Modeling EITC and SNAP Participation Rates with ACS PUMS

May 2017— Civis Analytics

CIVIS™
ANALYTICS

Building a Data-Driven World™

Civis Analytics & The Robin Hood Foundation

Closing the Participation Gap in Anti-Poverty Programs in New York City

Civis Analytics

Civis Analytics was formed out of the 2012 Obama for America Analytics team. We built a scientific understanding of each voter. Our data science influenced every strategy and tactic: voter targeting, messaging, media buys, and fundraising.

Since 2012, we have applied our methods across industries, including non-profit, government, and corporate clients.

The Client

Robin Hood Foundation has worked to lift people out of poverty in New York City since 1988. Robin Hood uses money raised to provide resources for soup kitchens, homeless shelters, schools, job-training programs, and other vital services to New Yorkers in need. They provide grants to organizations on the ground, and conduct research to collect data on poverty.

Closing the Participation Gap in Anti-Poverty Programs in New York City

The Problem

SNAP (food stamps) and the Earned Income Tax Credit (EITC) are two of the most effective anti-poverty programs in the US. One in five eligible people do not claim these benefits, though.

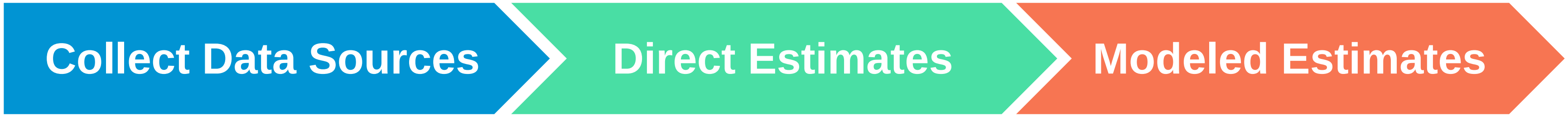
The Goal

Close the SNAP and EITC participation gaps in New York City by connecting eligible individuals to these benefits.

Our Challenges

There are lists of participating families, but there are no lists of eligible households. Information that does exist on participation trends is at the state level. Eligibility criteria are complicated.

We estimated participation rates in these programs by calculating direct estimates, then modeling indirect estimates



Collect Data Sources

Direct Estimates

Modeled Estimates

- **Participation data**
  - HRA for SNAP
  - IRS for EITC
- **Eligibility data**
  - ACS microdata
- **Enhancement data**
  - Census, ACS geo-data
  - NYC specific data sources

- Use ACS anonymized microdata as a proxy for eligibility requirements for both EITC and SNAP

- To improve the precision of direct estimates, “borrow strength” from other data sources by using information about each PUMA to run a regression and multilevel model on the direct estimates

