

# Big Data and Economics

## Neighborhoods and Upward Mobility

---

Kyle Coombs  
Bates College | EC/DCS 368

# Table of contents

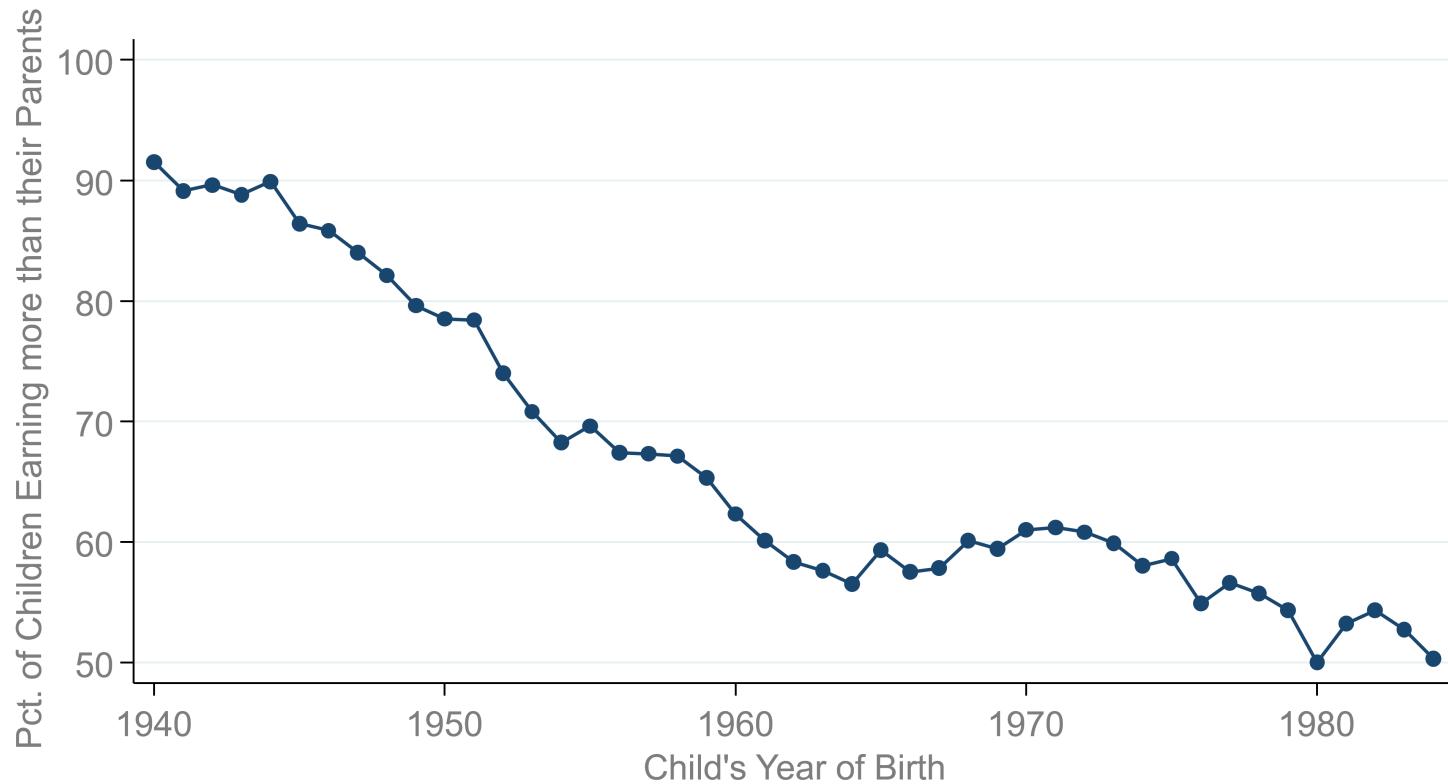
1. Prologue
2. Geographical Variation in Upward Mobility
3. Characteristics of High-Mobility Areas
4. Spatial Correlation and Decay

# Prologue

---

# Prologue

- Today's lecture is a little different than the last few
- We're talking about an application of big data to a big question: why do some people move up the income ladder and others don't?
- This is a big question in economics and public policy
- Chetty answered it using big data and spatial analysis
  - By big I mean: essentially all tax returns in the USA from 1989-2015
- He released summaries of the data publicly in 2018 as the Opportunity Atlas
- These show tons of descriptive measures of income mobility at various levels of geography: state, county, and Census Tract



Source: Chetty et al. (2014)

# Why is the "American Dream" Fading?

- Why are children's chances of climbing the income ladder falling in the USA?
  - What can be done to reverse this trend?
- Need to go beyond macroeconomic data to answer this question. Why?
  - Too many changes happening over time and across space to separate out the causal factors.
  - Also: only a handful of data points (classic macro problem)

# Enter the Opportunity Atlas

- Created in 2018, the Opportunity Atlas offers one measure of how income mobility differs by location in the USA
  - If some areas have more mobility than others, can we learn why and apply those lessons elsewhere?
- Data sources:
  - Anonymized Census data (2000, 2010 ACS) covering U.S. population
  - Federal income tax returns from 1989-2015.
- Method: Link parents based on dependent claiming on tax returns
- Target sample: Children born between 1978-1983 (U.S. citizens and authorized immigrants who arrived as children)

There's bound to be a mess with this much data, so they create an analysis sample

