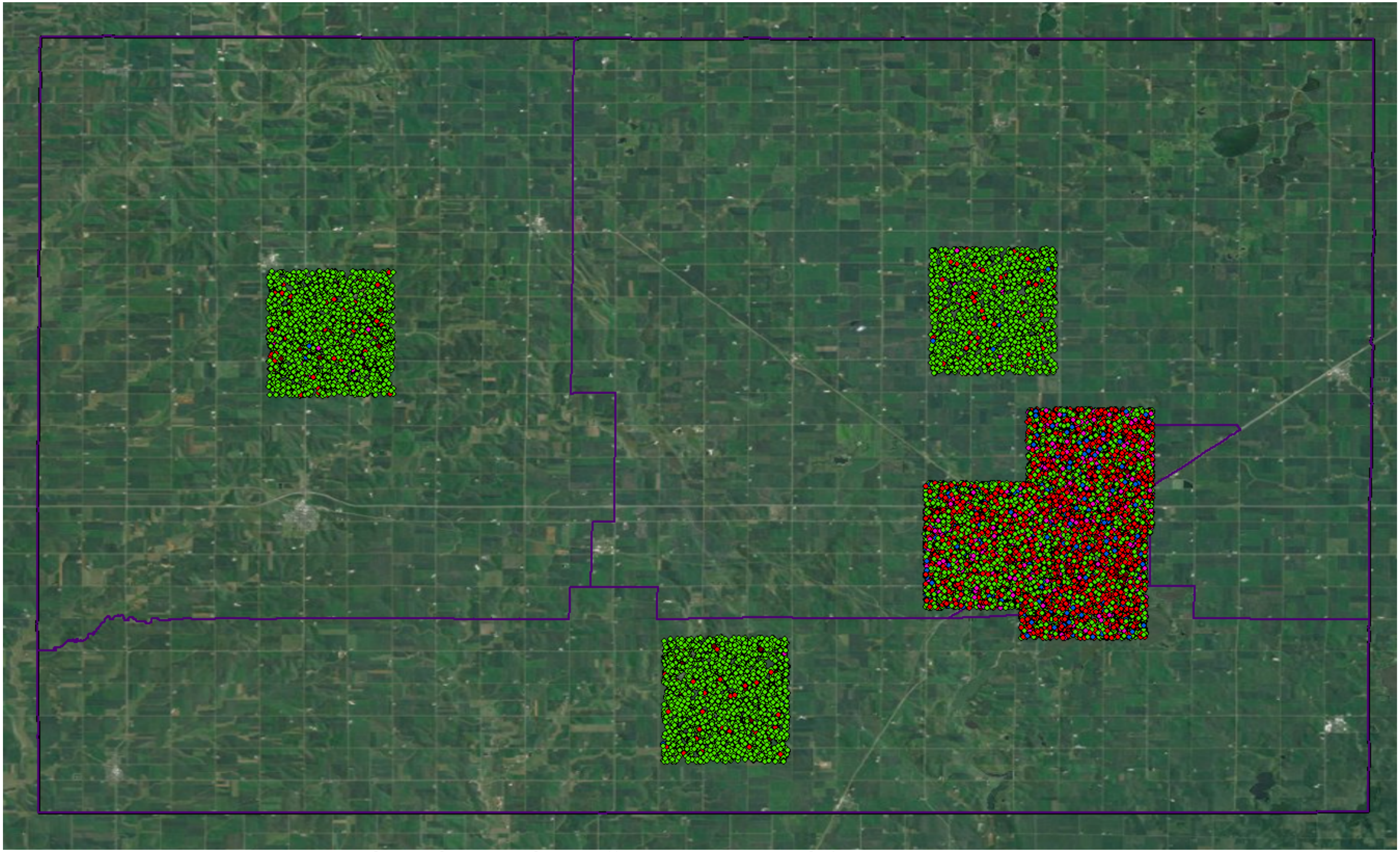
CBSA=Worthington, centered around CBSA centroid



