Getting Started with Using Amazon Mechanical Turk to Label Data

Brandon Kopp
Office of Survey Methods Research @ BLS

Government Advances in Statistical Processing
23 September 2019
Labeled Datasets

- One of the main chokepoints in machine learning is collecting high quality labeled data for training and testing.
- Getting expertly trained coders in your organization may be too expensive and may be low priority if those coders have more pressing production concerns.
- Crowdsourcing the data labeling task is one way to limit this expense if you can share the data that needs to be labeled and you can control the quality.
Our Use Case

- Summary of Benefits and Coverage (SBC) are mandated by the Affordable Care Act (ACA) to be available for every healthcare plan.
- The National Compensation Survey (NCS) sometimes receives these as part of data collection and they then need to be transcribed.

<table>
<thead>
<tr>
<th>Important Questions</th>
<th>Answers</th>
<th>Why This Matters:</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the overall deductible?</td>
<td>$500/person or $1,000/family</td>
<td>You must pay all the costs up to the deductible amount before this plan begins to pay for covered services you use. Check your policy or plan document to see when the deductible starts over (usually, but not always, January 1). The Common Medical Events chart below shows how much you pay for covered services after you meet the deductible.</td>
</tr>
<tr>
<td>Are there other deductibles for specific services?</td>
<td>Yes. $300 for prescription drug coverage. There are no other specific deductibles.</td>
<td>You must pay all of the costs for these services up to the specific deductible amount before this plan begins to pay for these services.</td>
</tr>
<tr>
<td>Is there an out-of-pocket limit on my expenses?</td>
<td>Yes. For participating providers $2,500/person or $5,000/family. For non-participating providers $4,000/person or $8,000/family</td>
<td>The out-of-pocket limit is the most you could pay during a coverage period (usually one year) for your share of the cost of covered services. This limit helps you plan for health care expenses.</td>
</tr>
</tbody>
</table>
Is it possible to automate the extraction of data used by the National Compensation Survey from the SBCs and code that information into tables?

Sampled 3,000 SBC PDFs from publically available set on data.healthcare.gov*

Used mTurk to collect transcriptions of the data we were interested in

---

<table>
<thead>
<tr>
<th>Common Medical Event</th>
<th>Services You May Need</th>
<th>Participating(In-Network) Provider (You will pay the least)</th>
<th>Non-Participating(Out of Network) Provider (You will pay the most)</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you have a test</td>
<td>Diagnostic test (x-ray, blood work)</td>
<td>0% coinsurance</td>
<td>30% coinsurance</td>
</tr>
<tr>
<td></td>
<td>Imaging (CT/PET scans, MRIs)</td>
<td>0% coinsurance</td>
<td>30% coinsurance</td>
</tr>
</tbody>
</table>

- Copay
- Coinsurance
- After Deductible
- Per
- Not Covered
- No Charge

* PDFs downloaded from URLs in the 2017 & 2018 Plan Attribute Public Use Files
Amazon Mechanical Turk (mTurk)

- A labor marketplace where those needing discrete, online tasks performed can find workers and run their task’s workflow.
- Often used for work that is simple for humans but difficult for computers (image tagging, audio transcription, etc.).
- Other, similar services:
  - RapidWorkers
  - microWorkers
  - ClickWorker
  - Figure-Eight
  - CrowdSource
mTurk Terminology

- Requestor – The person requesting a service on mTurk
- Worker – The person performing a service on mTurk
- HIT (Human Intelligence Task) – A discrete task that a Requestor wants performed

Does this image contain a cat?
- Yes
- No

Labeling 1,000 Images = 1,000 HITs
mTurk Terminology

**Assignment** – The number of Workers completing a HIT

**Reward** – Compensation for the completion of a single HIT

**Batch** – A grouping of HITs

**Qualification** – Either custom or mTurk provided filter for Workers who can complete (or even see) your HIT
Getting Started with Amazon Mechanical Turk
Two Major Questions Before You Start

- Are the data I want labeled releasable?
  - They will, by necessity, be viewed by people outside of your organization