- **Analysis sample:** 20.5 million children, 96% coverage of target sample

# Toolkit to use these data

- Data cleaning and wrangling
- Data visualization
- Spatial analysis
- Regression analysis

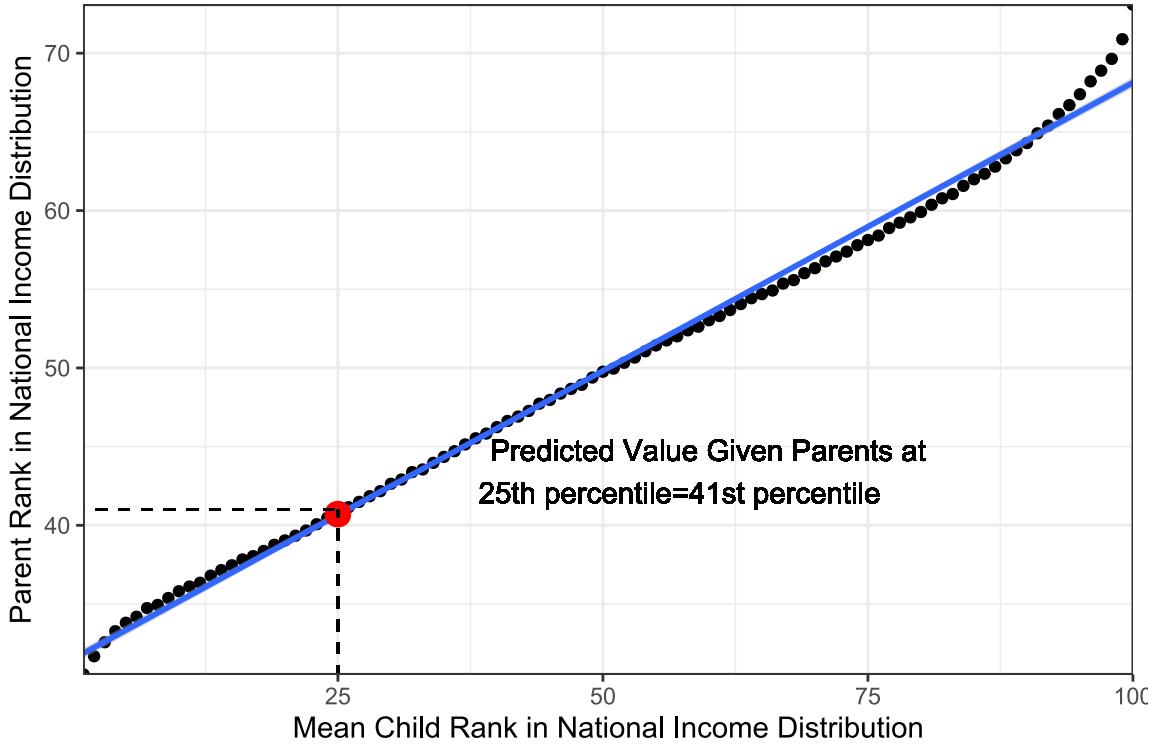
# Incomes in Tax Data

- Parent household incomes: average income reported on Form 1040 tax return from 1994-2000
- Children incomes measured from tax returns in 2014-15 (ages 31-37)
- But income levels differ over time! How do we compare them?
  - Use percentile ranks in the *national* distribution
  - Rank children relative to others born in same year and parents relative to other parents
- **What is a percentile?**

# Incomes in Tax Data

- Parent household incomes: average income reported on Form 1040 tax return from 1994-2000
- Children incomes measured from tax returns in 2014-15 (ages 31-37)
- But income levels differ over time! How do we compare them?
  - Use percentile ranks in the *national* distribution
  - Rank children relative to others born in same year and parents relative to other parents
- **What is a percentile?**
- **Income percentile:** The fraction of the national income distribution that a person's income exceeds
- Take average income percentile of children by parental income percentile

# Average Child Income Percentile by



Source: [The Opportunity Atlas](#)

