

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

Important Questions	Answers	Why This Matters:
What is the overall <u>deductible</u> ?	\$500 /person or \$1,000 /family Doesn't apply to <u>preventive care</u>	You must pay all the costs up to the <u>deductible</u> amount before this <u>plan</u> begins to pay for covered services you use. Check your policy or plan document to see when the <u>deductible</u> 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 <u>deductible</u> .
Are there other <u>deductibles</u> for specific services?	Yes. \$300 for <u>prescription drug coverage</u> . There are no other specific <u>deductibles</u> .	You must pay all of the costs for these services up to the specific <u>deductible</u> amount before this <u>plan</u> begins to pay for these services.
Is there an <u>out-of-pocket limit</u> on my expenses?	Yes. For participating <u>providers</u> \$2,500 /person or \$5,000 /family For non-participating <u>providers</u> \$4,000 /person or \$8,000 /family	The <u>out-of-pocket limit</u> 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.

Our Use Case

- 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

Common Medical Event	Services You May Need	What You Will Pay	
		Participating(In-Network) Provider (You will pay the least)	Non-Participating(Out of Network) Provider (You will pay the most)
If you have a test	Diagnostic test (x-ray, blood work)	0% coinsurance	30% coinsurance
	Imaging (CT/PET scans, MRIs)	0% coinsurance	30% coinsurance

- 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



Labeling 1,000 Images = 1,000 HITs



mTurk Terminology

HIT Groups (1-20 of 3866) ⊕ Show Details ⊖ Hide Details | Items Per Page: 20 ▼

Requester	Title	HITs ▼	Reward ▼	Created ▼	Actions	
⊕ Flow MTurk	Parking Sign Transcription	15,256	\$0.08	6d ago	Preview	🔒 Qualify
⊕ OpenLaw	Legal Entity Classification	10,343	\$0.01	11/2/2018	Preview	Accept & Work

- 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://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://requester sandbox.mturk.com>
- ▶ <https://worker sandbox.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

```
<iframe height="350px" src="
${sbc_url}" width="100%"></iframe>
```

HTML

How HIT
variables are
defined

	A	B
1	sbc_url	plan_name
2	https://s3.amazonaws.com/	Connected Care Ex
3	https://s3.amazonaws.com/	Florida Hospital Go
4	https://s3.amazonaws.com/	Health First GYM A

Input Data



Webpage

	AD	AE
1	Input.sbc_url	Input. plan_name
2	https://s3.amazonaws.	Connected Care Ex
3	https://s3.amazonaws.	Florida Hospital Go
4	https://s3.amazonaws.	Health First GYM A

Output Data

Collapsible
Instructions

Insurance Document Transcription Instructions (Click to expand)

Frame with
PDF

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

Coverage Period: Beginning on or after 01/01/2017
Coverage for: Individual or Family | Plan Type: PPO

! 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	Answers	Why this Matters:
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 .	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.	You don't have to meet deductibles for specific services, but see the chart starting on page 2 for other costs for

Open PDF in new tab/window and/or Expand the data entry section.

Popup
Instructions

Diagnostic Tests ?

Service Type:

Network Type:

Copay \$:

Coinsurance %:

Per:

Applied After
Deductible:

Notes:

Diagnostic Test (All) ▼

▼

▼

☐ Yes

☐ No Charge

☐ Not Covered

Diagnostic Test (All) ▼

▼

▼

☐ Yes

☐ No Charge

☐ Not Covered

Diagnostic Test (All) ▼

▼

▼

☐ Yes

☐ No Charge

☐ Not Covered

Issue With Diagnostic Tests:

▼

Notes About Diagnostic Tests:

Data Collection
Form

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

The screenshot displays the Amazon MTurk interface. At the top, there's a navigation bar with 'amazonmturk' logo and tabs for 'Create', 'Manage', and 'Developer'. Below this, there are links for 'New Project' and 'New Batch with an Existing Project'. The main section is titled 'Start a New Batch with an Existing Project' and contains a table of existing projects. A red arrow points from the 'Publish Batch' button in the table to the 'Publish Your Project' title. Another red arrow points from the 'Your project will then be visible to Workers' list item to the 'HIT Groups' section at the bottom.

