True Business Data

Data: Open for business
The U.S. census bureau has a responsibility to maintain the best possible business data

Census Bureau Mission
“To serve as the leading source of quality data about the nation's people and economy”

Census Bureau Goal
Our goal is to provide the best mix of timeliness, relevancy, quality and cost for the data we collect and services we provide.

http://www.census.gov/about/what.html
U.S. business data in a sorry siloed state

Slow, small, poorly structured and siloed

Current Census data
- Public, useful high level metric
- Every 5 years, at state level,
- Highly aggregated

State-level Govt. data
- Open source
- Unstructured and inconsistent across state boundaries

Proprietary business data
- Entity level, highly structured
- Inaccessible behind APIs: Google maps, Yelp, etc.
- Or firewalls e.g. CES Longitudinal business data
Solution: Better business data created from the web

Based on the Common Crawl

The Common Crawl is an open source snapshot of the public web (50Tb) updated ~monthly.

Using big data processing and machine learning techniques we’ve created a tool that enables rich, recent business data to be extracted by Zip Code.
Open business data presents a huge opportunity
The bottleneck to a richer understanding of U.S. business ecosystem

This project is a data product and not an interface, or a pipeline, or a classifier

A rich new resource for everyone interested in US business data, and the 6 million active small businesses that drive the US economy.

1SBA.gov, 2012 Census summary, NYTimes: government-incentives
Open business data presents a huge opportunity

Users

Business

Government

Academic

1 SBA.gov, 2012 Census summary, NYTimes: government-incentives
So what is True Business Data?

An open source set of data that provides listings of businesses and their locations created from the common crawl.

It currently contains:

- Address(es)
- Website URL
- Date

So what?
Why is it better?

- **Openness**
  - High level
  - Low level

- **Closed**

TBD
## Why is it better?

<table>
<thead>
<tr>
<th></th>
<th>Openness</th>
<th>Timeliness</th>
<th>Granularity</th>
<th>Cost</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States Census Bureau</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>OpenData</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Google</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
“If only I had local business data I could perform a better analysis, and give a better answer to my question!”

(This was us 12 weeks ago)
How do governmental agencies track the impact of $80bn of economic business incentives they spend each year?

True Business Data helps keep track of government expenditures by:

- Enabling granular tracking of thousands of business
- Enabling impact assessments of incentives across geographic areas
- Providing up-to date information to decision makers

Currently the government relies on tax forms, which are updated yearly.
Google’s Maps business relies on accuracy above all else to serve 1Bn users monthly.

How does True Business Data drive value for Google?

- A major unsolved problem: have you ever had an issue?
- Enabling validation and identification of new and removed businesses in a much rapid fashion
- Business openings and closing present an ongoing issue to this key metric

Currently Google relies on user reported information.
Academic use case

Where do academics get data to enable research on business ecosystems?

True Business Data is the best data set because it:

- Provides reliable snapshots of local economies across the United States
- Is publically released with no licensing fees or limitations
- Enables replication by providing common reference data

In Silicon Valley there is a sense that you prosper only because you’re surrounded by lots of resources that make it possible to succeed - beyond what your own entity controls.

Rosabeth Kanter
Harvard Business School
The list goes on….

Real estate valuation
Can the value of real estate be predicted from the local businesses?

Business expansion / creation
Where is the best place to start a new business? Is it dependent on what is in the area?

Supply chain management
What businesses in my area can help your business grow and thrive?

Advertising
Who can use your product, and how are you going to reach them?

Future capstone projects ;}
Project Stages

1. Reduce large dataset to something manageable
   a. Spent time figuring proper approach to tackle data size
   b. For proof of concept, focused on Berkeley businesses
   c. Final output of this stage was a 10 GB dataset, with 9108 websites and 865K web pages, containing:
      i. URL
      ii. Text Content

   - Scan common crawl for a list (and stats) of all websites with at least one Berkeley address
   - Manually remove a few excessively big websites and come up with appropriate filters to ignore certain web pages
   - Do a final pass on the common crawl and extract HTML content from all chosen web pages
2. Train classifier to detect business websites
   a. Used the output of the previous stage to iterate and find the best classification model
   b. Started by labeling close to 1K websites
   c. Best model uses logistic regression stacking ensemble
   d. After running classifier on full data, got 3.9K Berkeley businesses

3. Run across multiple snapshots to get monthly business list
   a. Used 25 VMs cluster and spent more than 1K CPU-hours processing data
   b. 3 MapReduce jobs per crawl
Future improvements

Expand True Business data nationwide, provide more snapshots and data access options

Areas for (even) further improvement:

- **Improve accuracy/precision**: add more labeled data.
- **Include new programmatic fields**: additional business metadata like phone number, email, business type.
- **Expand globally**: enable extensibility to cover other countries.
- **Expand methodology**: adapt method to create data for other areas.
Closing Thoughts

- Focused on generating an open source dataset not previously available
- Our intention is to spur other data science projects

“Data is the new oil” – Clive Humby
Our Team

● Michael
  ○ Ideas generator
  ○ Cloud resources
  ○ Multi-job Hadoop processing

● Stephen
  ○ Slides master extraordinaire
  ○ Web front genius
  ○ Web Classifier

● Jaime
  ○ EMR / AWS pipelines
  ○ Data exploration and munging
  ○ Web Classifier