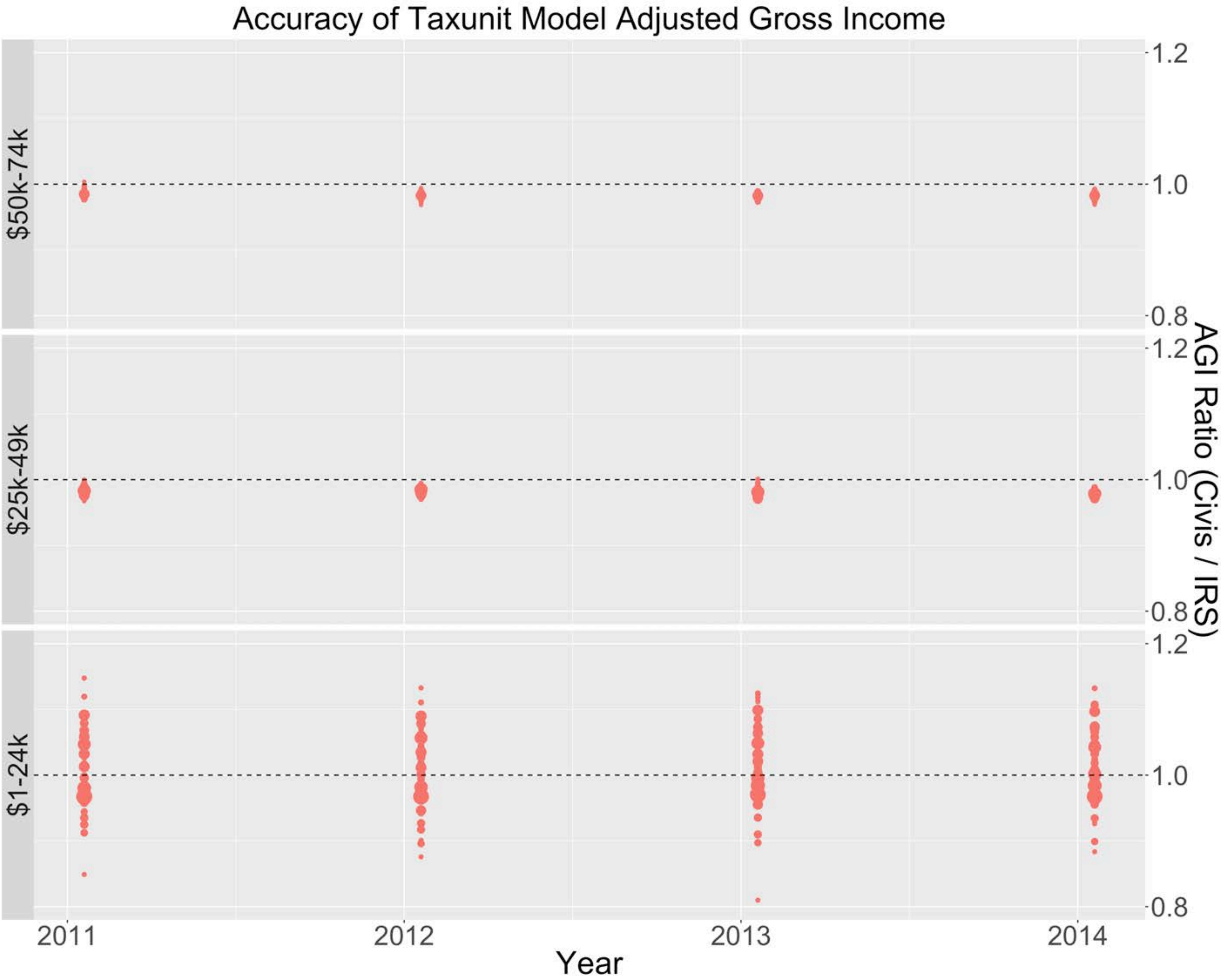
Reading EITC Eligibility Out Of the ACS

EITC eligibility depends on your:

1. Citizenship status
2. Length of domestic residence
3. Marital status
4. Number of dependents
5. Income
6. Age

Some of this information is explicitly in the ACS; some isn’t

1. Citizenship status
2. Length of domestic residence
3. Marital status
4. Number of dependents
5. Income
6. Age


Accuracy of Taxunit Model Adjusted Gross Income
$50k-74k
$25k-49k
$1-24k
1.2
1.0
0.8
1.2
1.0
0.8
1.2
1.0
0.8
AGI Ratio (Civis / IRS)
2011
2012
2013
2014
Year

Improving our ACS Estimates with Modeling

We modeled our PUMA-level estimates by using a feature-rich data set at the PUMA level

Our modeling process started with a feature-rich dataset with information about each PUMA, like average income and demographics, as well as additional features we obtained from the NYC Open Data portal and RHF partners.

This was our dependent variable →

Direct Estimate	PUMA	Avg. HH size	Pct. with college degree	+100s of features
0.23		1.7424	0.21	….
0.74		2.284	0.17	….
0.81		4.273	0.35	….
…	….	….	….	….

Direct estimates

Additional appended data

We modeled our PUMA-level estimates in two primary steps: feature selection and a multi-level linear model

We built two sets of models to **curate our feature set** and **predict participation rates.**

Our rationale was that the multi-level model would scale our direct estimates up/down by “drawing strength” from other variables in the data.

1 **We selected which features should be used in modeling by training a LASSO regression using cross-validation**.

2 **Using the features selected in step 1, we fit a multi-level Bayesian model to the predicted participation rates created by the LASSO regression**.

In practice, modeling tended to bring extreme direct estimates closer to the average participation rate.

We developed tract-level estimates to provide more granular information for our client

We built a **tract-level model** with PUMA estimates of non-participation density as the dependent variable

- We built tract-level features for our independent variables
- This model identifies tracts that look like tracts in PUMAs with high non-participation density

The predicted non-participation density in most tracts was **close to our PUMA level estimates**

- We deliberately took a conservative approach to avoid directing resources to places where we are less sure they will have an impact.

