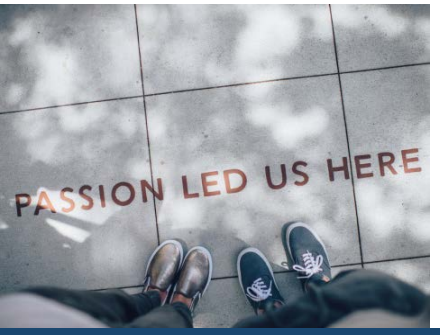


Using Big Data to Solve Economic and Social Problems

Professor Raj Chetty
Head Section Leader: Gregory Bruich, Ph.D.

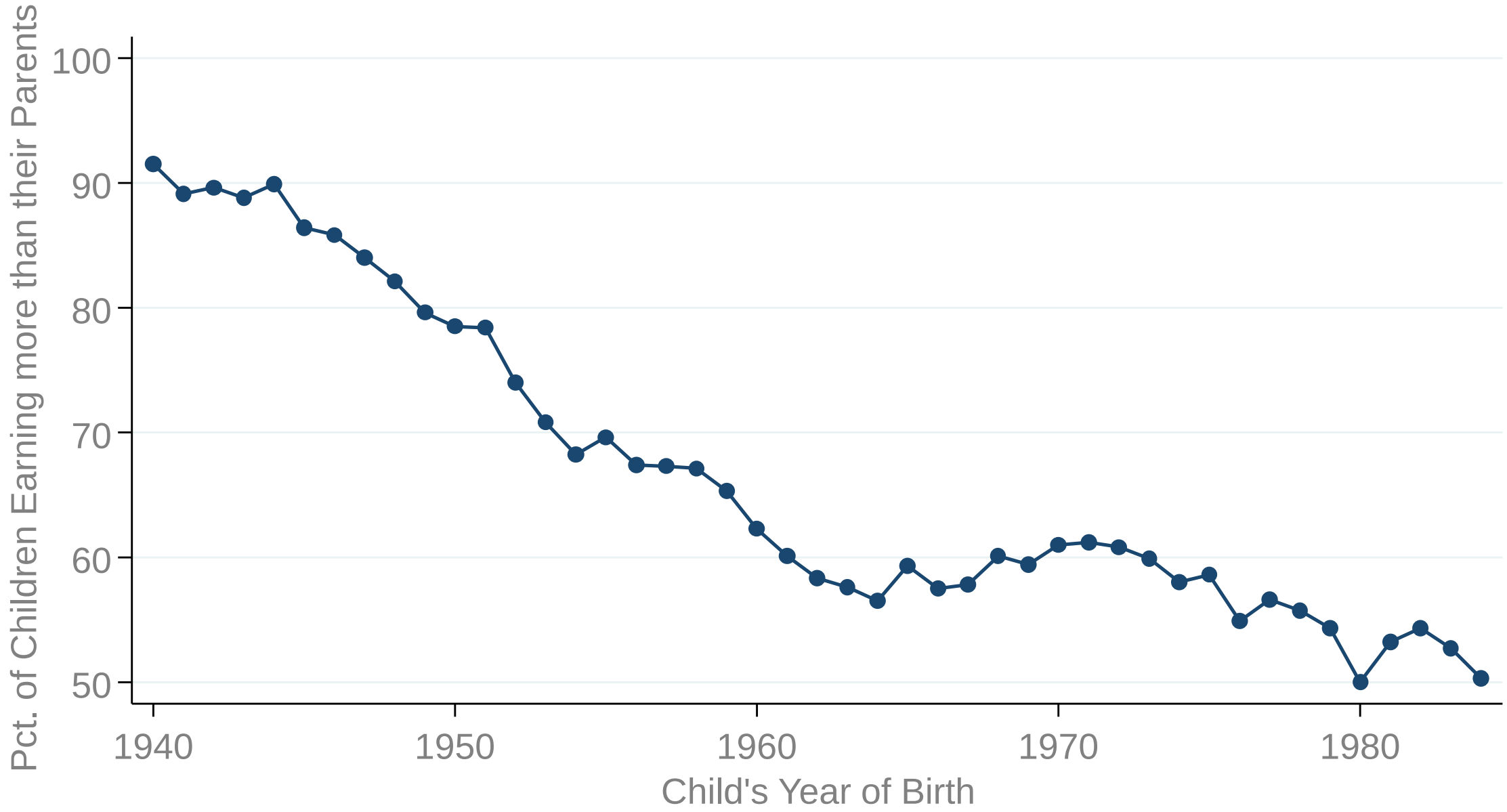
Spring 2019

HARVARD UNIVERSITY

OPPORTUNITY INSIGHTS

The Fading American Dream

Percent of Children Earning More than Their Parents, by Year of Birth



Source: Chetty, Grusky, Hell, Hendren, Manduca, Narang (Science 2017)

Why is the American Dream Fading?

