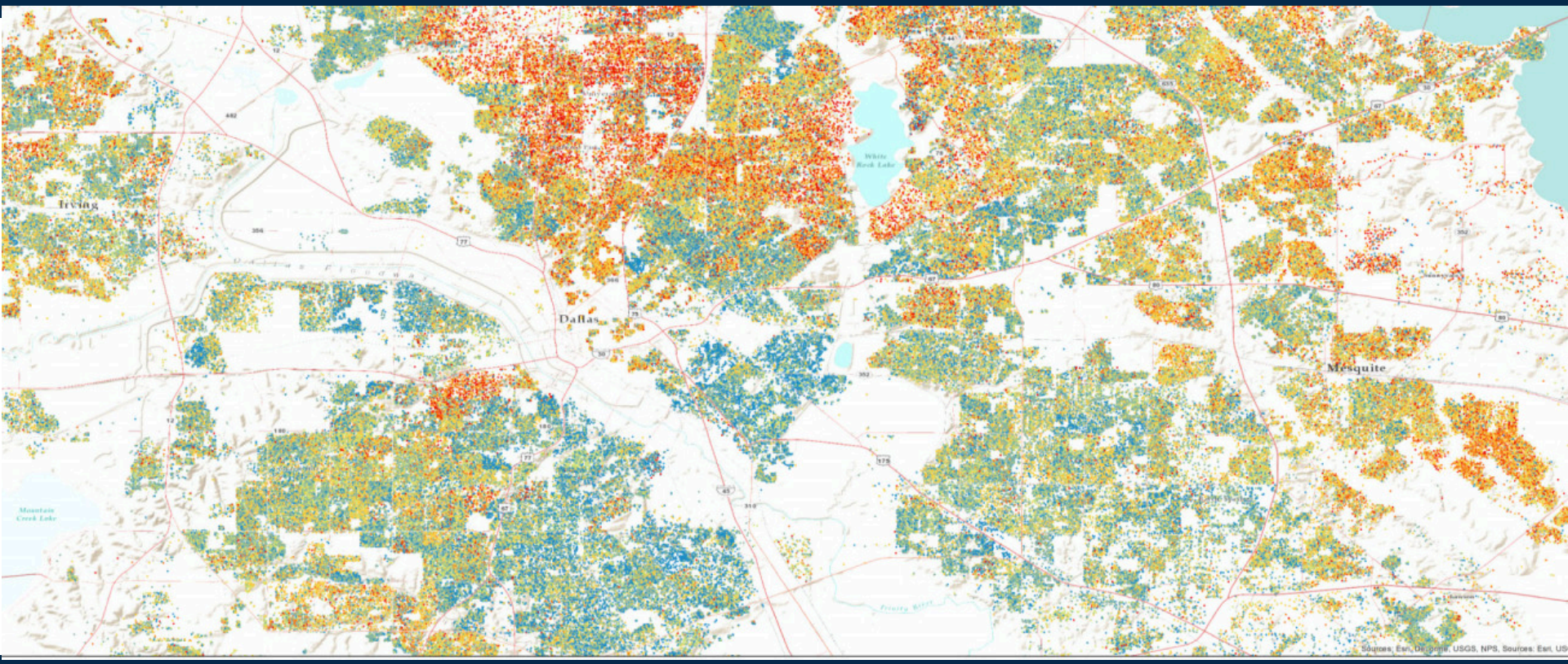
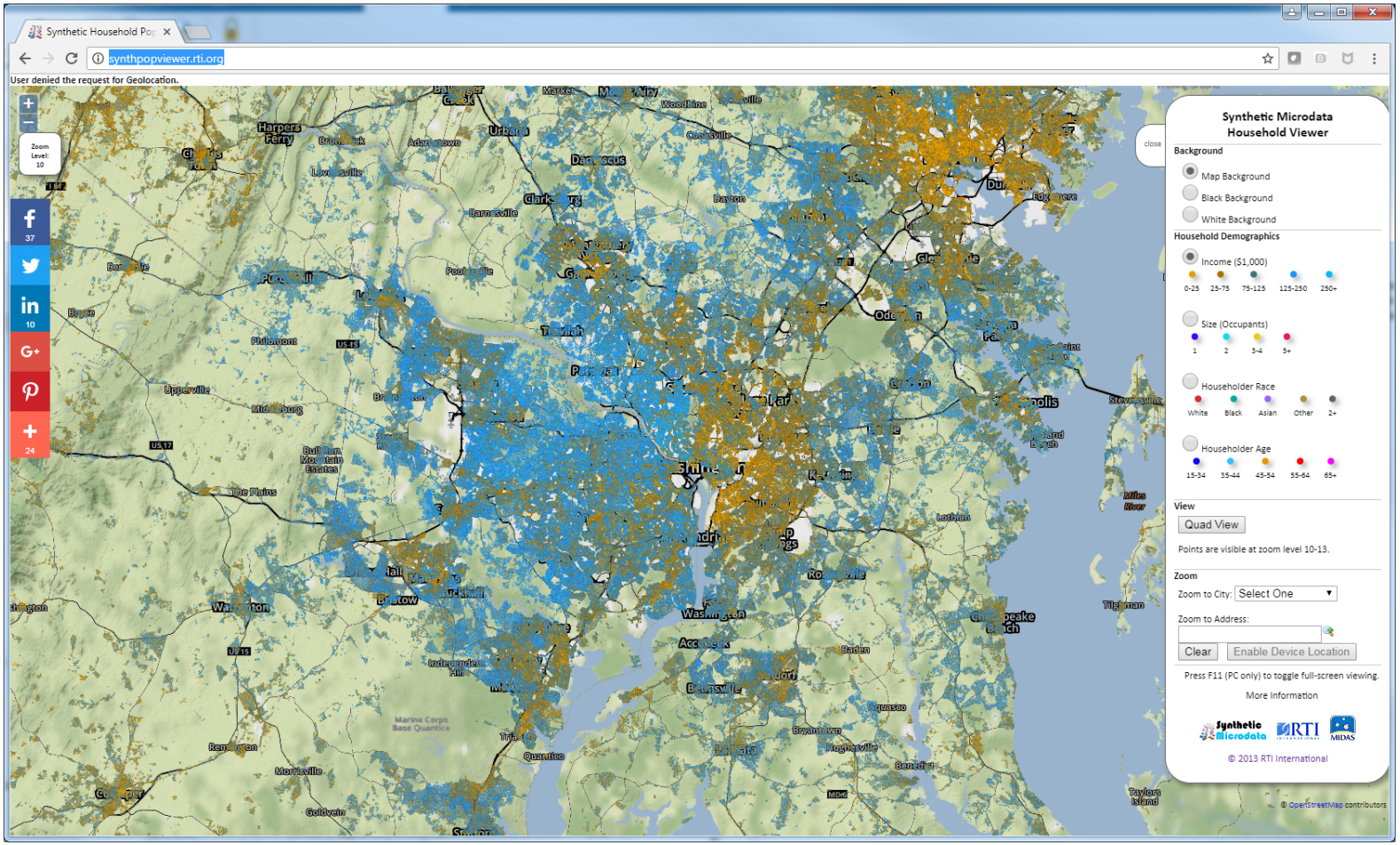
# Introduction to Synthetic Populations

