Lab 1

Data

Jan 24, 2013 – Michael Porath (@poezn)
Process
Process

Finding → Gathering → Transforming → Evaluating
Process

Finding → Gathering → Transforming → Evaluating
From (Idea To) Raw Data To Data
Finding data
The needle in the haystack

Open Government Data

- The World Bank
- DataSF
- DATA.gov
- OECD.StatExtracts
Finding data
The needle in the haystack

Open Government Data
- THE WORLD BANK
- DataSF
- DATA.gov
- OECD.Stat Extracts

Product APIs
- foursquare
- fitbit
- Developers
Finding data
The needle in the haystack

Open Government Data
- The World Bank
- DataSF
- DATA.gov
- OECD.StatExtracts

Product APIs
- foursquare
- fitbit
- Developers

Third Party Data Providers
- factual
- infochimps
- Socrata
Finding data
The needle in the haystack

Open Government Data
- The World Bank
- Data SF
- DATA.gov
- OECD.StatExtracts

Product APIs
- foursquare
- fitbit
- Developers

Third Party Data Providers
- factual
- infochimps
- Socrata

Anywhere
Gathering data
The tedious part

Scraping
Collecting manually
Crowdsourcing

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>name</td>
<td>address</td>
<td>neighborhood rating</td>
<td>rating_over</td>
<td></td>
<td>rating_espresso</td>
</tr>
<tr>
<td></td>
<td>Blue Bottle</td>
<td>66 Mint St. Soma</td>
<td>8.5</td>
<td>1064</td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blue Bottle</td>
<td>1 Ferry Build Embarcadero</td>
<td>8.1</td>
<td>1128</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blue Bottle</td>
<td>315 Linden S Hayes Valley</td>
<td>8.3</td>
<td>820</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coffee Bar</td>
<td>1890 Bryant Potrero Hill</td>
<td>8.45</td>
<td>1059</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Epicenter CAFE</td>
<td>764 Harrison SOMA</td>
<td>8.3</td>
<td>1121</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Four Barrel</td>
<td>C 375 Valencia Mission</td>
<td>8.45</td>
<td>1070</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mercury Cafe</td>
<td>201 Octavia Hayes Valley</td>
<td>8.2</td>
<td>1171</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ritual Coffee</td>
<td>1634 Jerrold Bayview</td>
<td>8.2</td>
<td>1016</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ritual Coffee</td>
<td>1026 Valenci Mission</td>
<td>8.45</td>
<td>843</td>
<td>8.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cafe Capriccio</td>
<td>2200 Mason North Beach</td>
<td>8.05</td>
<td>1127</td>
<td>8.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Special Xtra</td>
<td>46 Mission St. SOMA</td>
<td>7.55</td>
<td>1177</td>
<td>8.3</td>
<td></td>
</tr>
</tbody>
</table>
Transforming and evaluating data

Clean your data

- Missing data points
- Inconsistent formats

Jan 24, 2013
2013/01/24
24/01/13
24th January 2013
Transforming and evaluating data

Clean your data

Identify your target format

- Missing data points
- Inconsistent formats

Jan 24, 2013
2013/01/24
24/01/13
24th January 2013

SQL Database?
CSV?
Excel?
JSON?
API?
Example
Process

Finding → Gathering → Transforming → Evaluating
THE NEW FILM BY QUENTIN TARANTINO