Source: Simulated “Synthetic” Data created from Census PUMS Data

◆ Asian
◆ Black
◆ Hispanic
◆ Other
◆ White

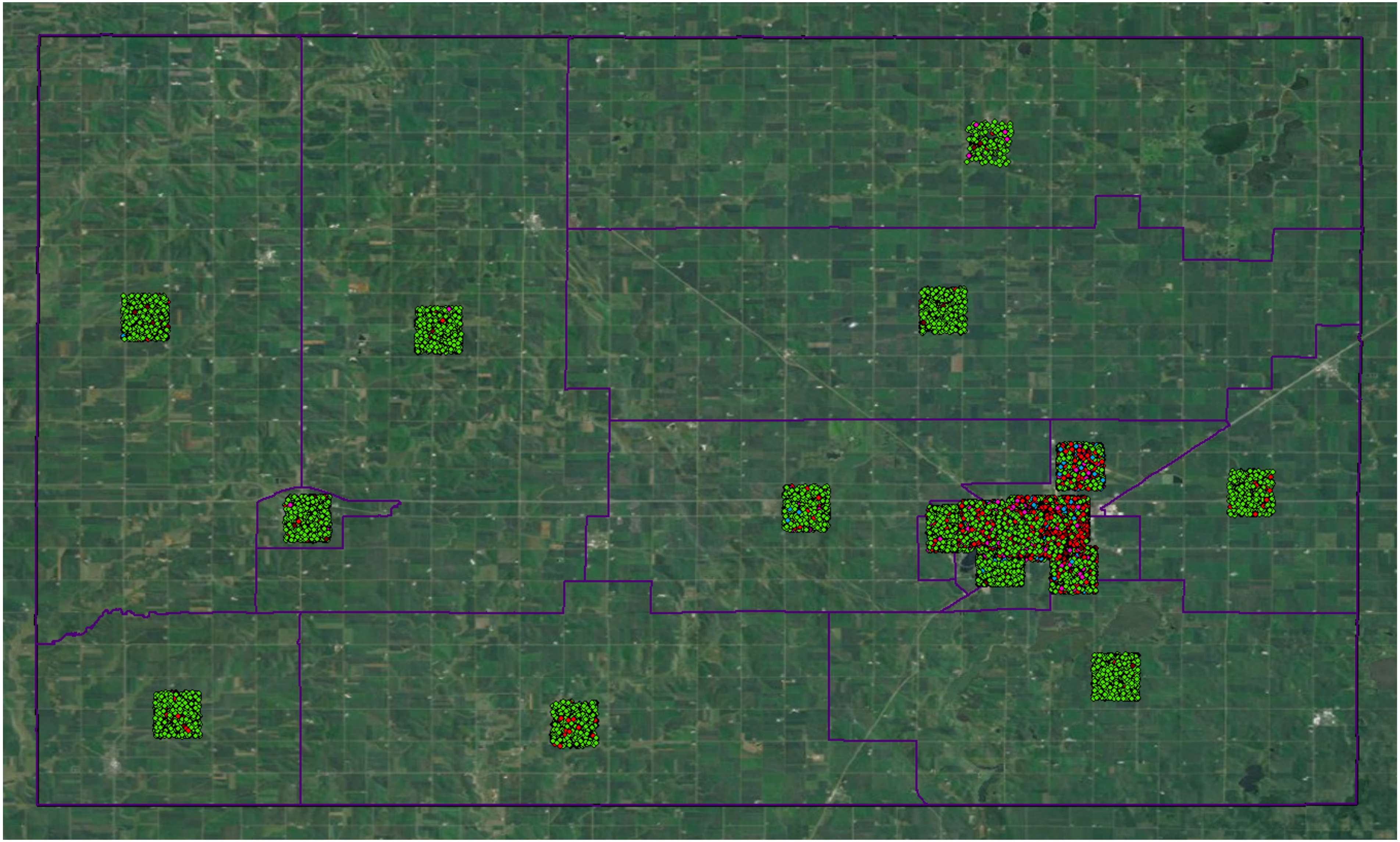
CBSA=Worthington, centered around Census Tract centroid



Source: Simulated “Synthetic” Data created from Census PUMS Data

- ◆ Asian
- ◆ Black
- ◆ Hispanic
- ◆ Other
- ◆ White

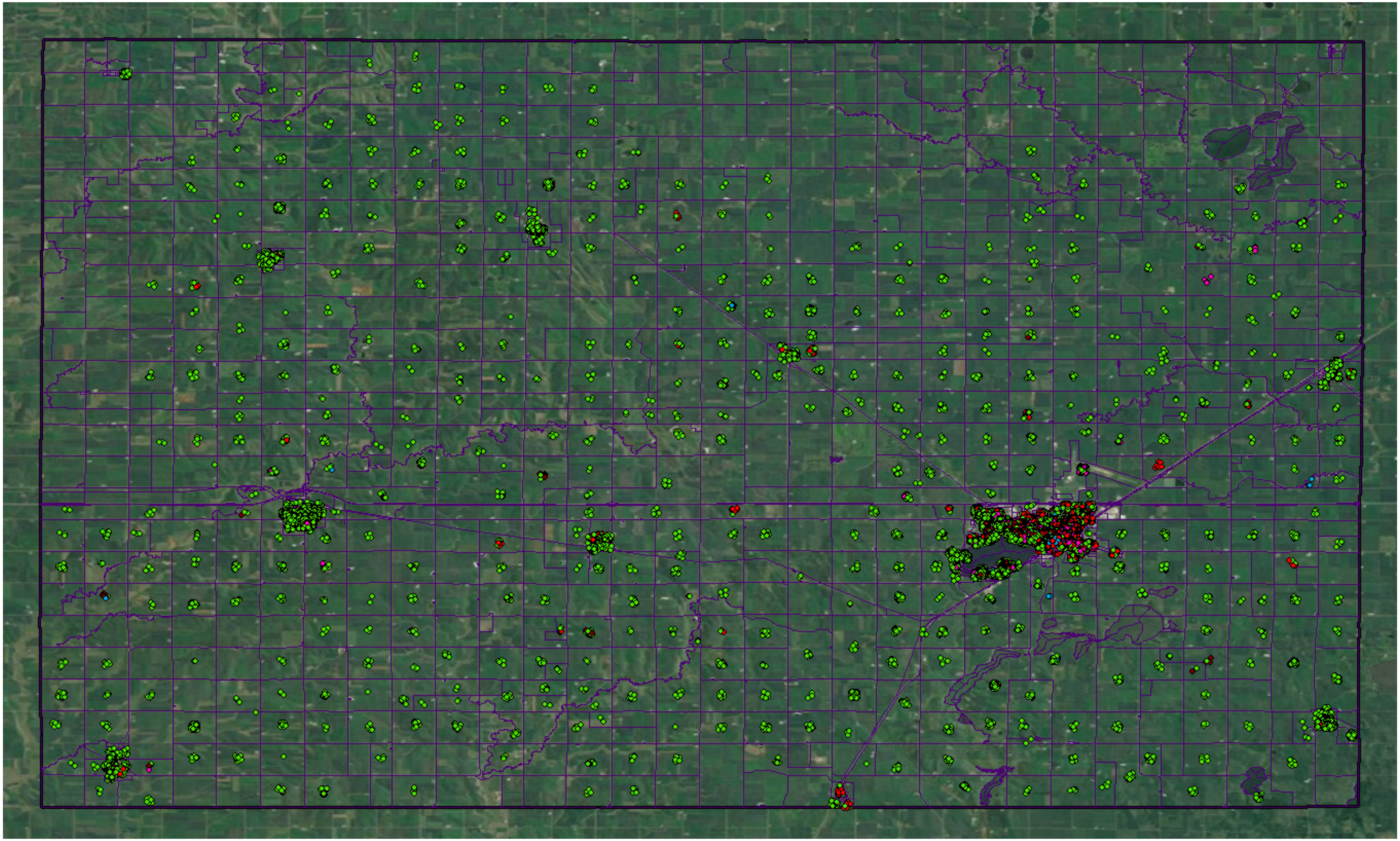
CBSA=Worthington, centered around Block Group centroid



