Naturally-Occurring Data (Big data, non-designed data)

Data created in the normal course of activities of households, businesses, and governments

Contrast:
Designed data, e.g., surveys or experiments
Naturally-Occurring Data, Examples

Transactional and account data

*Households*
• Bank accounts, credit cards

*Business*
• Scanner data: price and quantity
• Payroll: employment and wages
Naturally-Occurring Data, Examples

Government:
• New claims for unemployment insurance
• Tax, Social Security records
• Benefits data
  – Food stamps, Social Security, Medicare
• Federal Contracts
• Tolls
• Airplane flight and load data
• Import duties
• Regulatory filings
Naturally-Occurring Data, Examples

Web scrapping:

• Prices
• Help wanted advertising
• Textual analysis/document contents
Naturally-Occurring Data, Examples

Social media:

• Twitter
• Facebook
• Craig’s list
• Linked-in
• Search activity
Naturally-Occurring Data, Examples

Other:
- Press reports
- Fitbit, etc
- Health club utilization
- Images
- Satellite data (ground cover, light)
- Property sales and assessment
- Box office sales
- U-Haul long-distance rentals
Conventional sources

Enumeration

• Census of population
• CPI price observations
Conventional sources

**Government surveys**
- Current Population Survey (unemployment)
- Census of manufacturing and other businesses
- Consumer Expenditure Survey

**Academic surveys**
- Health and Retirement Study
- Panel Study of Income Dynamics
Hybrid: Survey and Admin data

Government statistics
• Census survey use tax records for imputation
• GDP uses many commercial sources

Academic surveys
• Health and Retirement Study links Social Security and Medicare records
Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (A) and Argentina (B), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street’s PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of 1 July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (18), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP]
Regional Mortgage Conditions

Percentage of delinquent mortgages

View the percentage of mortgages by delinquency status according to county and over time.

2013 DEC

Percent in Foreclosure by Zip Code

Quintiles *

1.0% 3.3% 5.5% 10.8%

* quintiles established on Dec. 2013 data for the displayed region
Note: Regions with insufficient data are shown in gray.

Source: CoreLogic LoanPerformance, Lender Processing Services Mortgage Performance
Light and economic activity

Source: Chen and Nordhaus PNAS
Data from Financial App

“How Harnessing Naturally Occurring Data to Measure the Response of Spending to Income.” Science (July 2014)

Leading Methods: Surveys and Administrative Records

Surveys of individuals are comprehensive but ... 
• Self-reported
• Typically low frequency  
  – Long, varying, and staggered reporting intervals  
  – Infrequent reports  
  – Published with considerable lag

Administrative records are accurate, high frequency, and timely but ... 
• Not comprehensive  
• Large fractions of expenditure, portfolio, or income are missing
Current system gets it wrong exactly when it matters

Key components of GDP extrapolated

![Real GDP (2007Q4=1)](chart)

- Preliminary
Current system gets it wrong exactly when it matters

Key components of GDP extrapolated

→ Huge miss in fourth quarter of 2008
Financial App

• App for mobile phones, tablets, and the web
• Has registered more than 10 million registered users 2007
  – Pilot sample of 75,000
  – Now following 1,000,000+ users
• Users can integrate information from nearly any financial account with a web-based portal
• Users provide app with the credentials necessary to access these portals and, every day, app automatically logs into and scrapes the associated webpages
Some Challenges of Data

• No direct information on demographics
• Data are raw, not organized for research
• Spending is not pre-categorized
• Sample is not randomly selected
Who is in App?

Comparison of demographics

• App: Third-party data based on email
• ACS: American Community Survey from Census Bureau
<table>
<thead>
<tr>
<th>Age</th>
<th>App</th>
<th>ACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–20</td>
<td>0.6</td>
<td>5.7</td>
</tr>
<tr>
<td>21–24</td>
<td>5.3</td>
<td>7.4</td>
</tr>
<tr>
<td>25–34</td>
<td>37.9</td>
<td>17.5</td>
</tr>
<tr>
<td>35–44</td>
<td>30.1</td>
<td>17.0</td>
</tr>
<tr>
<td>45–54</td>
<td>15.0</td>
<td>18.4</td>
</tr>
<tr>
<td>55–64</td>
<td>7.8</td>
<td>16.1</td>
</tr>
<tr>
<td>65+</td>
<td>3.5</td>
<td>18.0</td>
</tr>
<tr>
<td>Gender</td>
<td>App</td>
<td>ACS</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Male</td>
<td>59.9</td>
<td>48.6</td>
</tr>
<tr>
<td>Female</td>
<td>40.1</td>
<td>51.4</td>
</tr>
</tbody>
</table>
### Education

