Data Science for the Public Good

# USING SYNTHETIC POPULATIONS TO ASSESS IDENTIFIABILITY RISKS IN THE AMERICAN HOUSING SURVEY

Mark Almanza (Virginia Tech), Adrienne Rogers (Virginia Tech) with Josh Goldstein & Dave Higdon (SDAL),
Shawn Bucholtz (Housing and Urban Development) & Tamara Cole (Census Bureau)

## Overview

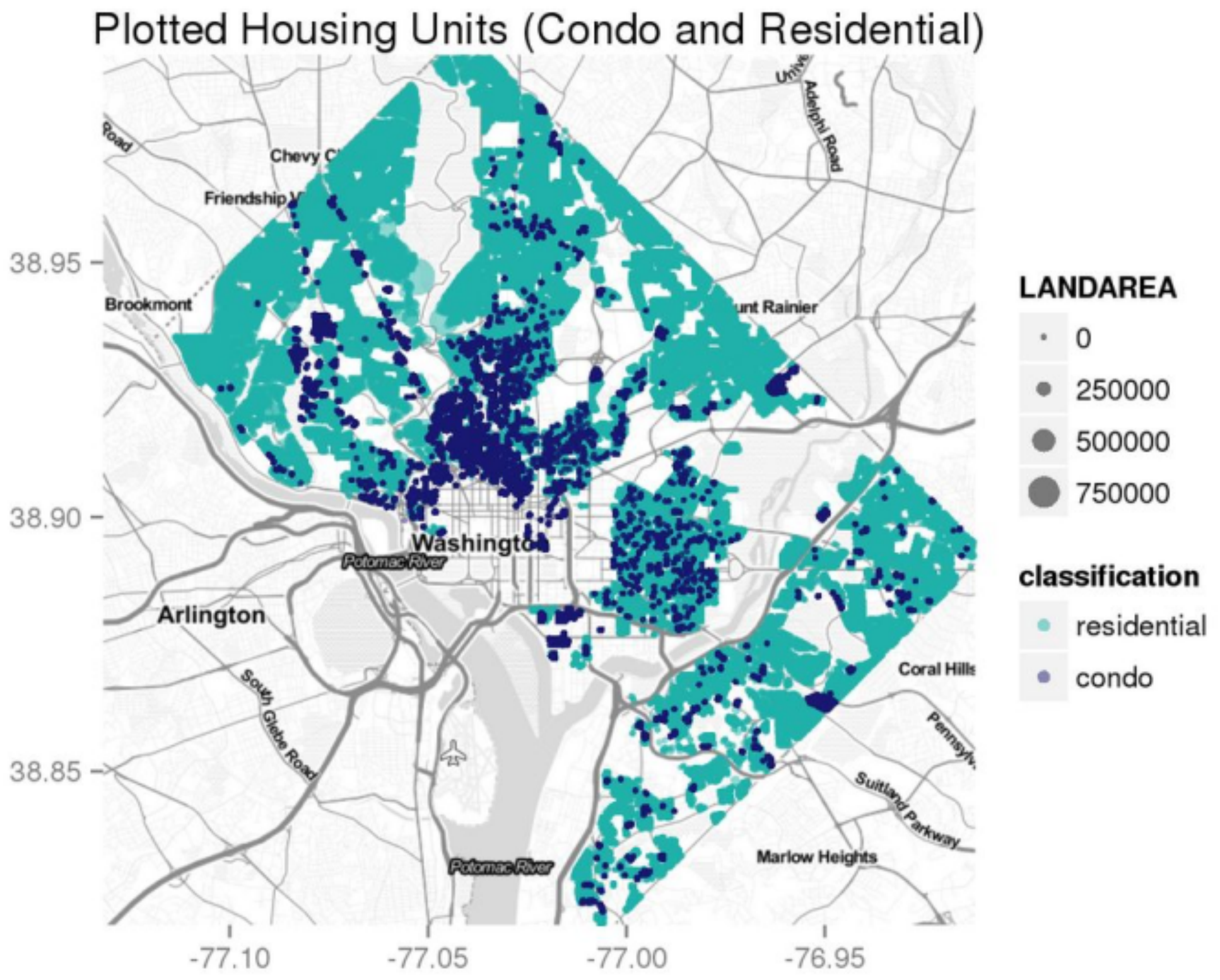
- SDAL is assessing the feasibility of matching public records with the 2013 American Housing Survey (AHS) sample housing units for the purpose of disclosure avoidance research.
- The AHS is a biennial survey of housing units conducted by the Census Bureau on behalf of the Department of Housing and Urban Development (HUD).
- Protecting confidentiality in Federal surveys is a legal requirement and necessary to ensure respondent participation.