- Central policy question: why are children's chances of climbing the income ladder falling in America?
  - And what can we do to reverse this trend…?
- Difficult to answer this question based solely on historical data on macroeconomic trends
  - Numerous changes over time make it hard to test between alternative explanations
  - Problem: only a handful of data points

Theoretical Social Science

- Until recently, social scientists have had limited data to study policy questions like this
- Social science has therefore been a *theoretical* field
  - Develop mathematical models (economics) or qualitative theories (sociology)
  - Use these theories to explain patterns and make policy recommendations, e.g. to improve upward mobility

Theoretical Social Science

- Problem: theories untested → five economists often have five different answers to a given question

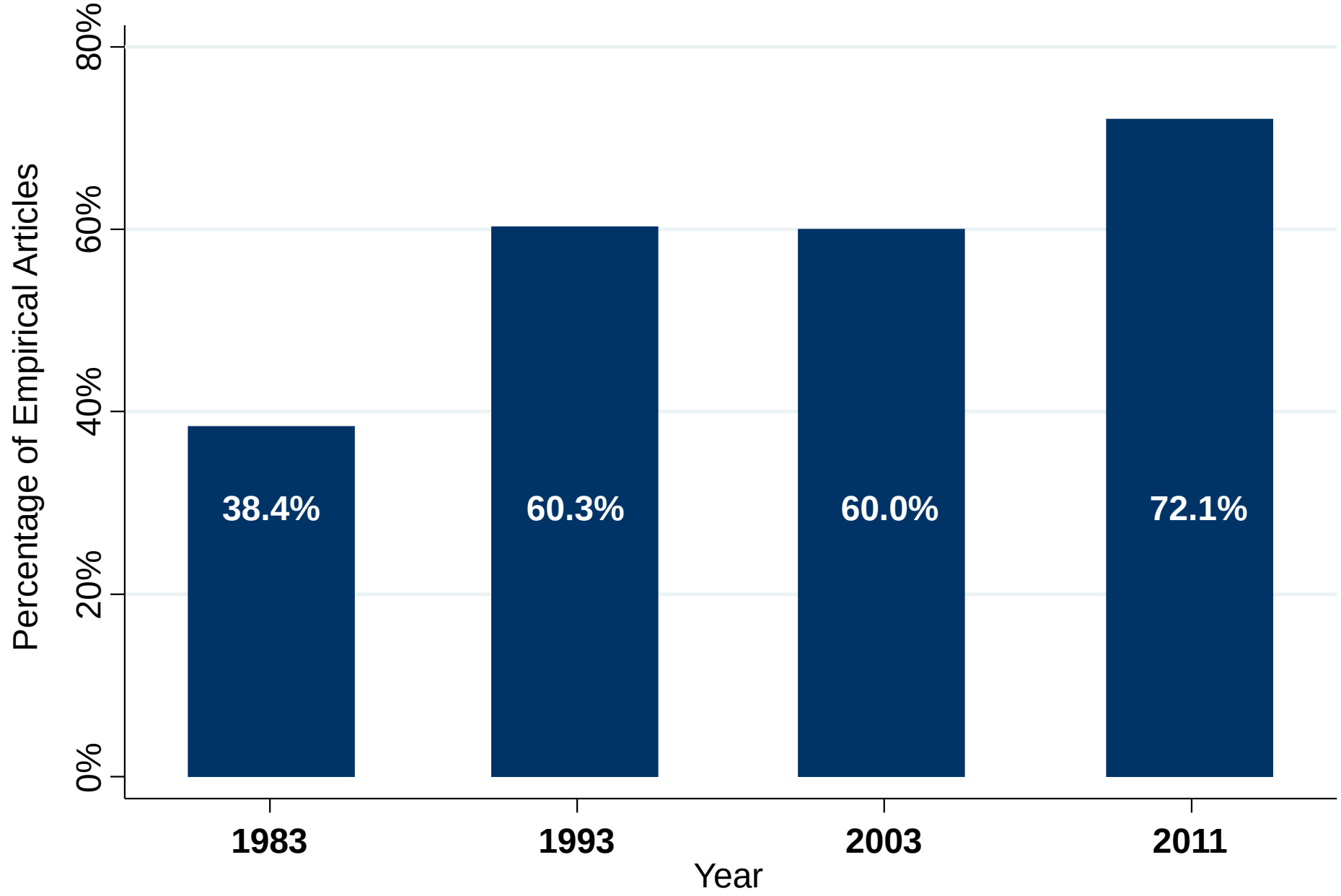
- Leads to a politicization of questions that in principle have scientific answers

  - Example: is Obamacare reducing job growth in America?