Source: Simulated “Synthetic” Data created from Census PUMS Data

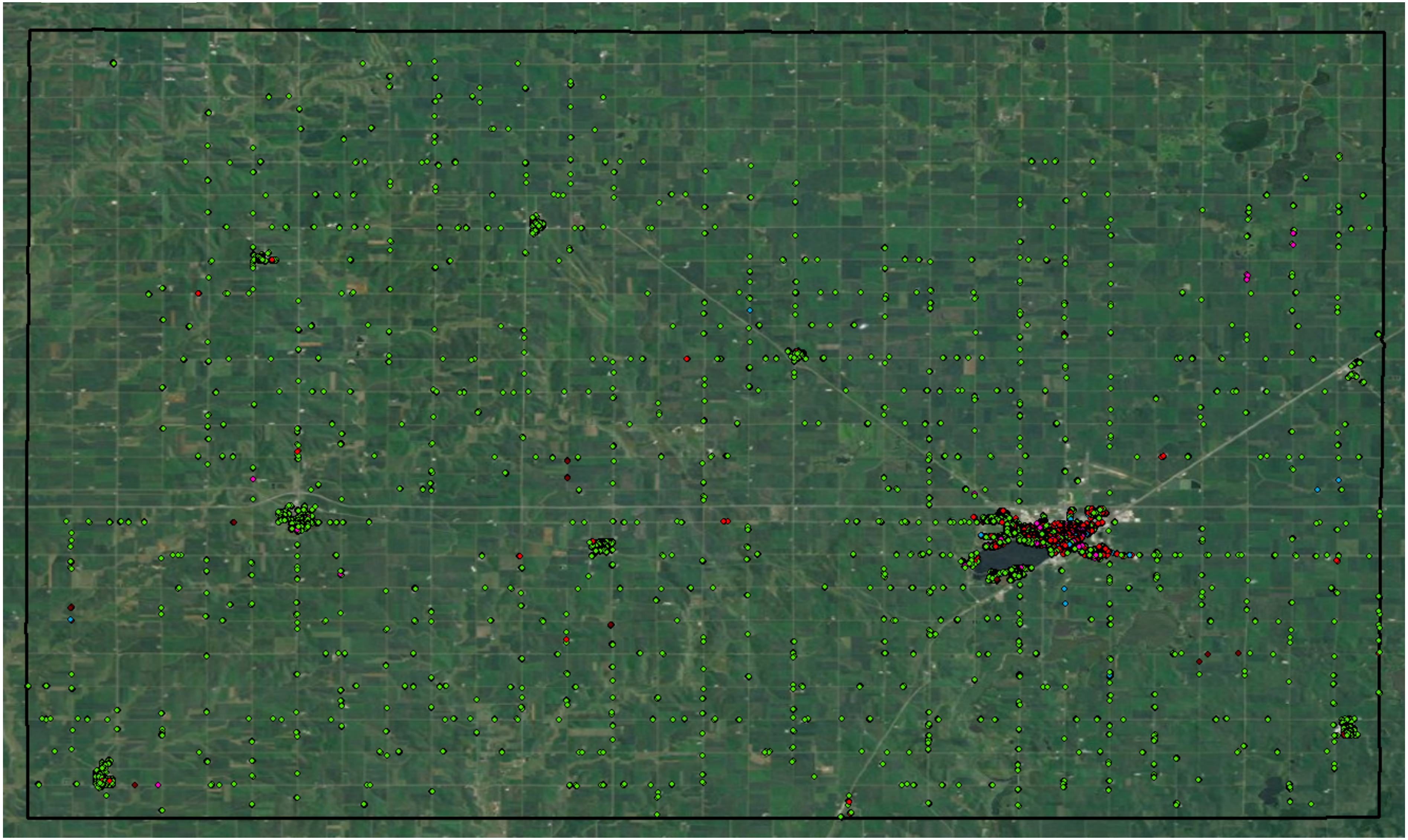
◆ Asian
◆ Black
◆ Hispanic
◆ Other
◆ White