- Is my use of mTurk more like contracting a service or more like survey data collection?
  - Talk to your agency’s OMB liaison
  - Talk to your procurement folks
Getting Used to mTurk

- **Live Site**
  - [https://requester.mturk.com](https://requester.mturk.com)
  - [https://worker.mturk.com](https://worker.mturk.com)

- **Sandbox Environment**
  - The sandbox has the exact same functionality as the live site, but your account isn’t charged when you “post” a HIT **be careful you’re on the right site**
  - [https://requestersandbox.mturk.com](https://requestersandbox.mturk.com)
  - [https://workersandbox.mturk.com](https://workersandbox.mturk.com)
Project Setup

Choose Parameters of Project

- Provide a title & description for Workers
- Specify number of assignments per HIT
- Specify how much you will pay per HIT
- Choose Worker qualifications

Design Your Collection Instrument

- They do not have a SurveyMonkey-style interface
- mTurk has templates to get you started, but you will almost certainly need to modify the one you select
- You have to know some HTML & CSS
Design Your Collection Instrument

**Input Data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sbc_url</td>
</tr>
<tr>
<td>2</td>
<td><a href="https://s3.amazonaws.com/Connected">https://s3.amazonaws.com/Connected</a> Care Ex</td>
</tr>
<tr>
<td>3</td>
<td><a href="https://s3.amazonaws.com/Florida">https://s3.amazonaws.com/Florida</a> Hospital Go</td>
</tr>
<tr>
<td>4</td>
<td><a href="https://s3.amazonaws.com/Health">https://s3.amazonaws.com/Health</a> First GYM A</td>
</tr>
</tbody>
</table>

**Output Data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input.sbc_url</td>
<td>Input. plan_name</td>
</tr>
<tr>
<td>2</td>
<td><a href="https://s3.amazonaws.com/Connected">https://s3.amazonaws.com/Connected</a> Care Ex</td>
</tr>
<tr>
<td>3</td>
<td><a href="https://s3.amazonaws.com/Florida">https://s3.amazonaws.com/Florida</a> Hospital Go</td>
</tr>
<tr>
<td>4</td>
<td><a href="https://s3.amazonaws.com/Health">https://s3.amazonaws.com/Health</a> First GYM A</td>
</tr>
</tbody>
</table>

**HTML**

```html
<iframe height="350px" src="${sbc_url}" width="100%"></iframe>
```
Insurance Document Transcription Instructions (Click to expand)

MEDICA WI Individual Choice Gold Copay Plus
Summary of Benefits and Coverage: What this Plan Covers & What it Costs

This is only a summary: If you want more detail about your coverage and costs, you can get the complete terms in the policy or plan document at www.medica.com or by calling 888-592-8211.

Important Questions

What is the overall deductible? $1,000 Individual/ $3,000 Family for in-network services, $10,000 Individual/ $20,000 Family out-of-network. If you have other family members on the policy, the overall family deductible must be met before the plan begins to pay. Deductible does not apply to preventive care or co-pay services from in-network providers.

Why this Matters:

You must pay all the costs up to the deductible amount before this plan begins to pay for covered services you use. Check your policy or plan document to see when the deductible starts over (usually, but not always, January 1st). See the chart starting on page 2 for how much you pay for covered services after you meet the deductible.

Are there other deductibles for specific?

No.

Why this Matters:

You don’t have to meet deductibles for specific services, but see the chart starting on page 2 for other costs for...

Diagnostic Tests

Service Type: Diagnostic Test (All)  Network Type:  Copay A:  Coinurance %:  Per:  

Applied After Deductible: Yes  No Charge  Not Covered

Notes: 

Issue With Diagnostic Tests: 

Notes About Diagnostic Tests: 
Publish Your Project

- You can then publish your project
- This is where you upload your input file
- You will see a printout of your costs.
- Your project will then be visible to Workers
Pilot