The Rise of Data and Empirical Evidence

- Today, social science is becoming a more empirical field thanks to the growing availability of data
  - Test and improve theories using real-world data
  - Analogous to natural sciences

Empirical (Data-Based) Articles in Leading Economics Journals, 1983-2011



Source: Hamermesh (JEL 2013)

Social Science in the Age of Big Data

- Recent availability of “big data” has accelerated this trend
  - Large datasets are starting to transform social science, as they have transformed business
- Examples:
  - Government data: tax records, Medicare
  - Corporate data: Google, Uber, retailer data
  - Unstructured data: Twitter, newspapers

Why is Big Data Transforming Social Science?

1. Greater reliability than surveys
2. Ability to measure new variables (e.g., emotions)
3. Universal coverage → can “zoom in” to subgroups
4. Large samples → can approximate scientific experiments

Why This Course?

- Companies like Amazon have succeeded in solving major *private market* problems using technology and big data

- Goal of this course: show how same skills can be used to address important *social* problems

  - We need more talent in this area given pressing challenges such as rising inequality and global warming

- To achieve this goal, provide an introduction to a broad range of topics, methods, and real-world applications

  - Start from the *questions* to motivate the methods rather than the traditional approach of doing the reverse

Overview of Topics

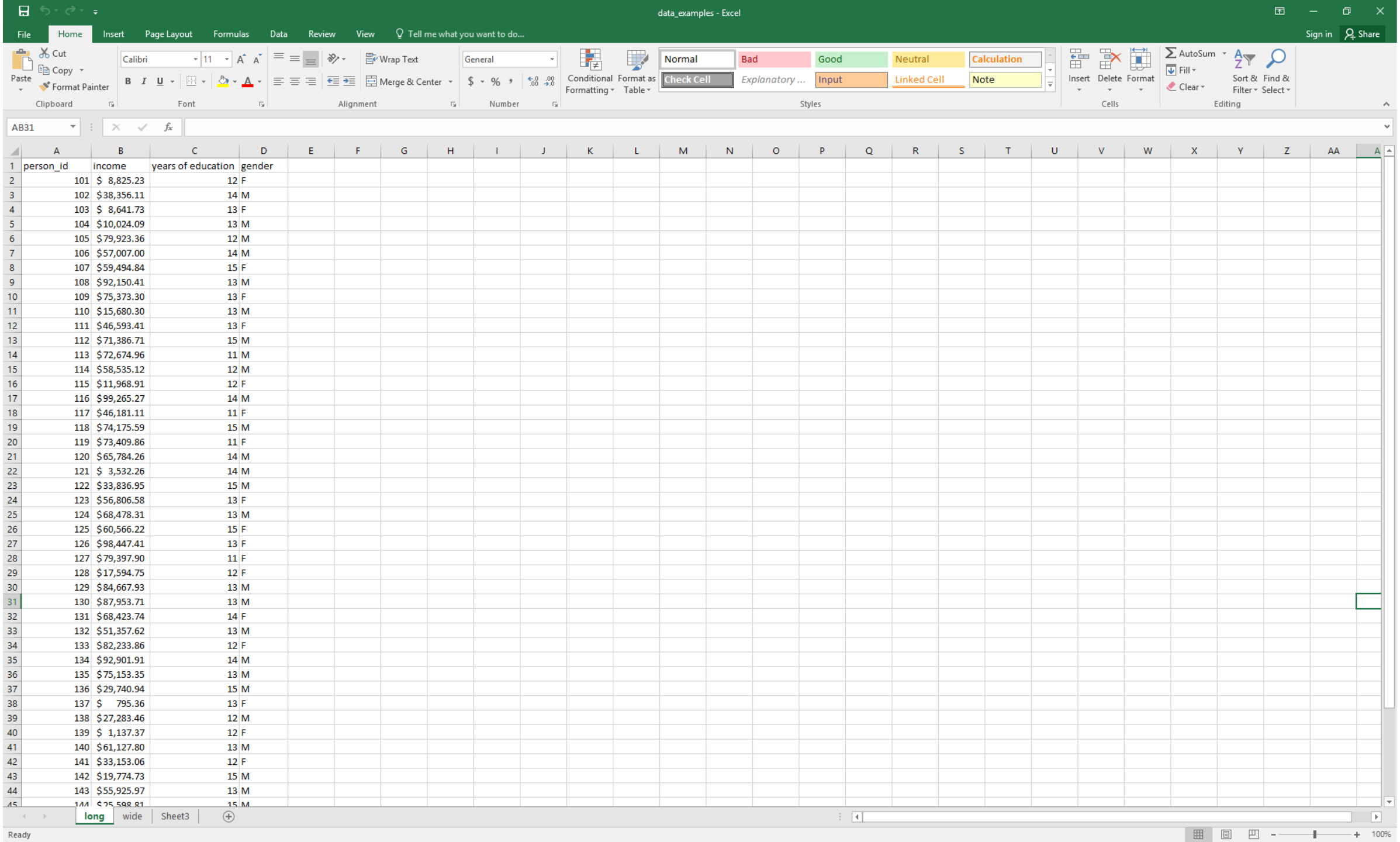
1. Equality of Opportunity
2. Education
3. Racial Disparities
4. Health
5. Criminal Justice
6. Tax Policy
7. Climate Change
8. Economic Development and Institutional Change