We also tested models that varied:

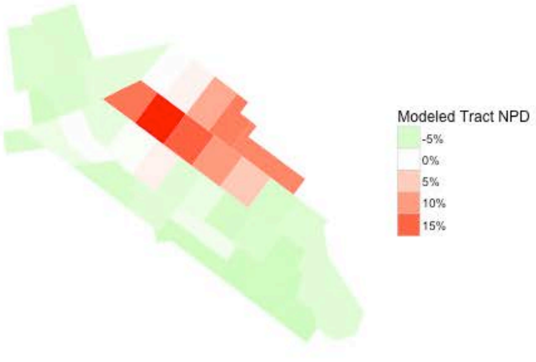
- the algorithm type (linear, tree-based, ensemble methods)
- whether we applied a transformation to the dependent variable, like a log
- how strongly we capped the tract features

We validated our methods by rolling up the tract estimates to the PUMA and city level, and by comparing them visually

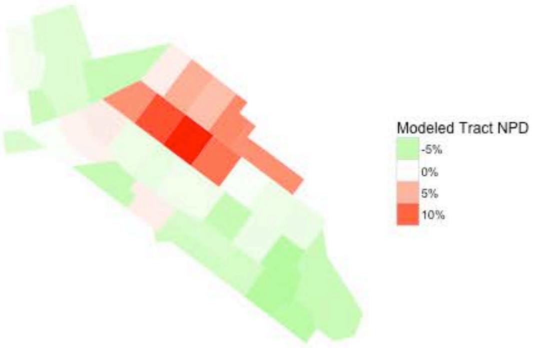
Maps of nonparticipation density at the tract-level in Bushwick

Bushwick, Brooklyn (PUMA 4002)
Tract NPD - PUMA Average NPD
(red areas indicate above average tract NPD)