CBSA=Worthington, centered around Census Block centroid



Source: Simulated “Synthetic” Data created from Census PUMS Data

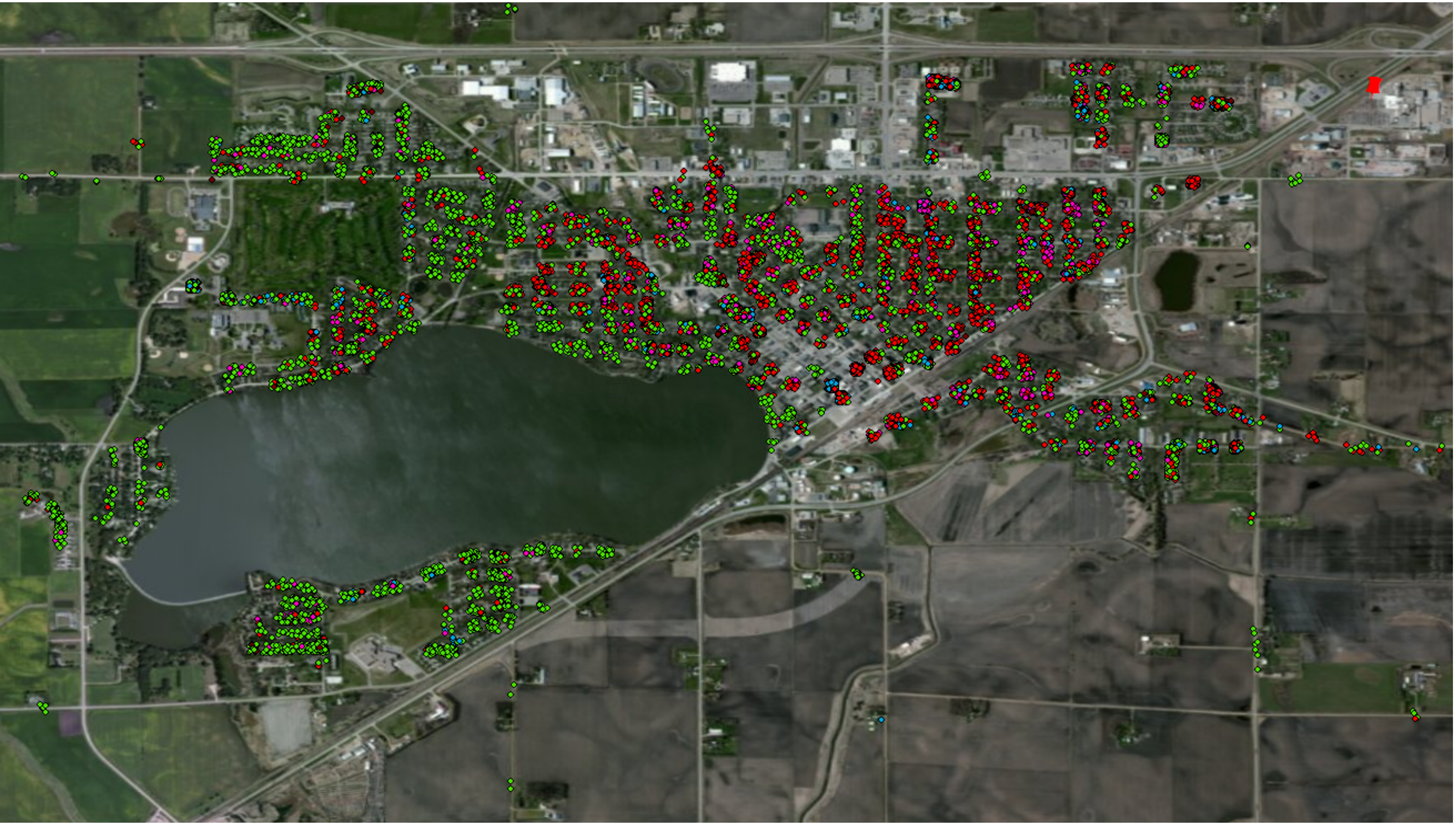
CBSA=Worthington, MN
original individuals by household



