1. Introduction and Motivation for the Course

- The American dream is a multifaceted concept that means different things to different people, but one central aspect of it is the idea that America is a country where if you work hard you'll have a chance of moving up in the income distribution relative to your parents.
- Are we actually living up to the ideal that America is a place where kids can rise up relative to their parents?
- To answer this question, we first compute a simple statistic: the fraction of children who go on to earn more than their parents did, measuring both the children's incomes and their parents' incomes in their mid-30s.
- We plot the statistic separately by the year in which the child was born, which is shown along the x-axis in the figure below. The point at the far left of the figure shows the data for children born in the 1940s. The point at the far right of the figure shows the data for children born in the 1980s, who are in their 30s today, when we are measuring their incomes.
- The figure shows that for children born in the middle of the last century in America, it was a virtual guarantee that they would achieve the American dream: 92% of children born in 1940 went on to earn more than their parents did. If you look at what has happened over time, you see a dramatic fading of the American dream.
This broad trend, shown in the figure above, is of great interest to economists in understanding why the economy has changed so much from the time that America was a place where we expected one generation to be more prosperous than the previous one.

2. The Rise of Data and Empirical Evidence

- Until recently, social scientists have had limited data to study big-picture policy questions. As a result, social science has primarily been a theoretical field.
- The problem with that purely theoretical approach, is that when your theories are untested, you often end up with the politicization of questions that, in principle, are fundamentally scientific questions with scientific answers.
- Social science today is becoming a much more empirical field thanks, fundamentally, to the growing availability of data. We can test and improve the theories that we've developed in the past using real-world data, and we can take an approach that's more analogous to the natural sciences, where we're running experiments or quasi-experiments to figure out what works, and making decisions on the basis of data, rather than speculation.

3. Social Science in the Age of Big Data

- The recent availability of big data has further accelerated that trend towards empirical research. Examples of data sources used in social science today include:
  - Government data. This includes information like tax records from the Internal Revenue Service (IRS) and health records from the Medicare system cover essentially the entire U.S. population. These data consist of huge samples of information that allow social scientists to study research questions in ways that were unprecedented in the past.
  - Corporate data. This includes datasets like Google search queries, data from Uber on where people are driving or where people are asking for cars, and scanner data from grocery retailers consisting of every product that people purchase.
  - Unstructured data. Examples include text from Twitter feeds or newspapers. They're in the form of text, but there's a tremendous amount of information there.

4. Why is Big Data Transforming Social Science?

- Traditionally, the approach to collect information from people was to survey them and ask them questions. The problem with those surveys is that both they are very expensive to field, and as a result, they're quite small, but also, they tend not to be very reliable. Data from administrative sources tend to be more reliable.
- There is also an ability to measure new variables that would be hard to quantify, like people's emotions or political beliefs. If you ask people about those things directly, it is hard to elicit truthful responses in a very clear way. In contrast, a big data approach might be to use Facebook likes or information from Twitter feeds to understand how people feel about different things in a scalable way.
- Large, administrative datasets often enable universal coverage of every person in the United States, or all people in the population you're interested in. Such large samples allow us to approximate scientific experiments.
5. Why This Course?

- The same types of skills that are used to solve private market issues using big data can be put to great use in tackling challenges like growing inequality and climate change.
- In order to achieve that goal, the idea of this class is to introduce a broad range of topics, methods, and real-world applications of these sorts of ideas. Fundamentally, we want to start from the questions that motivate the methods we teach in economics and social science, rather than the traditional approach, which is to do the reverse.

6. Overview of Course Topics

- In this class, we will cover eight topics: Equality of Opportunity, Education, Racial Disparities, Health, Criminal Justice, Climate Change, Tax Policy, Economic Development and Institutional Change

7. Statistical Methods You Will Learn in This Class

- We are going to take a topic-oriented focus in this class. In the context of those topics, we're going to introduce these methods that will be useful for studying these topics.
  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

8. Two Types of “Big Data”

- Big data is kind of a buzzword that everybody likes to talk about, but what does it actually mean?
- It is helpful to classify big data into two different types.
- The first is what you might think of as a long dataset, a dataset where there are many observations and only a few variables on each of those observations.
- The figure below is intended to give you a sense of the typical structure of a long dataset that people often work with in social science. You could imagine you have information on all these different people, how much they're earning, how many years of education they have (e.g., did they go to college), and their gender.
In contrast, often what people in the private sector mean when they talk about big data is what I would call a wide dataset. In a wide dataset, there are few observations relative to the number of variables on each of those observations.

