

Crowdsourcing for Statisticians

Designing Applications and Analyzing Data on MTurk

Adam Kapelner* and Lyle Ungar**

*Department of Statistics, The Wharton School, University of Pennsylvania

**Department of Computer Science, University of Pennsylvania

August 5, 2013
Joint Statistical Meeting Half Day Tutorial

Curriculum and Plan

Overview of Crowdsourcing

- Modern crowdsourcing with emphasis on microlabor for microtasks
- Amazon's Mechanical Turk (MTurk) from an employer and worker perspective
- What industry professionals and academics use MTurk for

Curriculum and Plan

Overview of Crowdsourcing

- Modern crowdsourcing with emphasis on microlabor for microtasks
- Amazon's Mechanical Turk (MTurk) from an employer and worker perspective
- What industry professionals and academics use MTurk for

Survey tasks

- Design and validity of MTurk surveys

Curriculum and Plan

Overview of Crowdsourcing

- Modern crowdsourcing with emphasis on microlabor for microtasks
- Amazon's Mechanical Turk (MTurk) from an employer and worker perspective
- What industry professionals and academics use MTurk for

Survey tasks

- Design and validity of MTurk surveys

Labeling tasks

- The design and analysis of MTurk "labeling" tasks (with case studies)
- Model building and machine learning

Curriculum and Plan

Overview of Crowdsourcing

- Modern crowdsourcing with emphasis on microlabor for microtasks
- Amazon's Mechanical Turk (MTurk) from an employer and worker perspective
- What industry professionals and academics use MTurk for

Survey tasks

- Design and validity of MTurk surveys

Labeling tasks

- The design and analysis of MTurk "labeling" tasks (with case studies)
- Model building and machine learning

Experimental tasks

- The design of and analysis of MTurk "experimental" tasks (with case studies)
- Drawing conclusions from crowdsourced experiments

Course Goals

What you should be able to do by noon today

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data
- D For crowdsourced randomized experiments:

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data
- D For crowdsourced randomized experiments:
 - Evaluate if the experiment is suitable for crowdsourcing

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data
- D For crowdsourced randomized experiments:
 - Evaluate if the experiment is suitable for crowdsourcing
 - Design the experiment

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data
- D For crowdsourced randomized experiments:
 - Evaluate if the experiment is suitable for crowdsourcing
 - Design the experiment
 - Analyze the resulting data

Course Goals

What you should be able to do by noon today

- A Know what MTurk is used for in academia and industry *and thereby* recognize where crowdsourcing could be useful in your work
- B Use MTurk for survey research
- C For crowdsourced labeling tasks:
 - Evaluate if the labeling task is suitable for crowdsourcing
 - Design the labeling task
 - Analyze the resulting data
- D For crowdsourced randomized experiments:
 - Evaluate if the experiment is suitable for crowdsourcing
 - Design the experiment
 - Analyze the resulting data

Our crowdsourcing experience. What is yours?



Lyle Ungar is a professor of Computer and Information Science at the University of Pennsylvania. His research group makes extensive use of crowdsourcing for labeling text including tweets, Facebook posts, and word senses to support research in natural language processing and psychology.



Adam Kapelner is a PhD student in Statistics at Wharton focused on crowdsourcing applications. He has published multiple papers using crowdsourced experimentation and labeling data in economics, psychology, and linguistics, including the first paper advocating crowd-sourced natural field experiments.

What is Crowdsourcing?

"Its not outsourcing; its crowdsourcing" Howe (2006)



What is Crowdsourcing?

"Its not outsourcing; its crowdsourcing" Howe (2006)



"the biggest paradigm shift in innovation since the Industrial Revolution" -Wendy Kaufman, NPR, 2008

Types: I - Wisdom of the Crowd

Types: I - Wisdom of the Crowd

- Wikipedia



Types: I - Wisdom of the Crowd

- Wikipedia



- Yahoo Answers



Types: I - Wisdom of the Crowd

- Wikipedia



- Yahoo Answers



- Stackoverflow



Types: II - Crowdfunding

Types: II - Crowdfunding

- Kickstarter / RocketHub / Fundly / IndieGoGo



Types: II - Crowdfunding

- Kickstarter / RocketHub / Fundly / IndieGoGo



- Realty Mogul



Types: III - Creative Work

Types: III - Creative Work

- 99 Designs, threadless



Types: III - Creative Work

- 99 Designs, threadless



- oDesk, eLance, rentacoder, guru



Types: III - Creative Work

- 99 Designs, threadless



- oDesk, eLance, rentacoder, guru



- IdeaBounty and Innocentive



Types: III - Creative Work

- 99 Designs, threadless



- oDesk, eLance, rentacoder, guru



- IdeaBounty and Innocentive



- Inducement Prize: Ansari-X, DARPA balloon challenge, Netflix Prize, the Nobel?



Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk

Type IV - Microwork / Microlabor

Type IV - Microwork / Microlabor

- Amazon's Mechanical Turk (MTurk)



Type IV - Microwork / Microlabor

- Amazon's Mechanical Turk (MTurk)



- Specialized: LiveOps, Clickworker, Ahn's ESP Game



Type IV - Microwork / Microlabor

- Amazon's Mechanical Turk (MTurk)



- Specialized: LiveOps, Clickworker, Ahn's ESP Game



- Intermediaries: Crowdflower, Crowdsource, Alegion, Smartsheet



smartsheet
THE POWER OF DONE

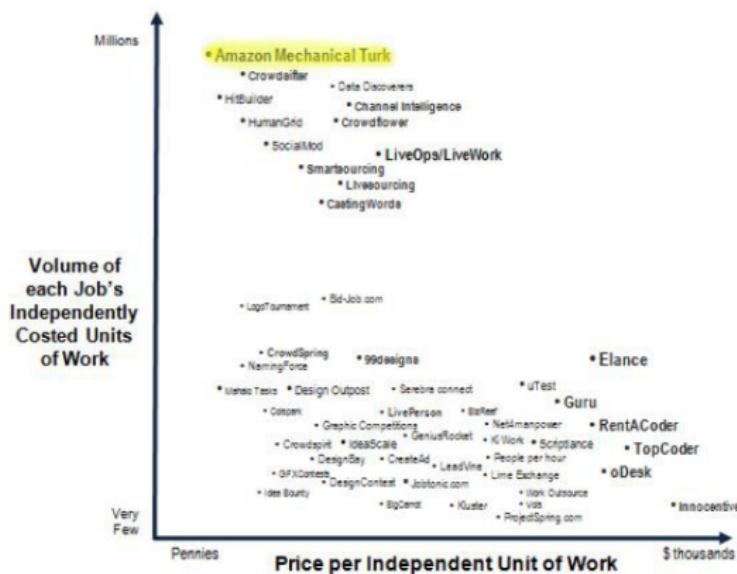
What is MTurk?

What is MTurk?

A marketplace where employers ("requesters") can post jobs and employees ("Turkers") can complete jobs. It is a labor market for microtasks: jobs are usually quick, simple, they can be "one-off", range from a few seconds to an hour or more.

What is MTurk?

A marketplace where employers (“requesters”) can post jobs and employees (“Turkers”) can complete jobs. It is a labor market for microtasks: jobs are usually quick, simple, they can be “one-off”, range from a few seconds to an hour or more.



Crowdsourcing platform by Volume of work by price (Frei, 2009)

Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk

The etymology of “Mechanical Turk”

A contraption built in 1770 for the empress of Austria, it beat many challengers including Napolean and Ben Franklin in chess, and it was exposed in 1820, destroyed in an 1854 Philadelphia fire. Below is an image of a mock reconstruction:



The etymology of “Mechanical Turk”

A contraption built in 1770 for the empress of Austria, it beat many challengers including Napolean and Ben Franklin in chess, and it was exposed in 1820, destroyed in an 1854 Philadelphia fire. Below is an image of a mock reconstruction:



Amazon calls it “artificial artificial intelligence.”

Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk



The “Magic Machine” is the crowdsourcing platform (courtesy of MicroTask.com).

MTurk was originally devised to “get people to perform tasks that are simple for humans but difficult for computers” (Callison-Burch, 2010a).

You can put any task in the box and someone will do it, somewhere, for a price you set.



History of Internet-based Crowdsourcing

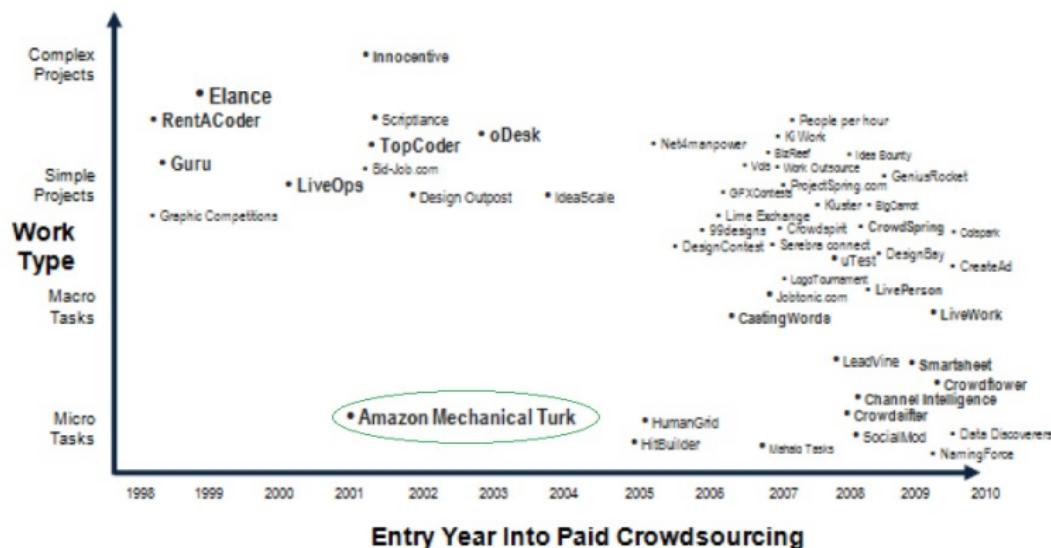


Figure : Crowdsourcing history (Frei, 2009)

A Typical Industry Task on MTurk

Tasks are dubbed “Human Intelligence Tasks” or HITs for short.
Note the requester is a crowdsourcing intermediary.