DJANGO UNCHAINED
“I bet that’s the most successful western so far”
Question

Which is the most successful Western so far?
## Western Movies

<table>
<thead>
<tr>
<th>Movie</th>
<th>Release Date</th>
<th>Distributor</th>
<th>MPAA Rating</th>
<th>Domestic Gross</th>
<th>Inflation-Adjusted Gross</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hondo</td>
<td>Jan 1, 1953</td>
<td>United Artists</td>
<td>PG</td>
<td>$8,200,000</td>
<td>$107,830,003</td>
</tr>
<tr>
<td>The Alamo</td>
<td>Oct 24, 1960</td>
<td>United Artists</td>
<td>PG-13</td>
<td>$7,900,000</td>
<td>$90,334,780</td>
</tr>
<tr>
<td>Major Dundee</td>
<td>Apr 7, 1965</td>
<td>Sony Pictures</td>
<td>PG-13</td>
<td>$14,873</td>
<td>$116,188</td>
</tr>
<tr>
<td>Bandolero!</td>
<td>Jan 1, 1968</td>
<td>20th Century Fox</td>
<td>PG-13</td>
<td>$12,000,000</td>
<td>$72,274,806</td>
</tr>
<tr>
<td>'C'era una volta il West</td>
<td>May 28, 1969</td>
<td>Paramount Pictures</td>
<td>PG-13</td>
<td>$5,321,508</td>
<td>$29,568,098</td>
</tr>
<tr>
<td>Big Jake</td>
<td>Jan 1, 1971</td>
<td>PG-13</td>
<td>$7,500,000</td>
<td>$35,863,640</td>
<td></td>
</tr>
<tr>
<td>Tombstone</td>
<td>Dec 25, 1993</td>
<td>Walt Disney</td>
<td>R</td>
<td>$56,505,065</td>
<td>$90,906,513</td>
</tr>
<tr>
<td>Lightning Jack</td>
<td>Mar 11, 1994</td>
<td>Savoy</td>
<td>PG-13</td>
<td>$16,821,273</td>
<td>$24,969,720</td>
</tr>
<tr>
<td>Bad Girls</td>
<td>Apr 22, 1994</td>
<td>20th Century Fox</td>
<td>R</td>
<td>$15,187,851</td>
<td>$25,815,733</td>
</tr>
<tr>
<td>Wyatt Earp</td>
<td>Jun 24, 1994</td>
<td>Warner Bros.</td>
<td>PG-13</td>
<td>$25,052,000</td>
<td>$38,006,035</td>
</tr>
<tr>
<td>Wagons East</td>
<td>Sep 9, 1994</td>
<td>Sony Pictures</td>
<td>PG-13</td>
<td>$4,358,940</td>
<td>$8,429,424</td>
</tr>
<tr>
<td>The Quick and the Dead</td>
<td>Feb 10, 1995</td>
<td>Sony Pictures</td>
<td>R</td>
<td>$18,552,460</td>
<td>$33,634,344</td>
</tr>
<tr>
<td>Tall Tale</td>
<td>Mar 24, 1995</td>
<td>Walt Disney</td>
<td>PG</td>
<td>$8,247,627</td>
<td>$14,939,186</td>
</tr>
<tr>
<td>Last of the Dogmen</td>
<td>Sep 8, 1995</td>
<td>Savoy</td>
<td>PG</td>
<td>$7,008,542</td>
<td>$12,666,709</td>
</tr>
<tr>
<td>Wild Bill</td>
<td>Dec 1, 1995</td>
<td>MGM</td>
<td>R</td>
<td>$2,169,373</td>
<td>$3,934,790</td>
</tr>
<tr>
<td>Dead Man</td>
<td>May 10, 1996</td>
<td>Miramax</td>
<td>R</td>
<td>$1,025,488</td>
<td>$1,830,567</td>
</tr>
<tr>
<td>Ride With the Devil</td>
<td>Nov 24, 1999</td>
<td>USA Films</td>
<td>R</td>
<td>$630,779</td>
<td>$967,211</td>
</tr>
<tr>
<td>Shanghai Noon</td>
<td>May 26, 2000</td>
<td>Walt Disney</td>
<td>PG-13</td>
<td>$56,932,305</td>
<td>$83,338,756</td>
</tr>
<tr>
<td>All the Pretty Horses</td>
<td>Dec 22, 2000</td>
<td>Miramax</td>
<td>PG-13</td>
<td>$15,527,125</td>
<td>$22,275,742</td>
</tr>
</tbody>
</table>
Process

- Finding
- Gathering
- Transforming
- Evaluating
Tip

HTML pages to Excel/CSV/Spreadsheet

Google Docs

=importHtml(\texttt{URL}, \texttt{element}, \texttt{index})

Tip

HTML pages to Excel/CSV/Spreadsheet

Google Docs

=importHtml(URL, element, index)

=importHtml(
    "http://www.the-numbers.com/movies/genre/Western",
    "table",
    4
)

Process

Finding → Gathering → Transforming → Evaluating
Data Cleaning

Stanford Text Wrangler
http://vis.stanford.edu/wrangler/app/

Open Refine
http://openrefine.org/
Process

Finding  
Gathering  
Transforming  
Evaluating
Let's try again

http://blogs.ischool.berkeley.edu/i247s13/lab-1-data/
Questions?
Aaaaaaargh!
I can’t code!
Start somewhere
Just do it!

Learn to code interactively, for free.

People all over the world are learning with Codecademy. Join in now!

Get Started!

http://www.codecademy.com/tracks/javascript
Reminder
Assignment 1

Task Find and document 2 visualizations
• 1 online
• 1 “in the wild”

Deliverable 2 visualizations; 1 page writeup

Due Tuesday Jan 29, 3:00PM

More information on the class blog
Tuesday
From Data to Visualization