Examples of Statistical Methods You Will Learn in this Class

1. Descriptive Data Analysis: correlation, regression, survival analysis
2. Experiments: randomization, non-compliance
3. Quasi-Experiments: regression discontinuity, difference-in-differences
4. Machine Learning: prediction, overfitting, cross-validation
5. Stata (or other) statistical programming language

Statistical Methods: Two Types of “Big Data”

- Big data can be classified into two types
  - “Long” data: many observations relative to variables (e.g., tax records)



data_examples - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do... Sign in Share

Cut Copy Format Painter Paste Clipboard

Calibri 11 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table

Normal Bad Good Neutral Calculation Check Cell Explanatory ... Input Linked Cell Note Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

AB31

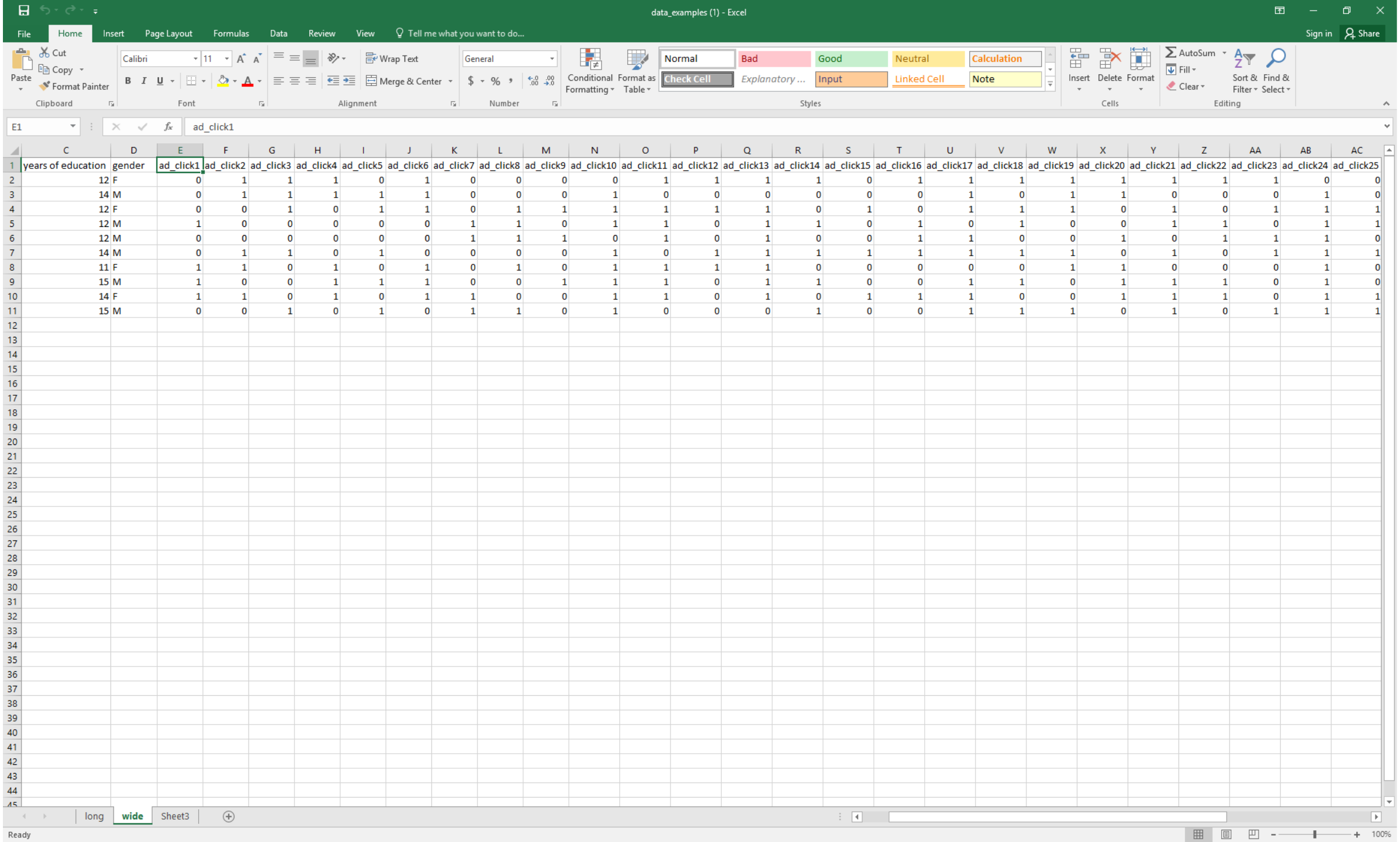
	A	B	C	D
1	person_id	income	years of education	gender
2	101	$ 8,825.23	12	F
3	102	$38,356.11	14	M
4	103	$ 8,641.73	13	F
5	104	$10,024.09	13	M
6	105	$79,923.36	12	M
7	106	$57,007.00	14	M
8	107	$59,494.84	15	F
9	108	$92,150.41	13	M
10	109	$75,373.30	13	F
11	110	$15,680.30	13	M
12	111	$46,593.41	13	F
13	112	$71,386.71	15	M
14	113	$72,674.96	11	M
15	114	$58,535.12	12	M
16	115	$11,968.91	12	F
17	116	$99,265.27	14	M
18	117	$46,181.11	11	F
19	118	$74,175.59	15	M
20	119	$73,409.86	11	F
21	120	$65,784.26	14	M
22	121	$ 3,532.26	14	M
23	122	$33,836.95	15	M
24	123	$56,806.58	13	F
25	124	$68,478.31	13	M
26	125	$60,566.22	15	F
27	126	$98,447.41	13	F
28	127	$79,397.90	11	F
29	128	$17,594.75	12	F
30	129	$84,667.93	13	M
31	130	$87,953.71	13	M
32	131	$68,423.74	14	F
33	132	$51,357.62	13	M
34	133	$82,233.86	12	F
35	134	$92,901.91	14	M
36	135	$75,153.35	13	M
37	136	$29,740.94	15	M
38	137	$ 795.36	13	F
39	138	$27,283.46	12	M
40	139	$ 1,137.37	12	F
41	140	$61,127.80	13	M
42	141	$33,153.06	12	F
43	142	$19,774.73	15	M
44	143	$55,925.97	13	M
45	144	$25,598.81	15	M