Find General Managers of Water Utilities		Not Qualified to work on this HIT (Why?) View a HIT in this group	
Requester:	CrowdFlower	HIT Expiration Date:	Jul 17, 2013 (6 days 5 hours)
		Reward:	\$0.07
		Time Allotted:	30 minutes
		HITs Available:	85
Description: Click on the Google search query link below and search for the general manager at the specified water utility.			
Keywords: mobmerge , builder , dolores , labs , crowd , flower , crowdflower , doloreslabs , doloreslabs , dolores , find , lead , generation , information , water , utilities , utility , fast , quick , good , pay , price , money , make , money , easy , quick , fast , hit , search , pay , price , info , money , web , link , contact			
Qualifications Required:		Your Value	
Total approved HITs is greater than 100	11	You do not meet this qualification requirement	
HIT approval rate (%) is greater than 96	100	You meet this qualification requirement	
Location is US	US	You meet this qualification requirement	

Note the qualifications, short description, keywords, time allotted, expiration, reward, number available.

The Turker Perspective

Find General Managers of Water Utilities

Requester: CrowdFlower

Reward: \$0.07 per HIT

HITs Available: 84

Duration: 30 minutes

Qualifications Required: Total approved HITs is greater than 100, HIT approval rate (%) is greater than 96, Location is US

Utility: Abilene in TX

Link: [Click here to Google search for the general manager!](#)

Alternatively, you can look on the [utility website](#), though the contact information is often not available.

Name:

Ex. John Appleseed

① Enter first and last name above (no middle initials please) for general manager at Abilene

Phone:

#####-####

Ext:

① Enter phone number above for general manager at Abilene

Email:

Ex: john@example.com

① Enter email above for general manager at Abilene

Check the box if you have tried both links and were unable to find this utility's general manager

Turker Tools

turkopticon.com, mturkforum.com, turkalert.com, turkrequesters blog, etc

Find General Managers of Water Utilities [View a HIT in this group](#)

Requester:	CrowdFlower	HIT Expiration Date:	Jul 17, 2013 (6 days 5 hours)	Reward:	\$0.07
communicativity:		2.35 / 5	30 minutes	HITs Available:	85
generosity :		2.44 / 5			
fairness :		3.18 / 5			
promptness :		3.36 / 5			

What do these scores mean?

Scores based on 625 reviews [Report your experience with this requester »](#)

s at Amazon | Developers | Press | Policies | Blog
amazon.com, Inc. or its Affiliates

An amazon.com company

All Reviews	Flagged By You	Your Reviews	Description
AMT Requester	Rating [info]		
CrowdFlower A2IR7ETVOIULZU 85219 HIT Group » Review Requester »	FAIR: 1 / 5 FAST: 1 / 5 PAY: 1 / 5 COMM: 1 / 5		Flags me and doesn't accept my hit after I spend 15 minutes doing it. I try another one and it does the same thing. Says no work available after I accept for other hits. Unable to successfully submit or try anything. Wastes lots of space on list when I'm looking for hits. Only seems to work for some people occasionally who must be incredibly lucky Jul 10 2013 ac flag comment
CrowdFlower A2IR7ETVOIULZU 85079 HIT Group » Review Requester »	FAIR: 1 / 5 FAST: 1 / 5 PAY: 1 / 5 COMM: 1 / 5		Every HIT lets me accept but then says there is no work available - it's driving up my returned rate for no reason. Jul 09 2013 asu...@g... flag comment
CrowdFlower A2IR7ETVOIULZU 84700 HIT Group » Review Requester »	FAIR: 1 / 5 FAST: 1 / 5 PAY: 1 / 5 COMM: 1 / 5		There work simply is never there. They need to fix their systems or stop using MTurk. Jul 07 2013 kkke...@h... flag comment
CrowdFlower	FAIR: 5 / 5		Their hits are sometimes broken.

Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk

The Requester Perspective

	A	B	C
1	<u>Num</u>	<u>Name</u>	<u>State</u>
2	1	Pacific Gas & Electric	CA
3	2	Southern California Edison	CA
4	3	Florida Power & Light	FL
5	4	Commonwealth Edison	IL
6	5	Consolidated Edison	NY
7	6	Georgia Power	GA
8	7	Dominion Resources	VA
9	8	Detroit Edison	MI
10	9	Public Service Enterprise Group	NJ
11	10	Energen Future Holdings	TX

- A spreadsheet CSV file — each row is one HIT.

The Requester Perspective

	A	B	C
1	<u>Num</u>	<u>Name</u>	<u>State</u>
2	1	Pacific Gas & Electric	CA
3	2	Southern California Edison	CA
4	3	Florida Power & Light	FL
5	4	Commonwealth Edison	IL
6	5	Consolidated Edison	NY
7	6	Georgia Power	GA
8	7	Dominion Resources	VA
9	8	Detroit Edison	MI
10	9	Public Service Enterprise Group	NJ
11	10	Eneray Future Holdings	TX

- A spreadsheet CSV file — each row is one HIT.
- # rows: large, done in parallel.