SICSS 2019

Kasey Jones & Jay Rineer

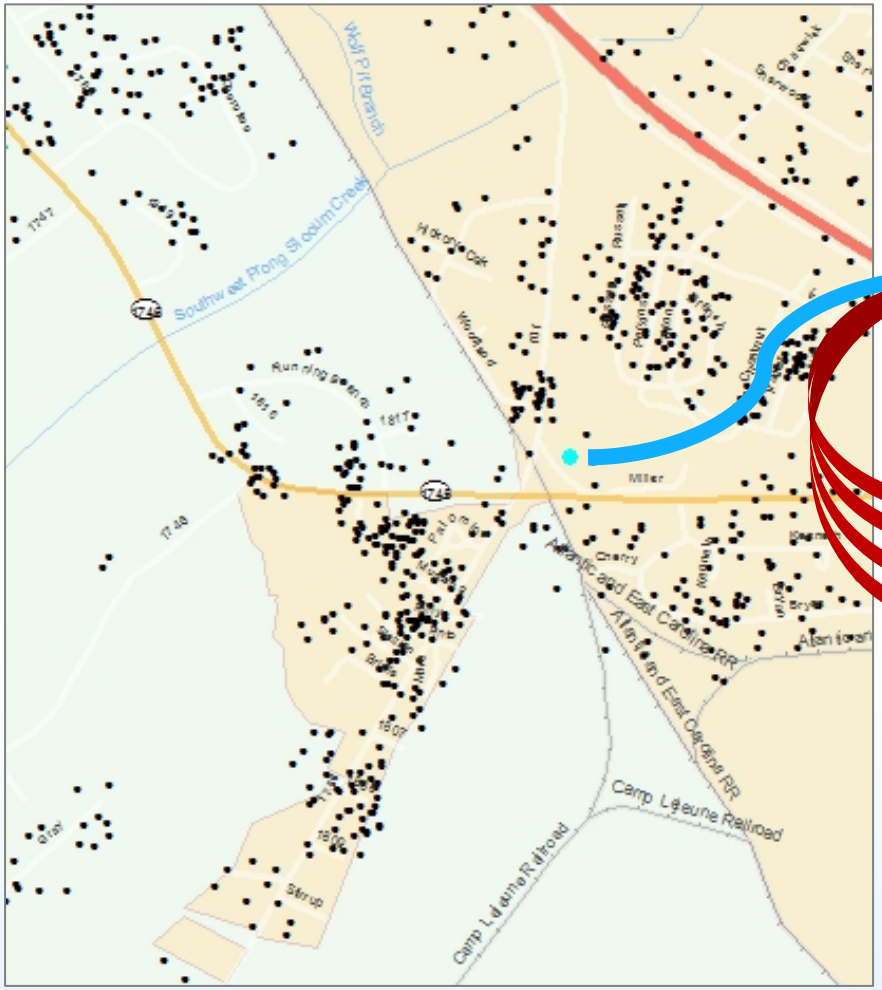


# Synthetic Populations




http://synthpopviewer.rti.org/

# Synthetic Populations




## Household

| ID | Persons | Age of HH | Income | Race of HH |
|---|---|---|---|---|
| 9903 | 4 | 47 | 55,000 | White |