long wide Sheet3

Ready 100%

Statistical Methods: Two Types of “Big Data”

- Big data can be classified into two types
  - “Long” data: many observations relative to variables (e.g., tax records)
  - “Wide” data: few observations relative to variables (e.g. Amazon clicks, newspapers)



data_examples (1) - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do... Sign in Share

Cut Copy Format Painter Paste Clipboard

Calibri 11 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table

Normal Bad Good Neutral Calculation Check Cell Explanatory ... Input Linked Cell Note Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

E1 ad_click1

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	years of education	gender	ad_click1	ad_click2	ad_click3	ad_click4	ad_click5	ad_click6	ad_click7	ad_click8	ad_click9	ad_click10	ad_click11	ad_click12	ad_click13	ad_click14	ad_click15	ad_click16	ad_click17	ad_click18	ad_click19	ad_click20	ad_click21	ad_click22	ad_click23	ad_click24	ad_click25
2	12	F	0	1	1	1	0	1	0	0	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0
3	14	M	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0
4	12	F	0	0	1	0	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1
5	12	M	1	0	0	0	0	0	1	1	0	1	1	0	1	1	0	1	0	1	0	0	1	1	0	1	1
6	12	M	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	0	1	0	1	1	1	0
7	14	M	0	1	1	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1
8	11	F	1	1	0	1	0	1	0	1	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0
9	15	M	1	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1	1	0	0
10	14	F	1	1	0	1	0	1	1	0	0	1	1	0	1	0	1	1	1	0	0	1	1	1	0	1	1
11	15	M	0	0	1	0	1	0	1	1	0	1	0	0	0	1	0	0	1	1	1	0	1	0	1	1	1

long wide Sheet3

Ready 100%

Statistical Methods: Two Types of “Big Data”

- Statistics/computer science has focused on “wide” data
  - Main application: *prediction*
  - Example: predicting income to target ads
- Social science has focused on “long” data
  - Main application: *identifying causal effects*
  - Example: effects of improving schools on income

Examples of Economic Concepts You Will Learn in this Class

1. Effects of price incentives
2. Supply and demand
3. Competitive equilibrium
4. Adverse selection
5. Behavioral economics vs. rational models

Two Types of Sections

- We recognize that not everyone taking this class has the same background in statistics and economics
  - Some students have taken many courses already, others are just starting
- Lectures will be structured so that everyone can follow them, with no prior knowledge assumed
- Sections will be divided into two types, based on whether students have prior coursework in statistics/econometrics
  - Please respond to emails you will receive this week asking about your prior coursework and preferences

Empirical Projects

- To help students learn, we will assign four empirical projects that will get you into the data
- Will focus on real-world questions and involve coding, reading papers, and writing
- For example, fourth project will be analogous to the “Netflix challenge” to predict the movies people will like
- We will have a “Social Mobility challenge” to identify predictors of mobility and neighborhood change