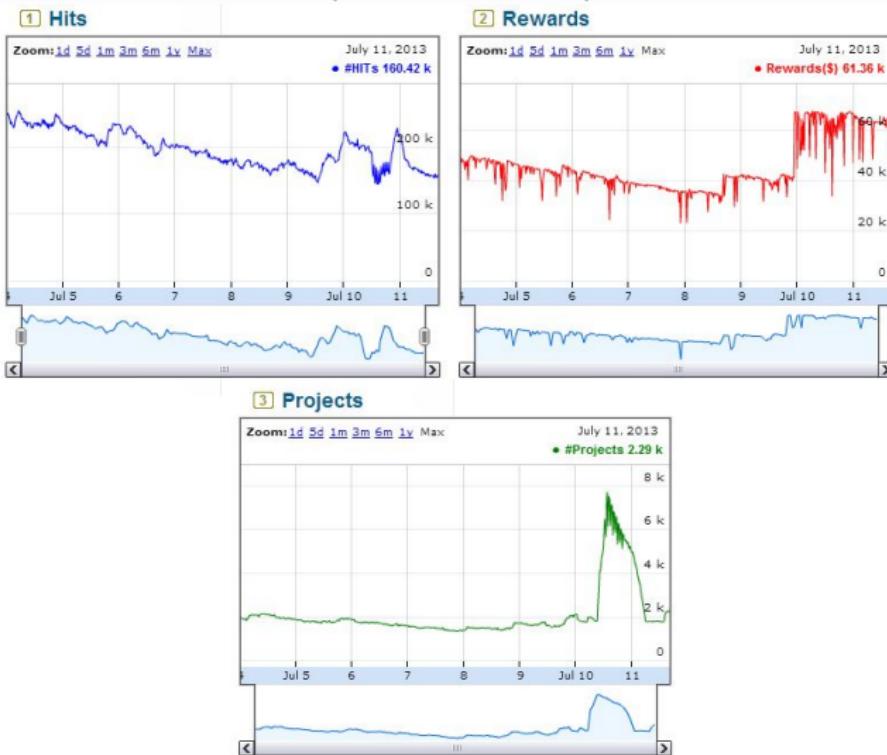
The Requester Perspective

A	B	C
1	Num Name	State
2	1 Pacific Gas & Electric	CA
3	2 Southern California Edison	CA
4	3 Florida Power & Light	FL
5	4 Commonwealth Edison	IL
6	5 Consolidated Edison	NY
7	6 Georgia Power	GA
8	7 Dominion Resources	VA
9	8 Detroit Edison	MI
10	9 Public Service Enterprise Group	NJ
11	10 Energy Future Holdings	TX

- A spreadsheet CSV file — each row is one HIT.
- # rows: large, done in parallel.
- Task creation is done through a simple system or through an iframe (explained in detail later).

What is available to a Turker?

From MTurk-tracker.com (Ipeirotis, 2010)



How many workers are there?

How many workers are there?

Amazon released this number in 2011:
≈ half million registered users in 190 countries.

How many workers are there?

Amazon released this number in 2011:
≈ half million registered users in 190 countries.



The rough distribution of Turker locations in the United States.
How regular are the Turkers? Unknown.

Who are these workers?

As of 2010, about 50% USA, 40% India, 10% everywhere else
(Ipeirotis, 2010a).

Who are these workers?

As of 2010, about 50% USA, 40% India, 10% everywhere else (Ipeirotis, 2010a). Workers in USA and India are paid in local currency but all others are paid in gift certificates.

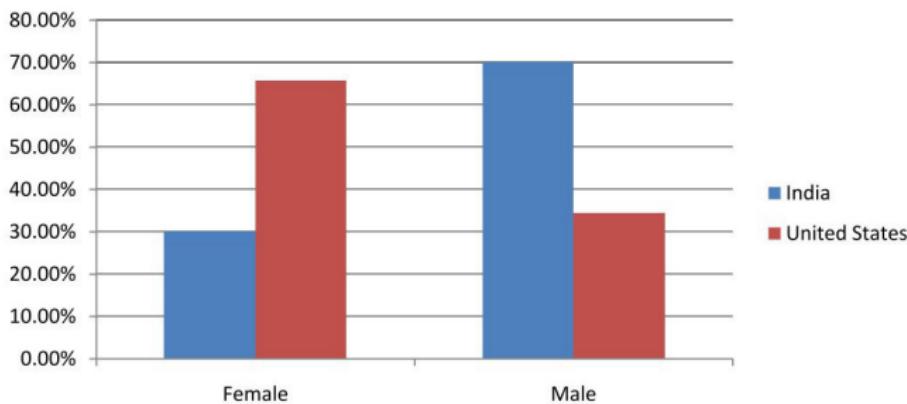
Who are these workers?

As of 2010, about 50% USA, 40% India, 10% everywhere else (Ipeirotis, 2010a). Workers in USA and India are paid in local currency but all others are paid in gift certificates. International usage has been growing since 2010 but there has not been an updated published survey.

Who are these workers?