# Geographic Variation in Upward Mobility

---

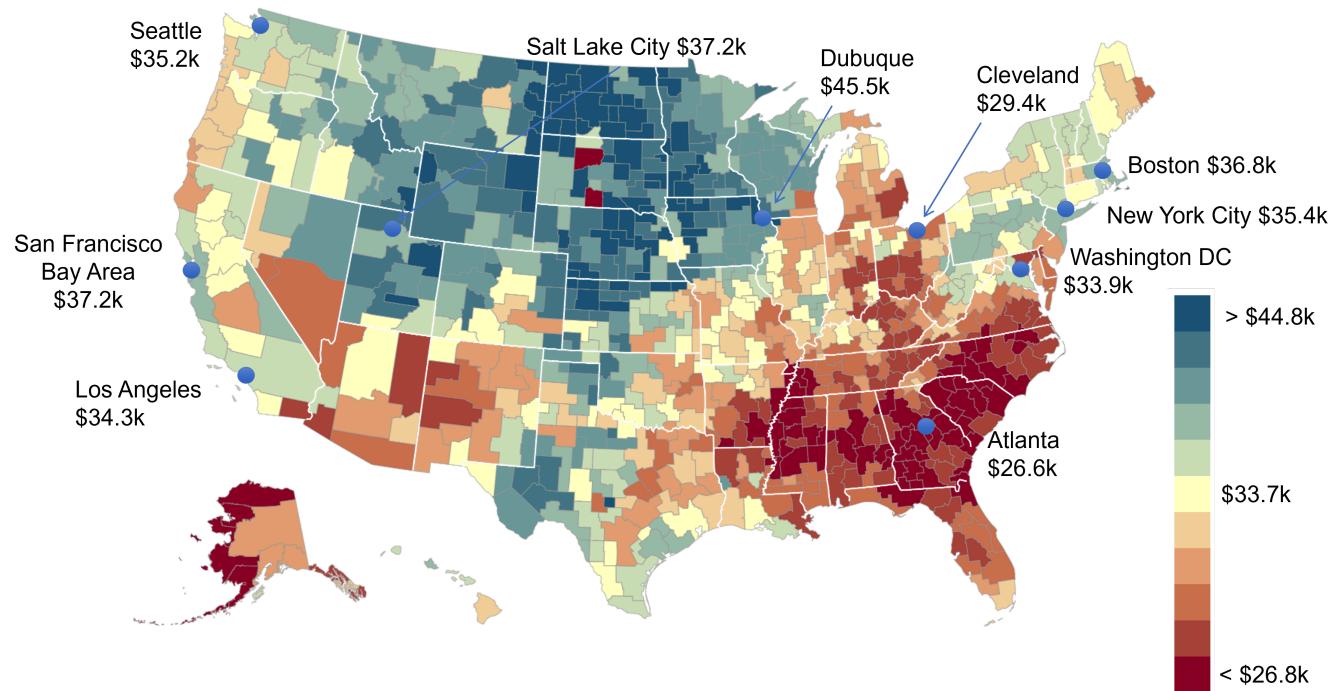
# What is mobility for a given area?

- Run this same regression of income ranks by Census tract, county, or commuting zone in the USA<sup>1</sup>
  - Census tracts are small geographic areas that contain 1,200-8,000 people
- For simplicity, Chetty et al. (2018) report the average income percentile of children whose parents were at the 25th percentile of the national income distribution
- This is a single measure of upward mobility that is easy to understand and compare across areas
  - It is not the only measure, but it is a good one
- **Big data tip:** Sensible summary statistics make big data more useful
  - The right statistic depends on the question you're asking
- **Where do you think has the lowest upward mobility? The highest?**

<sup>1</sup> Technical detail: Weight each child by fraction of childhood (up to 23) in a given area to account for movement across areas during childhood

## The Geography of Upward Mobility in the United States

Average Household Income for Children with Parents Earning \$27,000 (25<sup>th</sup> percentile)



Note: Blue = More Upward Mobility, Red = Less Upward Mobility Source: [The Opportunity Atlas](#)

# All that data and still limitations?

- They worked with the near universe of tax returns in the USA from 1989-2015
- Yet, they still have limitations
- What are a few?

# All that data and still limitations?

- They worked with the near universe of tax returns in the USA from 1989-2015
- Yet, they still have limitations
- What are a few?
- Underscores a key point: data limitations are a fact of life no matter how much data you have
- You are always simplifying the world to make it fit into data
- We use models to make sense of what those limitations are
- Even if you do not think you are using a model, you are

# What model? I didn't write one

- A model is a simplification of the world
- It outlines the variables that you assume are systematically related to each other
- e.g. Chetty et al. use tax data to measure income mobility
  - Unreported income is not included
  - Do you think unreported income is systematically underreported for some groups? In some areas?
  - To some extent this can be tested
- Can anyone think of examples of places where models are used to interpret data?