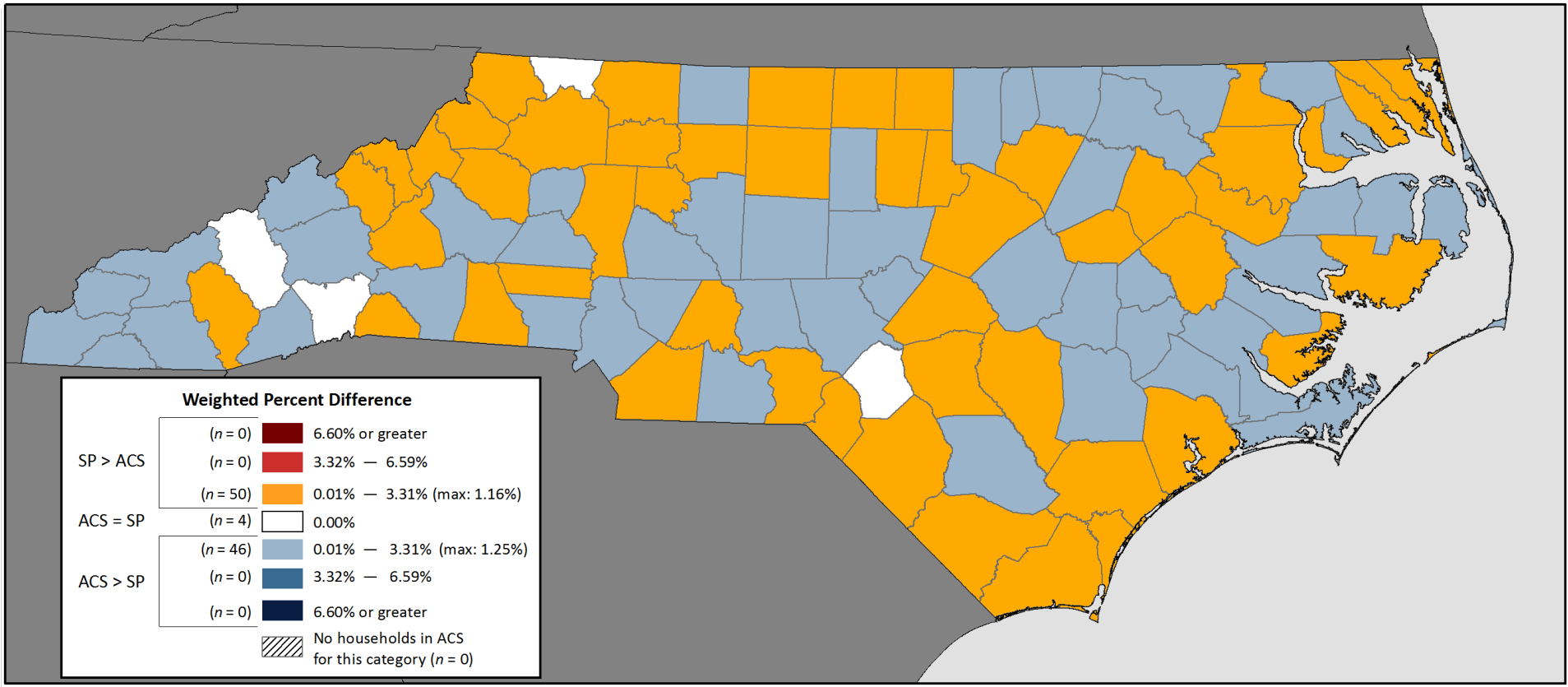
## Persons

| ID | Person Num | Age | Sex | school_ID | work_ID |
|---|---|---|---|---|---|
| 9903 | 1 | 47 | Male | N/A | 23401 |
| 9903 | 2 | 45 | Female | N/A | N/A |
| 9903 | 3 | 15 | Female | 18047 | N/A |
| 9903 | 4 | 10 | Female | 34789 | N/A |

