Big Data and Big Cities

Ed Glaeser
(joint work with Nikhil Naik, Mike Luca and Scott Kominers)
Change in Housing Prices, 2001-2006 vs. 2006-2011

- Detroit
- Houston
- Las Vegas
- New York
- Phoenix
- DC

-1  -.75  -.5  -.25  0  .25

Change in FHFA Price, 2001-2006

Change in Housing Prices, 2001-2006 vs. 2006-2011

Change in FHFA Price, 2006-2011

0  .2  .4  .6  .8

Change in FHFA Price, 2001-2006
Change in FHFA, 1996-2012
by Quintile of Population Density, 2010

Note: For MSAs with populations greater than 250,000 in 2010.
Average Population Change, 2000-2010

Average Per Capita Income, 2000

Population Density
Per Capita Income, 2000
Population Change, 2000-2010

Source: U.S. Census
Will the last person to leave Seattle (and Milan) please turn out the lights?

Photo by Postdil

Artist’s Impression by Daniel Libeskind Studio
Chinitz: Contrasts in Agglomeration: New York and Pittsburgh
Economic Growth and Firm Size

MSA Employment Growth (1977-2010) by Average Firm Size (1977) Quintiles

Smallest firms are in Quintile 1

- Quintile 1
- Quintile 2
- Quintile 3
- Quintile 4
- Quintile 5

Average Percent Growth in Employment, 1977-2010

1 2 3 4 5
Rich and Poor Innovation Districts
## Local Regulation: Chamber of Commerce Red Tape (Higher is Less)

<table>
<thead>
<tr>
<th>City</th>
<th>Starting Business</th>
<th>Dealing with Construction Permits</th>
<th>Registering Property</th>
<th>Paying Taxes</th>
<th>Enforcing Contracts</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dallas</td>
<td>80</td>
<td>92</td>
<td>98</td>
<td>100</td>
<td>77</td>
<td>89.5</td>
</tr>
<tr>
<td>St. Louis</td>
<td>98</td>
<td>90</td>
<td>100</td>
<td>56</td>
<td>83</td>
<td>85.2</td>
</tr>
<tr>
<td>Raleigh</td>
<td>85</td>
<td>99</td>
<td>76</td>
<td>63</td>
<td>45</td>
<td>73.7</td>
</tr>
<tr>
<td>Boston</td>
<td>75</td>
<td>86</td>
<td>74</td>
<td>50</td>
<td>81</td>
<td>73.3</td>
</tr>
<tr>
<td>Atlanta</td>
<td>86</td>
<td>65</td>
<td>77</td>
<td>58</td>
<td>78</td>
<td>72.7</td>
</tr>
<tr>
<td>Detroit</td>
<td>98</td>
<td>60</td>
<td>51</td>
<td>64</td>
<td>53</td>
<td>64.9</td>
</tr>
<tr>
<td>Chicago</td>
<td>11</td>
<td>61</td>
<td>71</td>
<td>38</td>
<td>83</td>
<td>52.9</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>99</td>
<td>53</td>
<td>52</td>
<td>0</td>
<td>35</td>
<td>47.9</td>
</tr>
<tr>
<td>San Francisco</td>
<td>99</td>
<td>0</td>
<td>51</td>
<td>10</td>
<td>46</td>
<td>41.3</td>
</tr>
<tr>
<td>New York City</td>
<td>29</td>
<td>72</td>
<td>0</td>
<td>53</td>
<td>20</td>
<td>34.7</td>
</tr>
</tbody>
</table>
Subjective Well-Being Across Space

City and Rural Area Happiness Controlling for Characteristics

Area happiness

- 0.06 - 0.13
- 0.03 - 0.05
- 0.01 - 0.02
- 0.01 - 0.00
- 0.01 - 0.00
- 0.03 - 0.02
- 0.05 - 0.04
- 0.05 - 0.06
Subjective Well-Being and Population Growth

