Using Big Data To Solve Economic and Social Problems

Raj Chetty
The American Dream?

- Chance that a child born to parents in the bottom fifth of the income distribution reaches the top fifth:
The American Dream?

- Chance that a child born to parents in the bottom fifth of the income distribution reaches the top fifth:

  - **USA**: Chetty, Hendren, Kline, Saez 2014, 7.5%
  - **UK**: Blanden and Machin 2008, 9.0%
  - **Denmark**: Boserup, Kopczuk, and Kreiner 2013, 11.7%
  - **Canada**: Corak and Heisz 1999, 13.5%
The American Dream?

- Chance that a child born to parents in the bottom fifth of the income distribution reaches the top fifth:
  - USA: 7.5% (Chetty, Hendren, Kline, Saez 2014)
  - UK: 9.0% (Blanden and Machin 2008)
  - Denmark: 11.7% (Boserup, Kopczuk, and Kreiner 2013)
  - Canada: 13.5% (Corak and Heisz 1999)

→ Chances of achieving the “American Dream” are almost two times higher in Canada than in the U.S.
Why is Upward Mobility Lower in America?

- Central policy question: why are children’s chances of escaping poverty so low in America?
  - And what can we do to improve their odds…?

- Difficult to answer this question based solely on country-level data
  - Numerous differences across countries makes 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 $\rightarrow$ 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)
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: Facebook, 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?

- Silicon Valley has been very successful in solving private-sector problems using technology and big data.

- Goal of this course: show how same skills can be used to address important social and economic 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.
Overview of Topics

1. Equality of Opportunity
2. Education
3. Health
4. Environment
5. Criminal Justice and Discrimination
6. Political Polarization
Overview of Methods

1. Descriptive Data Analysis
2. Experiments
3. Quasi-Experiments
4. Machine Learning
5. Stata programming
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)
<table>
<thead>
<tr>
<th>person_id</th>
<th>income</th>
<th>years of education</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>$8,820,23</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>102</td>
<td>$8,385,11</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>103</td>
<td>$8,641,73</td>
<td></td>
<td>13 F</td>
</tr>
<tr>
<td>104</td>
<td>$10,024,09</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>105</td>
<td>$79,923,36</td>
<td></td>
<td>12 M</td>
</tr>
<tr>
<td>106</td>
<td>$57,007,00</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>107</td>
<td>$59,494,84</td>
<td></td>
<td>15 F</td>
</tr>
<tr>
<td>108</td>
<td>$52,159,41</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>109</td>
<td>$75,373,30</td>
<td></td>
<td>13 F</td>
</tr>
<tr>
<td>110</td>
<td>$15,680,30</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>111</td>
<td>$46,591,41</td>
<td></td>
<td>13 F</td>
</tr>
<tr>
<td>112</td>
<td>$71,396,71</td>
<td></td>
<td>15 M</td>
</tr>
<tr>
<td>113</td>
<td>$72,674,96</td>
<td></td>
<td>11 M</td>
</tr>
<tr>
<td>114</td>
<td>$58,335,12</td>
<td></td>
<td>12 M</td>
</tr>
<tr>
<td>115</td>
<td>$33,908,91</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>116</td>
<td>$99,201,27</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>117</td>
<td>$46,181,11</td>
<td></td>
<td>11 F</td>
</tr>
<tr>
<td>118</td>
<td>$74,175,59</td>
<td></td>
<td>15 M</td>
</tr>
<tr>
<td>119</td>
<td>$77,409,86</td>
<td></td>
<td>11 F</td>
</tr>
<tr>
<td>120</td>
<td>$65,784,26</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>121</td>
<td>$3,332,26</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>122</td>
<td>$33,836,95</td>
<td></td>
<td>15 M</td>
</tr>
<tr>
<td>123</td>
<td>$56,806,58</td>
<td></td>
<td>15 F</td>
</tr>
<tr>
<td>124</td>
<td>$66,478,31</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>125</td>
<td>$60,566,22</td>
<td></td>
<td>15 F</td>
</tr>
<tr>
<td>126</td>
<td>$59,447,41</td>
<td></td>
<td>13 F</td>
</tr>
<tr>
<td>127</td>
<td>$79,397,90</td>
<td></td>
<td>11 F</td>
</tr>
<tr>
<td>128</td>
<td>$17,594,75</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>129</td>
<td>$84,067,93</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>130</td>
<td>$87,953,71</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>131</td>
<td>$88,423,74</td>
<td></td>
<td>14 F</td>
</tr>
<tr>
<td>132</td>
<td>$31,357,62</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>133</td>
<td>$82,233,86</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>134</td>
<td>$92,901,31</td>
<td></td>
<td>14 M</td>
</tr>
<tr>
<td>135</td>
<td>$75,153,35</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>136</td>
<td>$29,740,94</td>
<td></td>
<td>15 M</td>
</tr>
<tr>
<td>137</td>
<td>$795,20</td>
<td></td>
<td>15 P</td>
</tr>
<tr>
<td>138</td>
<td>$27,283,46</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>139</td>
<td>$1,137,37</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>140</td>
<td>$61,127,80</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>141</td>
<td>$31,153,06</td>
<td></td>
<td>12 F</td>
</tr>
<tr>
<td>142</td>
<td>$19,774,73</td>
<td></td>
<td>15 M</td>
</tr>
<tr>
<td>143</td>
<td>$55,925,97</td>
<td></td>
<td>13 M</td>
</tr>
<tr>
<td>144</td>
<td>$11,346,84</td>
<td></td>
<td>14 F</td>
</tr>
</tbody>
</table>
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)
<table>
<thead>
<tr>
<th>years of education</th>
<th>gender</th>
<th>ad_click1</th>
<th>ad_click2</th>
<th>ad_click3</th>
<th>ad_click4</th>
<th>ad_click5</th>
<th>ad_click6</th>
<th>ad_click7</th>
<th>ad_click8</th>
<th>ad_click9</th>
<th>ad_click10</th>
<th>ad_click11</th>
<th>ad_click12</th>
<th>ad_click13</th>
<th>ad_click14</th>
<th>ad_click15</th>
<th>ad_click16</th>
<th>ad_click17</th>
<th>ad_click18</th>
<th>ad_click19</th>
<th>ad_click20</th>
<th>ad_click21</th>
<th>ad_click22</th>
<th>ad_click23</th>
<th>ad_click24</th>
<th>ad_click25</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 F</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>12 M</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12 M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12 M</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11 F</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15 M</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14 F</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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
Lecture 1: Equality of Opportunity