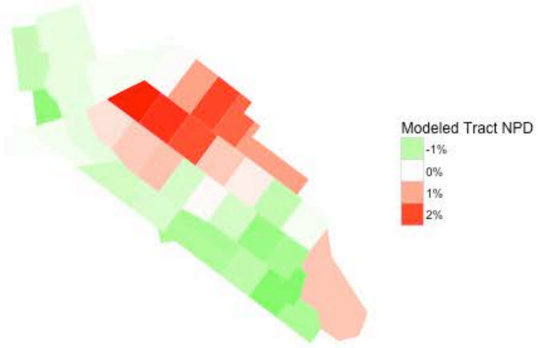
Tract-level, logit NPD,
no PUMA dummies



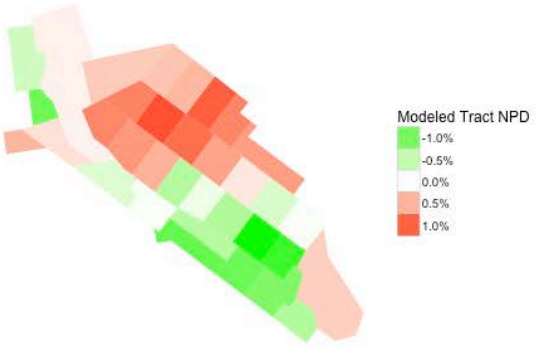
Tract-level, logit NPD,
PUMA dummies



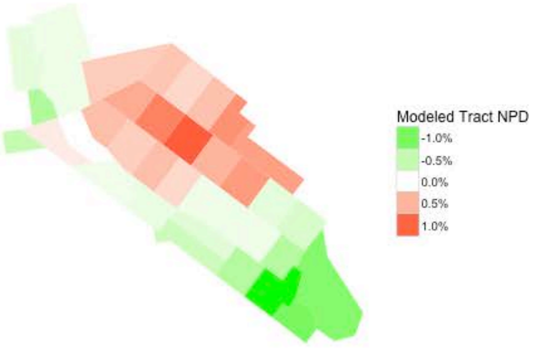
PUMA-level, raw NPD,
no capping



PUMA-level, raw NPD,
scored at 3-97

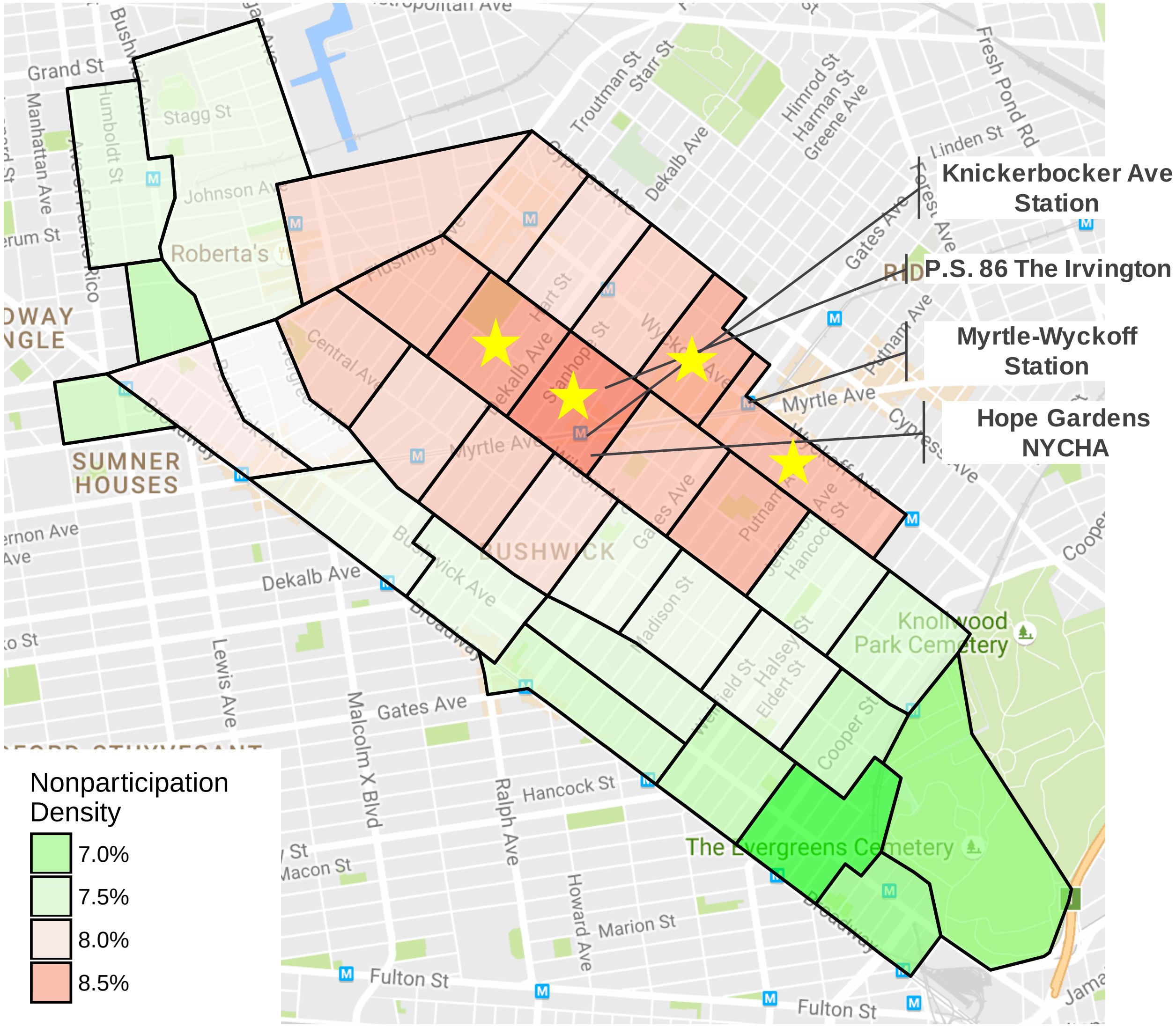


Tract-level, raw NPD,
PUMA dummies



We selected a tract-level model that balanced performance and consistency with our PUMA estimates.