The AHS collects the following information about each occupied housing unit:

- Ownership (owner's name)
- Lot characteristics (size, zoning, sewer connection)
- Structure characteristics (type, size, stories)
- Physical characteristics (rooms, bathrooms, size, year built, heating and cooling equipment)
- Demographic profile of the housing unit occupants
- Financial characteristics (mortgage, owner occupied only)
- Property Tax and other assessment information

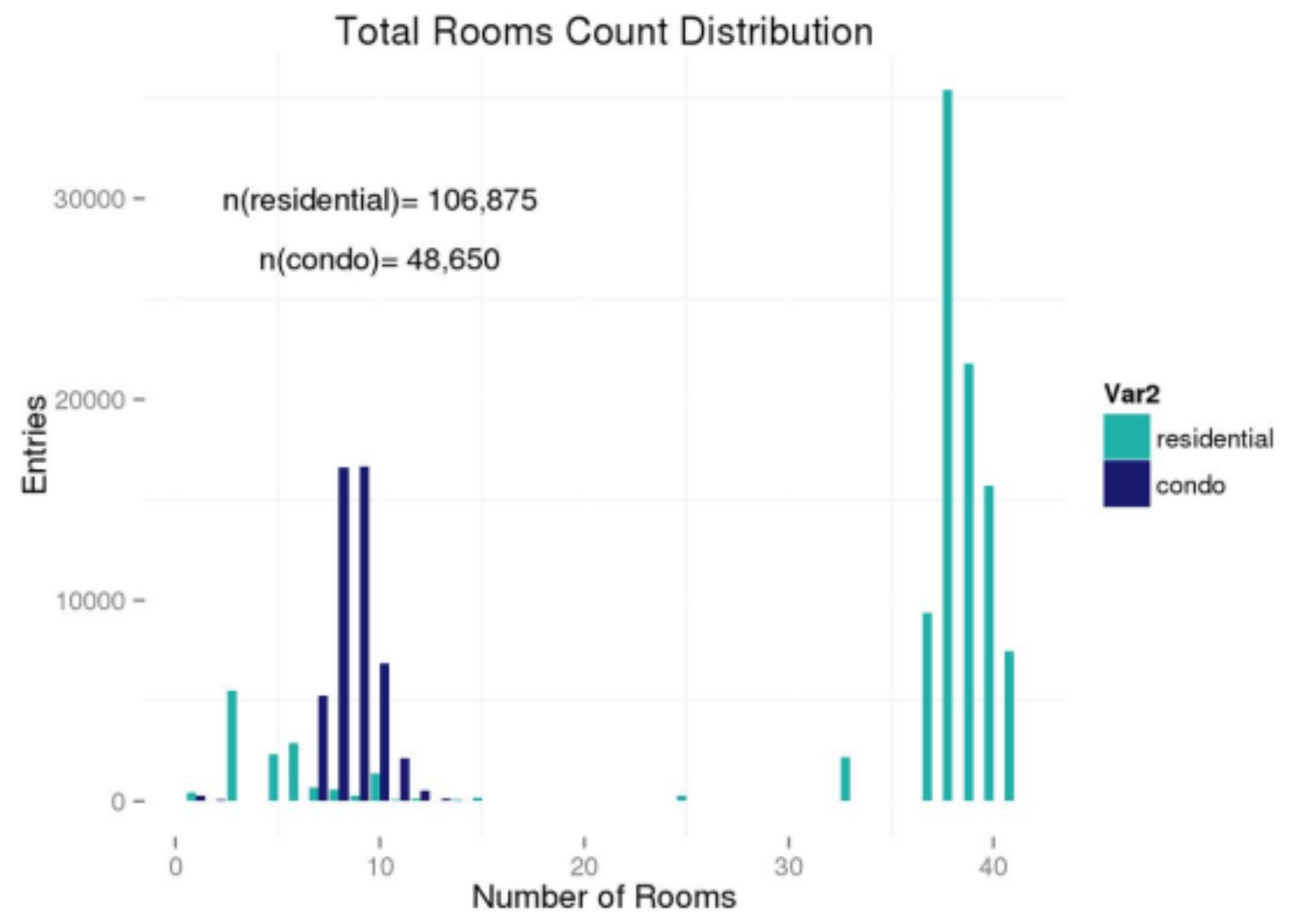
- A standard product for the AHS is public use microdata files (PUFs). PUFs contain answers to survey questions at the individual, household, and housing unit level.
- Some of the information in these PUFs is replicated in publicly available data sources – tax records, mortgage information, geographical information. CoreLogic claims to supply such information for 80% of homes in the US.
- SDAL will evaluate the risk of disclosure when linkages between the AHS PUFs and publicly available data are made.
- We are assessing this risk by:
  1. Building a synthetic population of households and occupants to mimic those of Arlington, Fairfax, and DC;
  2. Linking these synthetic households to publically available data.