<table>
<thead>
<tr>
<th></th>
<th>App</th>
<th>ACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than college</td>
<td>70.0</td>
<td>62.9</td>
</tr>
<tr>
<td>College</td>
<td>24.1</td>
<td>26.2</td>
</tr>
<tr>
<td>Graduate school</td>
<td>6.0</td>
<td>10.9</td>
</tr>
</tbody>
</table>
## Region

<table>
<thead>
<tr>
<th>Region</th>
<th>App</th>
<th>ACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>20.6</td>
<td>17.8</td>
</tr>
<tr>
<td>Midwest</td>
<td>14.6</td>
<td>21.5</td>
</tr>
<tr>
<td>South</td>
<td>36.7</td>
<td>37.4</td>
</tr>
<tr>
<td>West</td>
<td>28.1</td>
<td>23.4</td>
</tr>
</tbody>
</table>
## Transactions and accounts

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$P_5$</th>
<th>$P_{25}$</th>
<th>$P_{50}$</th>
<th>$P_{75}$</th>
<th>$P_{95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily transactions</td>
<td>4.54</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Credit card</td>
<td>1.23</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Checking account</td>
<td>3.03</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Saving account</td>
<td>0.22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Accounts</td>
<td>5.84</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Credit card</td>
<td>3.58</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Checking account</td>
<td>1.35</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Saving account</td>
<td>0.79</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Notes: In total, the 57,731,354 transactions are generated from 72,902 unique users over the study period.
## Account balance

### Panel (a): Bank

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$P_5$</th>
<th>$P_{25}$</th>
<th>$P_{50}$</th>
<th>$P_{75}$</th>
<th>$P_{95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>$14,415</td>
<td>$100</td>
<td>$700</td>
<td>$2,200</td>
<td>$7,900</td>
<td>$55,400</td>
</tr>
<tr>
<td>Checking</td>
<td>$6,969</td>
<td>$100</td>
<td>$500</td>
<td>$1,400</td>
<td>$3,800</td>
<td>$23,100</td>
</tr>
<tr>
<td>Saving</td>
<td>$6,476</td>
<td>$0</td>
<td>$0</td>
<td>$400</td>
<td>$2,500</td>
<td>$25,200</td>
</tr>
<tr>
<td>Money Market</td>
<td>$12,076</td>
<td>$0</td>
<td>$100</td>
<td>$900</td>
<td>$7,700</td>
<td>$57,400</td>
</tr>
<tr>
<td>C.D.</td>
<td>$12,734</td>
<td>$0</td>
<td>$0</td>
<td>$500</td>
<td>$4,000</td>
<td>$39,200</td>
</tr>
</tbody>
</table>

### Panel (b): Credit Card

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>$P_5$</th>
<th>$P_{25}$</th>
<th>$P_{50}$</th>
<th>$P_{75}$</th>
<th>$P_{95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance</td>
<td>$7,228</td>
<td>$200</td>
<td>$1,400</td>
<td>$3,600</td>
<td>$8,500</td>
<td>$26,100</td>
</tr>
<tr>
<td>Credit Limit</td>
<td>$23,019</td>
<td>$800</td>
<td>$4,200</td>
<td>$11,900</td>
<td>$29,500</td>
<td>$81,800</td>
</tr>
<tr>
<td>Utilization Ratio</td>
<td>0.48</td>
<td>0.02</td>
<td>0.15</td>
<td>0.45</td>
<td>0.78</td>
<td>1.00</td>
</tr>
<tr>
<td>Revolving Debt</td>
<td>$5,828</td>
<td>$1,200</td>
<td>$2,100</td>
<td>$3,500</td>
<td>$6,700</td>
<td>$18,000</td>
</tr>
<tr>
<td>APR</td>
<td>18.46%</td>
<td>10%</td>
<td>15%</td>
<td>18%</td>
<td>23%</td>
<td>27%</td>
</tr>
</tbody>
</table>
Measuring income and expenditures from transactions

**Total Income**

**Paycheck and Social Security**
Weekly Spending: Total

(a) Total (Full sample)

(b) Total (Linked users)
Weekly Spending: Non-Recurring

(c) Non-Recurring (Full sample)

(d) Non-Recurring (Linked users)
Weekly Spending: Coffee Shop and Fast Food

(e) Fast Food and Coffee Shop (Full sample)

(f) Fast Food and Coffee Shop (Linked users)
Using the data to answer macroeconomic questions

What is the MPC from predictable income, from unpredictable income?

- Does the standard lifecycle/permanent income fit the data?
- What MPC should be used to calibrate models?
- How will households respond to economic stimulus payments such as tax rebate?
Using the data to answer macroeconomic questions

Many individuals have very low liquidity.