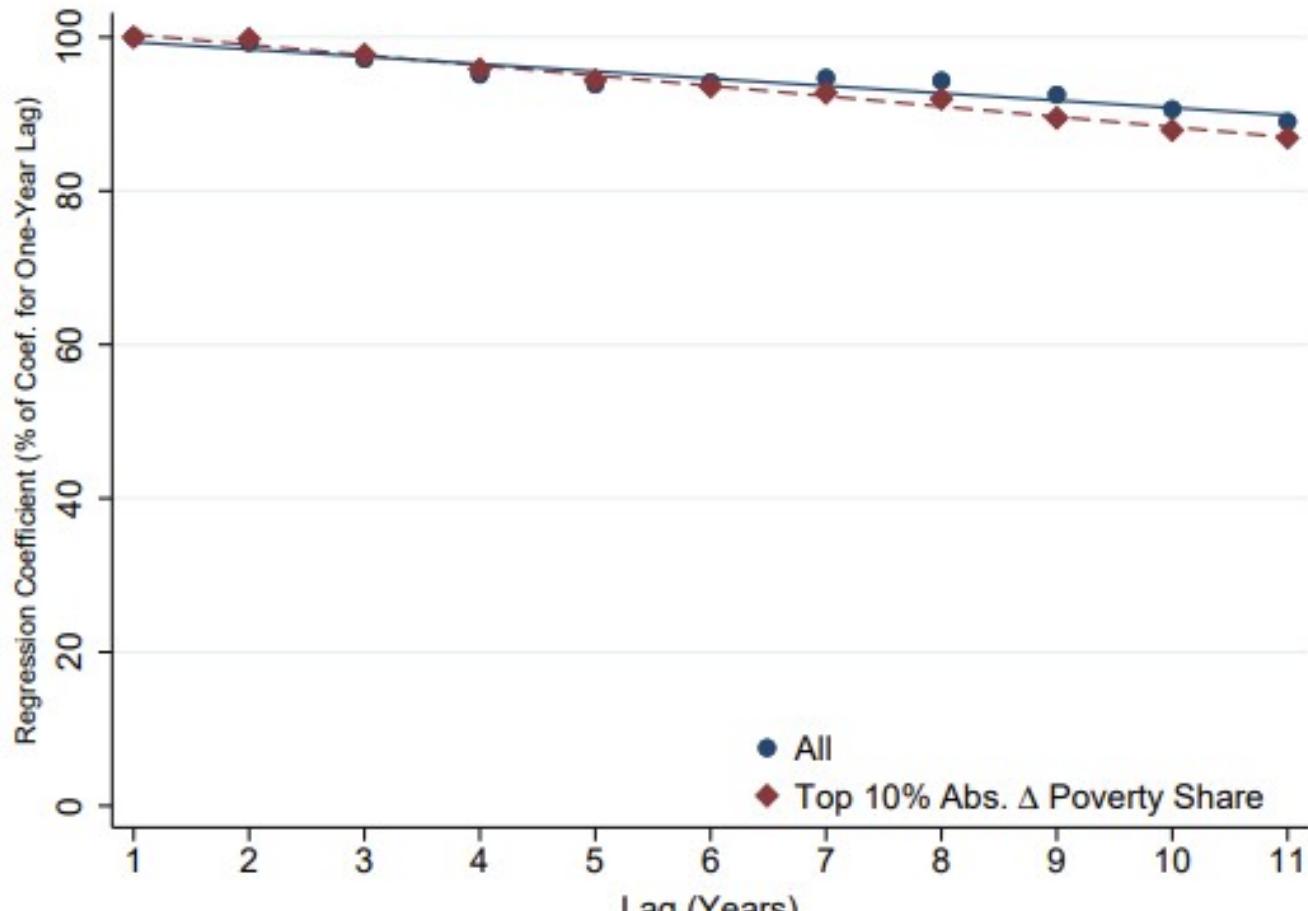
# What model? I didn't write one

- A model is a simplification of the world
- It outlines the variables that you assume are systematically related to each other
- e.g. Chetty et al. use tax data to measure income mobility
  - Unreported income is not included
  - Do you think unreported income is systematically underreported for some groups? In some areas?
  - To some extent this can be tested
- Can anyone think of examples of places where models are used to interpret data?
- Pre-1950s housing as a proxy for lead exposure

# Inferences about today

Chetty et al. extrapolate from cohorts born in the 80s to make inferences about today

- Assumption: mobility is not systematically changing over time, but it may lose precision
- Tests for "correlation" between mobility measures for cohorts born earlier in history