Discussions with Leading Experts on Real-World Applications

1. Affordable Housing: Shaun Donovan
2. College Completion: Timothy Renick
3. Food Stamps Programs: Jesse Shapiro
4. Health and Criminal Justice: Lynn Overmann
5. Poverty in Developing Countries: Esther Duflo

Important Note: Guest discussants are generously providing their time to us → attendance is mandatory and will count toward your grade

Topic I
Equality of Opportunity

Lecture 1 Outline

1. Geographical Variation in Upward Mobility in America

2. Causal Effects of Places vs. Sorting

- Lecture 1 is based primarily on the following paper:

  Chetty, Friedman, Hendren, Jones, Porter. “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility” NBER wp, 2018

Part 1
Geographical Variation in Upward Mobility

Differences in Opportunity Across Local Areas

- How do children's chances of moving up vary across areas in America?
  - Are there some areas where kids do better than others? If so, what lessons can we learn from them?

- Recent studies have used big data to measure how upward mobility varies based on where children *grow up*

The Opportunity Atlas
Data Sources and Sample Definitions

- Data sources: Anonymized Census data (2000, 2010, ACS) covering U.S. population linked to federal income tax returns from 1989-2015
- Link children to parents based on dependent claiming on tax returns
- Target sample: Children in 1978-83 birth cohorts who were born in the U.S. or are authorized immigrants who came to the U.S. in childhood
- Analysis sample: 20.5 million children, 96% coverage rate of target sample

Measuring Parents' and Children's Incomes in Tax Data

- Parents' household incomes: average income reported on Form 1040 tax return from 1994-2000

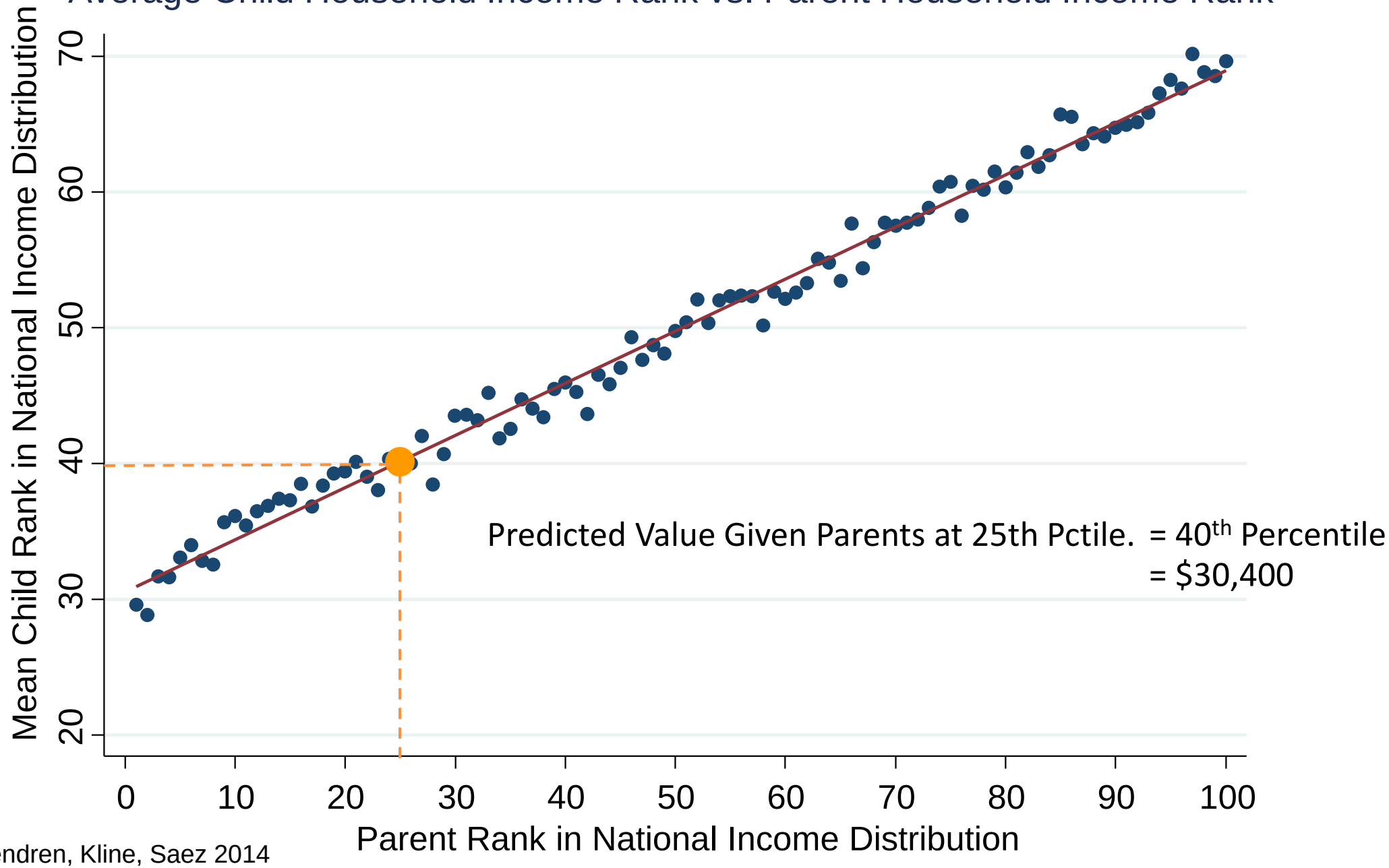
- Children's incomes measured from tax returns in 2014-15 (ages 31-37)