We presented results at the tract-level as a difference from the PUMA level estimate rather than using direct predictions



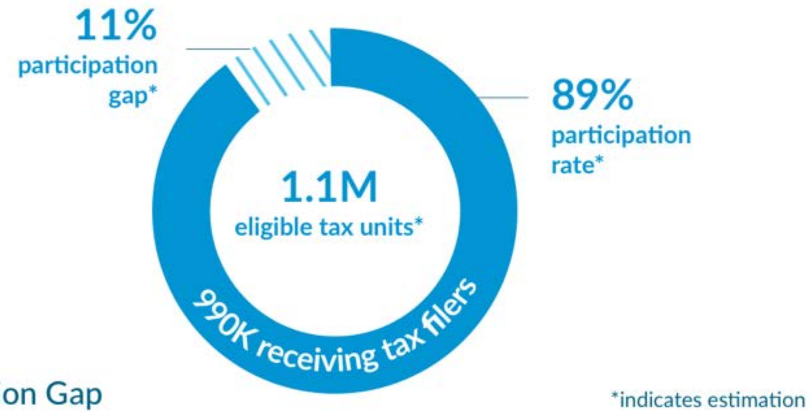
Key Insights

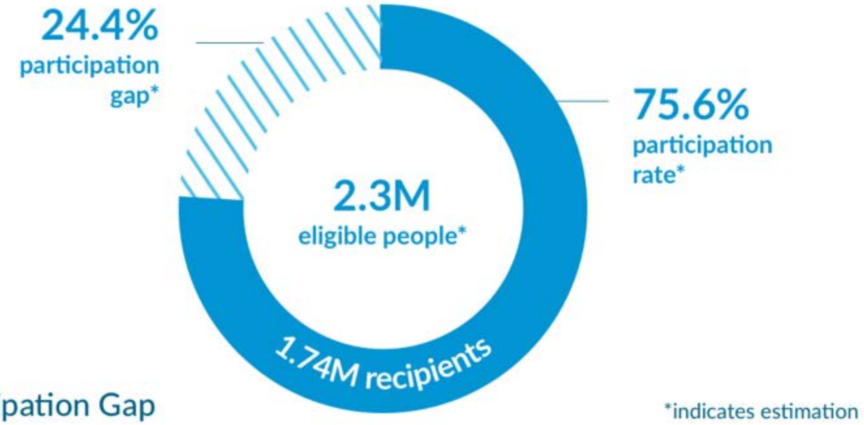
We also provided RHF with information on key service locations located inside of “hotspots” in PUMA.

To better understand the patterns of these hotspots, we profiled each tract grouping using data from the ACS.