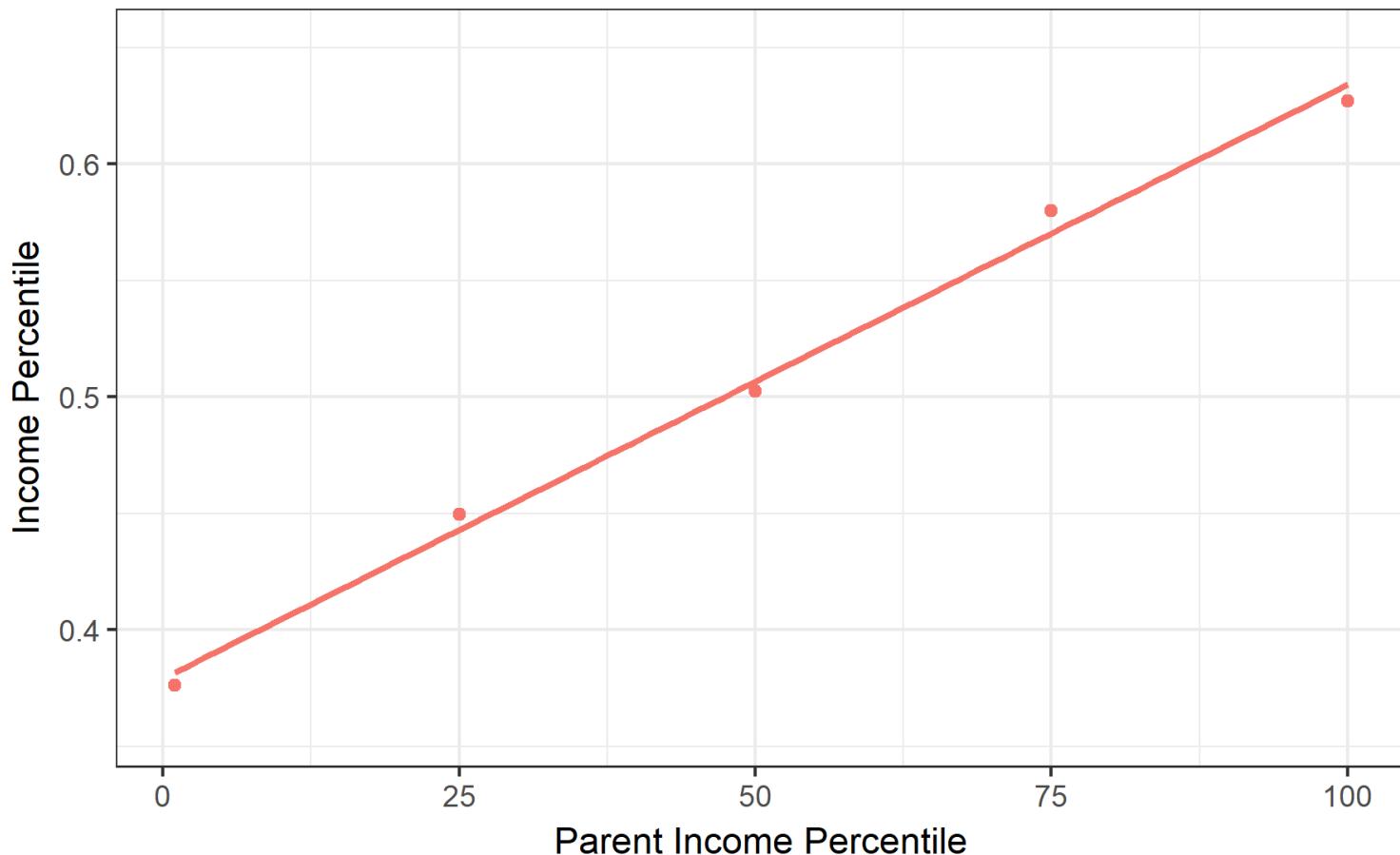
# Sometimes more data is the answer

- After a few years, the Opp Atlas team (Chetty, Dobbie, Goldman, Porter, and Yang 2024) updated with more data
- They repeat the same analysis for cohorts born between 1978 and 1992<sup>2</sup>
  - Measure adulthood income from 2005 to 2019
- Able to look at how mobility has changed over time by cohort, location, other demographics
- Indications that mobility has increased over time for some, but not all groups

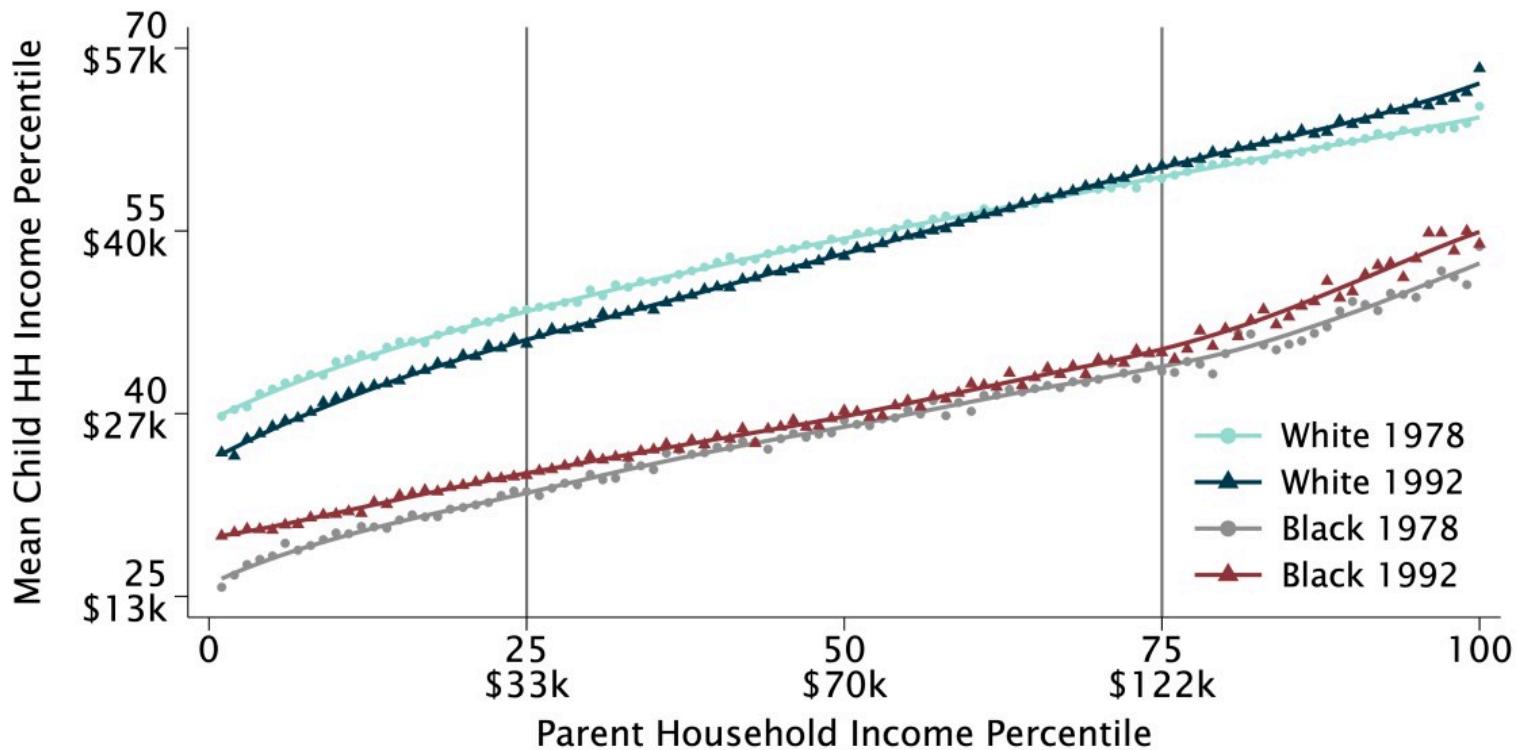
<sup>2</sup> I'm in the data!