## Overview of DC Data

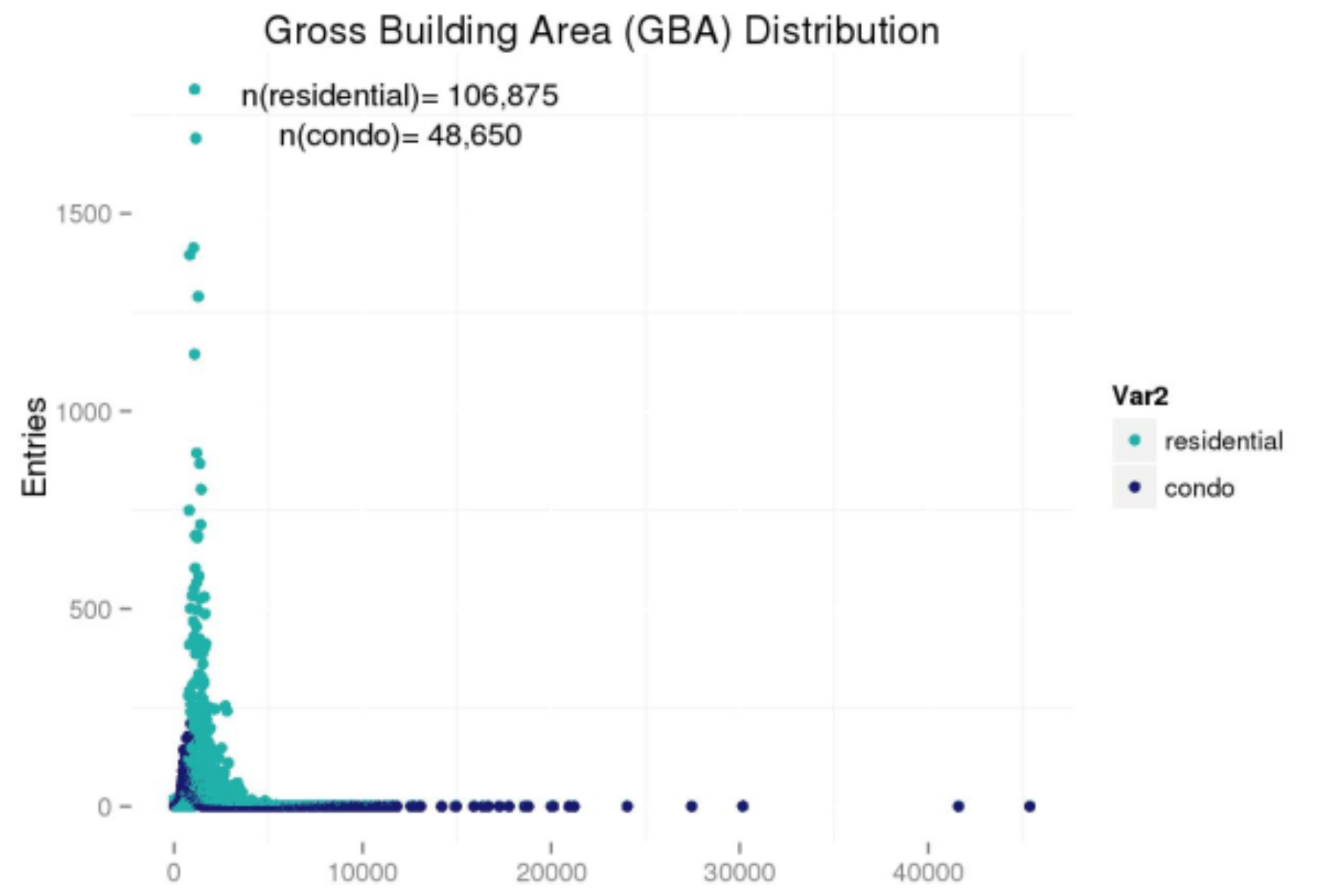


The DC Computer Assisted Mass Appraisal (CAMA) data contains residential, condominium, and commercial property data. For the purpose of this project, just the residential and condominium are used.