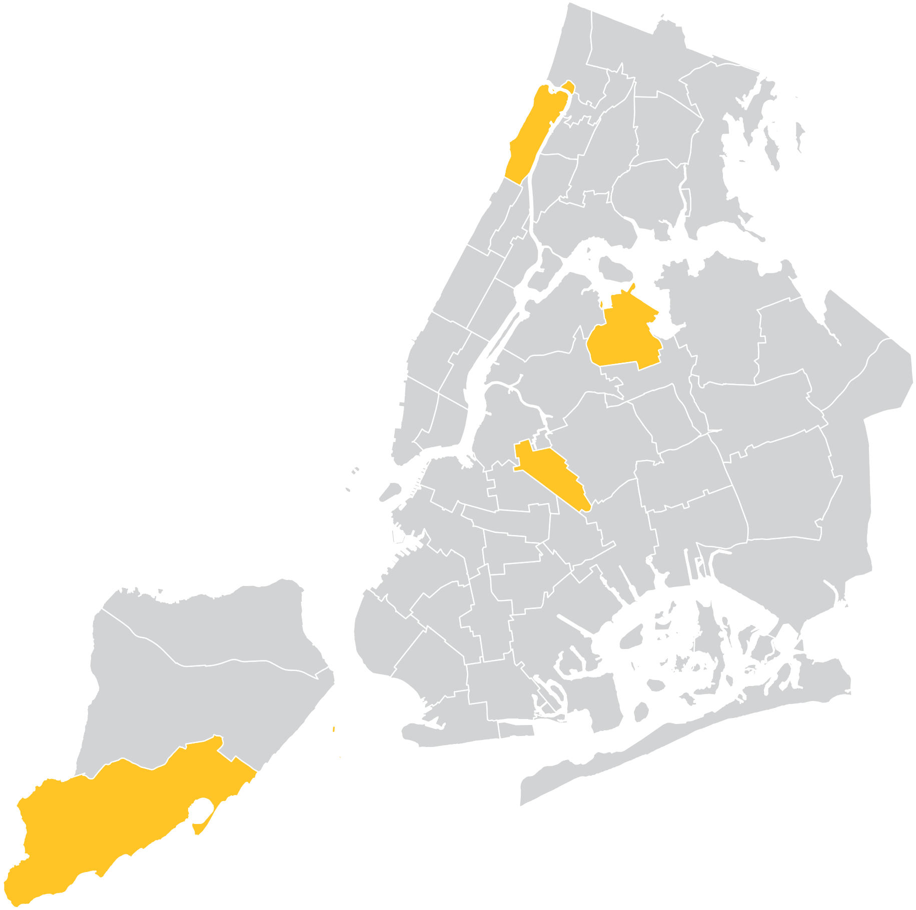
Results and Implementation

Citywide, 89% of eligible people received EITC in 2013 and 76% received SNAP in 2014





The EITC participation gap is closely associated with an area's economic condition



Target Communities

We recommend targeting communities for EITC that have a high nonparticipation density.

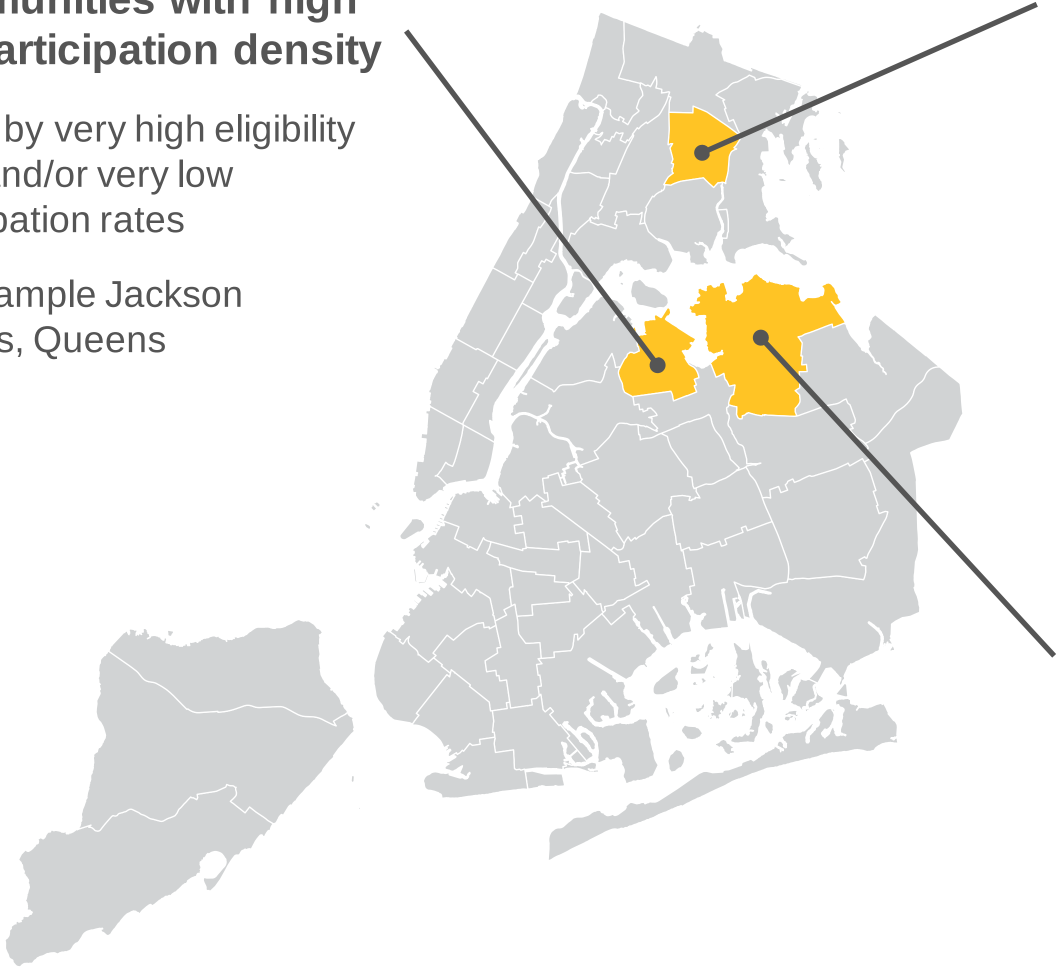
Our target communities were not focused on specific racial or ethnic subgroups, because participation rates by race and ethnicity were fairly similar.