- Focus on percentile ranks in **national** distribution:

  - Rank children relative to others born in the same year and parents relative to other parents

Intergenerational Income Mobility for Children Raised in Chicago

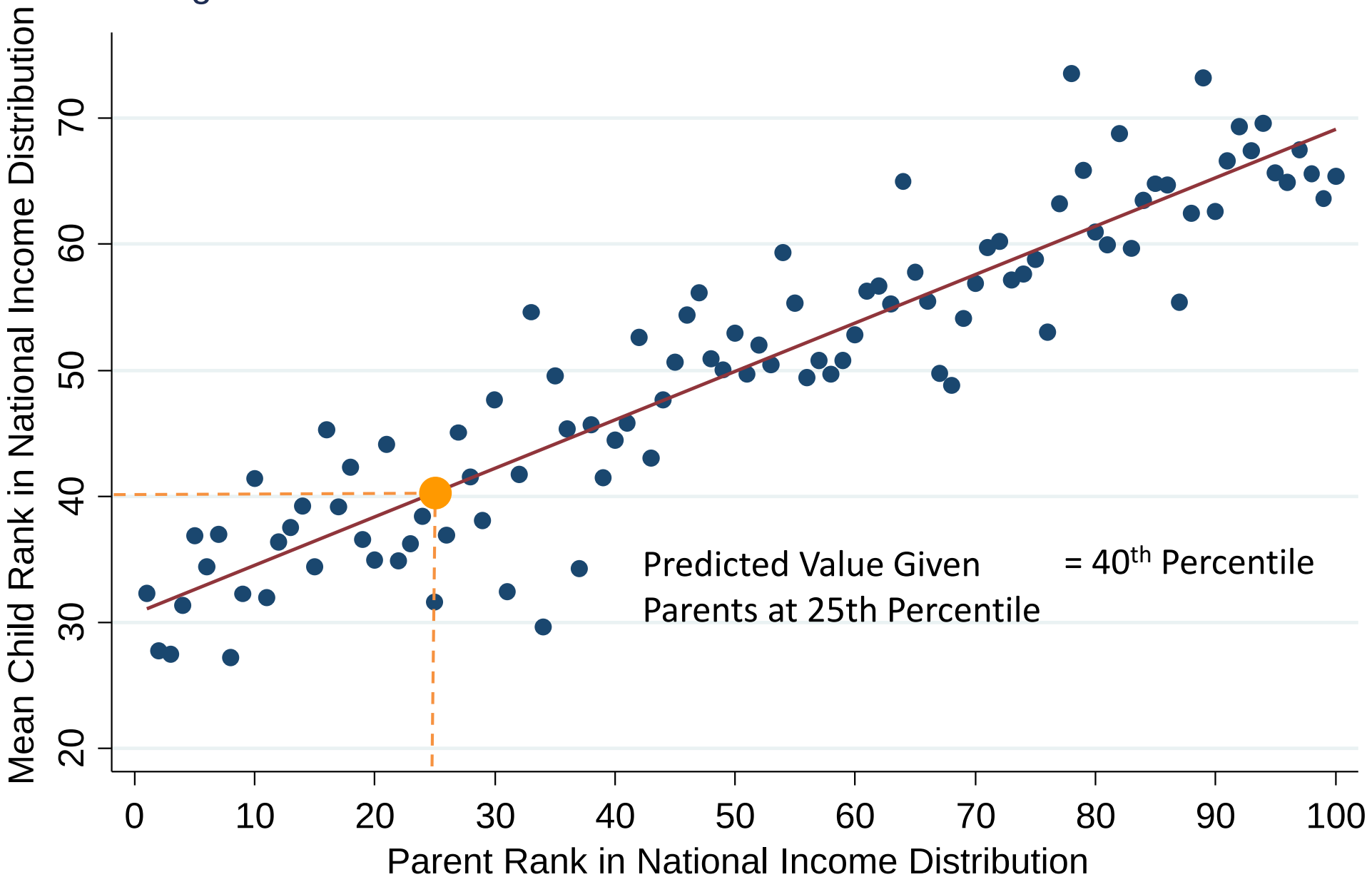
Average Child Household Income Rank vs. Parent Household Income Rank



