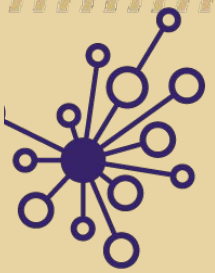


UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

UNIVERSITY of WASHINGTON

Building Households and Families out of Individual Level Administrative Data

Data Science for Social Good (DSSG) Summer 2022

The Team

DSSG Fellows:

Zhaowen Guo

Betelhem Aklilu Muno

İhsan Kahveci

Eliot Stanton

Project Lead:

Jennie Romich

Data Scientist:

Jessica Godwin

Agenda

- Introduction
- The Point-in-Time Approach
  - Co-residence
  - Last names
- Results and Evaluation
- The Longitudinal Approach
- Future Work

PURPOSE: Building households and families out of individual-level data

$15,000

$100,000

PURPOSE: Building households and families out of individual-level data

If they are unrelated individuals

> under poverty line

> non-poor

PURPOSE: Building households and families out of individual-level data

If they are related

> non-poor

CONCEPTS

Household Definition

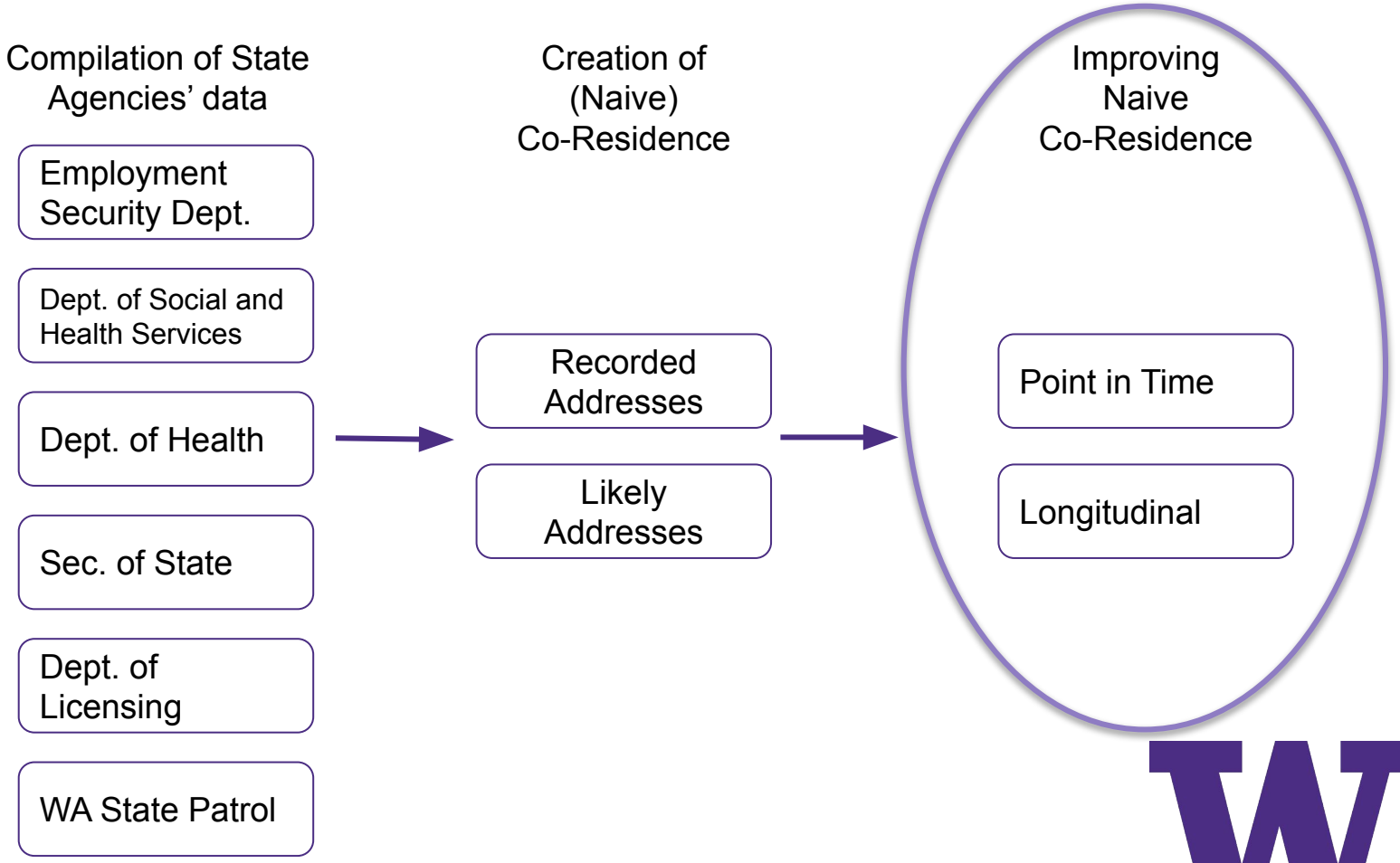
- Census definition
- Social definition(s)
- Overlapping concepts
  - residence
  - family
  - household

CREATION OF WMLAD

Washington Merged Longitudinal Administrative Dataset

10 million people

96 months (2010-2017)



ETHICS

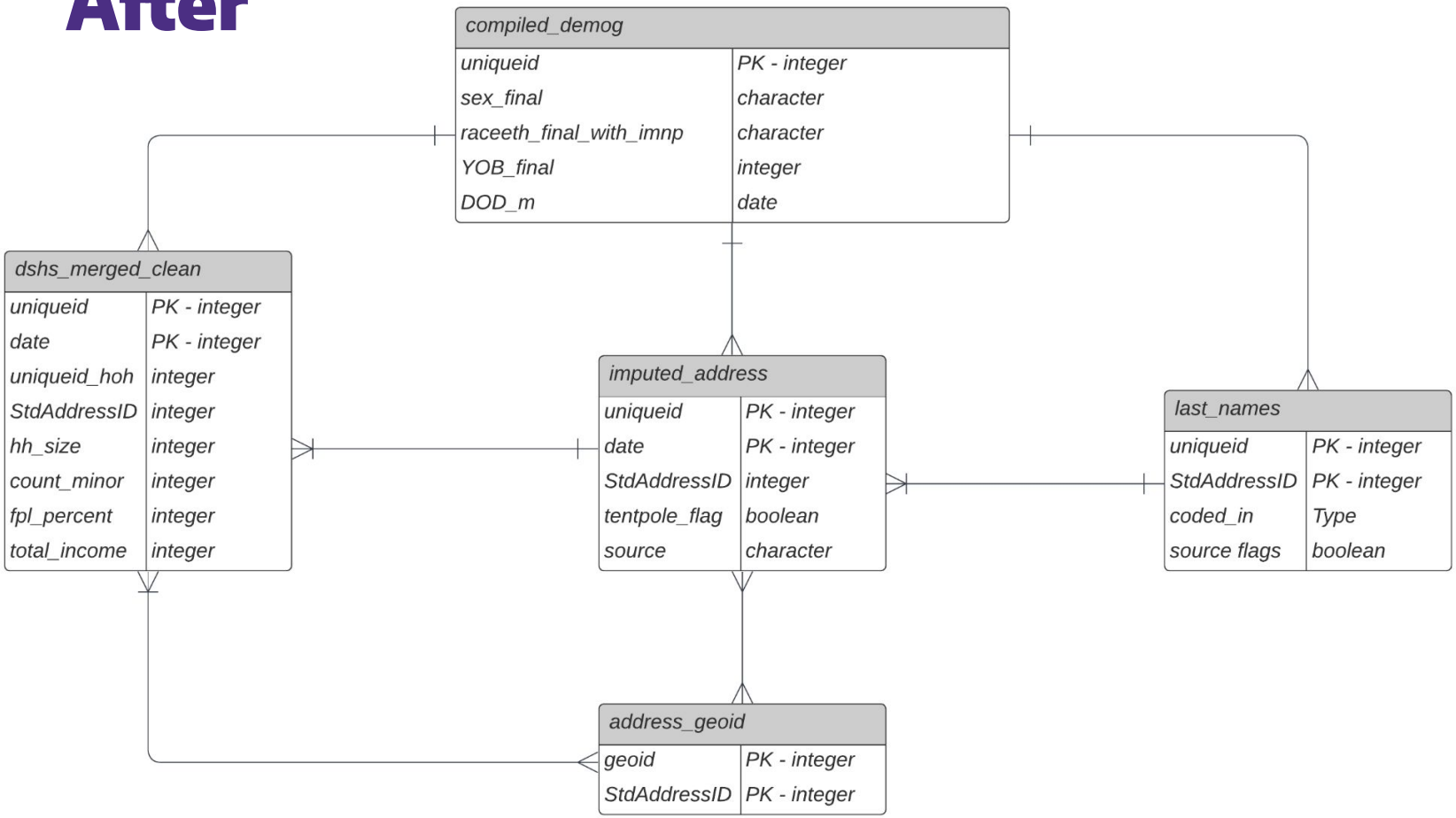
- Stabilizing social constructs
  - different definitions
- Privacy and consent
  - secure data enclave
  - layers of anonymity
- Reproducibility and accountability
  - detailed documentation
  - public good