# Mobility trends shift over time

Cohort: 1978



# Racial differences



# Characteristics of High-Mobility Areas

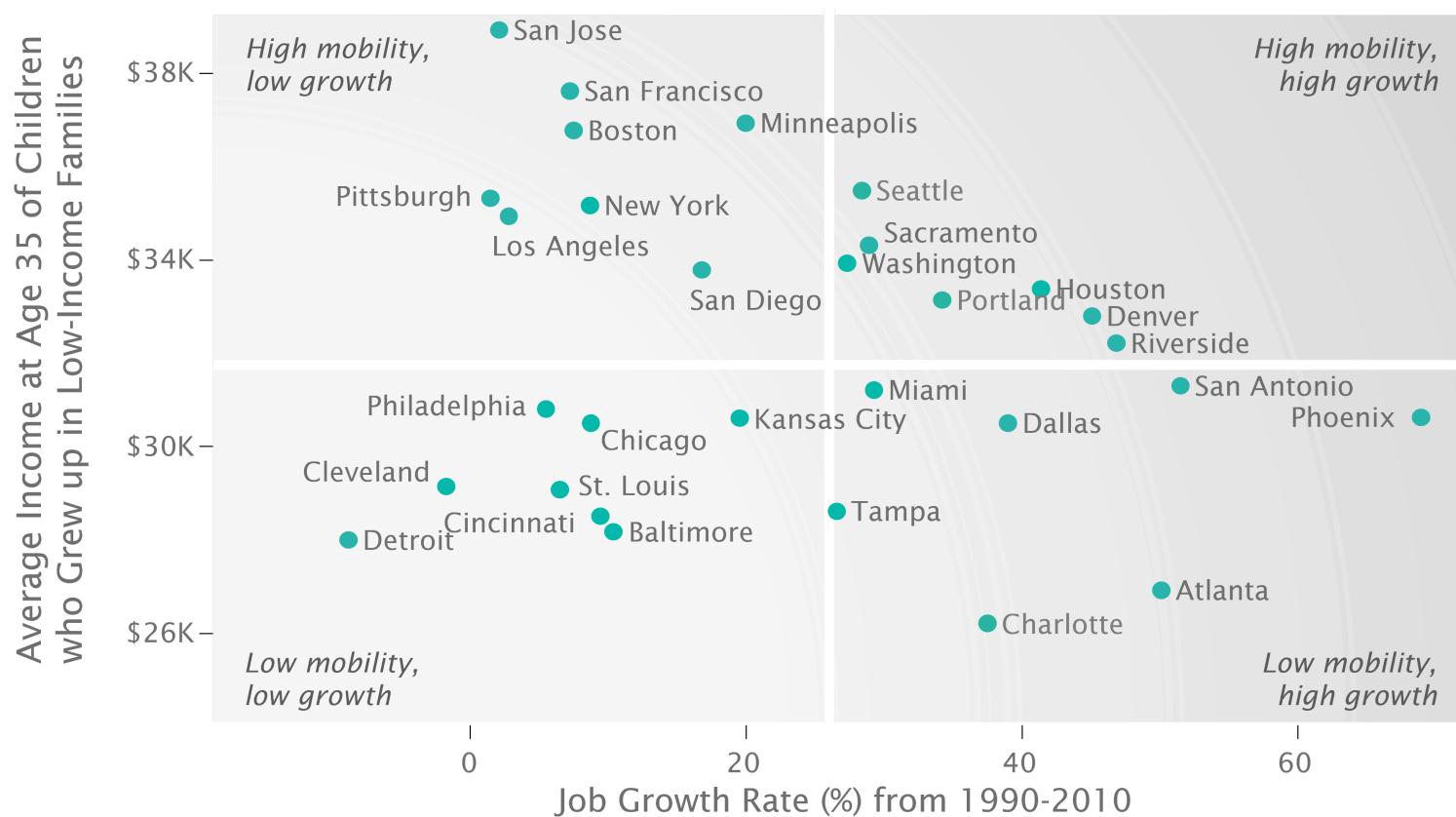
---

# Why does upward mobility differ?

Armed with a summary measure of upward mobility, we can ask:

- Why do some areas have more upward mobility than others?
- Spatial and correlational analysis is a good place to start
- What are potential characteristics of high mobility areas?
  - Better jobs?
  - Better schools?
  - Institutional differences?
  - Culture?

# Upward Mobility vs. Job Growth



# How to calculate a correlation

- Quick review: what is a correlation?
- Mathematically:

$$\text{Correlation} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

- Intuitively, what is it?