1. Local Area Differences in Upward Mobility within America

2. Geographical Variation: Causal Effects of Places or Sorting?

3. Characteristics of Low vs. High Mobility Areas

- Lecture 1 is based primarily on two papers:
  
  Chetty, Hendren, Kline, Saez. “Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the U.S.” QJE 2014

  Chetty and Hendren. “The Effects of Neighborhoods on Children’s Long-Term Outcomes” 2017a, b
Part 1
Local Area Variation
Differences in Opportunity Across Local Areas

- Chetty et al. (2014) use “big data” to measure upward mobility for every metro and rural area in the U.S.
  - De-identified tax records on all children born in America between 1980-1982 (10 million children)

- Classify children into locations based on where they grew up

- Rank children in *national* income distribution (not local distribution) when computing rates of upward mobility
The Geography of Upward Mobility in the United States
Chances of Reaching the Top Fifth Starting from the Bottom Fifth by Metro Area

San Jose 12.9%
Salt Lake City 10.8%
Atlanta 4.5%
Washington DC 11.0%
New York City 10.5%
Chicago 6.5%
Minneapolis 8.5%
Charlotte 4.4%

Note: Lighter Color = More Upward Mobility
Download Statistics for Your Area at www.equality-of-opportunity.org
The Geography of Upward Mobility in the Bay Area
Chances of Reaching the Top Fifth Starting from the Bottom Fifth by County

San Francisco 18.5%
San Mateo 17.4%
Santa Clara 17.7%
Alameda (Oakland) 11.4%

Lighter Color = More Upward Mobility
Download Statistics for Your Area at www.equality-of-opportunity.org
The Geography of Upward Mobility in the New York Area
Chances of Reaching the Top Fifth Starting from the Bottom Fifth by County

Brooklyn 10.6%
Bronx 7.3%
Manhattan 9.9%
Queens 16.8%
Suffolk 16.0%
Ulster 10.6%
New Haven 9.3%
Monroe 14.1%
Ocean 15.1%

Part 2

Causal Effects of Neighborhoods
Causal Effects of Neighborhoods vs. Sorting

- Two very different explanations for variation in children’s outcomes across areas:
  1. Sorting: different people live in different places
  2. Causal effects: places have a \textit{causal} effect on upward mobility for a given person
Identifying Causal Effects of Neighborhoods

- Ideal experiment: randomly assign children to neighborhoods and compare outcomes in adulthood

- We approximate this experiment using a quasi-experimental design
  - Study 7 million families who move across counties in observational data
  - Key idea: exploit variation in age of child when family moves to identify causal effects of environment

Source: Chetty and Hendren 2017
Oakland ($30,000)

Earnings Gain from Moving to a Better Neighborhood
Earnings Gain from Moving to a Better Neighborhood

San Francisco ($40,000)

Oakland ($30,000)
Earnings Gain from Moving to a Better Neighborhood

San Francisco ($40,000)

Oakland ($30,000)

Move at age 9 → 54% of gain from
growing up in San Francisco since birth
Earnings Gain from Moving to a Better Neighborhood