Before

Name	Date modified	Type	Size
dol_address	4/21/2021 6:38 AM	SAS Data Set	21,698,880...
uw_person_hh_merged060819	4/21/2021 2:31 AM	Stata Dataset	14,079,841...
voter_address_dedup1	4/20/2021 10:42 P...	SAS Data Set	10,918,144...
voter_address	4/20/2021 9:46 PM	SAS Data Set	9,844,032 ...
uw_person_hh_merged072419	4/20/2021 6:24 PM	Stata Dataset	7,890,903 ...
uw_person_final	4/20/2021 5:32 PM	Stata Dataset	7,117,279 ...
geography_census	4/20/2021 5:22 PM	SAS Data Set	7,115,208 ...
geography_census	4/20/2021 4:49 PM	Stata Dataset	7,010,934 ...
uw_hh_final	4/20/2021 3:52 PM	SAS Data Set	6,049,664 ...
uw_hh_final	4/20/2021 3:42 PM	Stata Dataset	6,032,592 ...
child_support_data	4/20/2021 3:13 PM	SAS Data Set	5,944,384 ...
child_support_study_v2_2	4/20/2021 3:18 PM	SAS Data Set	5,944,384 ...
uw30_hh_final	4/20/2021 1:55 PM	SAS Data Set	5,417,664 ...
uw_person_final	4/20/2021 12:43 P...	SAS Data Set	4,973,120 ...
uw_person_finalSORTED	4/20/2021 12:36 P...	Stata Dataset	4,951,153 ...
uw_person_hh_merged_month2	4/20/2021 12:15 P...	Stata Dataset	4,642,409 ...
child_support_data	4/20/2021 11:59 A...	Stata Dataset	4,572,310 ...
child_support_study_v2_2	5/2/2021 3:47 PM	Stata Dataset	4,572,310 ...
dol_address	4/20/2021 10:59 A...	Stata Dataset	4,062,563 ...
csdb_elig	4/20/2021 10:31 A...	SAS Data Set	3,877,760 ...
mi_txneed	4/20/2021 10:34 A...	SAS Data Set	3,877,760 ...
sud_txneed	4/20/2021 10:38 A...	SAS Data Set	3,877,760 ...
uw40_person_final	4/20/2021 10:14 A...	SAS Data Set	3,873,280 ...
csdb_dxrx_trans	4/20/2021 10:05 A...	SAS Data Set	3,872,448 ...
csdb_dxrx_trans	4/20/2021 9:07 AM	Stata Dataset	3,750,196 ...
homeless_array	4/20/2021 5:44 AM	SAS Data Set	2,890,752 ...
snap_array_LONG	4/20/2021 5:23 AM	Stata Dataset	2,736,849 ...
uw_hh_finalSORTED	4/20/2021 3:56 AM	Stata Dataset	2,262,225 ...
uw_hh_final072419	4/20/2021 3:52 AM	Stata Dataset	2,262,222 ...
uw30_hh_final	4/20/2021 3:48 AM	Stata Dataset	2,250,171 ...
wsp_arrest_uniqueid	4/20/2021 3:24 AM	SAS Data Set	2,057,408 ...
esd_2017	4/20/2021 3:13 AM	SAS Data Set	1,951,296 ...
esd_2015	4/20/2021 3:09 AM	SAS Data Set	1,923,008 ...
esd_2014	4/20/2021 3:04 AM	SAS Data Set	1,867,776 ...

Name	Date modified	Type	Size
WA_census_blocks	8/8/2022 4:04 PM	Microsoft Excel Com...	180,617,357 ...
DSHSmerged_geocodes_monthly_clean	7/24/2022 7:57 PM	Microsoft Excel Com...	156,934,423 ...
DSHSmerged_geocodes_monthly_clean_essenti...	8/3/2022 1:46 AM	Microsoft Excel Com...	15,669,941 KB
DSHS_hh_selected.RDS	7/19/2022 3:58 PM	RDS File	3,671,178 KB
coded_ln_unique	7/25/2022 12:57 PM	Stata Dataset	1,224,530 KB
tentpole_addresses_pivoted	7/24/2022 3:37 PM	Microsoft Excel Com...	1,038,826 KB
tentpole_sources_pivoted	7/24/2022 7:50 PM	Microsoft Excel Com...	1,033,927 KB
cleaned_names	7/29/2022 4:36 PM	Microsoft Excel Com...	561,687 KB
cleaned_names.RDS	7/29/2022 4:04 PM	RDS File	331,368 KB
coded_ln_unique.RDS	7/29/2022 11:36 AM	RDS File	126,113 KB
WA_tracts.RDS	8/3/2022 12:53 PM	RDS File	63,062 KB
small_imputed_address_lastnames.RDS	8/5/2022 6:38 PM	RDS File	59,627 KB
tracts_hhsize_all.RDS	8/10/2022 6:16 PM	RDS File	57,304 KB
small_imputed_address.RDS	8/5/2022 6:28 PM	RDS File	54,816 KB
imputed_hh_add_4.RDS	7/18/2022 5:42 PM	RDS File	53,190 KB
imputed_hh_census_month.RDS	7/18/2022 12:47 PM	RDS File	47,834 KB
DSHS_154_hh.RDS	7/19/2022 3:32 PM	RDS File	26,946 KB
WA_tracts_no_water.RDS	8/10/2022 3:57 PM	RDS File	10,575 KB
dshs_hh_census_month.RDS	7/19/2022 1:44 PM	RDS File	10,379 KB
HH_ln_GB.RDS	8/12/2022 3:26 PM	RDS File	4,773 KB
small_hh_demog_withFlags.RDS	8/10/2022 6:02 PM	RDS File	3,920 KB
small_hh_demog.RDS	8/8/2022 12:10 PM	RDS File	3,703 KB
HH_with_Cointoss.RDS	8/12/2022 5:45 PM	RDS File	2,650 KB
seattle_city_tracts	8/10/2022 4:02 PM	Microsoft Excel Com...	2,329 KB
small_compiled_demog.RDS	8/5/2022 6:37 PM	RDS File	2,298 KB
post-flip_pre-GBLN.RDS	8/12/2022 1:50 PM	RDS File	2,245 KB
Cointoss_Census_HHs_pre_GBLN.RDS	8/12/2022 3:32 PM	RDS File	2,007 KB
HH_with_Cointoss_num.RDS	8/12/2022 5:45 PM	RDS File	1,459 KB
WA_geo_tracts.RDS	8/3/2022 12:53 PM	RDS File	1,129 KB
census_variables	8/10/2022 4:05 PM	Microsoft Excel Com...	1,105 KB
tract_hh	7/28/2022 4:07 PM	Microsoft Excel Com...	576 KB
hh_changes_tract.RDS	8/3/2022 9:59 AM	RDS File	332 KB
census_tracts_hhsize.RDS	8/10/2022 6:13 PM	RDS File	180 KB
tract_hh.RDS	7/27/2022 2:50 PM	RDS File	160 KB

After

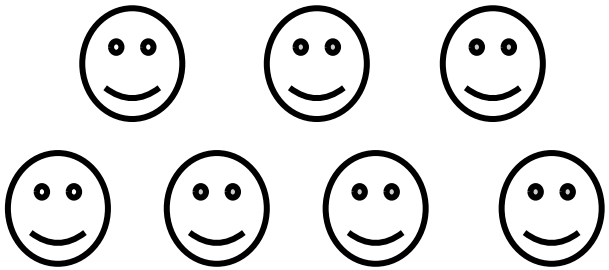

Tracking Poverty ER Diagram
compiled_demog
uniqueid PK - integer
sex_final character
raceeth_final_with_imnp character
YOB_final integer
DOD_m date
dshs_merged_clean
uniqueid PK - integer
date PK - integer
uniqueid_hoh integer
StdAddressID integer
hh_size integer
count_minor integer
fpl_percent integer
total_income integer
imputed_address
uniqueid PK - integer
date PK - integer
StdAddressID integer
tentpole_flag boolean
source character
last_names
uniqueid PK - integer
StdAddressID PK - integer
coded_in Type
source flags boolean
address_geoid
geoid PK - integer
StdAddressID PK - integer

Methods

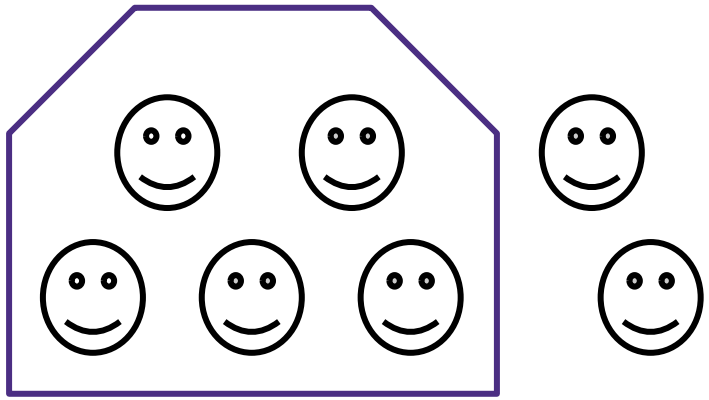
- The point-in-time approach
  - Naive Co-Residence
  - Last Names
    - Deterministic
    - Probabilistic
- The longitudinal approach

The Point-in-Time Approach

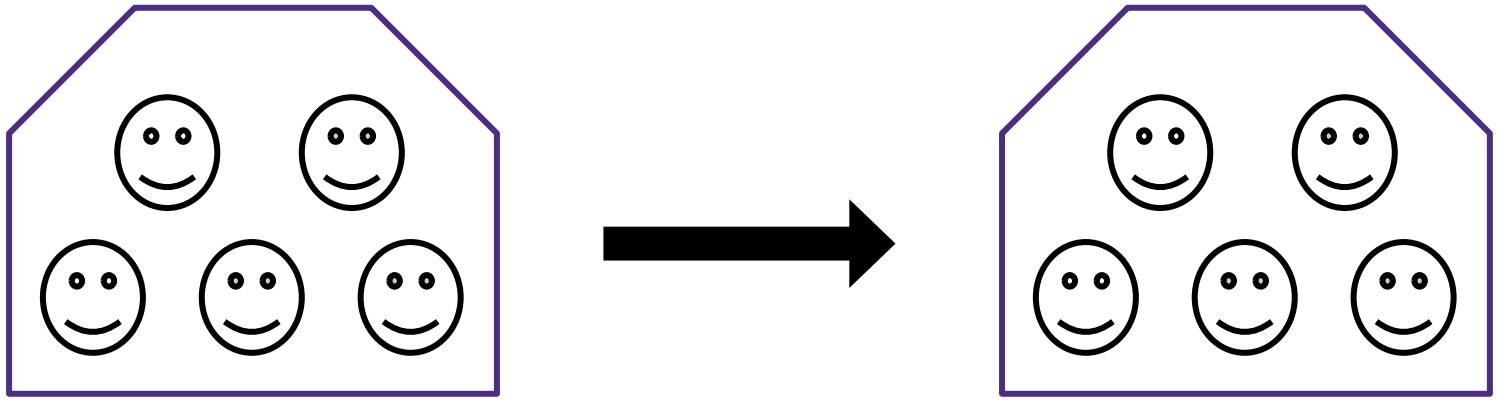
Creating Households



Creating Households



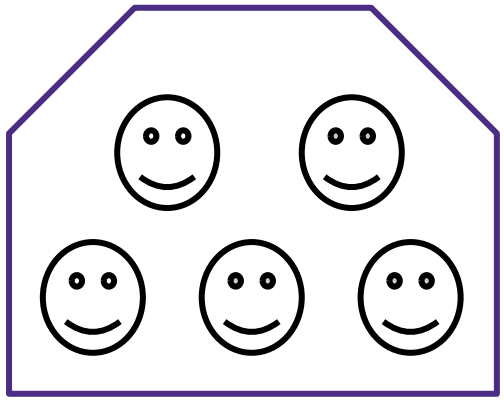
Naive Co-Residence Households



Inferring Relations from Data

What additional information can we consider?