Source: Simulated “Synthetic” Data created from Census PUMS Data

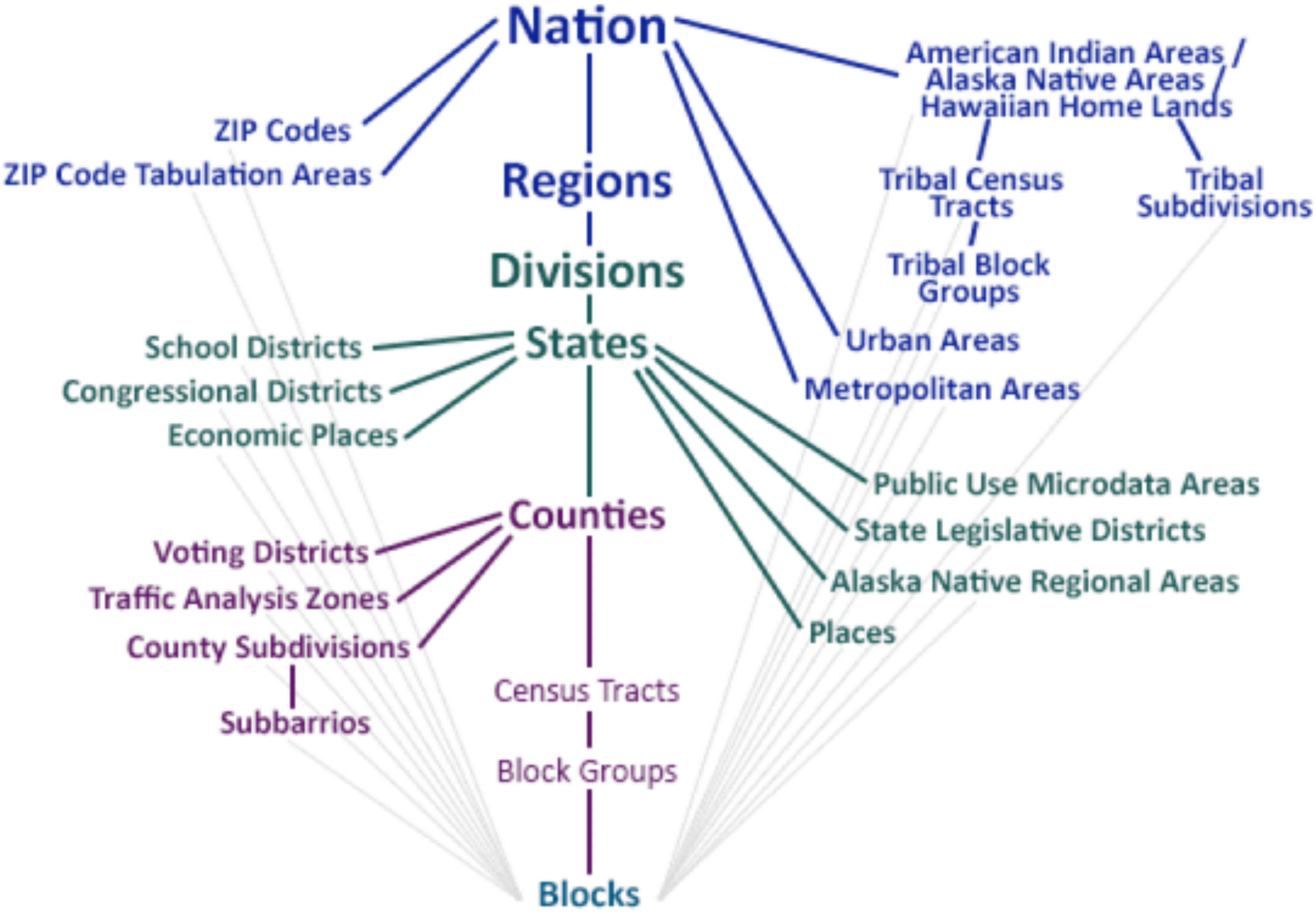
- ● Asian
- ● Black
- ● Hispanic
- ● Other
- ● White

CBSA=Worthington, original individuals by household, focusing on the city of Worthington