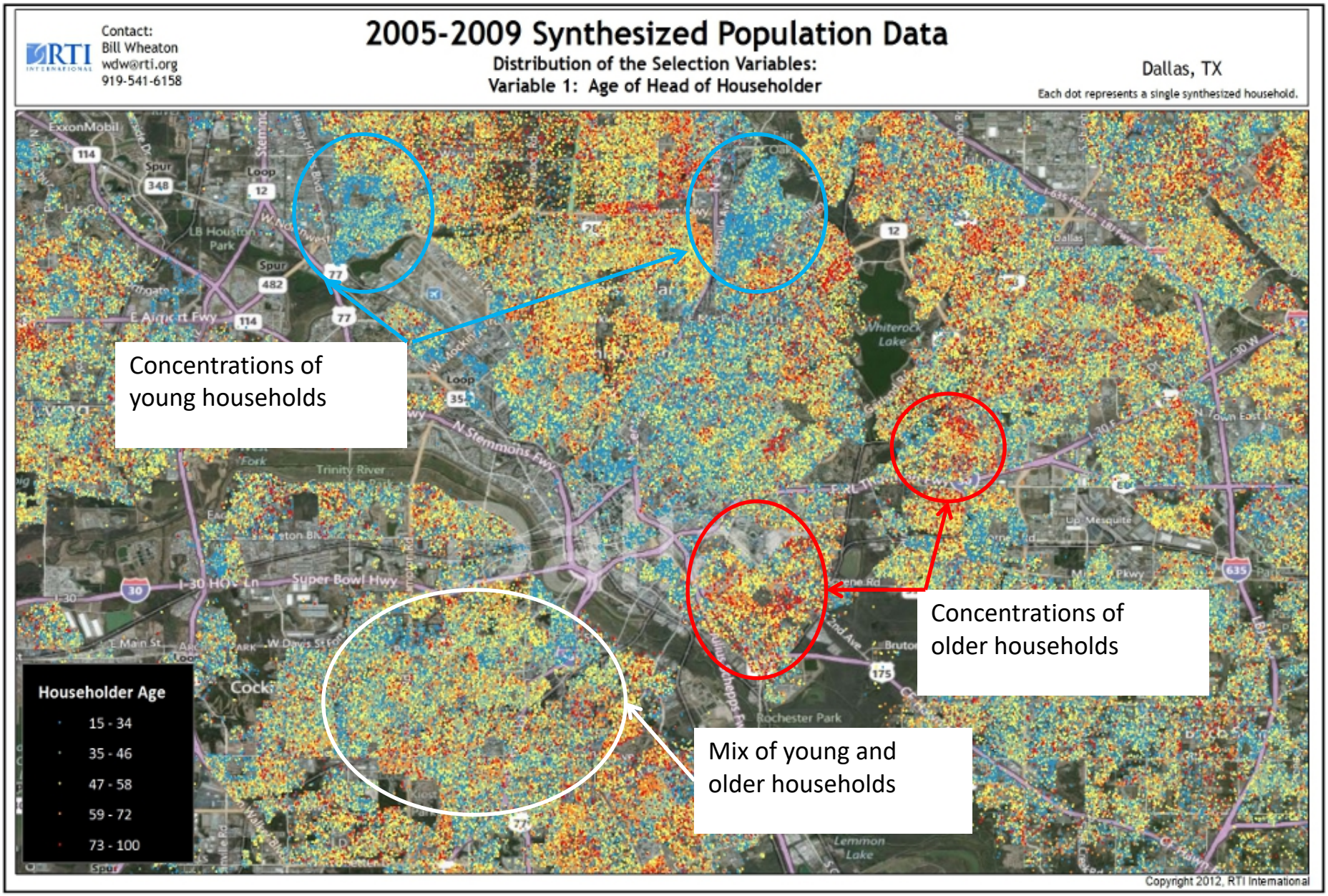
# Creating a Synthetic Population: Data Inputs and Techniques

- Block-group Level Demographics
  - American Community Survey 5 year samples
    - ACS, 2007-2011 (2009) … 2011-2015 ACS 5-year Estimates (2013)
- Public Use Microdata (PUMS)
  - Individual ACS responses
  - Household and individual level data
  - Family structure maintained
  - 5% Sample within Public Use Microdata Areas (PUMAs)
  - PUMAs contain about 100,000 persons
- Household Locations
  - Placed based on high resolution ICLUS population data (EPA)
  - Population at 90-meter grid cell resolution
- Iterative Proportional Fitting (IPF)
  - Creates a synthetic population that matches ACS counts at a block-group level by selecting PUMS microdata records.