The SNAP gap is greater in areas with significant ethnic communities where English is not the only language spoken

Communities with high nonparticipation density

Driven by very high eligibility rates and/or very low participation rates

For example Jackson Heights, Queens



Communities where African Americans or Hispanics are a minority

Nonparticipation across entire community is low but appears concentrated among a racial / ethnic minority

For example Pelham Parkway, Bronx

Plurality Asian communities

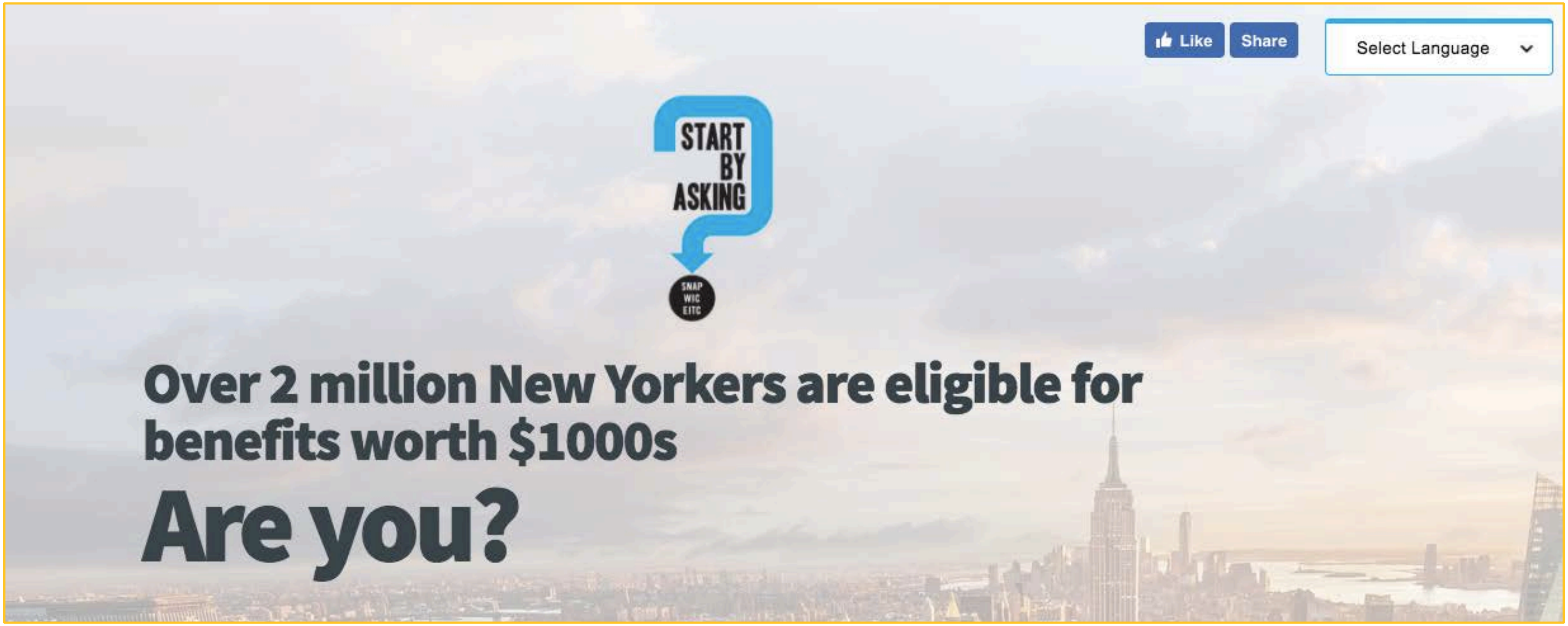
Above average eligibility and below average participation

For example Flushing, Queens

Robin Hood Foundation is rolling out a citywide ad campaign to drive enrollment in these programs

Our work shows RHF areas where their ads could be most helpful.

We provided recommendations for future data collection, so that their partner organizations on the ground may gain a better understanding of who is (or is not) enrolling in these benefits programs.



Over 2 million New Yorkers are eligible for benefits worth $1000s

Are you?

Thank You