- **limited options**
- **ethics of race and sex**

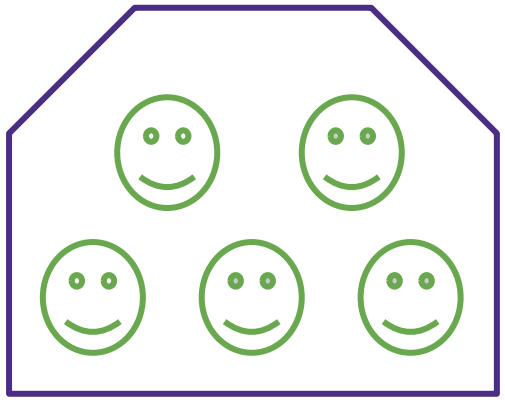


Inferring Relations from Data

What additional information can we consider?

Last names

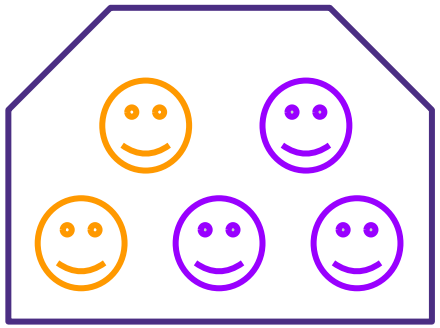
- **logic**
- **coverage**



63% of naive cases

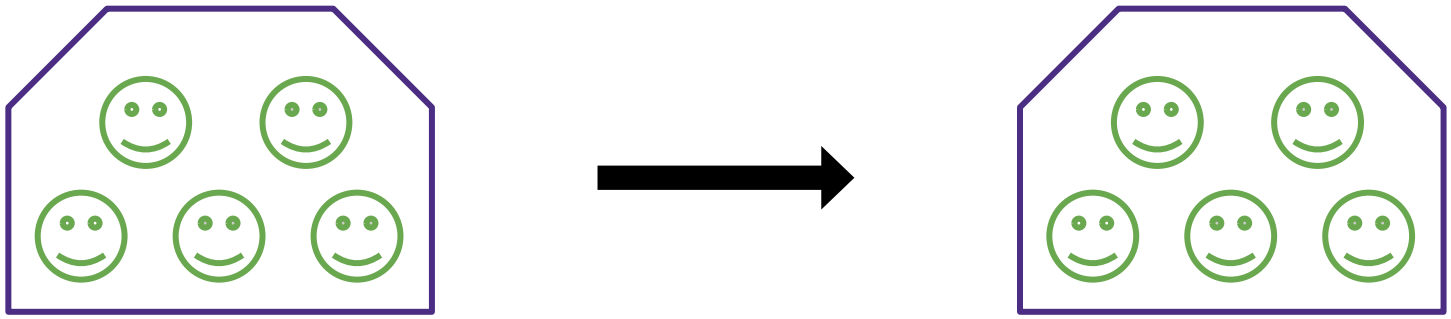
Last Name Households

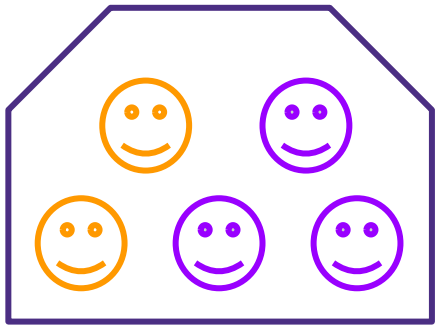
Color = Last Name



Last Name Households

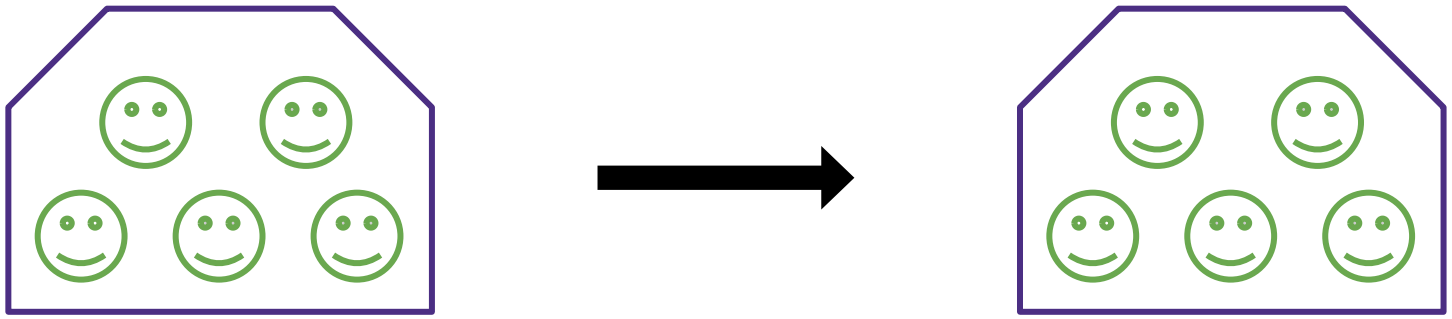
Color = Last Name

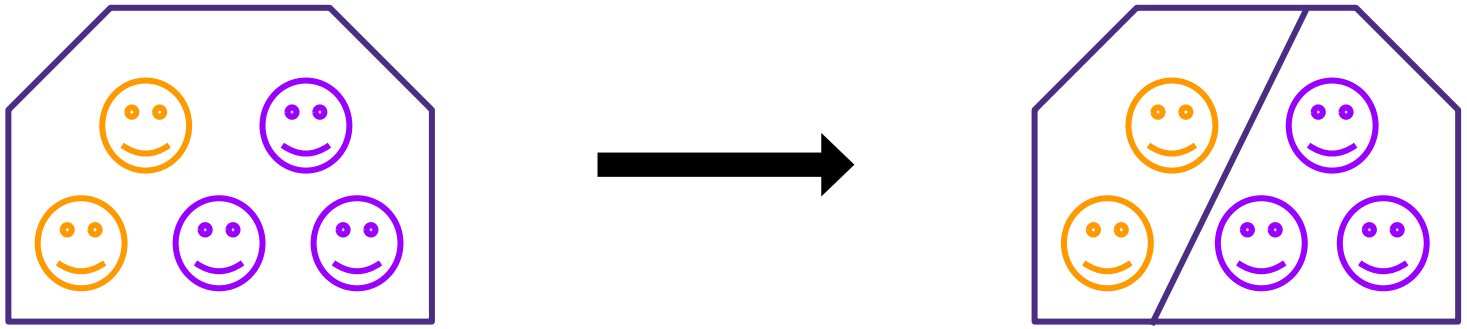




Last Name Households (Deterministic)

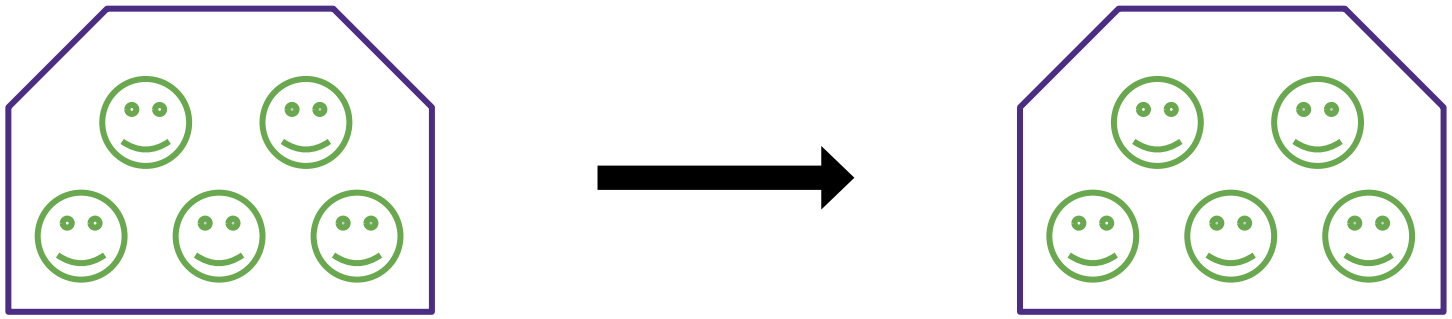
Color = Last Name

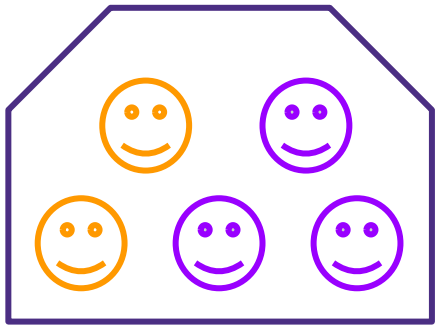




Last Name Households (Probabilistic)