Source: Simulated “Synthetic” Data created from Census PUMS Data

"Off Spline" Geographies and Subtraction Geographies

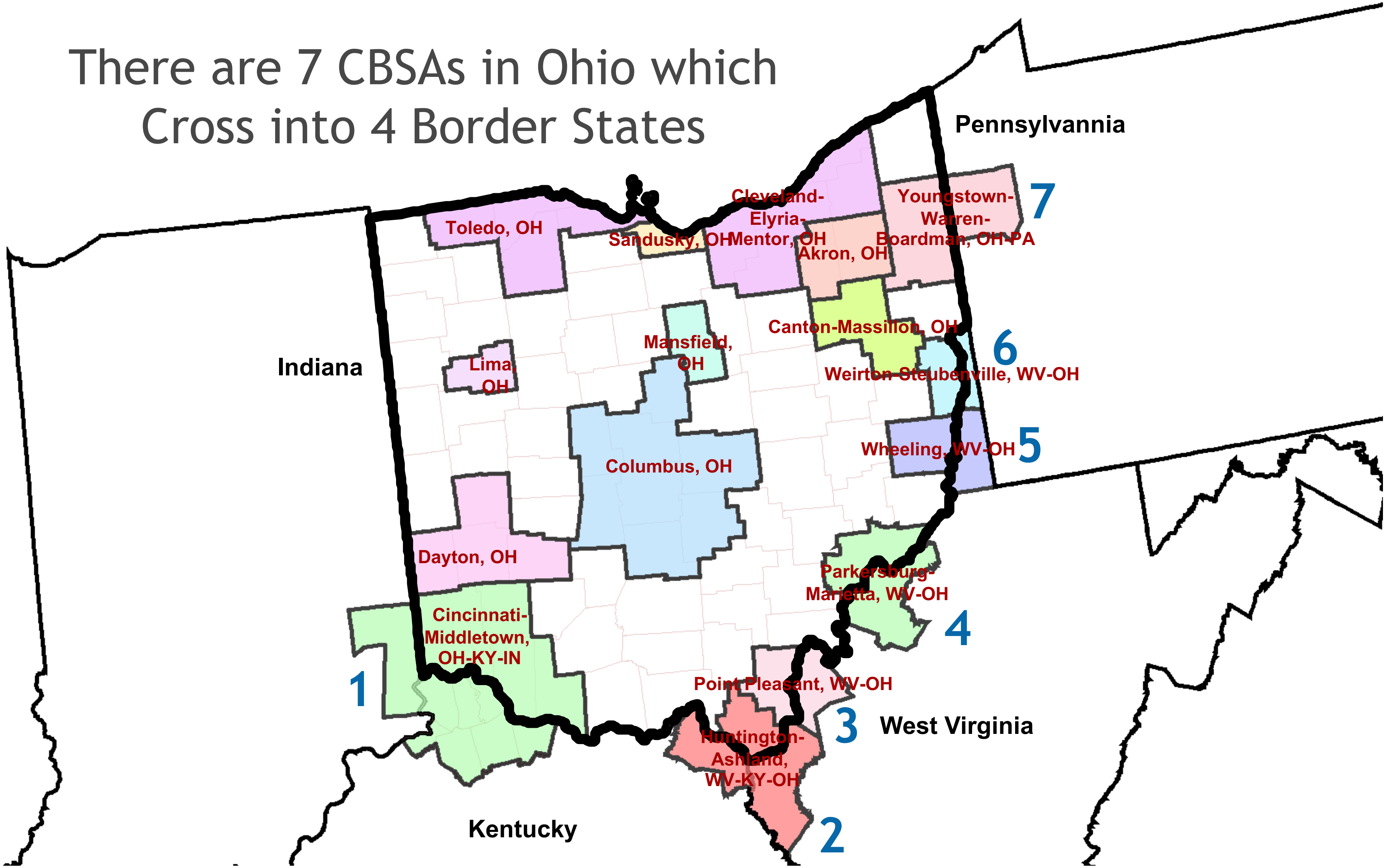


Challenge: Subtraction Geography (i.e., Geographical Differencing)

- Challenge: Data recipients often request reporting on more than one geography (e.g., both State and 3 digit Zip code).
- ***Subtraction Geography*** creates disclosure risk problems when more than one geography is reported for the same area and the geographies overlap.
- Also called ***geographical differencing***, this problem occurs when the multiple overlapping geographies are used to reveal smaller areas for re-identification searches.

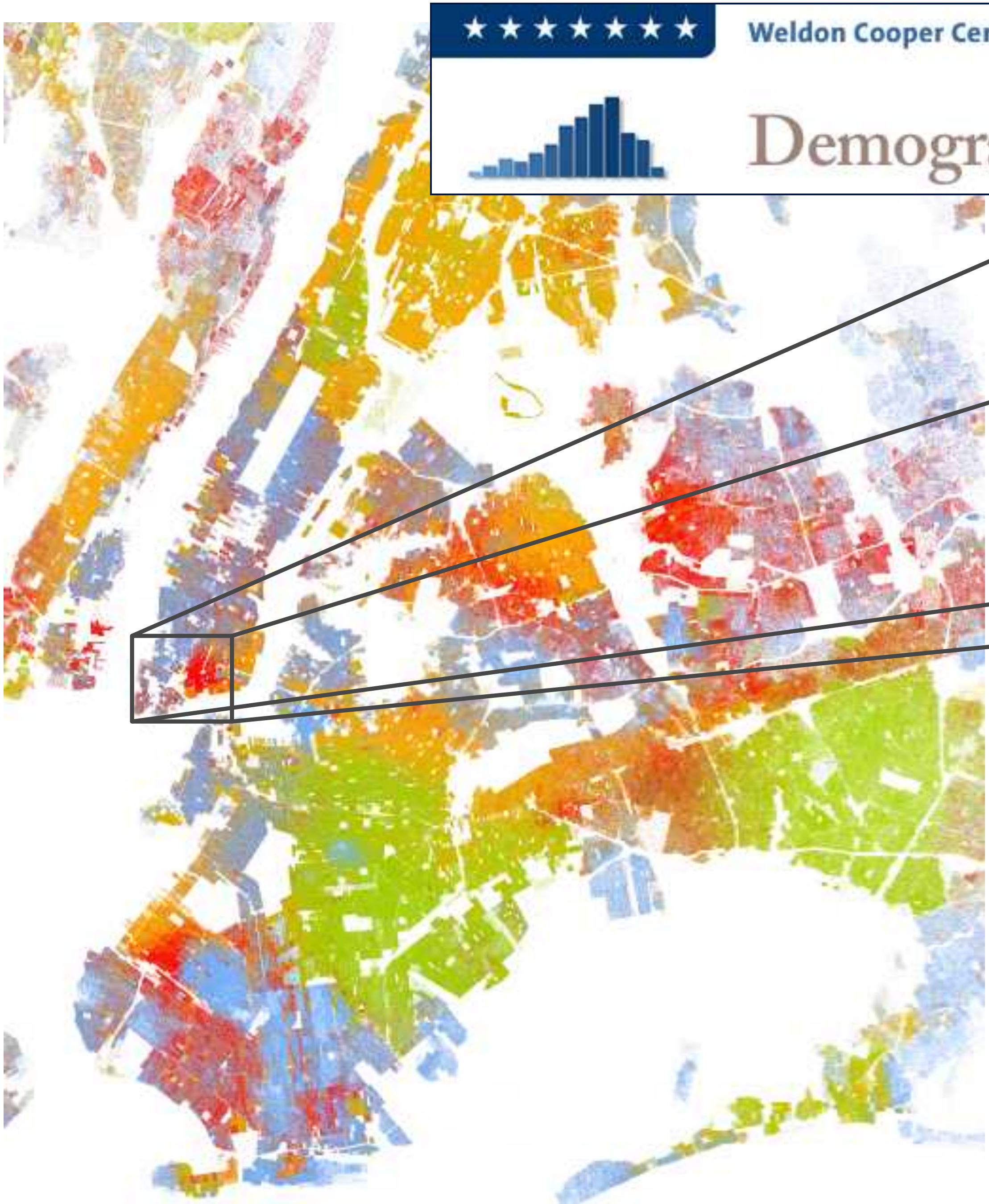
Example: OHIO Core-based Statistical Areas