In statistics and computer science, the focus is typically on these wide datasets. The main application that they focus on is prediction. Social scientists typically focus on long datasets because we are primarily interested in identifying causal effects. We don't just want to predict what's going to happen, we want to be able to change what is going to happen.

9. Economic Concepts You Will Learn in this Class

- Examples of economic concepts you will learn in this class include:
  1. Effects of price incentives
  2. Supply and demand
  3. Competitive equilibrium
  4. Adverse selection
  5. Behavioral economics vs. rational models

10. Empirical Projects

- An important component of the class is four empirical projects where you will get hands-on experience with some of the core methods in the class. The projects are listed below.
  1. Stories from the Atlas: Describing Data using Maps, Regressions, and Correlations
  2. Do Smaller Classes Improve Test Scores? Evidence from a Regression Discontinuity Design
  3. The Creating Moves to Opportunity (CMTO) Experiment
  4. Using Google DataCommons to Predict Social Mobility
• Each of the projects focuses on real-world questions and different core methods that I hope you learn in the class.

11. Geographical Variation in Upward Mobility

• This first lecture is based primarily on a recent paper we put out just a couple of months ago called, “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.”
• The question we want to ask first is how do children's chances of moving up vary across areas in America?

12. Description of the Data used to Construct the Opportunity Atlas

• How are we actually going to measure upward mobility separately by geographic area in the United States? Let me start by describing how we do this in the paper, “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.”
• We take data from the 2000 and 2010 Censuses, and we link that to information from federal income tax returns. We use tax return data from 1989 to 2015. Linking those datasets yields information on essentially every American between 1989 and 2015, including how much they are earning, where they live, the dependents they have, and other information, year by year.
• In that dataset, we want to study economic opportunity across generations. In order to link parents to their children, we use information from dependency claims on tax returns. (In order to receive a tax deduction, parents must enter their child’s Social Security Number on their tax returns.) We’re able to use this information to link 99% of kids in America back to their parents, thereby generating an intergenerational sample where you can study income inequality and mobility across generations.
• This 8-billion row dataset ends up covering 20.5 million children born between 1978 and 1983, representing 96% of our target population. We analyze children born during those particular years because we need the children to be old enough that we can measure their earnings reliably.
• We’re interested in people who were born in the U.S. or are authorized immigrants who came to the U.S. in childhood. We look at authorized immigrants because these datasets don’t go a great job of covering undocumented immigrants.
• There are some kids who you can’t link to their parents and people you can’t link the census form to the tax form.

13. Measuring Parents’ and Children’s Incomes in Tax Data

• We measure incomes using information from the anonymized tax return data. For parents, we use average income between 1994 and 2000 reported on Form 1040, the main tax return in the U.S. Similarly, for kids, we measure average income in 2014 and 2015, the last two years of the data we were working with. That is when the children are in their mid-30s.
• Using this information, we’re going to focus on percentile ranks in the national distribution. What that means, concretely, is that we’re going to rank kids relative to all the other kids born in the same year, and parents relative to all other parents.
• We’re comparing kids to other kids of the same age. Then likewise, we compare parents to other parents. We do that because we want to adjust for the fact that as people grow older, their incomes tend to rise.

• The chart below was constructed using data for kids who were raised in the Chicago metro area, which consists of Chicago and the surrounding suburbs.

![Intergenerational Income Mobility for Children Raised in Chicago](source: Chetty, Hendren, Kline, Saez 2014)

On the x-axis, we’re showing the parent rank in the national income distribution. There are a hundred dots here, one corresponding to each percentile of the distribution. Then in each of those hundred bins, we’re plotting the average ranking of the child in the national income distribution.

Now as you go to the right, you’re looking at kids from richer and richer families, and you see that there’s a very strong upward-sloping pattern. This reflects the simple fact that if you were born to a richer family in America, you yourself tend to be richer in adulthood.

I’m going to find the line that fits that data most accurately using a method called regression. Then I’m going to focus on the value of this line, called the predicted value, at the 25th percentile of the parent income distribution. This allows me to essentially construct a digestible single statistic to summarize what upward mobility looks like in each place.