Project Name	Title	Created	Last Edited	
SBC-Transcription	Collect data from insurance coverage PDFs (~3 minutes)	February 6, 2018	September 16, 2019	Publish Batch Edit Copy Delete
SBC-Qualification	Qualification for "Collect data from insurance coverage PDFs" (~15 minutes)	October 17, 2018	March 7, 2019	Publish Batch Edit Copy Delete

HIT Groups

[+ Show Details](#) [- Hide Details](#) Items Per Page:

Requester	Title	HITs	Reward	Created	Actions
+ Brandon Kopp	Collect data from insurance coverage PDFs (~3 minut...	3	\$0.30	22s ago	Preview Accept & Work





9 Workers



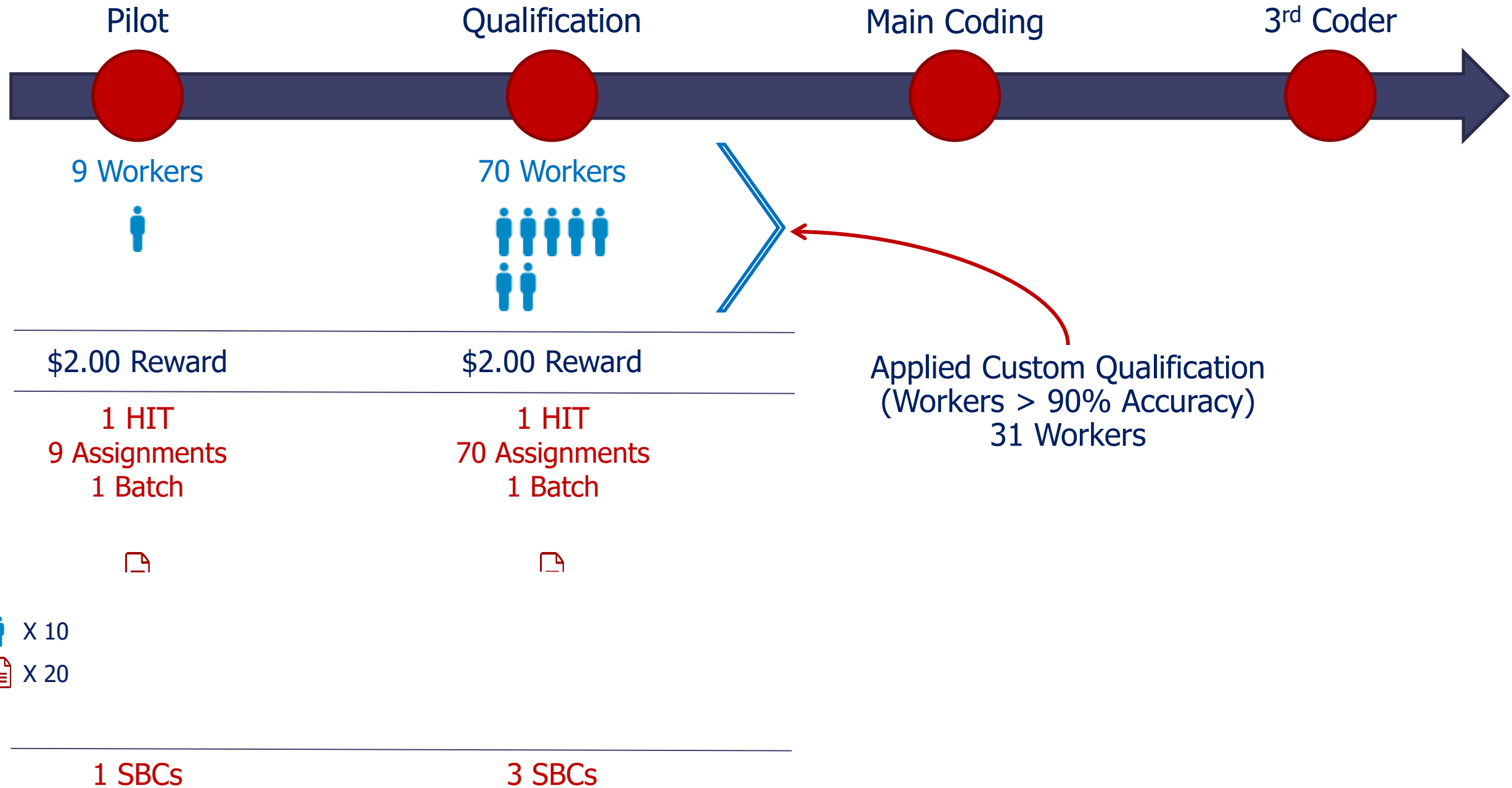
\$2.00 Reward

1 HIT
9 Assignments
1 Batch



 X 10
 X 20

1 SBCs





9 Workers



70 Workers



24 Workers



\$2.00 Reward