Qualification

Main Coding

3rd Coder

9 Workers

$2.00 Reward

1 HIT
9 Assignments
1 Batch

X 10

X 20

1 SBCs
Pilot Qualification

9 Workers

$2.00 Reward

1 HIT
9 Assignments
1 Batch

Qualification

70 Workers

$2.00 Reward

1 HIT
70 Assignments
1 Batch

Main Coding

Applied Custom Qualification (Workers > 90% Accuracy)
31 Workers

3rd Coder

1 SBCs

3 SBCs
### Pilot Qualification

- **9 Workers**

### Main Coding

- **24 Workers**

### 3rd Coder

**3,000 HITs**
- 2 Assignments per HIT
- 6 Batches

<table>
<thead>
<tr>
<th>Task</th>
<th>Workers</th>
<th>Reward</th>
<th>HITs</th>
<th>Assignments</th>
<th>Batches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot Qualification</td>
<td>9</td>
<td>$2.00 Reward</td>
<td>1 HIT</td>
<td>9 Assignments</td>
<td>1 Batch</td>
</tr>
<tr>
<td>Main Coding</td>
<td>24</td>
<td>$0.30 Reward</td>
<td>3,000 HITs</td>
<td>2 Assignments per HIT</td>
<td>6 Batches</td>
</tr>
</tbody>
</table>

**Total SBCs**
- 1 SBCs (Pilot Qualification)
- 3 SBCs (Main Coding)
- 3,000 SBCs (3rd Coder)
**Pilot Qualification**
- 9 Workers
- 1 SBCs
- $2.00 Reward
- 9 Assignments
- 1 Hit
- 1 Batch

**Main Coding**
- 24 Workers
- 3,000 SBCs
- $0.30 Reward
- 70 Assignments
- 1 Hit
- 1 Batch

**3rd Coder**
- 18 Workers
- 1,015 SBCs
- $0.30 Reward
- 24 Workers
- 24 SBCs
- Same HITs as Main Coding
- Added 1 Assignment
- 1 Batch
Pilot Qualification Main Coding 3rd Coder

9 Workers 70 Workers 24 Workers 18 Workers

1 SBCs 3 SBCs 3,000 SBCs 1,015 SBCs

1 HIT 1 HIT 3,000 HITs 1 HIT
9 Assignments 2 Assignments per HIT Added 1 Assignment
1 Batch 1 Batch 1 Batch 1 Batch

$2.00 Reward $2.00 Reward $0.30 Reward $0.30 Reward

1d 4hr 2d 6hr 11d 1hr 1d 3hr

$22 $196 $2,197 $460

1d 4hr 2d 6hr 11d 1hr 1d 3hr

$22 $196 $2,197 $460

1 SBCs 3 SBCs 3,000 SBCs 1,015 SBCs

$2.00 Reward $2.00 Reward $0.30 Reward $0.30 Reward

1d 4hr 2d 6hr 11d 1hr 1d 3hr

$22 $196 $2,197 $460

1 SBCs 3 SBCs 3,000 SBCs 1,015 SBCs

$2.00 Reward $2.00 Reward $0.30 Reward $0.30 Reward

1d 4hr 2d 6hr 11d 1hr 1d 3hr

$22 $196 $2,197 $460

1 SBCs 3 SBCs 3,000 SBCs 1,015 SBCs
mTurk API

- Some tasks are not easy with the tools on mturk.com
- We used the Python version of the mTurk API to:
  - Communicate with Workers
  - Add assignments to HITs
  - Extend expiration time for HIT
- Can test things out in Sandbox
- **Be careful, there are no guard rails**

```python
import boto3

client = boto3.client(
    service_name= 'mturk',
    region_name= 'us-east-1',
    aws_access_key_id= [YOUR ACCESS KEY],
    aws_secret_access_key= [YOUR SECRET KEY],
    endpoint_url= [URL FOR API])
```

- API Documentation
- Connecting to the mTurk API
- Assigning permission to use API
Takeaways

- Start by talking to your OMB liaison and contracting folks
- You can get a lot of work done in a short amount of time
- You can control quality through qualifications, releasing in batches, and worker monitoring
- It helps to know HTML, CSS, and Python (or some other language that can access the API)
- The Sandbox environment helps alleviate some anxiety about getting started
Contact Information

Brandon Kopp
Research Psychologist
OSMR/BLS
bls.gov/osmr
202-691-7514
kopp.brandon@bls.gov