In Chicago, on average, kids who start out in families at the 25th percentile end up at the 40th percentile. Kids growing up in low-income families in Chicago, roughly speaking, earn about $30,000, on average, when they’re adults.

We can’t directly use the value of the dot on the above chart at the 25th percentile. Instead we use a regression line. This is because there is noise and random variation in the data, specifically with smaller samples of people.

When working with small samples, it starts to become very important to fit that regression line—in other words, to use the discipline of a statistical model. That’s the core idea of statistical models, to take the underlying data and represent it in a way that is more stable.
The conversion to percentiles is very important here. If we did this analysis in dollars, that relationship is very far from linear. It is very curved, which makes it harder to fit systematically with a statistical model.

- To construct the Opportunity Atlas, we fit line like this to the kids who grew up in every different census tract in America.

- A Census tract is a small definition of a neighborhood that the Census Bureau has created. There are 70,000 Census tracts in America, each of which has about 4,200 people.

- In order to handle children who might have moved while they were kids, we weigh children by the fraction of their childhood that they spent in each area.

14. Geographic Variation in Upward Mobility by Commuting Zone

- The map below plots average household earnings of children who grew up in low-income families. The map presents this statistic separately for each of the 741 commuting zones (CZs) in the United States. CZs are aggregations of counties based on commuting patterns that are similar to metro areas but cover the entire United States.

- Note that the map shows household income in dollars, but the underlying statistic is based on the predicted percentile rank defined earlier. The ranks have been converted to dollars because it’s more intuitive and concrete.

![Map of Geographic Variation in Upward Mobility](image)

- In the map, blue colors depict areas with high levels of upward mobility and red colors depict areas with low levels of upward mobility. The map shows broad geographic variation. One of the most interesting features of this map is that the highest upward mobility areas in America are the Great Plains, the rural parts of the country in the center of the country.

- Charlotte is one of the cities in America with the highest rates of job growth in the United States. Yet, somehow remarkably, for low-income kids who grow up in Charlotte, they do not have very good chances of moving up.
The map shows that in the current generation, there are some parts of America where kids' chances of moving up still look fantastic—actually better than any other country in the world. Then there's some places, like in much of the industrial Midwest, where your odds of climbing up look worse than any country for which we currently have data. America is a land of tremendous variability in opportunity.

This map shows nominal incomes, meaning it does not adjust for differences in cost of living. You can redraw this map, adjusting for differences in cost of living. When you do that, you get a map that looks almost identical to the one that I'm showing you here. To put it more precisely, the correlation between that data and these data is .9, meaning that it looks essentially the same.

We're focusing specifically here on kids growing up in low-income families. If you look at kids growing up in middle-class families, it's broadly similar. If you look at kids growing up in high-income families, you see that there's significantly less variation across areas for kids growing up in very-high-income families.

15. Local Area Variation in Upward Mobility: Los Angeles, CA

This geographic variation in upward mobility is not just about broad regional variation, but it's actually about extremely local variation. We can use the Opportunity Atlas to visualize the data.

The Opportunity Atlas starts out with the national map of the same statistics by commuting zone that we were looking at before. However, it allows us to zoom in to areas of specific interest.

Let us focus on one particular example: Nickerson Gardens in Los Angeles, CA, which is a public housing project in Watts. Let's look at black men growing up in the lowest-income families in the bottom 1% of the income distribution, which is actually representative of the incomes of the families living in this public housing project. The average household income of black men who grew up in the poorest families in Watts is just $3,300 a year. It has to be the case that lots of people are basically not working at all. You can see that in a very direct way in these data because we're able to look not just at income, but a variety of other outcomes, including incarceration.

Focusing on incarceration rates, you will see a really shocking and disturbing statistic about the United States, and this area in particular, which is that 44% of the black men who grew up in these lowest-income families are incarcerated on a single day, the date of the 2010 census.

If you go down to Compton, you see incarceration rates of 6.2%, which is a factor of 10 smaller than the 44% that we were seeing in Watts for black men growing up in low-income families. Compton is a different neighborhood than Watts, it’s not exactly the same, but I don’t think anybody from L.A. would have predicted that Compton would have drastically different outcomes like this from Watts.

That shows you that you can go two miles away and just have a dramatically different picture in terms of what kids' life trajectories look like. We see that in the stark example here within Los Angeles, but we see that sort of thing more broadly across the United States.