# How to calculate a correlation

- Quick review: what is a correlation?
- Mathematically:

$$\text{Correlation} = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

- Intuitively, what is it?
- It is a measure of how two variables move together normalized to be between -1 and 1
- What are some ways to calculation a correlation in R?

```
corr ← cor(outcomes$kfr_p25, outcomes$kfr_p75)
print(paste("This correlation between 25th and 75th percentile mobility is:", corr))
```

```
## [1] "This correlation between 25th and 75th percentile mobility is: -0.684480995406698"
```

# Correlations using regression

- One handy way to calculate a correlation is to use regression, exploiting the formula for the coefficient

In a regression, the coefficient on  $X$ :

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Subtract means and divide by the standard deviation, or "z-score," we'll calculate the correlation with a regression coefficient
- Why do this?
  - It is easy to interpret a coefficient
  - It is easy to get a confidence interval
  - It is easy to control for other variables
  - Great way to normalize wildly different variables

# Let's test it out

```
# library(fixest); library(broom) # already loaded
# Finish this code!
outcomes <- mutate(outcomes,
  kfr_p25_demean=,
  kfr_p75_demean=)

results <- feols(kfr_p75_demean ~ kfr_p25_demean, data=outcomes)
diff <- tidy(results) %>% filter(term='kfr_p25_demean') %>% pull(estimate) - corr

etable(results,note=paste('Correlation function and z-scored regression approach are within',
  round(diff,16), 'of each other'))
```

# Actual correlates

1. Segregation: Greater racial and income segregation associated with lower levels of mobility
2. Income Inequality: Places with smaller middle class have less mobility
3. School Quality: Higher expenditure, smaller classes, higher test scores correlated with more mobility
4. Family Structure:
  - Areas with more single parents have lower mobility
  - Strong correlation even for kids whose *own* parents are married
5. Social Capital
  - It takes a village to raise a child
  - Chetty et al. (2023) leveraged Facebook Data to create the Social Capital Atlas

# Why do we care about correlation?

- We all know correlation is not causation
- We'll discuss this in-depth after break if you don't believe me
- So why are we talking about correlation at all?
- One of the first steps in understanding a complex system is to understand how variables are related
- This is especially true when we have a lot of data
- Plus, almost ever causal relationship is just a correlation with a story
  - Story might be: I ran an experiment and found a correlation with a randomly assigned treatment
  - But the story might be: I assume some natural variation in the data is like a random assignment and I found a correlation

# Spatial Correlation and Decay

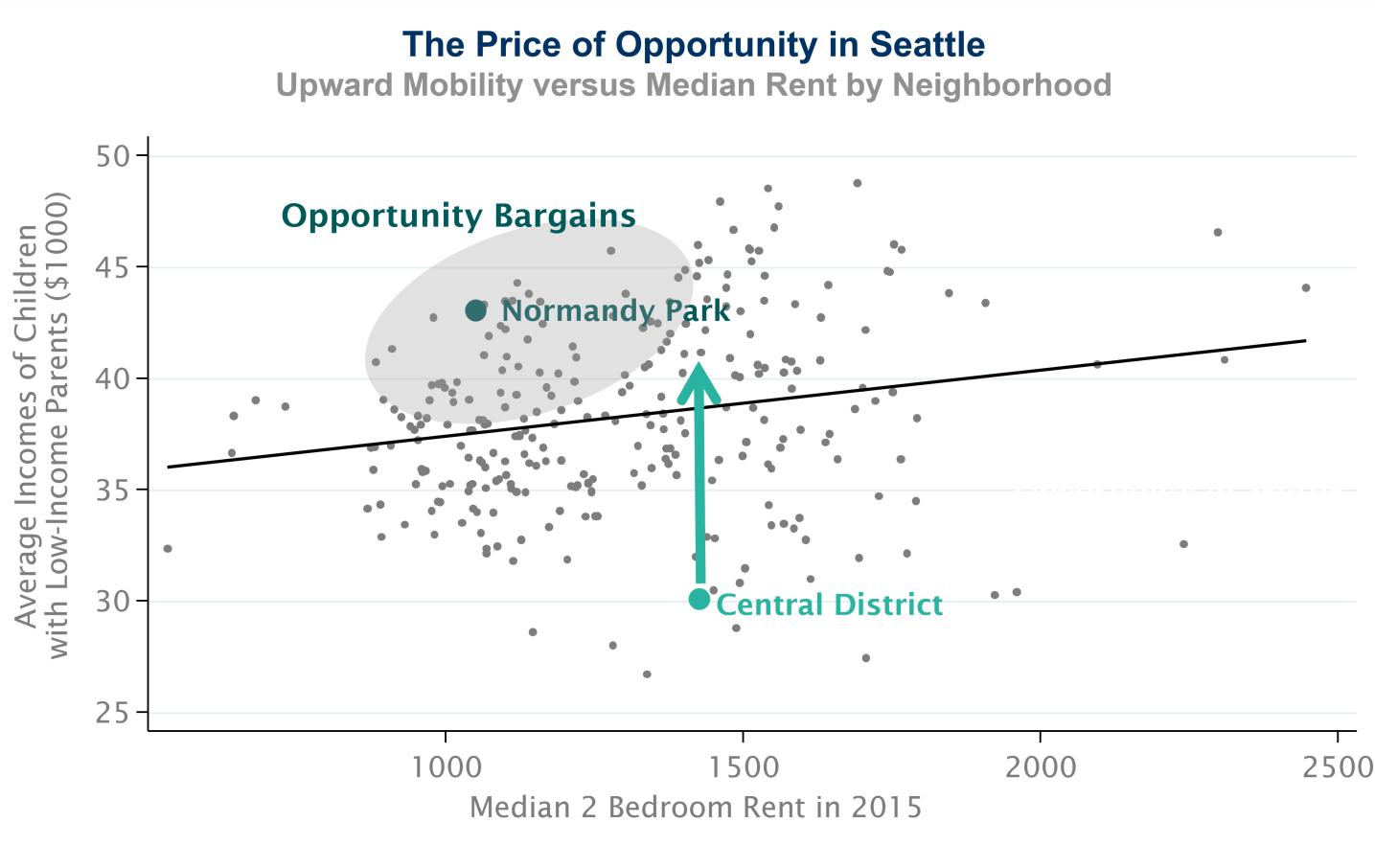
---

# Big question: why don't people move?

- If some areas have more mobility than others, why don't people move to those areas?
- Is it rent?