To better understand the data, basic data profiling occurred. The main goal was to understand how these data match the CoreLogic Data. This same step was used for all other administrative datasets included in the project.



Variables such as total room count, bedroom count, bathroom count, etc. help to understand the population needed for proper simulation.

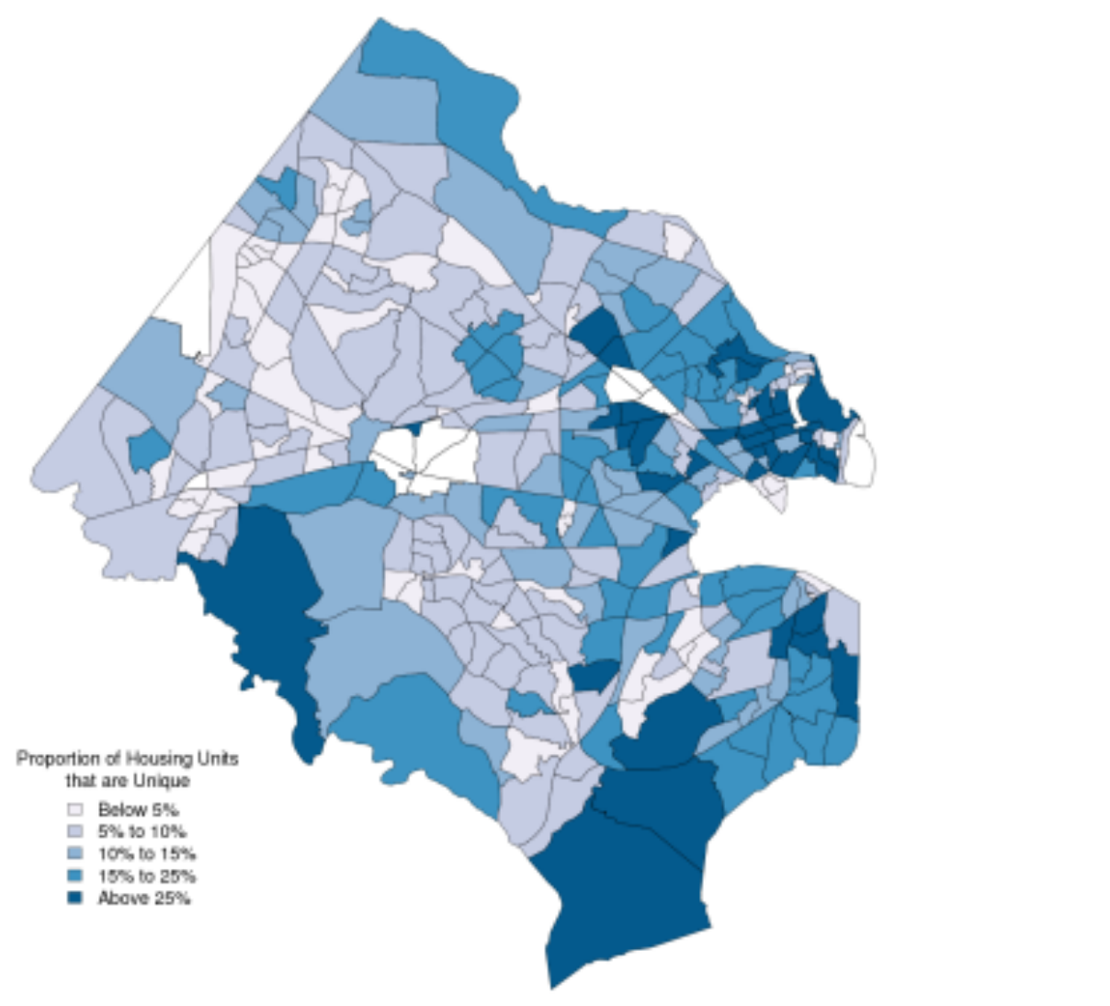


Structural variables such as Gross Building Area (GBA), Land Area, Heat, AC, Fireplace count, Exterior Wall type etc., are used to help identify singular locations.

## Assessing Identifiability Risk

- We measure identifiability through population uniqueness; the proportion of housing units that have unique values on a set of key variables.
- The variables used are ones that can be found in both the CoreLogic tax records and AHS data.

- **Continuous variables** (e.g. year built) are binned into quartiles.
- **Count variables** (e.g. number of bedrooms) are topcoded.
- Decreasing the bin size increases the proportion of unique values.



Proportion of unique values for single family properties by Census tract.

- The study area includes Arlington County, Fairfax County, Falls Church, and Fairfax City.
- 46,702 single family housing units (12% of the population) are uniquely identified. 81,916 units (21%) are identified to within a group of three units.