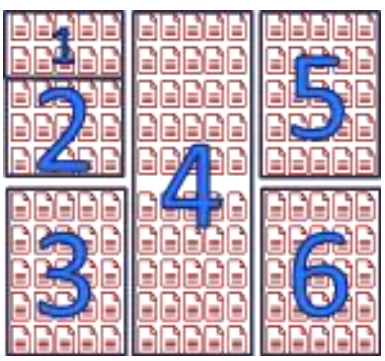
\$2.00 Reward



\$0.30 Reward

1 HIT
9 Assignments
1 Batch

1 HIT
70 Assignments
1 Batch

3,000 HITs
2 Assignments per HIT
6 Batches

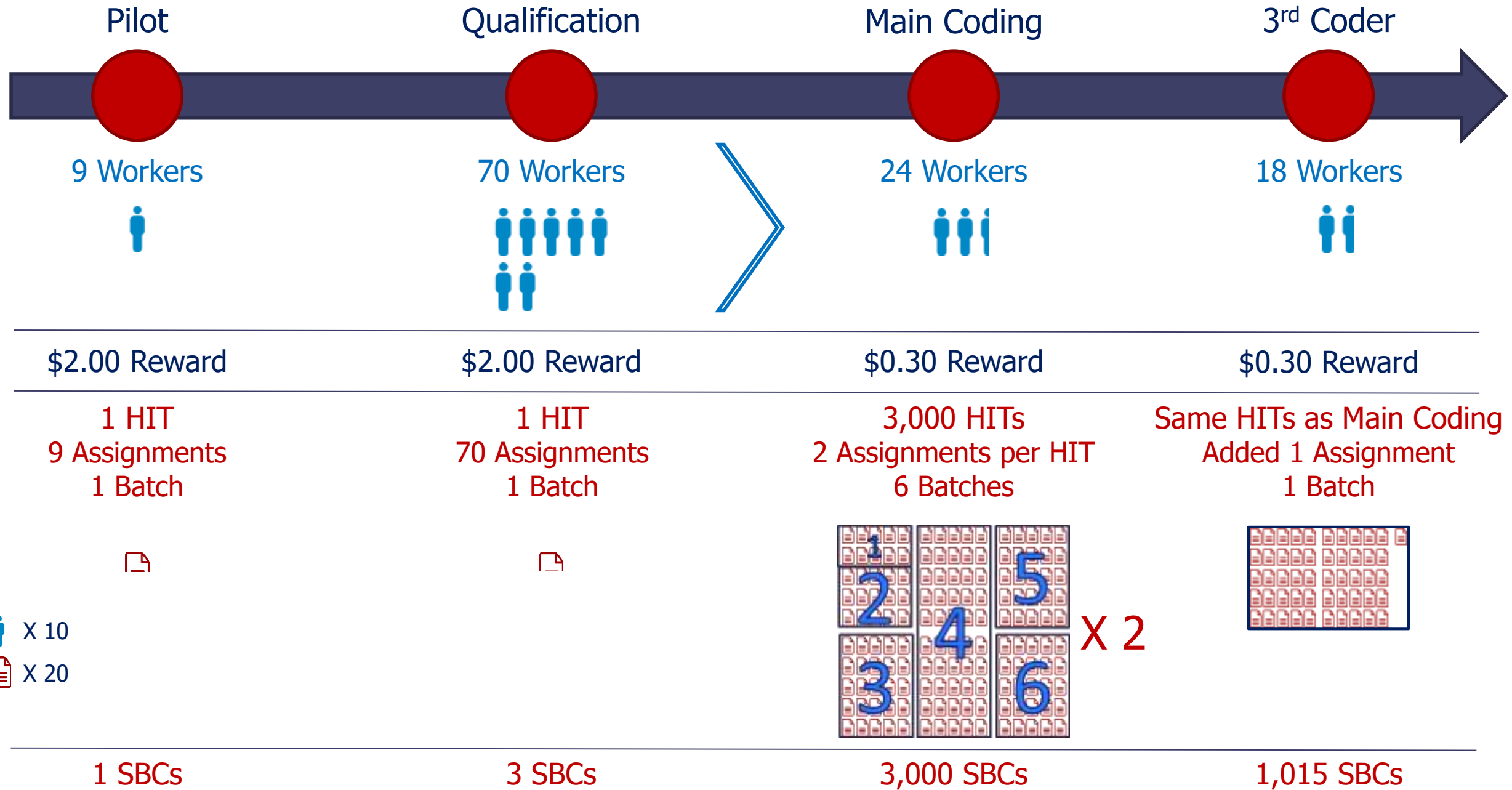


 X 10
 X 20

1 SBCs

3 SBCs

3,000 SBCs





9 Workers



70 Workers



24 Workers



18 Workers



\$2.00 Reward

1 HIT
9 Assignments
1 Batch

1d 4hr

\$22

X 10
 X 20

1 SBCs

\$2.00 Reward

1 HIT
70 Assignments
1 Batch

2d 6hr

\$196

3 SBCs

\$0.30 Reward

3,000 HITs
2 Assignments per HIT
6 Batches

11d 1hr

\$2,197

X 2

3,000 SBCs

\$0.30 Reward

Same HITs as Main Coding
Added 1 Assignment
1 Batch

1d 3hr

\$460

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****

```
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