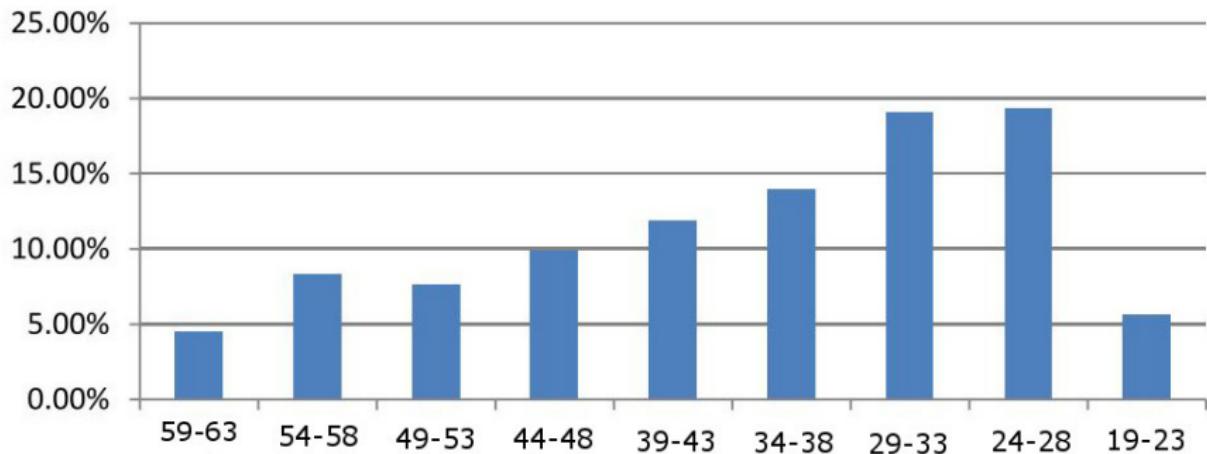
As of 2010, about 50% USA, 40% India, 10% everywhere else (Ipeirotis, 2010a). Workers in USA and India are paid in local currency but all others are paid in gift certificates. International usage has been growing since 2010 but there has not been an updated published survey.

Gender Breakdown



Who are these workers?

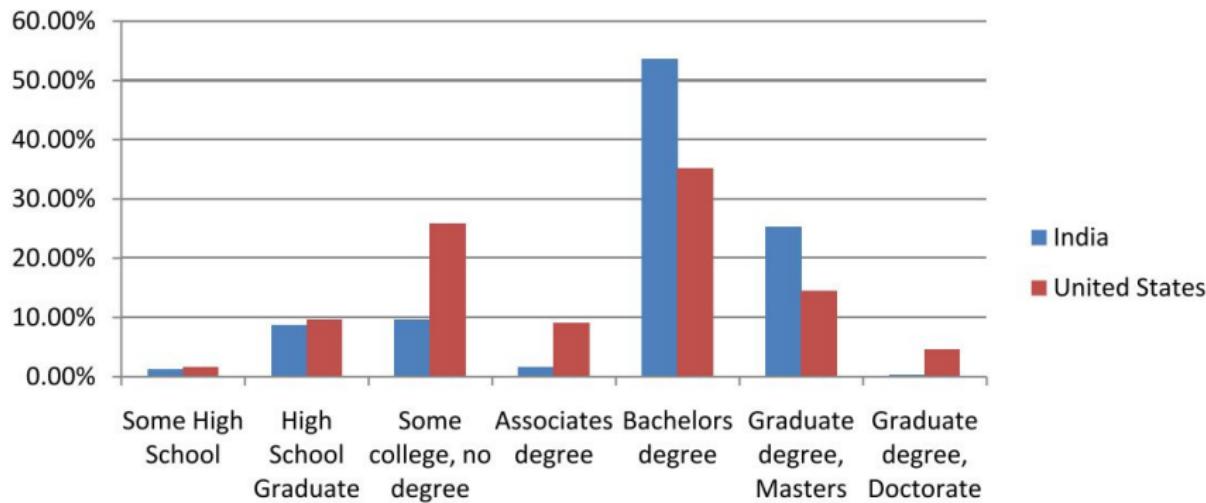
Age Distribution of American Turkers



(Indian workers shifted younger)

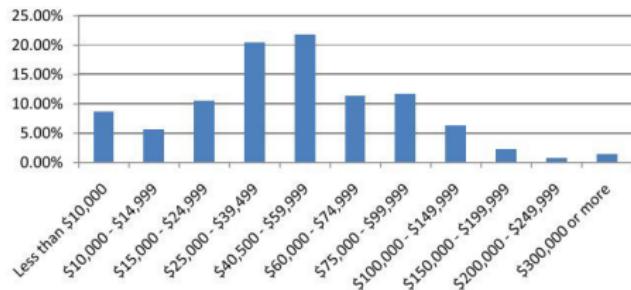
Who are these workers?

Education Level

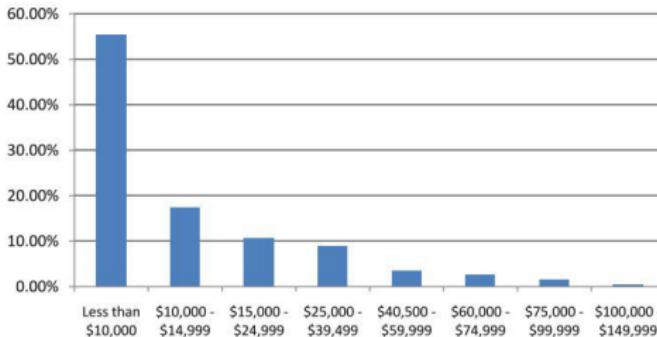


Who are these workers?