Source: Chetty, Hendren, Kline, Saez 2014

Intergenerational Income Mobility for Children Raised in a Hypothetical Census Tract

Average Child Household Income Rank vs. Parent Household Income Rank



Estimating Children's Average Outcomes by Census Tract

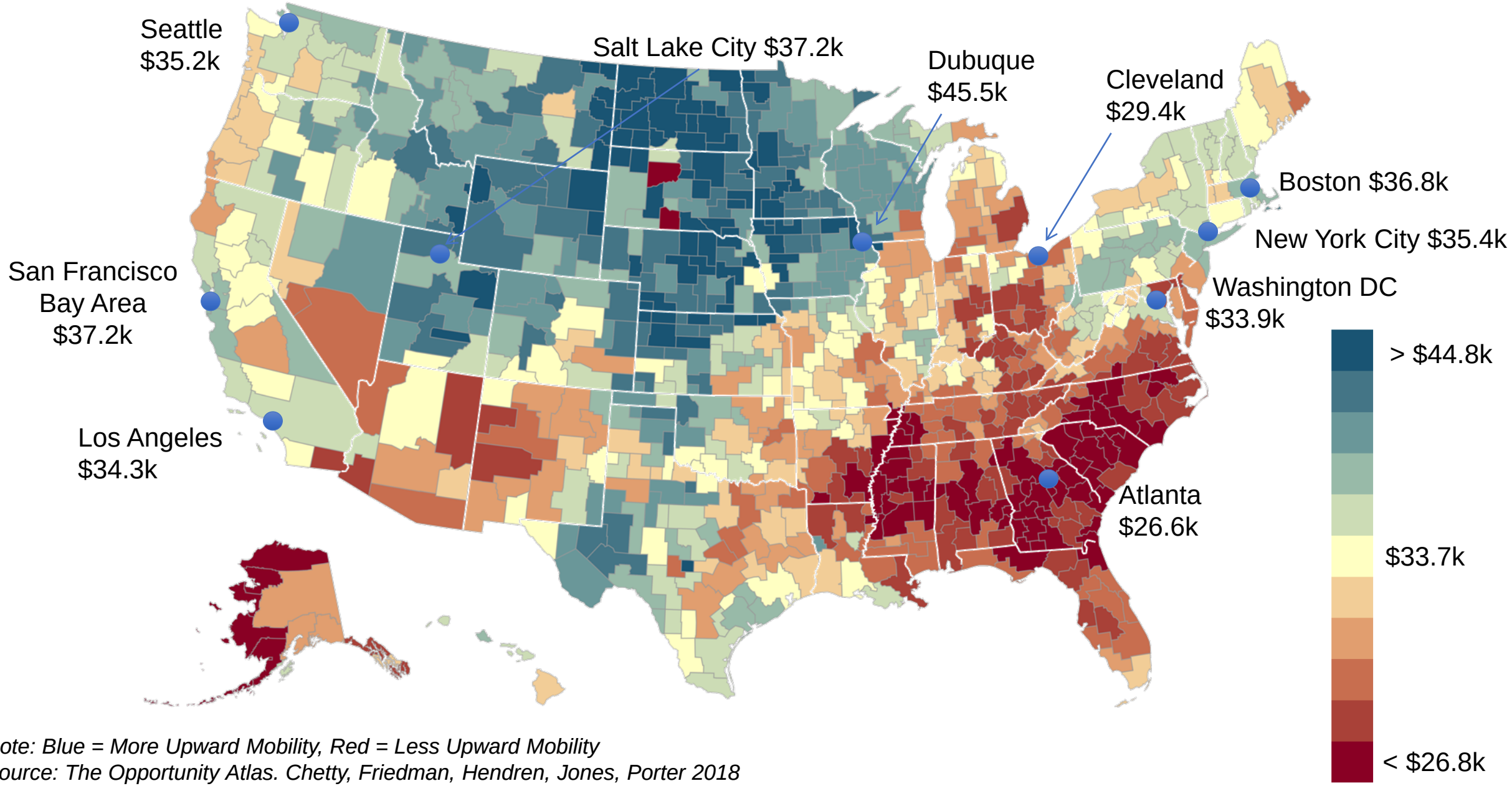
- Run a separate regression using data for children who grow up in each Census tract in America

- In practice, many children move across areas in childhood

  - Weight children by fraction of childhood (up to age 23) spent in a given area

The Geography of Upward Mobility in the United States

Average Household Income for Children with Parents Earning $27,000 (25th percentile)



Note: Blue = More Upward Mobility, Red = Less Upward Mobility

Source: The Opportunity Atlas. Chetty, Friedman, Hendren, Jones, Porter 2018