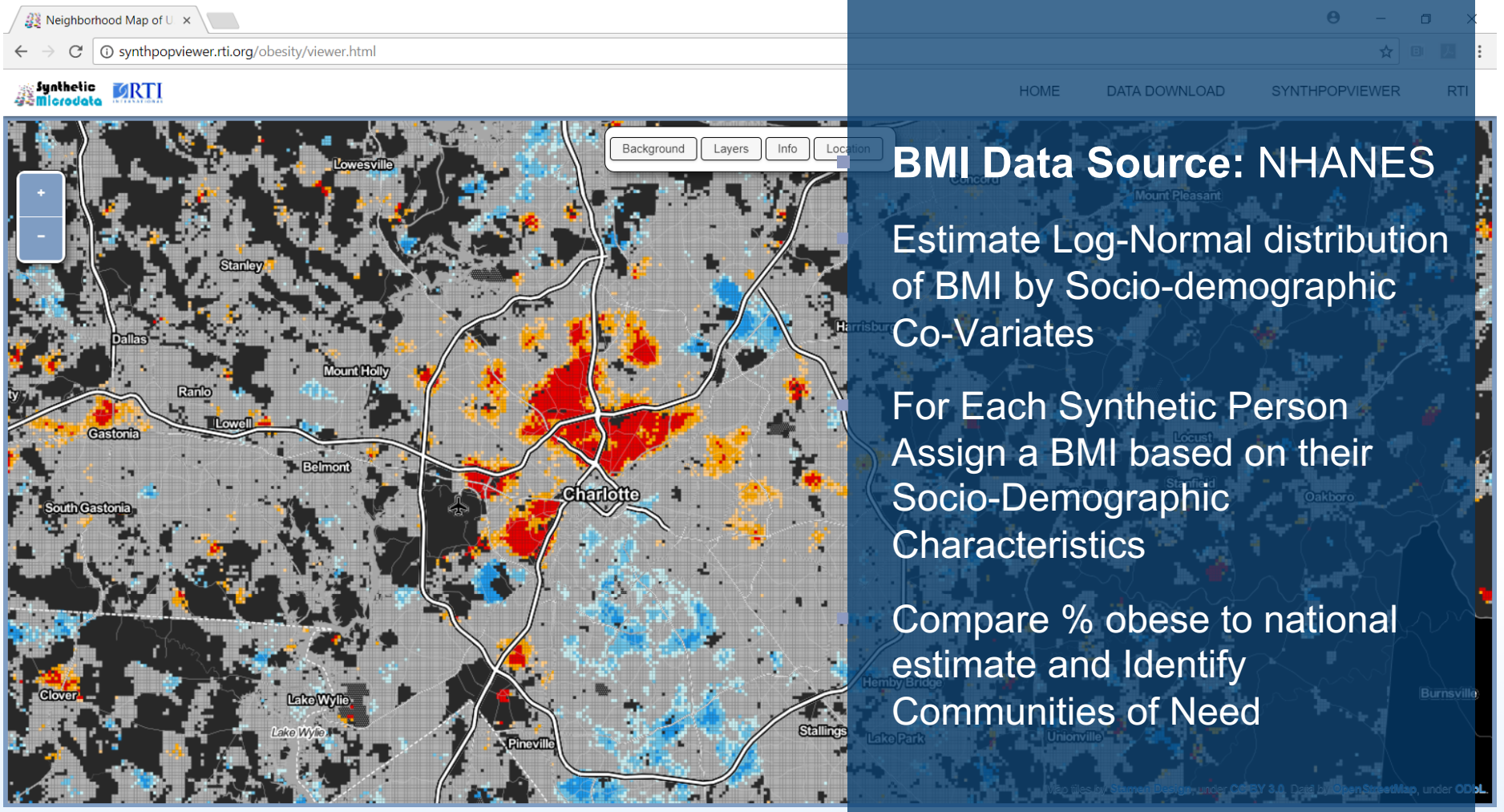
# Synthpop Accuracy, Mapped



# Householder Age



# Synthetic Populations - Obesity




**BMI Data Source:** NHANES

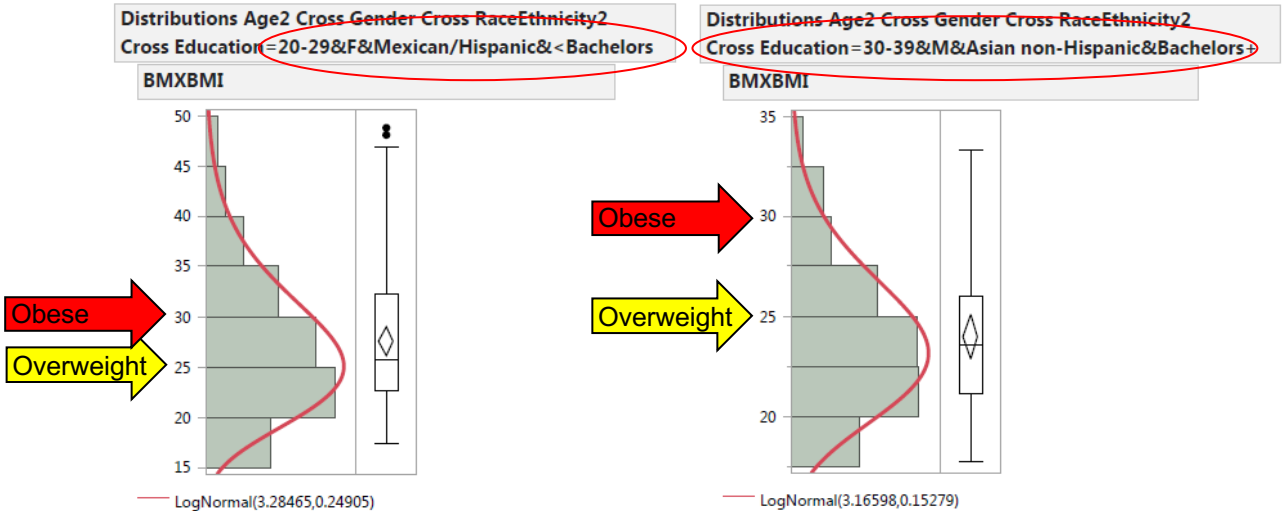
Estimate Log-Normal distribution of