Household Income for US workers

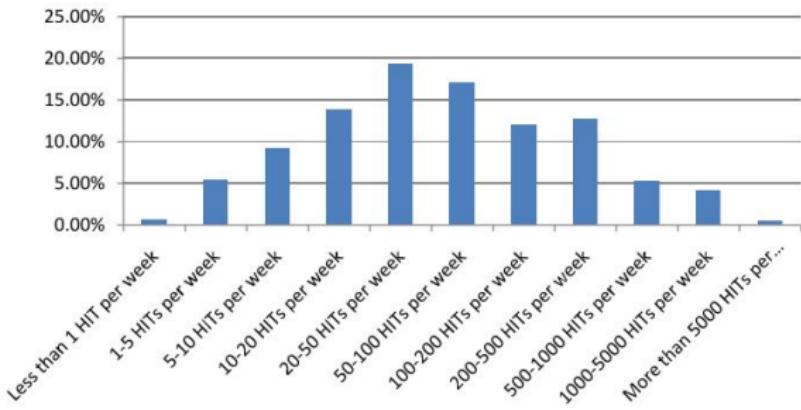


Household Income for Indian workers



Who are these workers?

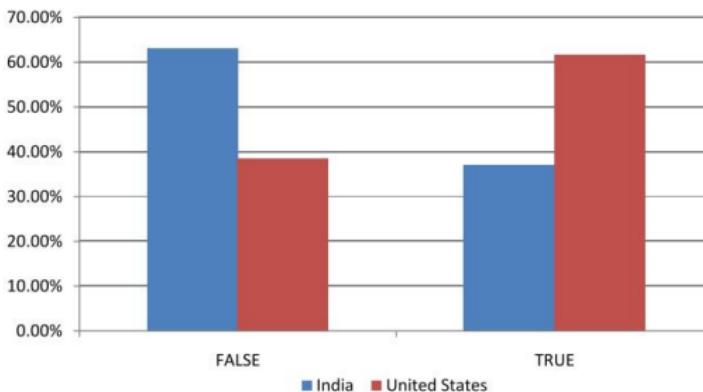
Number of HITs completed per week



(America and India)

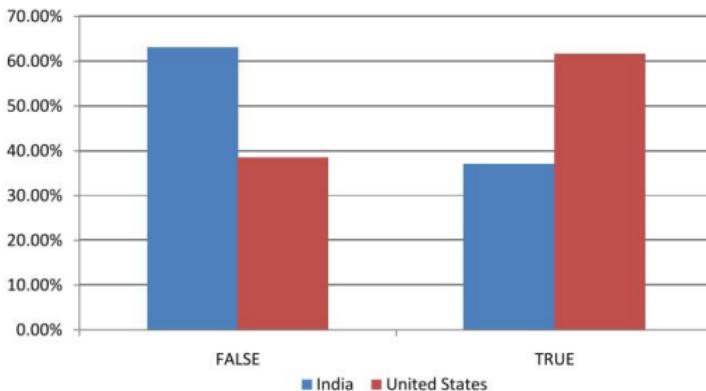
Who are these workers?

Mechanical Turk is my secondary source of income,
pocket change (for hobbies, gadgets, going out)



Who are these workers?

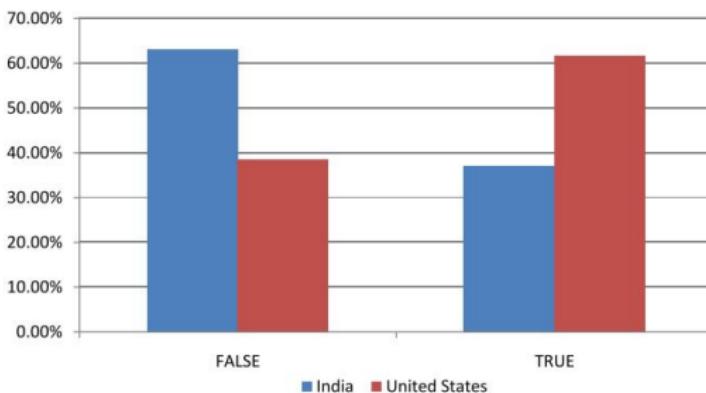
Mechanical Turk is my secondary source of income,
pocket change (for hobbies, gadgets, going out)



Undisplayed results: about 30% of Turkers use MTurk just to “kill time,”
41% use it for “entertainment,”

Who are these workers?

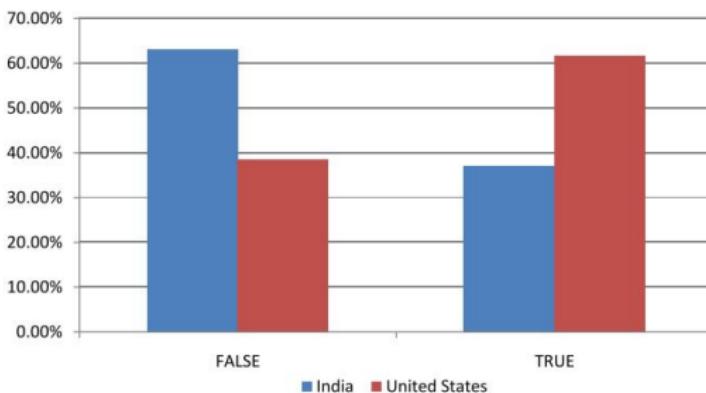
Mechanical Turk is my secondary source of income,
pocket change (for hobbies, gadgets, going out)



Undisplayed results: about 30% of Turkers use MTurk just to “kill time,” 41% use it for “entertainment,” about 15% use it as a primary source of income (25% in India), about 30% are unemployed or part-time employed.