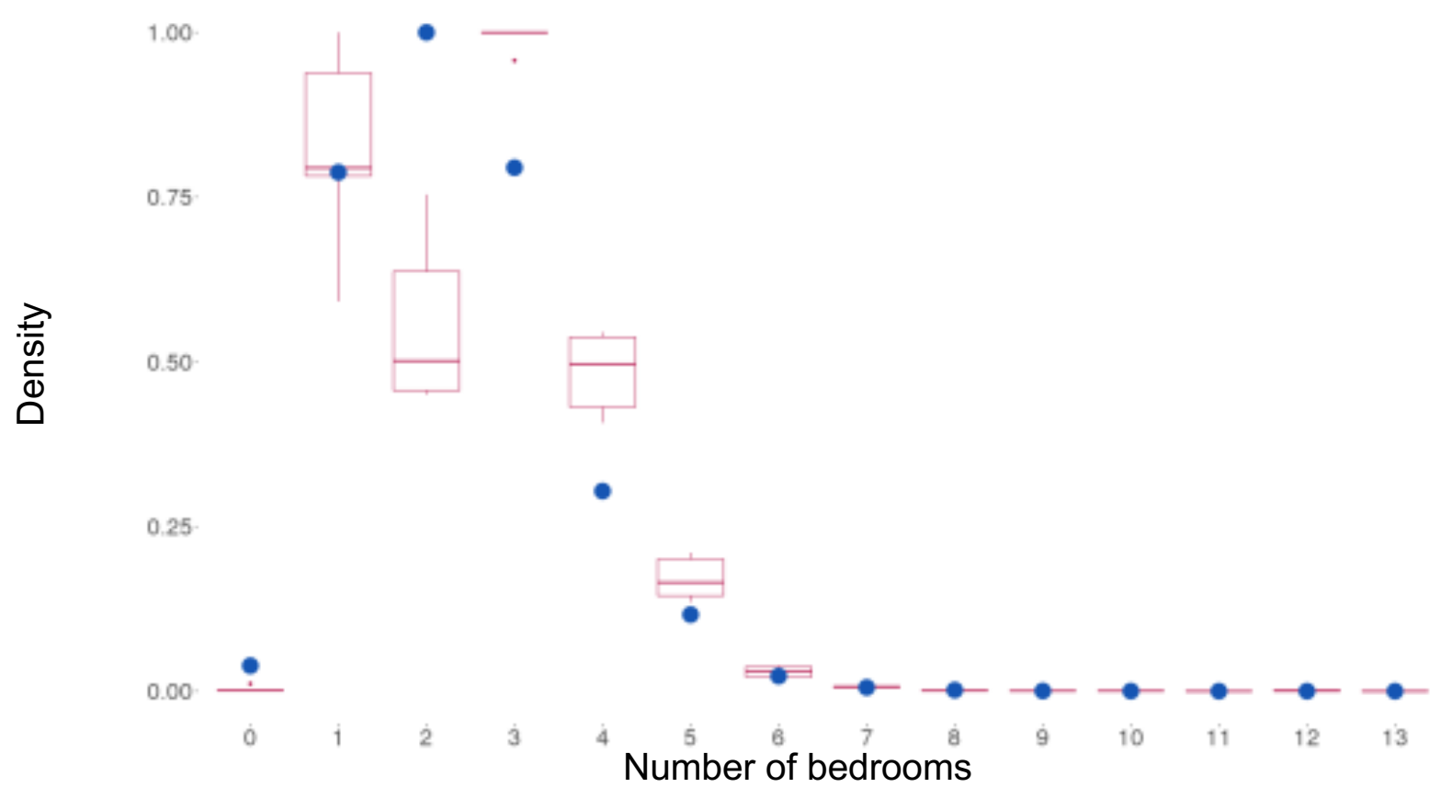
- Holdout method (right) measures uniqueness when a single variable is removed.
- Ex: Removing PUMA indicator reduces the proportion of uniquely identified units from 12% to 7%.

| Variable Removed | # Unique Units | % Unique Units |
|---|---|---|
| None | 46,702 | 12% |
| Parking | 46,702 | 12% |
| New Construction | 46,249 | 12% |
| Garage | 45,898 | 12% |
| Heating | 44,007 | 11% |
| Air Conditioning | 42,759 | 11% |
| Value | 40,399 | 10% |
| Number of Stories | 39,917 | 10% |
| Fireplace | 39,763 | 10% |
| Basement | 39,641 | 10% |
| Year Built | 37,461 | 9% |
| Absentee Owner | 36,346 | 9% |
| Real Estate Taxes Paid | 35,573 | 9% |
| Lot Size | 34,114 | 9% |
| Living Area | 33,608 | 8% |
| Bathrooms | 31,481 | 8% |
| Bedrooms | 28,528 | 7% |
| PUMA | 28,209 | 7% |