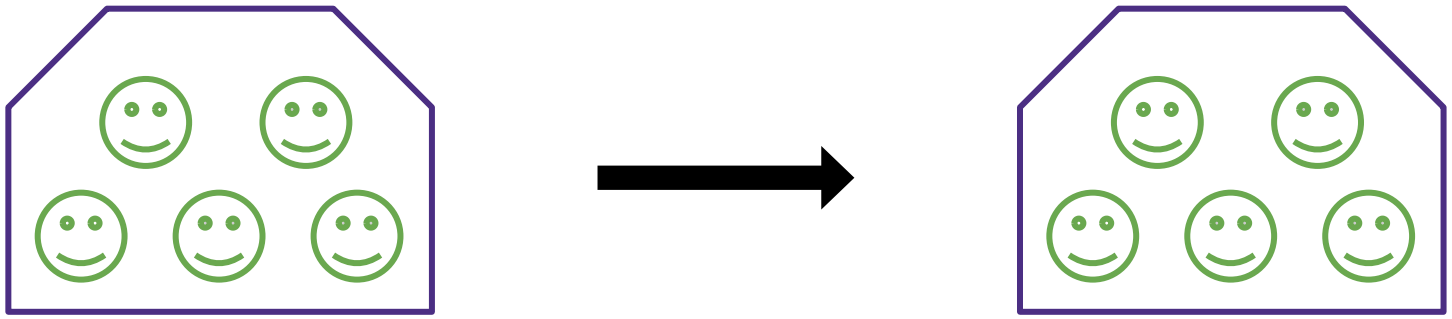
Color = Last Name

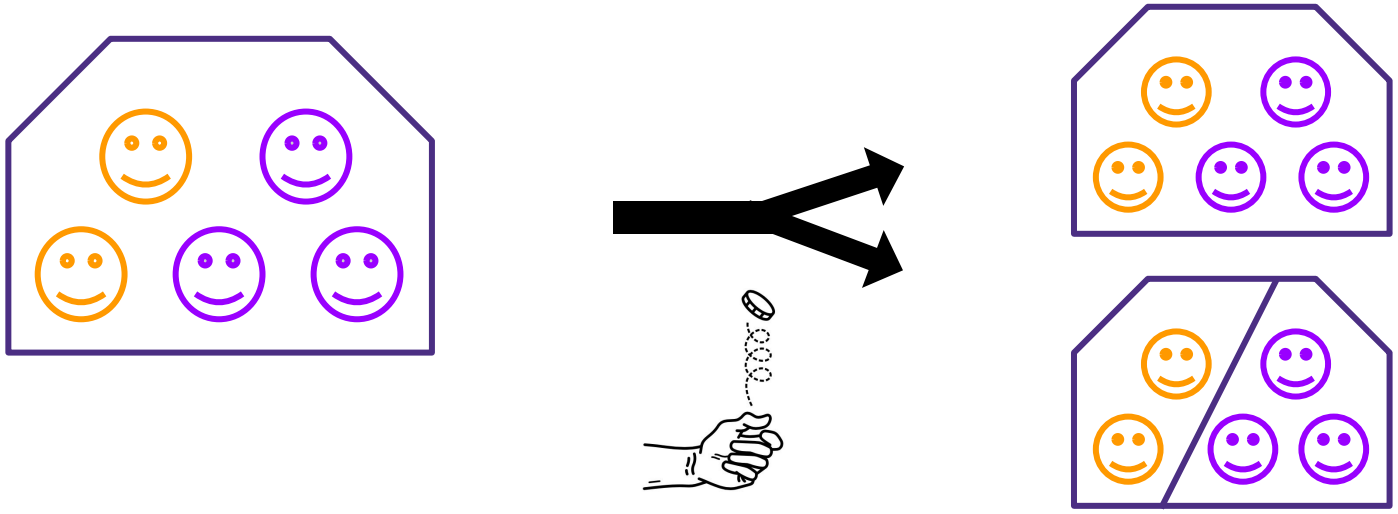




Last Name Households (Probabilistic)

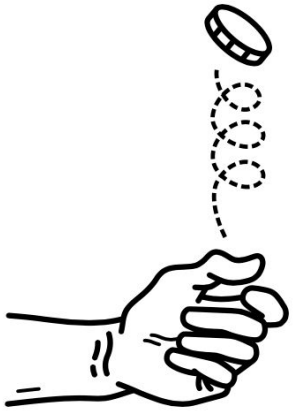
Color = Last Name





Last Name Households (Probabilistic)

Color = Last Name



County A: 80% families

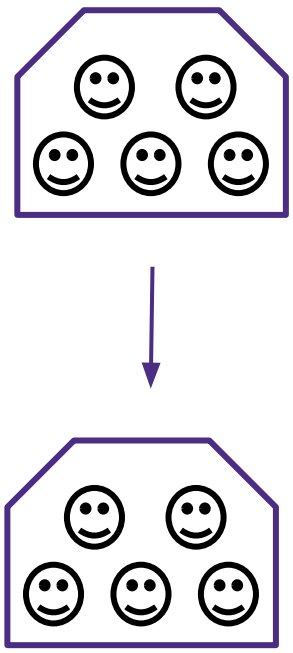
County B: 20% families

coin weighted based on Census probabilities of:

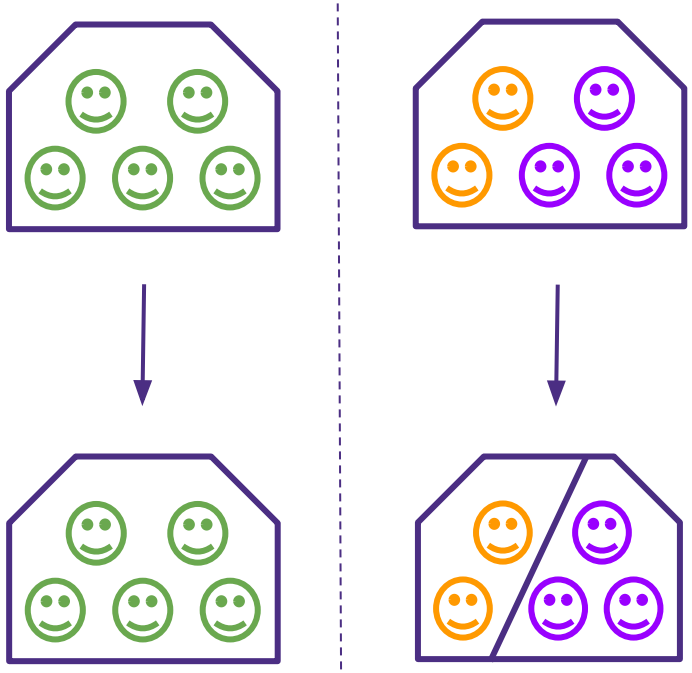
- family
- non-family

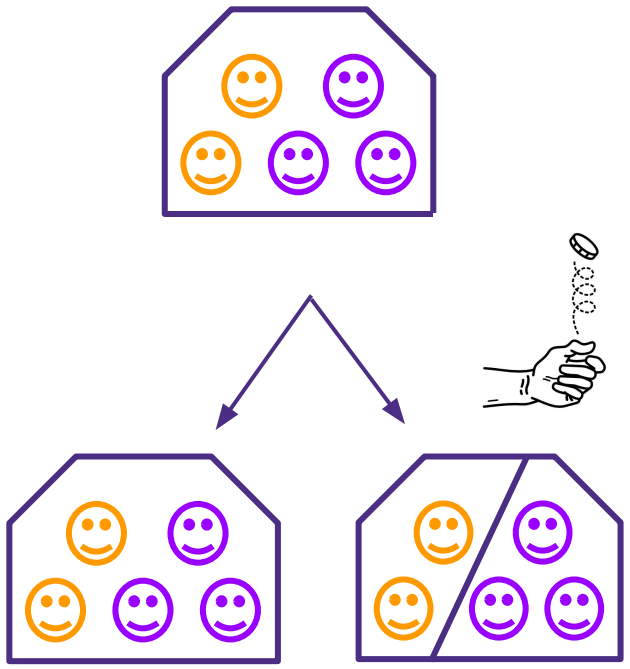
DEFINITIONS USED

Naive Co-Residence



Last Names: Deterministic Probabilistic





RESULTS

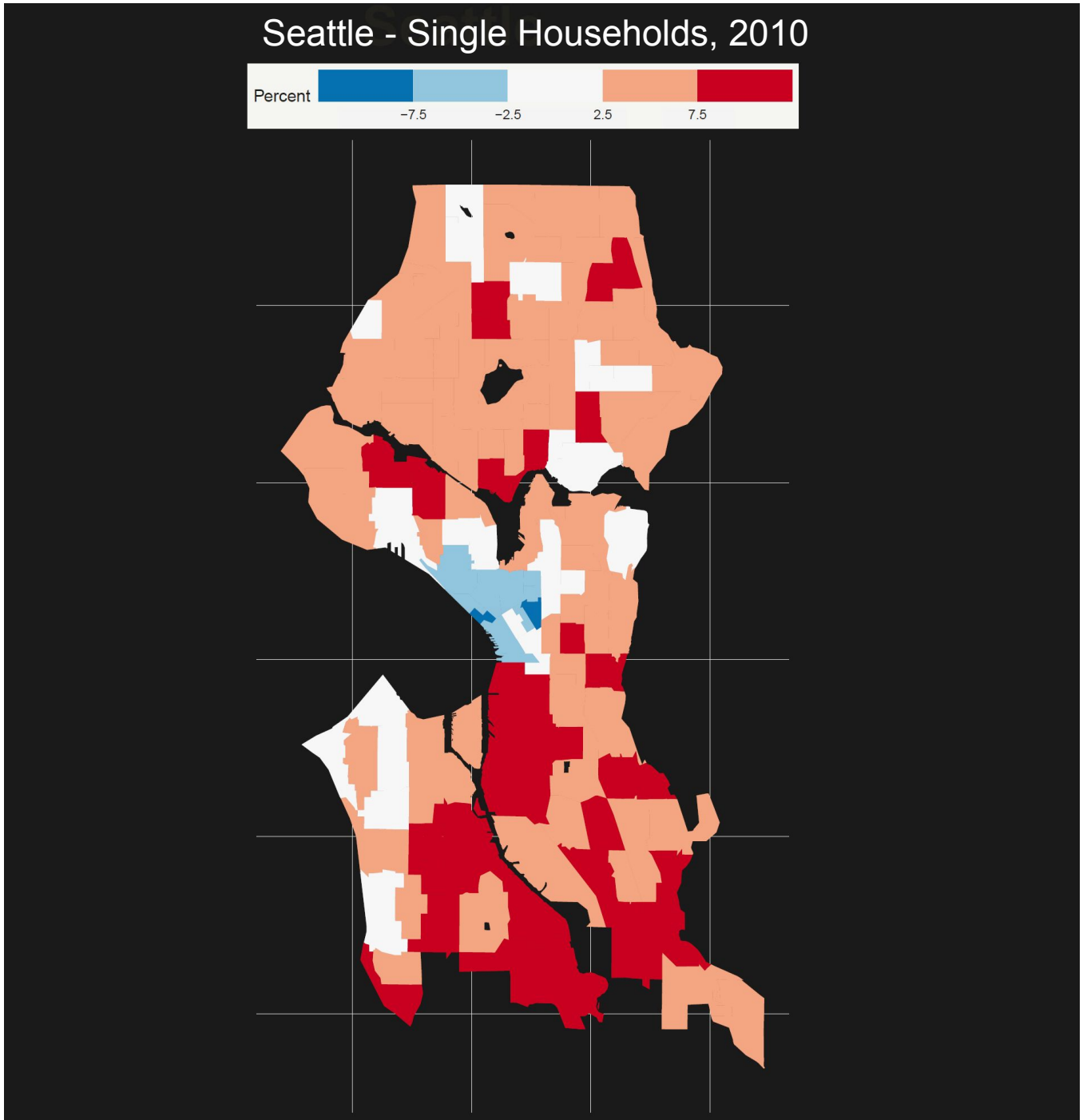
Residence Comparison

Household Size	Census	Naive Co-Residence
1	711619 (27.2%)	968066 (37.5%)
2	904232 (34.5%)	862607 (33.5%)
3	406397 (15.5%)	398067 (15.4%)
4	338260 (12.9%)	198363 (7.69%)
5	151893 (5.8%)	82175 (3.19%)
6	62772 (2.4%)	34342 (1.33%)
7+	44903 (1.7%)	35126 (1.36%)

RESULTS

Family (Economic Units) Comparison

Household Size	Census	Last Name (Deterministic)	Last Name (Probability)
2	728493 (43.2%)	850661 (70.3%)	327872 (39.7%)
3	378423 (22.4%)	235669 (19.5%)	239779 (29%)
4	327428 (19.4%)	90610 (7.5%)	131923 (16%)
5	148156 (8.78%)	24668 (2.04%)	64108 (7.76%)
6	61229 (3.63%)	6123 (0.5%)	299512 (3.62%)
7+	43726 (2.6%)	2551 (0.2%)	32894 (3.98%)


Seattle - Single Households, 2010
Percent
−7.5
−2.5
2.5
7.5

The Longitudinal Approach

Motivation

Comparison between census and naive co-residence

	Census	Naive co-residence
Household size	Percentage	Percentage
1	0.27	0.38
>1	0.73	0.62

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A				 	

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A				 	 

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A				 	 

Voter registration
Driver's license

Why over-estimation of one-person residences?

Scenario: Children interact with government agencies less frequently than adults

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Voter registration
Driver's license

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A		 			

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A		 			

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Frequent movements into and out of the same address are likely to represent an imputation error

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

No revision

Why over-estimation of one-person residences?

Scenario: Members of a household moved at the same time, but their addresses were not updated accordingly

	Month 1	Month 2	Month 3	Month 4	Month 5
Address A					
Address B					

Summary of Scenarios

- Lack of interactions
- Imputation error
- Move in and out

Results

Comparison between census, naive co-residence, and modified co-residence

	Census	Naive co-residence	Modified co-residence
Household size	Percentage	Percentage	Percentage
1	0.27	0.38	0.31
>1	0.73	0.62	0.69

CONTRIBUTIONS

- Create a relational database for future WMLAD users
- Apply both point-in-time and longitudinal approach to derive and improve household and family identifiers

FUTURE WORK

- Apply the longitudinal approach to households of larger sizes
- Integrate point-in-time family formation and longitudinal corrections
- Incorporate other information (i.e. social networks, demographics, anti-poverty programs) to capture a diversity of households

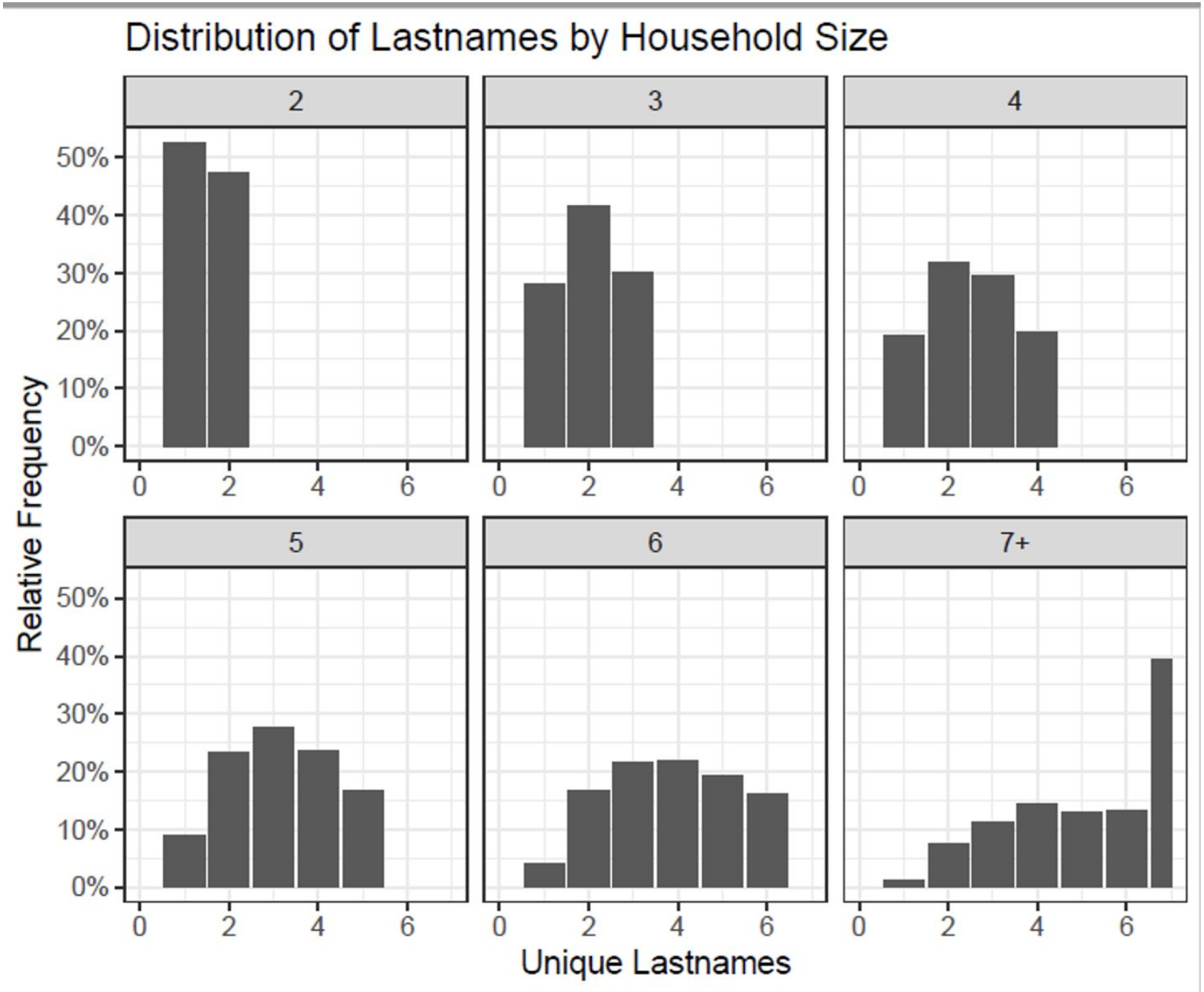
UNIVERSITY of WASHINGTON

eScience Institute

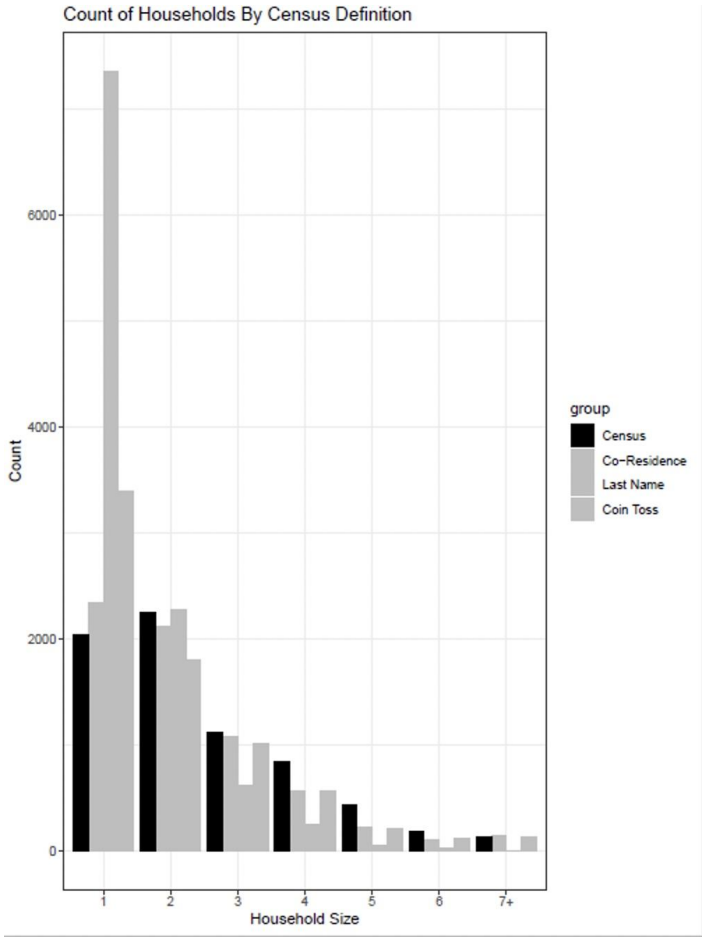
WMLAD funders: Arnold Ventures, Center for Equitable Growth, JPB Foundation/Institute for Poverty Research, City of Seattle, WorkRise, UW Center for Studies in Demography & Ecology