Who are these workers?

Mechanical Turk is my secondary source of income,
pocket change (for hobbies, gadgets, going out)



Undisplayed results: about 30% of Turkers use MTurk just to “kill time,” 41% use it for “entertainment,” about 15% use it as a primary source of income (25% in India), about 30% are unemployed or part-time employed.

Summary: 60% say it's a good way to earn money, 70% think it's a good way to burn time.

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
------	----------------------

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%
Images related	22%

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%
Images related	22%
Text related & OCR	17%

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%
Images related	22%
Text related & OCR	17%
Audio related	14%

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%
Images related	22%
Text related & OCR	17%
Audio related	14%
Video related	3%

Types of tasks

Top requesters (save intermediaries) taken from mturk-tracker.com July 4 - July 11, 2013 (Ipeirotis, 2010) have the following tasks categories:

Type	Estimated Proportion
Web scouring	42%
Images related	22%
Text related & OCR	17%
Audio related	14%
Video related	3%
Testing / Quality Assurance	3%

Web Scouring

- Verify a restaurant listing
- Match my products to Amazon products
- Is this a web page? Easily decide if the image is a webpage.
- Find Official Government Websites for Places
- Find the Email addresses for wedding venues
- Find the school website and locate the current school supply list
- Find Yelp Reviews for Businesses
- Categorize a Twitter search query
- Locate company website
- Find a company name from an email domain
- Find main title, subtitle, and authors for a book
- Web Page Categorization
- Find the Official Social Media Accounts for a Website

Images related

- Select images from a piece of adult content
- Choose all the options this category belongs under (Amazon)
- Label simple images
- Categorize ad images
- Find the bird in each image (research: Columbia)
- Name the color of this rug
- Estimate the age of the person in the image
- Clickable Image Tagging

Text related / OCR

- Judge the quality of similar sentences (research: CC Burch, UPenn)
- Word Alignment (research: CC Burch, UPenn)
- Categorizing buying behaviour on twitter
- Please transcribe the following words from image
- Classify Messages about Nissan
- Data Transcription from a PDF

Audio related

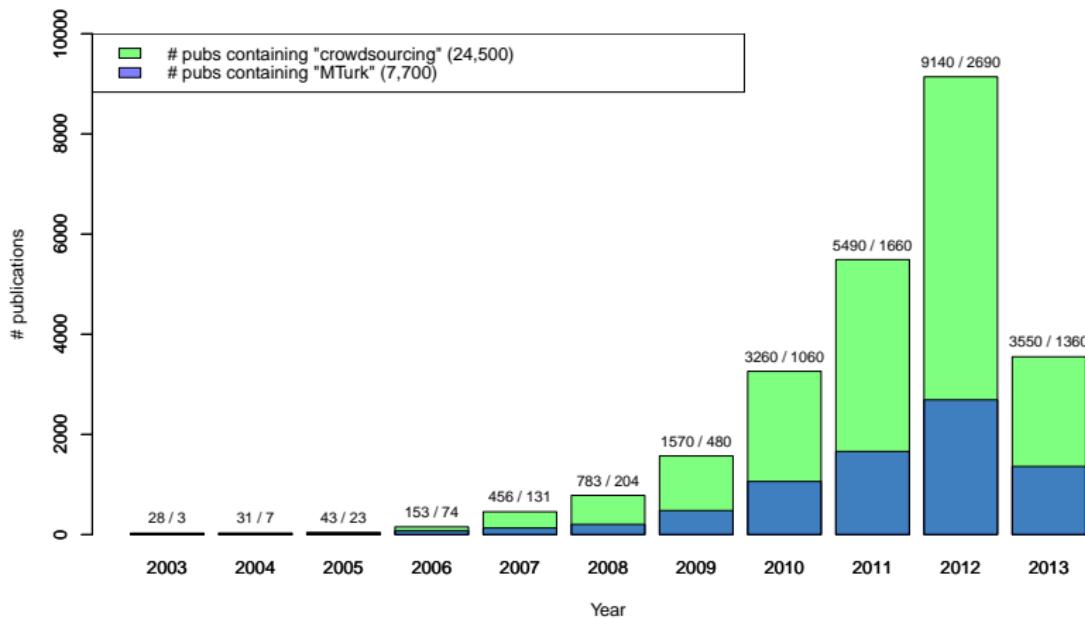
- Transcribe a ≈10sec Audio Clip
- Transcribe Recording
- record a word or phrase in your own voice (research, unknown)
- Audio transcription

Video related, Product testing

- Watch Video Snippets and Tell Us about the Tags
-

- Test an automated tourist information service

MTurk and Academia



Explosive Growth over the past few years!

What do researchers use MTurk for?

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation
 - Image analysis: trace objects, detect 3d planes, object similarity rating

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation
 - Image analysis: trace objects, detect 3d planes, object similarity rating
 - Audio: classify musical melodies

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation
 - Image analysis: trace objects, detect 3d planes, object similarity rating
 - Audio: classify musical melodies
- Running randomized controlled experiments testing an economic, behavioral, or psychological hypothesis of interest

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation
 - Image analysis: trace objects, detect 3d planes, object similarity rating
 - Audio: classify musical melodies
- Running randomized controlled experiments testing an economic, behavioral, or psychological hypothesis of interest

Other uses:

- Studying MTurk from an Economic and Sociological Perspective

What do researchers use MTurk for?