BMI by Socio-demographic Co-Variates

For Each Synthetic Person Assign a BMI based on their Socio-Demographic Characteristics

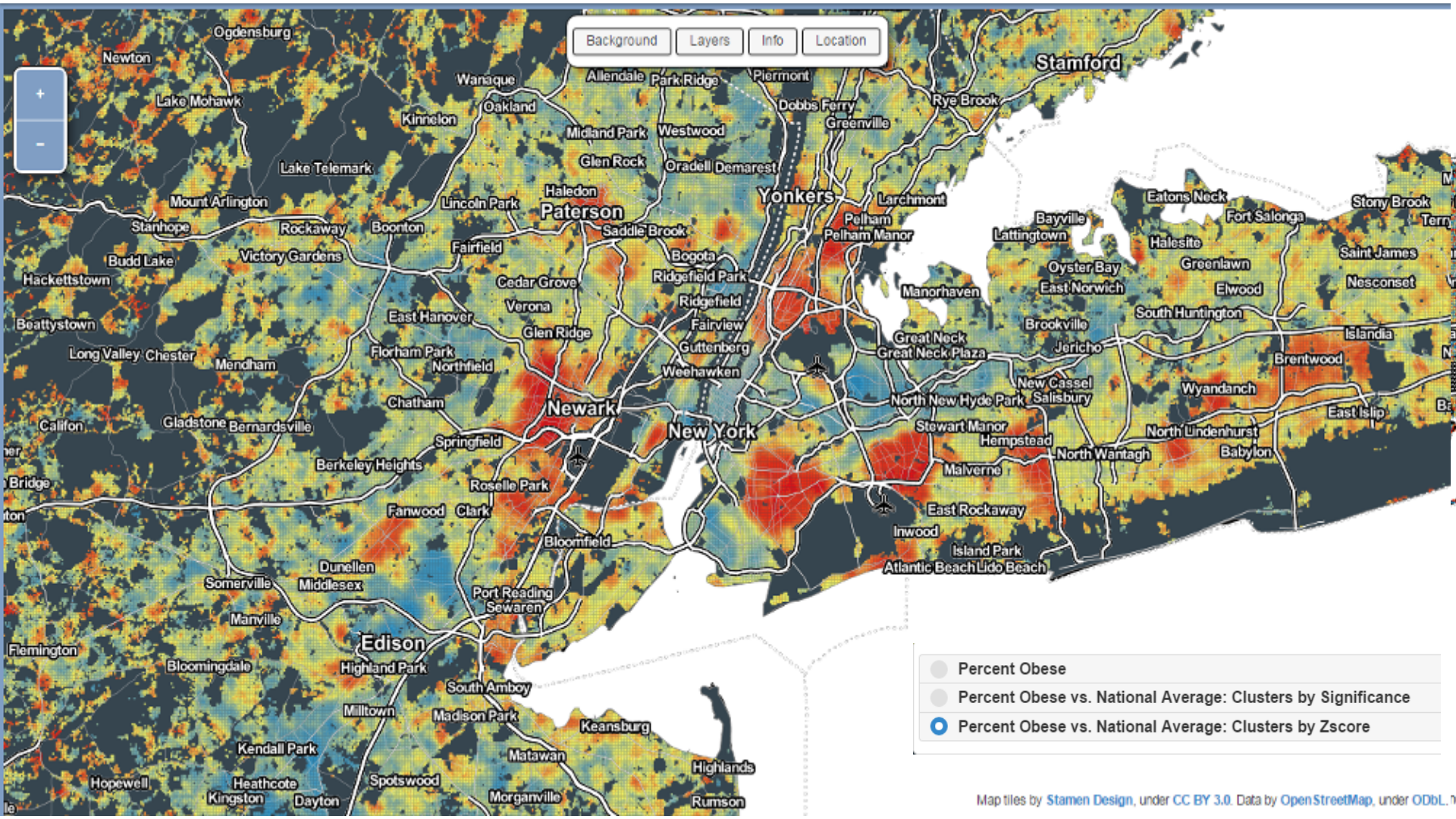
Compare % obese to national estimate and Identify Communities of Need