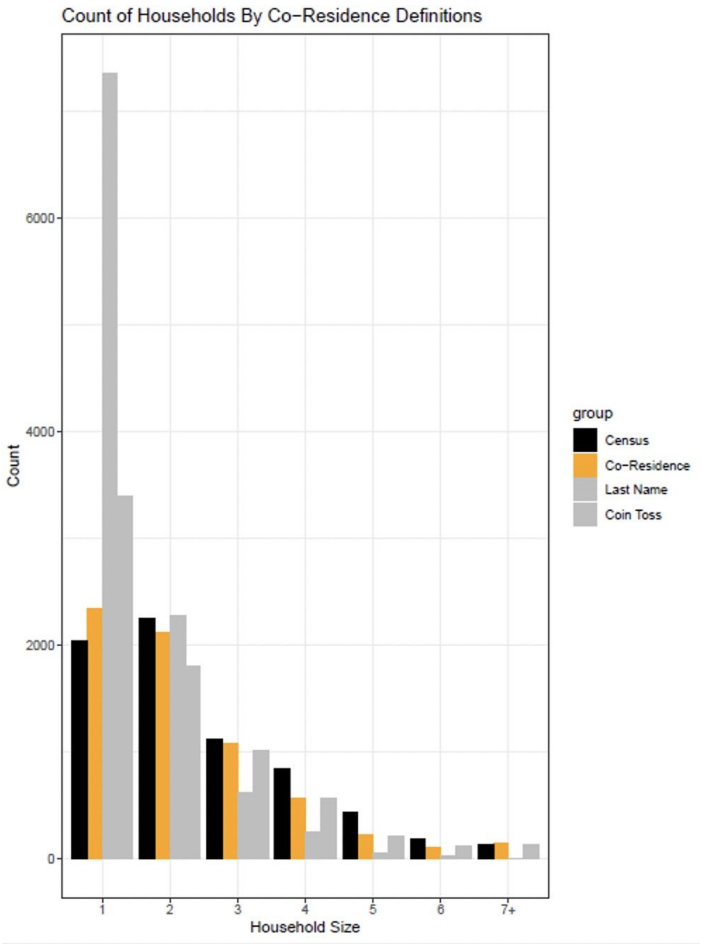
Thank you to Jim Mayfield, Taylor Danielson, and Lisa Nicoli, Washington State DSHS; Callie Freitag and Lizzy Pelletier, UW; Nicole Keenan, researcher and activist; and Phil Hurvitz, UW Data Collaborative.

Appendix

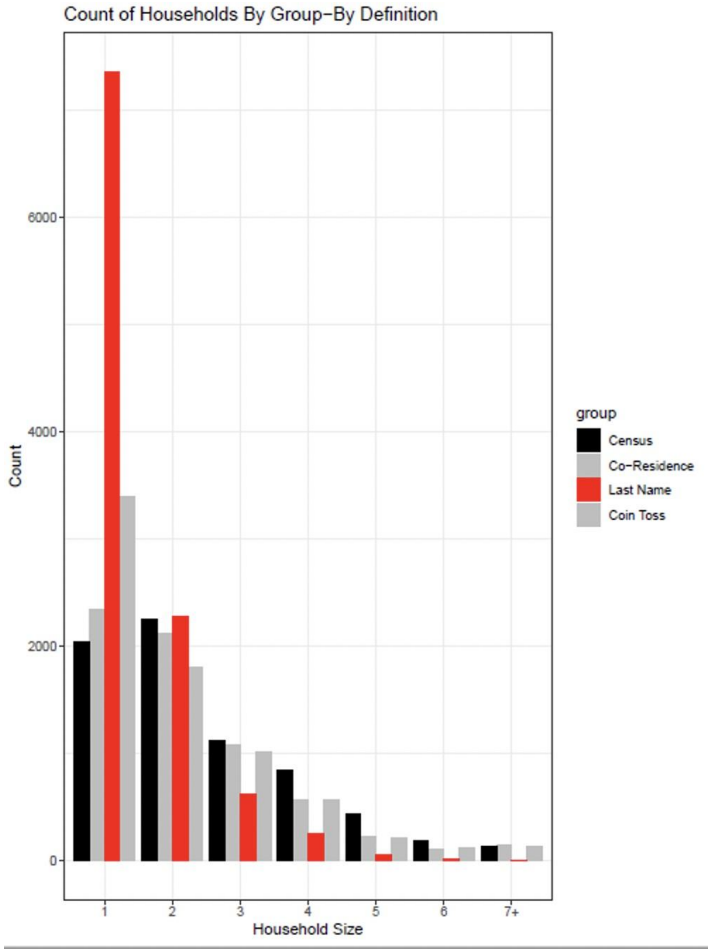

Distribution of Lastnames by Household Size
2
3
4
5
6
7+
Relative Frequency
50%
40%
30%
20%
10%
0%
0
2
4
6
Unique Lastnames

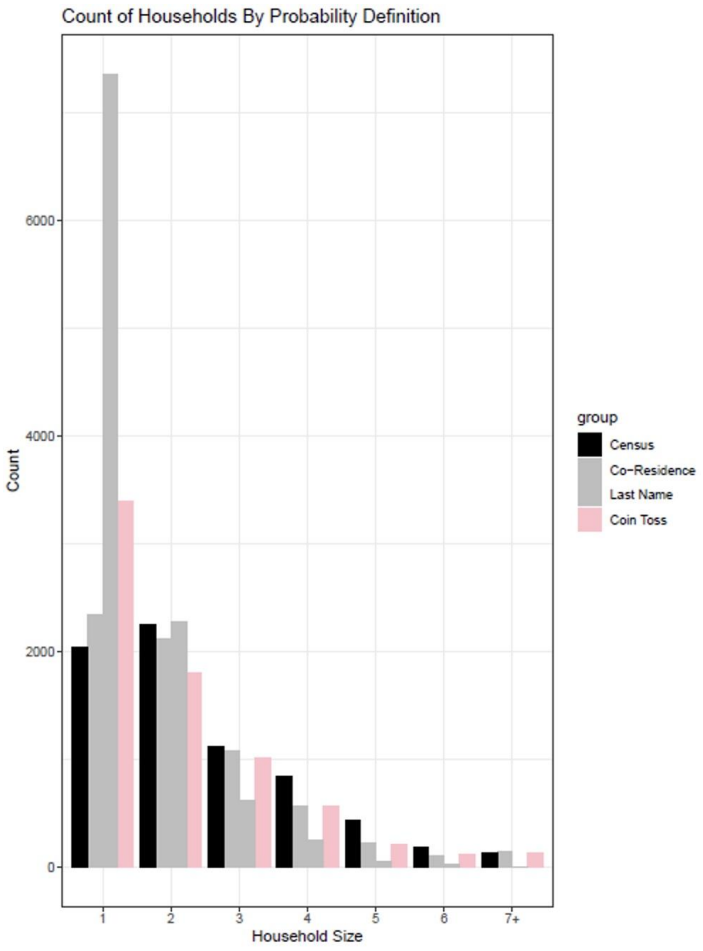
APRIL 2010 (Count)



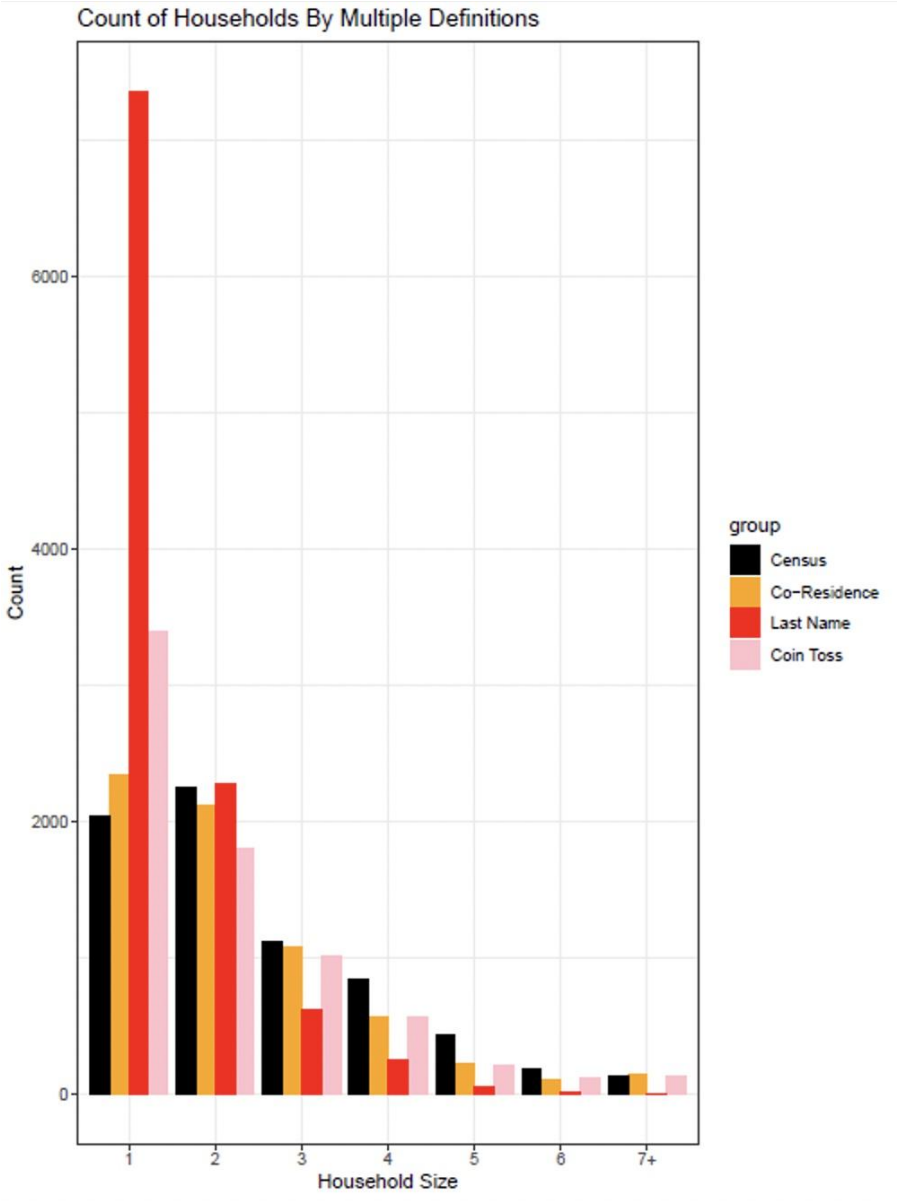


APRIL 2010 (Count)