- San Francisco ($40,000)
- Oakland ($30,000)

Gain from Moving to a Better Area vs. Age of Child when Parents Move
Earnings Gain from Moving to a Better Neighborhood

San Francisco ($40,000)

Oakland ($30,000)

Age of Child when Parents Move

Gain from Moving to a Better Area
Identifying Causal Effects of Neighborhoods

- Key assumption: *timing* of moves to a better/worse area unrelated to other determinants of child’s outcomes

- This assumption might not hold for two reasons:
  1. Parents who move to good areas when their children are young might be different from those who move later
  2. Moving may be related to other factors (e.g., change in parents’ job) that affect children directly
Identifying Causal Effects of Neighborhoods

Two approaches to evaluating validity of this assumption:

1. Compare siblings’ outcomes to control for family effects
Identifying Causal Effects of Neighborhoods

Two approaches to evaluating validity of this assumption:

1. Compare siblings’ outcomes to control for family effects

2. Use differences in neighborhood effects across subgroups to implement “placebo” tests
   - Ex: some places (e.g., low-crime areas) have better outcomes for boys than girls
   - Move to a place where boys have high earnings → son improves in proportion to exposure but daughter does not
Causal Effects of Neighborhoods: Summary

- Key lesson of this section: 70-80% of the variation in children’s outcomes across areas is due to *causal effects*

- This result has refocused public discussion on improving upward mobility in America to a local level
An Atlas of Upward Mobility Shows Paths Out of Poverty

By DAVID LEONHARDE, AMANDA COX and CLAIRE CAIN MILLER  MAY 4, 2015

In the wake of the Los Angeles riots more than 20 years ago, Congress created an anti-poverty experiment called Moving to Opportunity. It gave vouchers to help poor families move to better neighborhoods and awarded them on a random basis, so researchers could study the effects.

The results were deeply disappointing. Parents who received the vouchers did not seem to earn more in later years than otherwise similar adults, and children did not seem to do better in school. The program’s apparent failure has haunted social scientists and policy makers, making poverty seem all
San Francisco County is about average for income mobility for children in poor families. It is better than about 42 percent of counties.

Location matters — enormously. If you’re poor and live in the San Francisco area, it’s better to be in Contra Costa County than in San Francisco County or Alameda County. Not only that, the younger you are when you move to Contra Costa, the better you will do on average. Children who move at earlier ages are less likely to become single parents, more likely to go to college and more likely to earn more.
A Wake-Up Call for Charlotte-Mecklenburg

Land of opportunity? Not by a long shot

Charlotte is nation’s worst big city for climbing out of poverty

Over the last several decades, Charlotte-Mecklenburg has transformed from a small southern town to one of the country’s largest and most dynamic communities. We continue to attract people—nearly 50 a day—who move here to take advantage of our strong business climate, favorable weather and geographic location, and our reputation as a great place to live and raise a family. Accolades from the outside regularly tell us how tall we stand among other communities. As recently as February 7, 2017, U.S. News and World Report ranked us as the 14th best place to live in the country.1

Yet, in 2013 when the headline broke about the Harvard University/UC Berkeley study that ranked Charlotte-Mecklenburg 50th out of 50 in upward mobility2 for children born into our lowest income quintile, many in our community responded with disbelief. How, on the one hand, can we be such a vital and opportunity-rich community, and on the other, be ranked dead last in the odds that our lowest income children and youth will be able to move up the economic ladder as they become adults?
Part 3
Characteristics of High-Mobility Areas
Why Does Upward Mobility Differ Across Areas?

- Why do some places produce much better outcomes for disadvantaged children than others?

- Begin by characterizing the features of areas with high rates of upward mobility
Five Strongest Correlates of Upward Mobility

1. Segregation
   - Greater racial and income segregation associated with lower levels of mobility
Racial Segregation in Atlanta
Whites (blue), Blacks (green), Asians (red), Hispanics (orange)

Source: Cable (2013) based on Census 2010 data
Racial Segregation in Sacramento
Whites (blue), Blacks (green), Asians (red), Hispanics (orange)

Source: Cable (2013) based on Census 2010 data
Five Strongest Correlates of Upward Mobility

1. Segregation

2. Income Inequality
   - Places with smaller middle class have much less mobility
Five Strongest Correlates of Upward Mobility

1. Segregation

2. Income Inequality

3. School Quality
   - Higher expenditure, smaller classes, higher test scores correlated with more mobility
Five Strongest Correlates of Upward Mobility

1. Segregation

2. Income Inequality

3. School Quality

4. Family Structure
   - Areas with more single parents have much lower mobility
   - Strong correlation even for kids whose own parents are married
Five Strongest Correlates of Upward Mobility

1. Segregation
2. Income Inequality
3. School Quality
4. Family Structure
5. Social Capital

- “It takes a village to raise a child”