There are 7 CBSAs in Ohio which Cross into 4 Border States



Weldon Cooper Center for Public Service • University of Virginia

Demographics Research Group



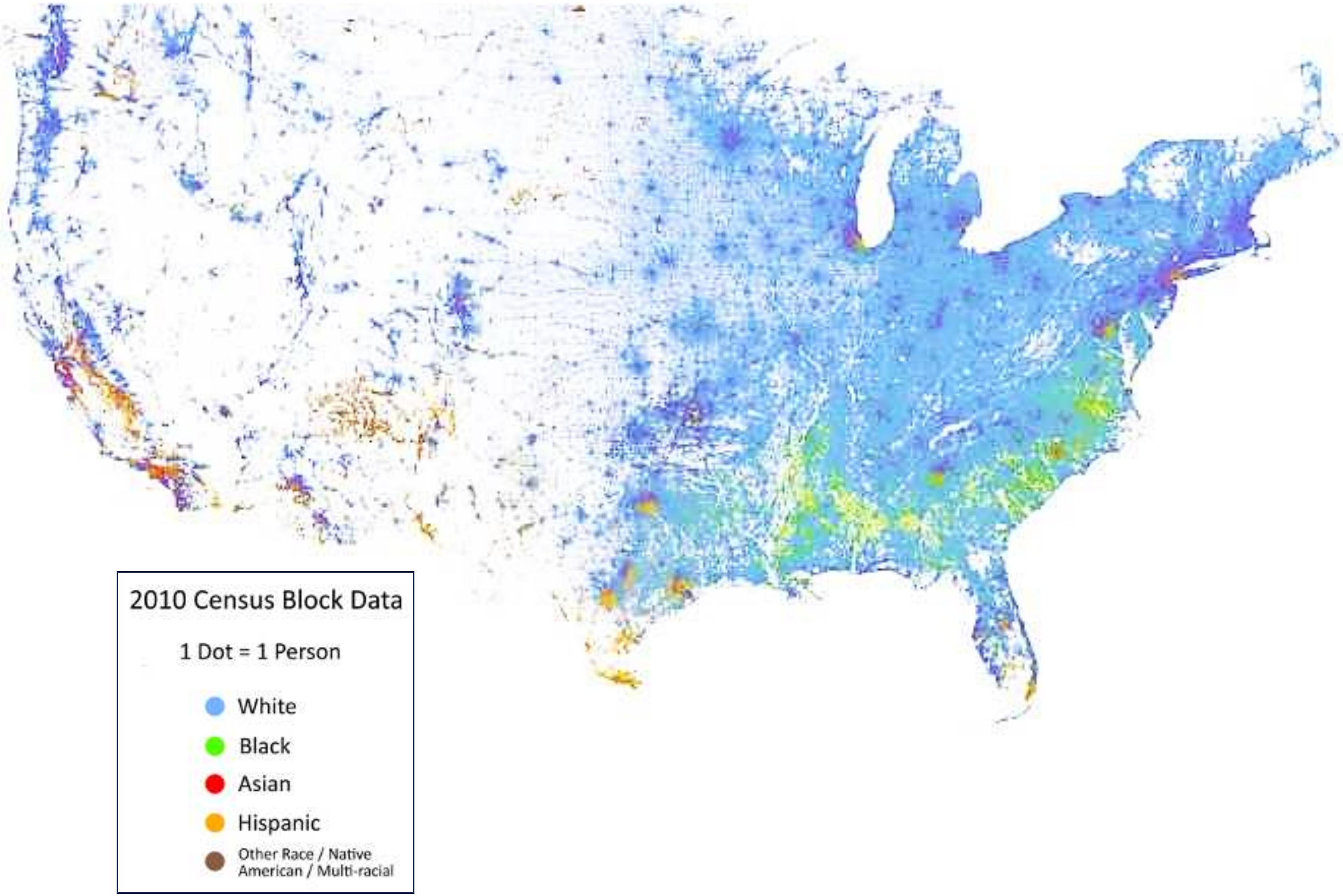
The Racial Dot Map

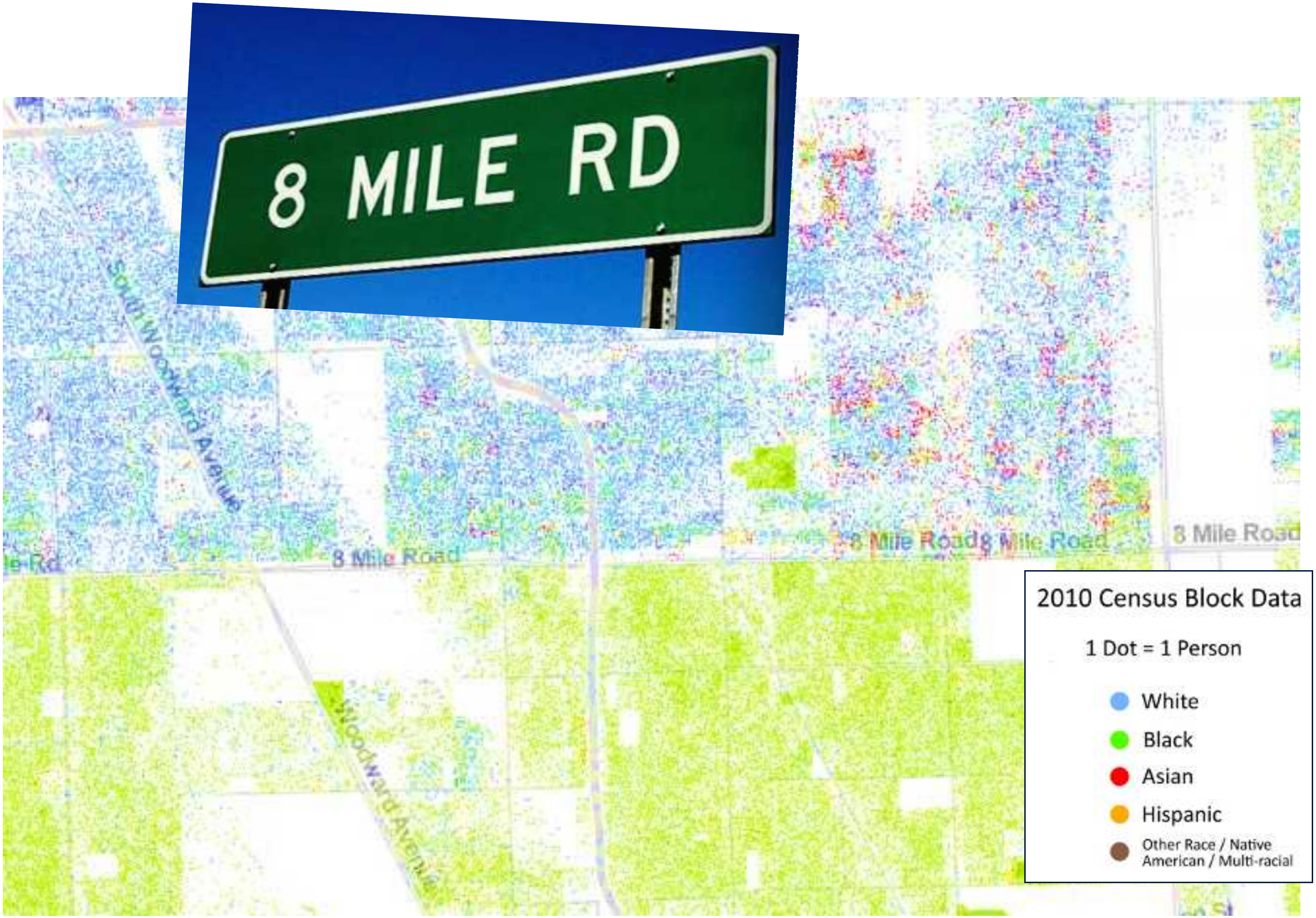
One Dot Per Person for the Entire United States

Created by Dustin Cable, July 2013

This is the most comprehensive map of race in America ever created.

http://demographics.coopercenter.org/DotMap/index.html




8 MILE RD
South Woodward Avenue
8 Mile Road
8 Mile Road
8 Mile Road
8 Mile Road
Woodward Avenue
2010 Census Block Data
1 Dot = 1 Person
● White
● Black
● Asian
● Hispanic
● Other Race / Native American / Multi-racial