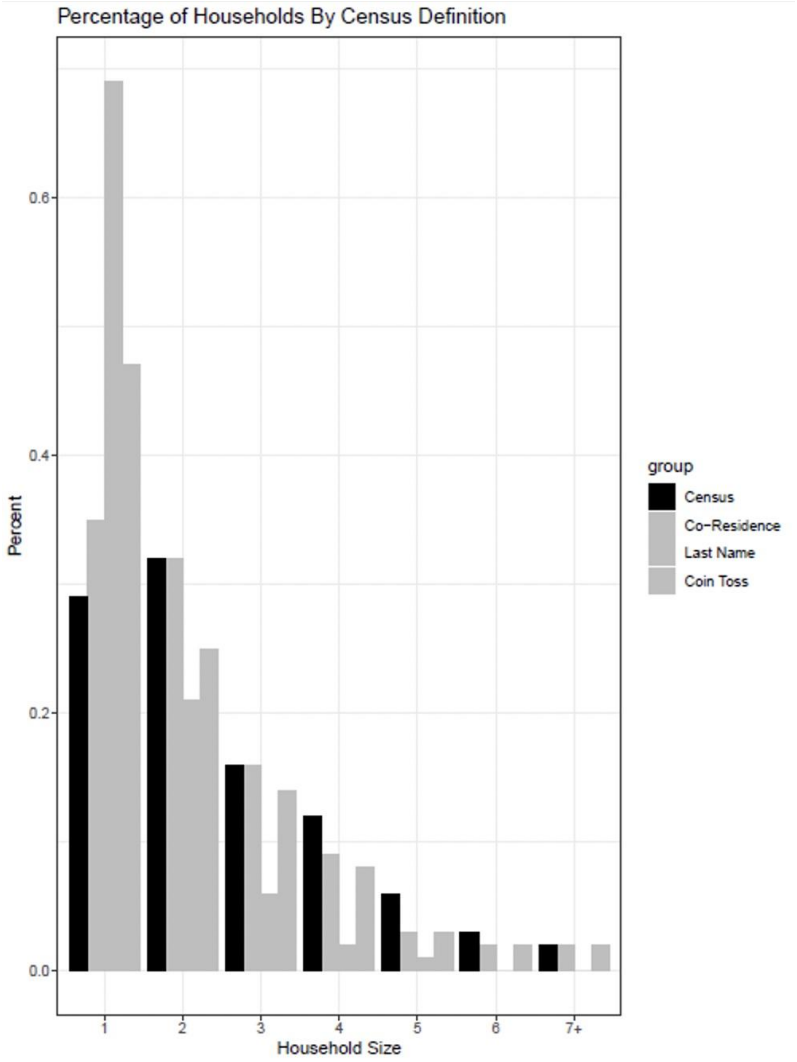


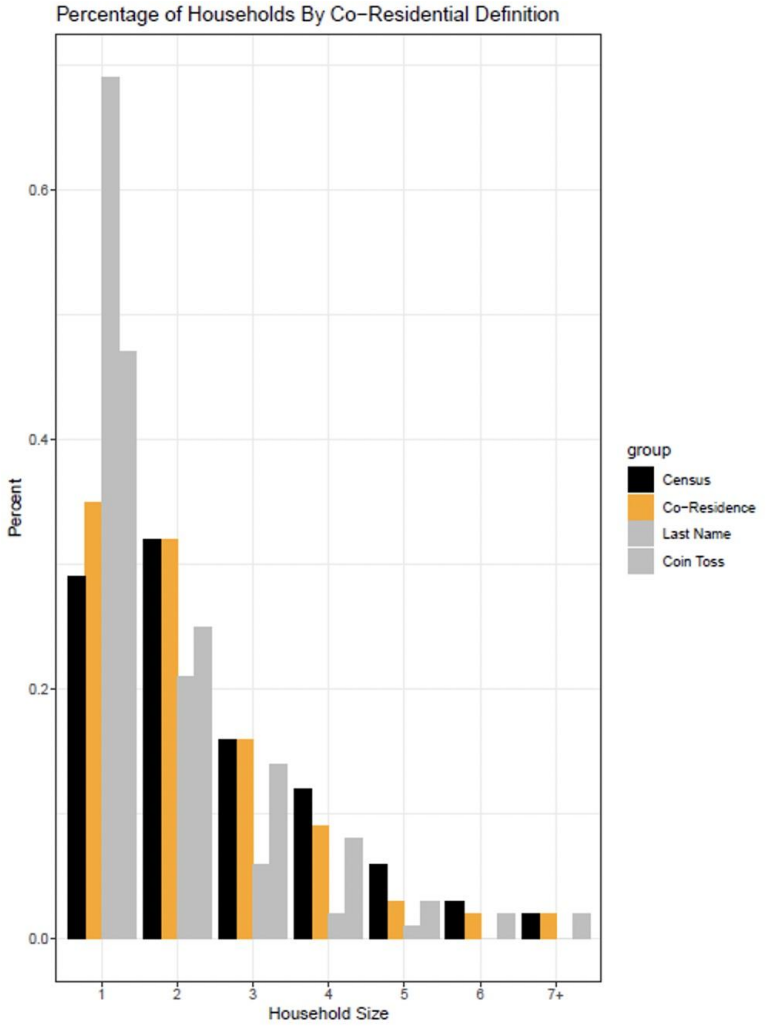


APRIL 2010 (Count)