Happiness after exogenous demographic controls, 2005-2010

Change in Log Population, 1950-2000
Using Big Data To Solve City Problems

• The Economic Agenda: Education and Entrepreneurship, and Incomes.
• The Demons of Density: Contagious Disease, Crime, Congestion and High Housing Prices.
• The Forms of Big Data
  – Much finer geographic records (the IRS data)
  – Similar data from private providers (corelogic)
  – Novel data sets on traditional outcomes (Zoona)
  – Novel data sets on relatively new things (Yelp)
  – Completely different data on things we had barely thought about before (Google Streetview)
What’s It Good For

• Big data does not intrinsically solve any of the causal inference issues that we have long worried about.

• It does make it possible to measure more things (hygiene, streetscapes) in more places in more ways.

• IRS records provide the mother-of-all-panel sets, which is particularly useful for spatial interventions
  – The right way to judge empowerment zones, for example, would be to use the panel structure
Led Astray By “Bigger” Data (.3)

Figure 1: Per Capita Payroll and Density Across New York City Zip Codes
Big Data and Education in the US

• Early Childhood Interventions (Heckman)

• Teacher Quality (Chetty, Friedman, Rockoff)

• Charter Schools (Angrist, Pathak, Walters)

• Science and Math (Joshua Goodman)
Impact of High Value-Added Teacher Entry on Cohort Test Scores

\[ \Delta \text{Score} = 0.036 \]
\[ \Delta \text{TVA} = 0.038 \]

\[ p [\Delta \text{score} = 0] < 0.001 \]
\[ p [\Delta \text{score} = \Delta \text{TVA}] = 0.76 \]

Number of Events = 1692

Year Relative to Entry of High Value-Added Teacher

Score in Current Grade

Score in Previous Grade
New York City’s Department of Health shows the timeline of the city’s mortality rate, which sharply dropped with the provision of clean water in the nineteenth century.

New York City Department of Health and Mental Hygiene
Zoonia in Zambia
Table 2: Effect of Water Supply Complaints on Diarrheal Disease

<table>
<thead>
<tr>
<th>Panel A: Any Water Supply Complaints</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable:</strong></td>
</tr>
<tr>
<td><strong>Age:</strong></td>
</tr>
<tr>
<td>Any Complaint</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Mean of DV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Day-Weighted Supply Complaints</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variable:</strong></td>
</tr>
<tr>
<td><strong>Age:</strong></td>
</tr>
<tr>
<td>ln(# of Complaints)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Mean of DV</td>
</tr>
<tr>
<td>District FEs</td>
</tr>
<tr>
<td>Week FEs</td>
</tr>
</tbody>
</table>

Notes: ***indicates significance at 1% level, ** at 5% level, * at 10% level.
**Table 3: Effect of Water Supply Complaints on Zoono Transactions**

### Panel A: Transactions Payed Out

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Number of Transactions</th>
<th>Transaction Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day-Weighted Complaints</td>
<td>-.03</td>
<td>-7.92</td>
</tr>
<tr>
<td></td>
<td>(.01)***</td>
<td>(3.28)**</td>
</tr>
<tr>
<td>Observations</td>
<td>16800</td>
<td>16800</td>
</tr>
<tr>
<td>Mean DV</td>
<td>34.5</td>
<td>9,983</td>
</tr>
</tbody>
</table>

### Panel B: Transactions Received

<table>
<thead>
<tr>
<th>Dep. Var.</th>
<th>Number of Transactions</th>
<th>Transaction Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day-Weighted Complaints</td>
<td>-.03</td>
<td>-9.08</td>
</tr>
<tr>
<td></td>
<td>(.01)**</td>
<td>(3.82)**</td>
</tr>
<tr>
<td>Observations</td>
<td>16800</td>
<td>16800</td>
</tr>
<tr>
<td>Mean DV</td>
<td>32.3</td>
<td>10,858</td>
</tr>
<tr>
<td>District FEs</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Week FEs</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Notes: ***indicates significance at 1% level, ** at 5% level, * at 10% level.
Restaurant hygiene inspections