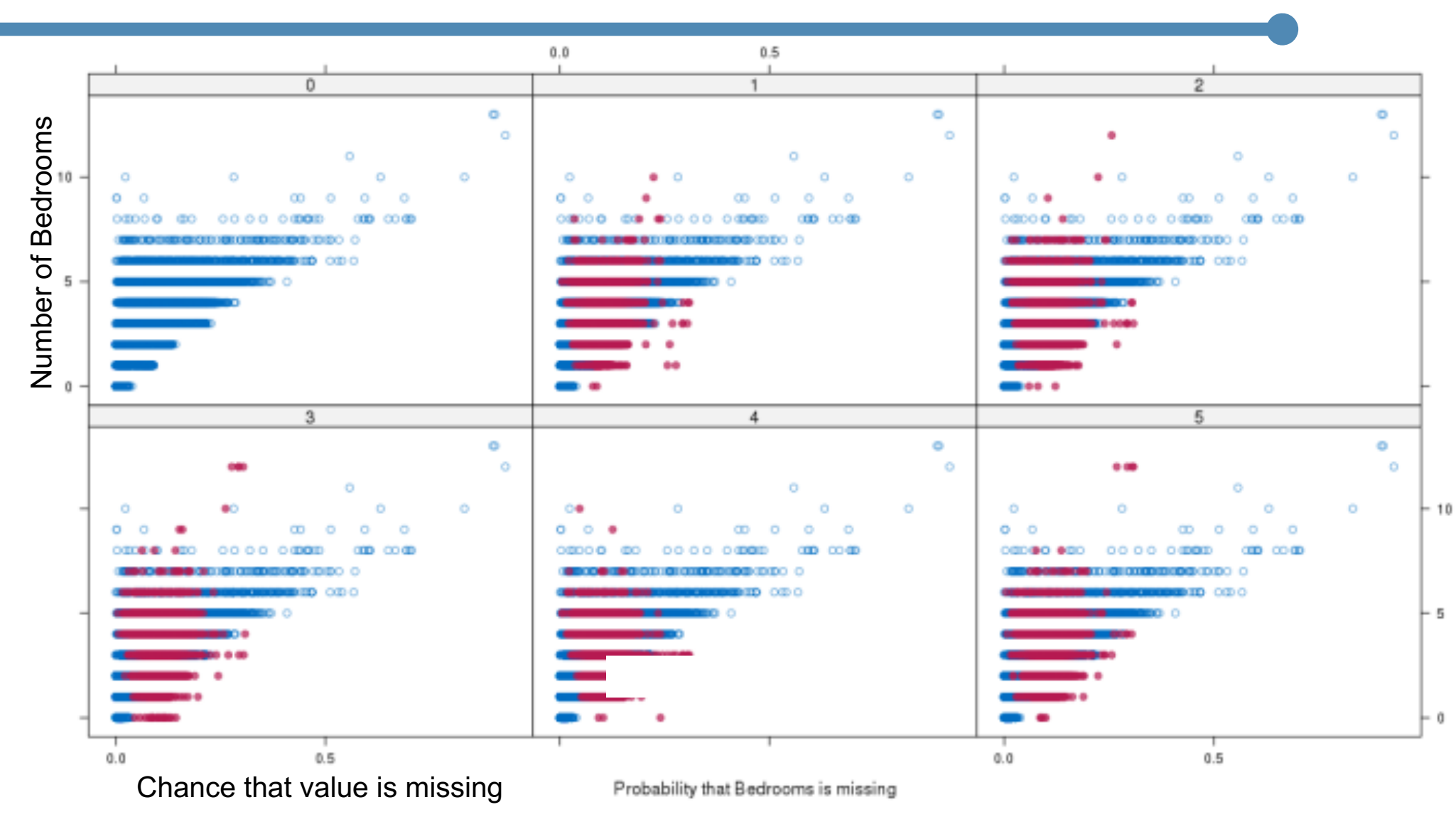
## Imputing Properties for Multifamily Housing Units

- A problem in estimating identifiability risk for all of Arlington County is the lack of data on multifamily units in the county.
- CoreLogic treats multifamily buildings as one entry, losing household by household information.