# Obesity Map Processing

- Example of ‘Data Enrichment’
- Calculate Shape of BMI Distribution by Age, Gender, Race/Ethnicity, Educational Attainment
- Source NHANES



# Obesity Map Processing



# Opioid Mapping via RTI's Synthetic Population (PRELIMINARY)

James Rineer, Georgiy Bobashev, Kasey Jones

# Goals

- Develop model for opioid misuse age >= 18
- Apply the model to RTI's existing Synthetic Population dataset
  - Compute a misuse label (misuser or not) for each 300million+ persons in the US
- Weight the misuse date using county OD estimates
- Visualize the data at the household level by raw count and then by percent misuse for grid area
- Aggregate the data and identify statistically significant spatial clusters

# Progress

- We've completed an initial national model and approach for adjusting with county level information

**Painkiller misuse among US household residents age 18+**

The formula for predicting individual probabilities of Pain-killer misuse at the national level is:

P=Y/(1+Y),

Where Y=exp(b0+b1X1+b2X2+…+b9X9)

The parameters are below:

| | Estimate |
|---|---|
| b0 (Intercept) | -2.11211 |
| b1 AGE3 (26-34/ ref:18-25) | -0.20780 |
| b2 AGE4 (35-49/ ref:18-25) | -0.59634 |
| b3 AGE5 (50-64/ ref:18-25) | -1.03172 |
| b4 AGE6 (65+ / ref:18-25) | -2.09052 |
| b5 SEX2 (female / ref: male) | -0.28753 |
| b6 (less HS/ref:HS+) | -0.29243 |
| b7 RACE1 (white /ref:other) | 0.25812 |
| b8 RACE2 (black /ref:other) | -0.05229 |
| b9 RACE3 (Hispanic /ref:other) | -0.02074 |

Assuming that the rate of OD deaths is proportional to the use of the painkillers we adjust probabilities of misuse to correspond to the distribution of painkiller OD deaths at the county level provided by the CDC.

Given the distribution of county-level *Pj* and the CDC's distribution of OD probabilities *Dj* create weights $W_j = (D_j*Sum(Pj))/(Sum(Dj)_j*P_j)$. We will also need to make sure that the totals are the same: $Sum(P_{ij})=Sum(W_j*P_{ij})$

So all probabilities adjust by an adjustment factor $A= Sum(P_{ij})/Sum(W_j*P_{ij})$

Finally the new probabilities will be $P_{jnew} = A*W_j*P_{ij}$

# Visualization of Results (not OD weighted yet)

- Standard deviation counts in 500m grid overlay. This will track with population density