# The Price of Opportunity in Seattle

Upward Mobility vs Median Rent by Neighborhood



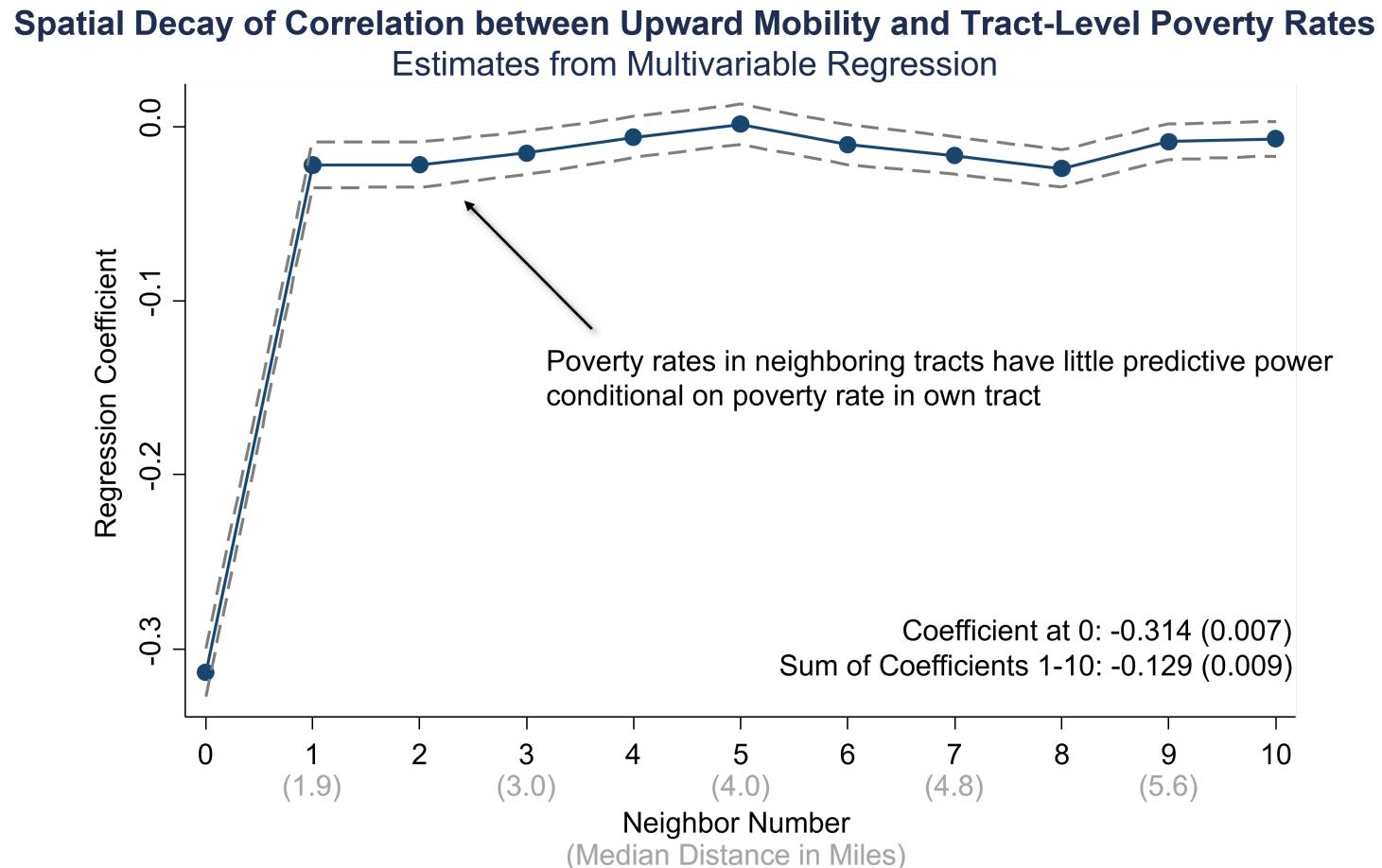
# Big question: why don't people move?

- Initial experiments indicate benefits exist from moving (we'll see later)
- If some areas have more mobility than others, why don't people move to those areas?
- Is it rent?
- Other costs of moving?
- Maybe they do not want to move as far?
- Overall, this is not a highly effective approach

# Well what if we invest locally?

- What if we invest in the areas that have low mobility? (place-based approach)
- Would there be spillovers between locations?
  - It is tough to improve one neighborhood (e.g. a tract), let alone many at once
  - Do we have to improve them all at once to help people?
- The answer to this question changes the policy approach

# Spatial decay suggests localized effects



# Overall Takeaways

- Correlation evidence is suggestive, but not causal
- Causality requires a more focused approach
- We will build this toolkit in the next few lectures

# Next lecture: Spatial Analysis and Opportunity Atlas

---