- Are they vulnerable to shocks?
- How do they manage an adverse shock, e.g., a car repair or temporary income loss?
- What buffers consumption for those with low cash-on-hand?
- Is low cash-on-hand a mistake?
Using the data to answer macroeconomic questions

Application 1:
How does spending respond to **predictable, recurring income** (e.g., paycheck)?

Application 2:
How does spending respond to a **temporary shock to liquidity** (government shutdown)?
Application 1: Excess sensitivity to paycheck and Social Security

Data allow

• Identification of payments

• Estimate of response by
  – Type novel classification of spending
  – By liquidity, etc
Economic Model

Paycheck/Social Security predictable, so receipt of income should not affect timing of spending
Econometric model

\[ x_{ict} = \sum_{j=Mon.}^{Sun.} \delta_{jc} + \sum_{k=-7}^{6} \beta_{kc} I_i(Paid_{t-k}) + \varepsilon_{ict}, \]

\( x_{ict} \) = daily spending/average daily spending, 
\( i \) = individual, \( c \) = type of spending, \( t \) = time

\( I_i(Paid_{t-k}) \) = dummy for getting paid on date \( t - k \)
Response of spending to paycheck

Total

Non-recurring

Coffee shop and fast food
Response of spending to paycheck
Non-recurring spending by liquidity

Low liquidity | Medium liquidity | High liquidity

Fraction of daily average spending vs. Days since check arrival
Application 1. Conclusions

• Much of high-frequency “excess sensitivity” owes to rational timing of payments
• Low liquidity individuals do display some excess sensitivity
Application 2: 2013 US government shutdown

Workers subject to shutdown

- lost 40% of pay in one pay period
- reimbursed in next pay period

Distinctive experiment: *Timing* of income only

- Liquidity shock
- No wealth effects
# Government Shutdown of Oct 2013

<table>
<thead>
<tr>
<th>Pay period</th>
<th>Sunday</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 22</td>
<td></td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td></td>
<td></td>
<td>Oct 1 Shutdown Begins</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>First pay date affected by shutdown</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Shutdown Ends</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typical pay date after shutdown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aug 21 - 80 House Republicans sign letter urging defunding of Obamacare

Sep 30 - Senate rejects House bill with measures to defund healthcare

Oct 5 - Bill compensating furloughed employees passed

Sep 20 - House Republicans approve legislation to defund healthcare while keeping government open

Oct 1 - Shutdown begins

Oct 17 - Obama signs bill ending shutdown
Treatment and control

Treatment

• Federal worker (paycheck memo), and
• Decline in paycheck consistent with shutdown

Control

• Other worker on same biweekly pay schedule as government
Pre-Shutdown Median Liquidity over the Paycheck Cycle

![Graph showing liquidity ratio over days since paycheck for different liquidity levels: Low, Med, High.](image-url)
“Seasonal” interactions

• Day of week
  – Spending
  – Clearing of payments
• Beginning of pay period effects
• Seasonal/holiday/macro effects

→ Having controls with same pay schedule valuable
Average weekly spending
Diff-in-diff effects of shutdown

\[ y_{i,t} = \sum_{k=1}^{T} \delta_k \times Week_{i,k} + \sum_{k=1}^{T} \beta_k \times Week_{i,k} \times Shut_i + \Gamma'X_{i,t} + \epsilon_{i,t} \]

Specification:

LHS = variable of interest (income, category of spending)
Normalized by average individual spending (daily rate)
Diff-in-diff effects of shutdown

Paycheck income

Total spending
Diff-in-diff effects of shutdown

Non-recurring spending

Recurring spending
Diff-in-diff effects of shutdown

Credit card spending

Credit card balance payments
Diff-in-diff effects of shutdown

Coffee shop and fast food
Diff-in-diff effects of shutdown, Credit Card Balance by “Liquidity Risk”

Credit Card Balance, Accounts at Liquidity Risk

Credit Card Balance, Accounts Not at Liquidity Risk

Notes: Horizontal axis is days since August 2013
Revolvers only
Data are by individual credit card account levels
“At risk” accounts have payment due dates in pay period affected by shutdown
Diff-in-diff effects of shutdown, Credit Card Balance by “Liquidity Risk”

Credit Card Balance, Accounts at Liquidity Risk

Credit Card Balance, Accounts Not at Liquidity Risk

Notes: Horizontal axis is days since August 2013
Revolvers only
Data are by individual credit card account levels
“At risk” accounts have payment due dates in pay period affected by shutdown
Lessons from shutdown

Puzzle for standard models:
• Very sizeable spending response to a two-week delay in pay

Success for standard models:
• Rearrangement of payments, not consumption

New data essential