- Imputation of missing values can help create the data needed for the study of indentifiability risk.
- Multiple imputation, as shown here, enables the use of one or more variables to predict values for these missing data.



Distribution of observed values for number of bedrooms (blue) and the distribution of imputed values (red)



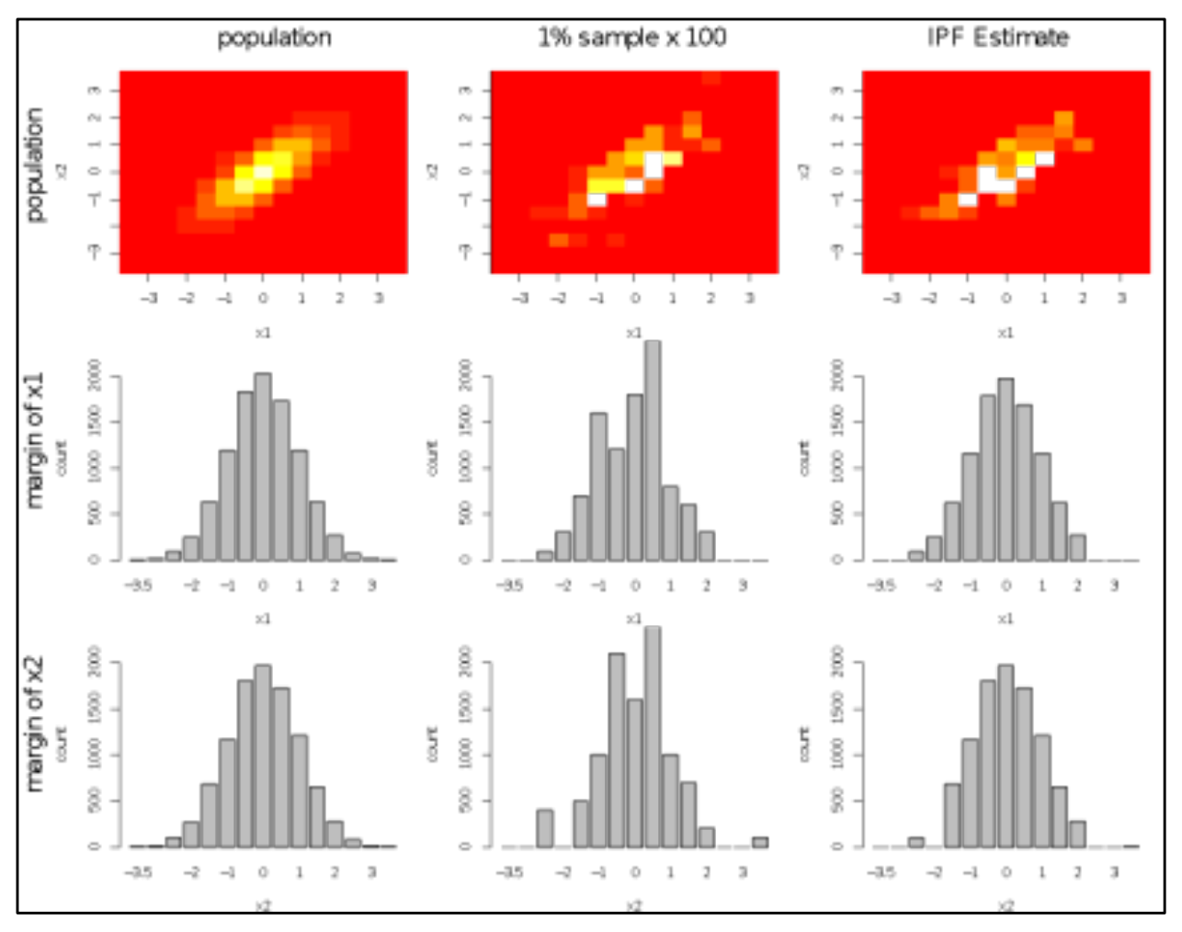
Scatterplots showing the probability that any given datum is missing from the observed values, along with distributions for each imputation