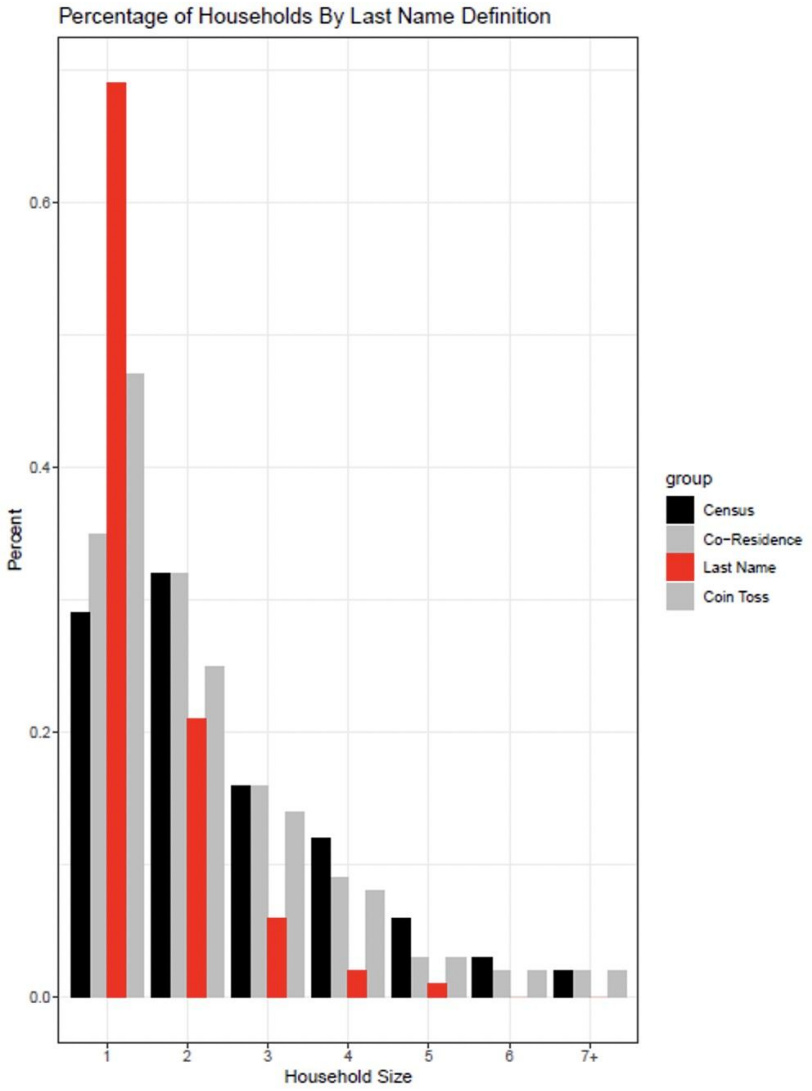


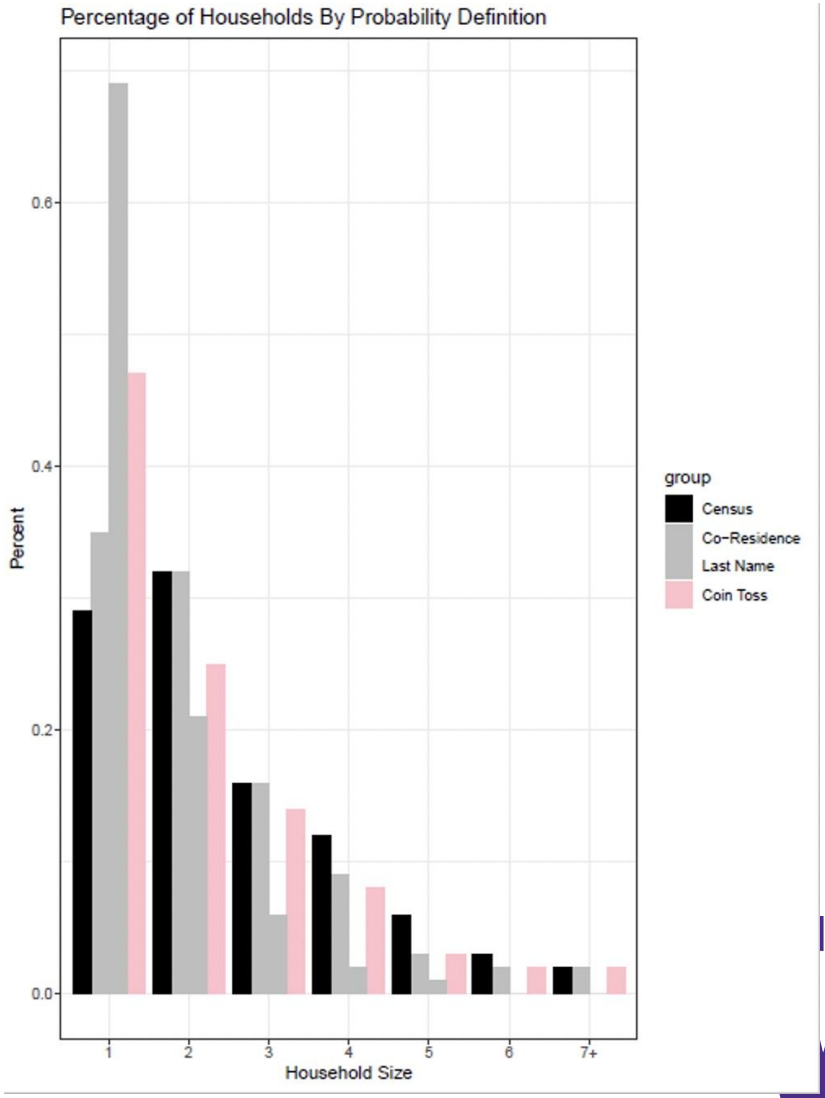
APRIL 2010 (Percentage)



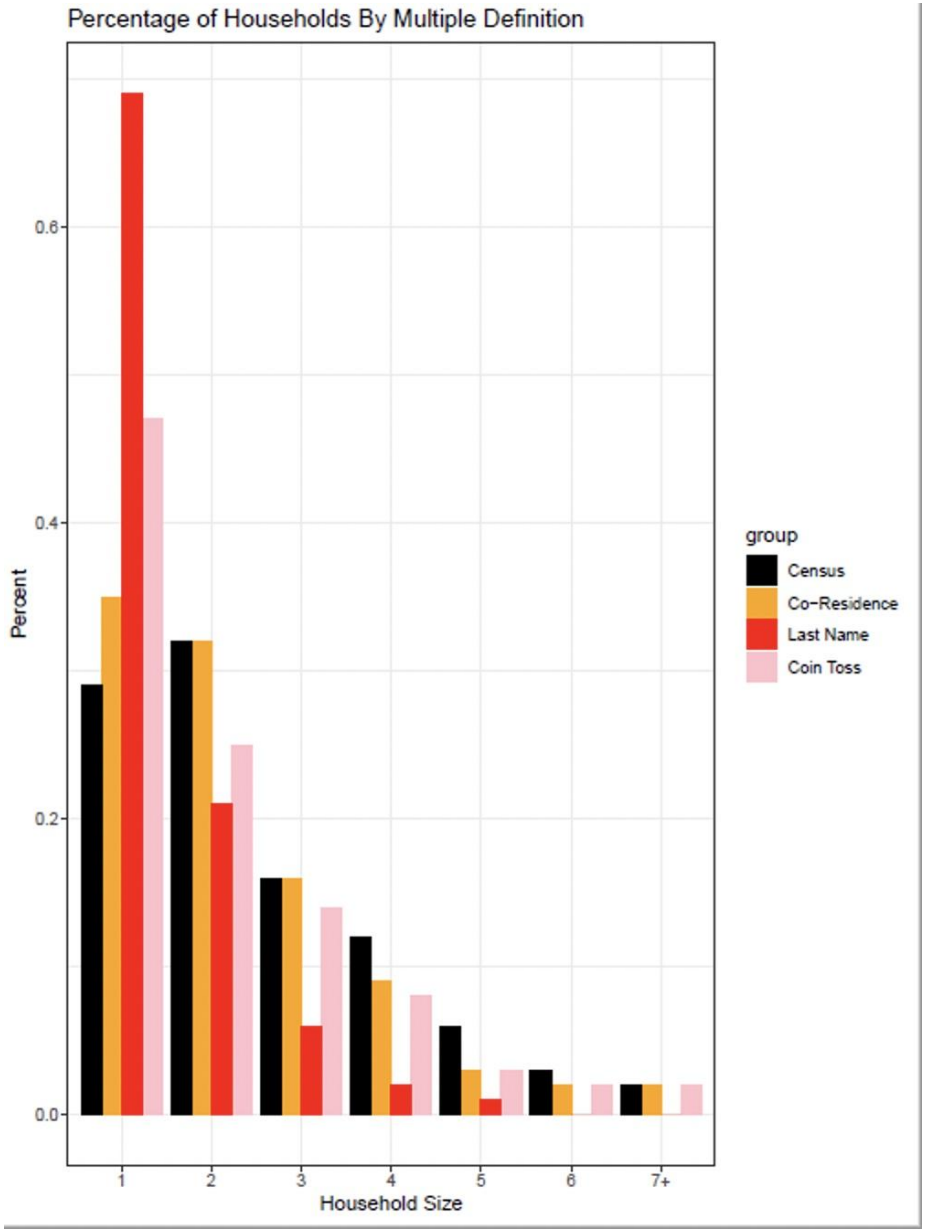


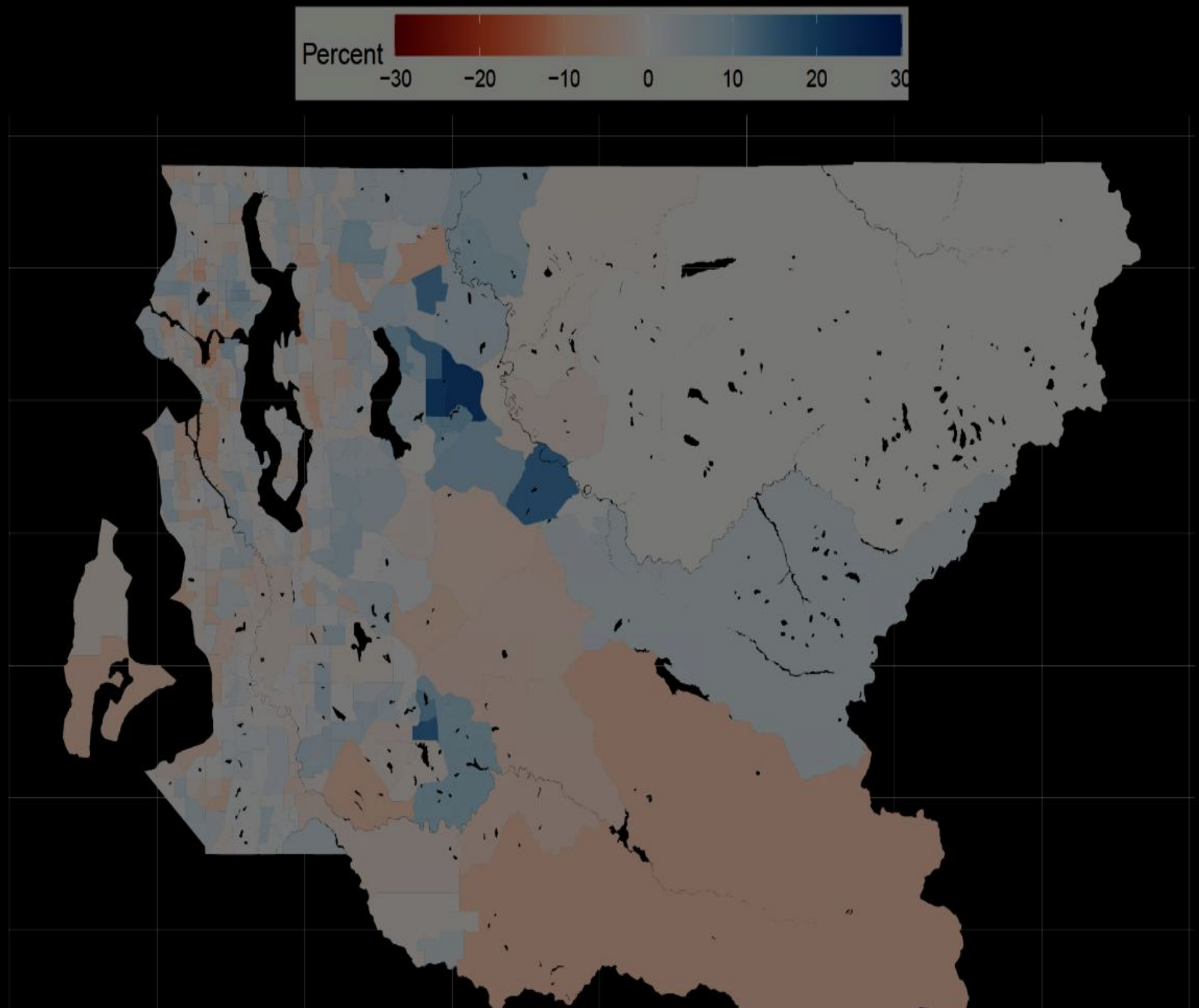
APRIL 2010 (Percentage)





APRIL (Percentage)




Percent
−30
−20
−10
0
10
20
30