• Data and technology have changed
  – Policy has remained the same
• Disclosure side
  – Market with very little information
  – Early success story of disclosure (Jin and Leslie 2003), so known potential impact
• Ideal setting for information design questions
  – What conditions cause posting to work?
  – What are the behavioral factors underlying customer response?
• Scope for improving policy
  – Dai and Luca 2016
Tournaments and Hygiene Inspections

• Process and scoring varies (sometimes a lot) by city
• In SF:
  – restaurants inspected roughly 2X per year.
  – violations classified as major (lots of rats) and minor (a rat)
  – final score between 0 and 100
• In Boston:
  – Restaurants inspected at least once per year
  – Violations classified as minor, major, and severe
  – Until now, no grades
• Goal:
  – Identify risks
  – Shut down worst offenders, enforce clean up
Essentially a prediction problem

• *Which* restaurant is most likely to have a violation?

• By targeting inspections, can be more efficient:
  – Identify more risks, or,
  – Reduce number of inspections

• Eg: 1 random annual inspection for each restaurant, plus targeted
Tournament:

• Cosponsored with Yelp
• Supported by City of Boston
• Combined Yelp data with Boston inspection results:
  – Objective to predict violations.
  – Weights chosen by city (minor = 1, major = 2, severe = 5).
  – Evaluated using RMSLE
Yelp Ratings Predict Hygiene Scores

![Graph showing the relationship between Average Yelp Rating and Hygiene Score, with a trend line indicating a positive correlation.]
Tournament: Rewards

<table>
<thead>
<tr>
<th>Place</th>
<th>Prize Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>$3,000</td>
</tr>
<tr>
<td>2nd</td>
<td>$1,000</td>
</tr>
<tr>
<td>3rd</td>
<td>$1,000</td>
</tr>
</tbody>
</table>

Prize money provided by Yelp
Competition Process

Train your models on historical data and submit predictions for the Phase I leaderboard.

We release new data and you predict for a period 6 weeks into the future.

You must have your predictions for the next 6 weeks submitted by this date.

No new submissions are allowed. As inspectors report their findings, we compare your predictions against what actually happens.

Final winner is determined based on comparison of predictions to actual inspection results during the PHASE II evaluation period.
Results

• > 500 signups

• Development phase:
  – ~55 completed at least one entry
  – ~450 sets of predictions

• Evaluation phase:
  – 23 submitted final algorithms
  – During this time, Boston inspected 364 restaurants
Gains for Boston: ~40%

To catch 3,604 weighted violations, inspect this many restaurants:

- 7/7 - 8/18: 364
- 2nd Place: 249
- 1st Place: 219
Crime: NYC Homicides per 100,000

Figure 9: Homicides in New York City
Engineering vs. Economics
Detroit tried to reverse its decline with foolish investments like its People Mover, which here glides over essentially empty streets.

*Dennis MacDonald/ World of Stock*
The Curitiba Innovation

Picture by Mariordo
Bottom Up Innovation: Zipcar

Photo by Mario Roberto Duran Ortiz
The Physical City: NIMBYism vs. Monumentalism

Mumbai has recently begun building up, but the city is still short, expensive, and congested because of decades of overrestricting height. Scott Eels / Bloomberg / Getty Images

Astana by ChelseaFunNumberOne -
The great urbanist Jane Jacobs looks none too happy with the tall buildings surrounding her. She argued vigorously against such high-rises and in favor of a low-slung cityscape like that of New York’s Greenwich Village. Her arguments have not all proven correct.

*Bob Gomel/ Time & Life Pictures/Getty Images*
Training Sample – New York Income

Performance on the Training Sample

R^2 = 0.85
Testing Sample – New York Income

(b) Performance on the Testing Sample

$R^2 = 0.81$