## Constructing the Synthetic Population

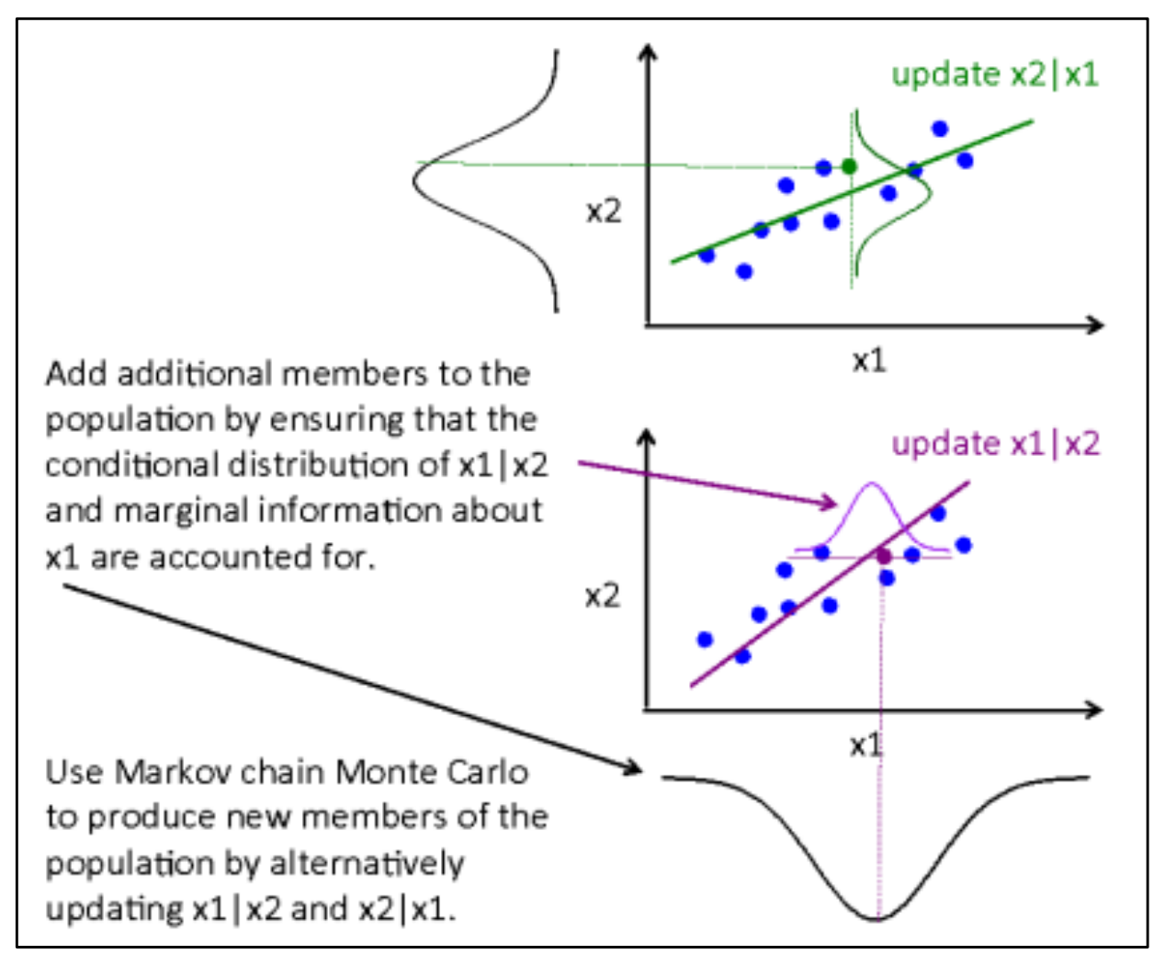
- SDAL seeks to construct a synthetic population that matches the variation of occupants/families in households for Arlington.
- Standard methods of generating synthetic populations (e.g. Beckman et al., 1996) replicate microsample files to match marginal distributions of key variables (householder age, householder income, household size) at the block group level.
- This replication is not suitable for identifiability studies since variations in income, age, size, tenure, etc., are what make certain respondents identifiable.

**Example**: A population (N=10000) with two continuous characteristics (x1 and x2) is surveyed, producing:

1. 2 tables summarizing marginals of the full population.
2. Micro data for a SRS of 100 respondents.



- SDAL has borrowed concepts from Bayesian multiple imputation (Raghunathan, Reiter and Rubin, 2003) to construct a population with realistic individual-level variation that matches margins of variables supplied by Census summary tables.
- The resulting population is produced by treating the full population as unknowns, and producing posterior realizations of the population using Markov Chain Monte Carlo.
- Simple 2-D studies will be adapted to produce synthetic households for Arlington and Fairfax counties.



Add additional members to the population by ensuring that the conditional distribution of x1|x2 and marginal information about x1 are accounted for.

Use Markov chain Monte Carlo to produce new members of the population by alternatively updating x1|x2 and x2|x1.

Virginia Tech
Biocomplexity Institute

## References

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415-429.
Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*(1), 1.

bi.vt.edu/sdal