The goal of this tutorial is to teach you how to analyze data for these “main uses:”

- Collecting survey responses
- Collecting labels (sometimes for building a machine learning system)
 - Solving the “noisy labeler” problem: multiple workers
 - Computational linguistics: translation, document relevance, word-sense disambiguation
 - Image analysis: trace objects, detect 3d planes, object similarity rating
 - Audio: classify musical melodies
- Running randomized controlled experiments testing an economic, behavioral, or psychological hypothesis of interest

Other uses:

- Studying MTurk from an Economic and Sociological Perspective
- Creative extensions to MTurk and plugins to MTurk (mostly computer scientists)

The three MTurk applications we will focus on

Surveys

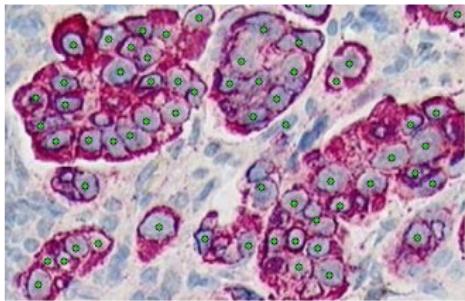


The three MTurk applications we will focus on

Surveys



Labelings

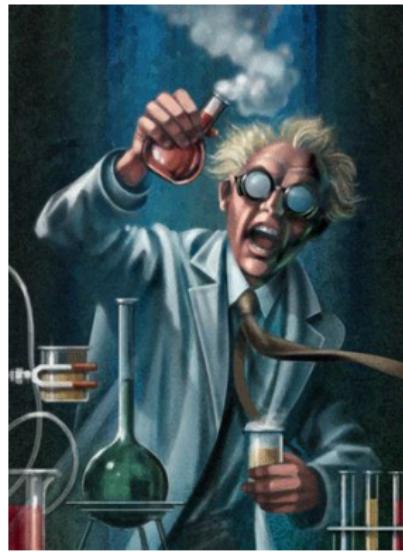


The three MTurk applications we will focus on

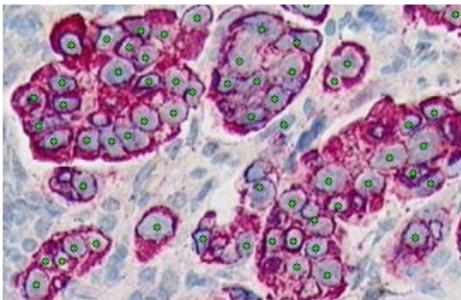
Surveys



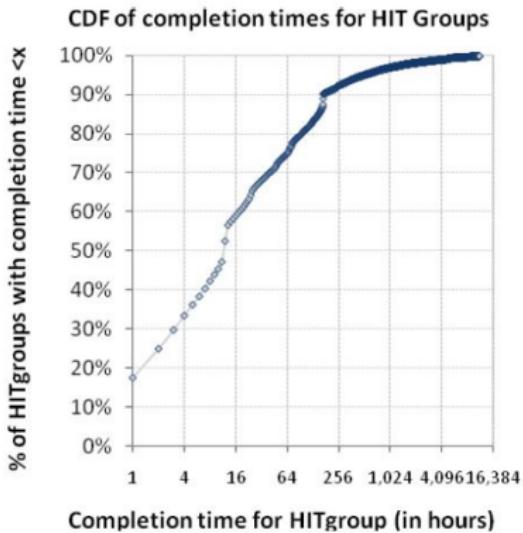
Experiments



Labelings

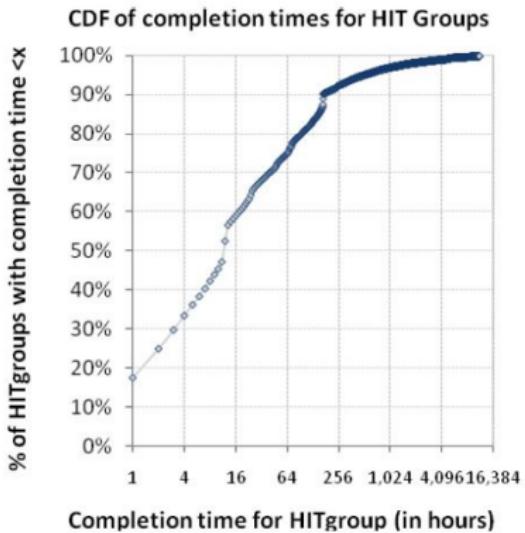


The Market: HIT Completion Times



(Ipeirotis, 2010c)

The Market: HIT Completion Times



(Ipeirotis, 2010c)

Design implications?

The Market: HIT Wages



The Market: HIT Wages



There is no magic formula for HIT wage, time to completion, etc. It's all experienced trial and error. "MTurk is not a M/M/1 queue."

Microlabor Incentives

Traditional economic theory: higher wage \Rightarrow higher performance.

Microlabor Incentives

Traditional economic theory: higher wage \Rightarrow higher performance.
Large body of research in economics and psychology that shows
this isn't the case, and not only on MTurk.

Microlabor Incentives

Traditional economic theory: higher wage \Rightarrow higher performance.
Large body of research in economics and psychology that shows
this isn't the case, and not only on MTurk. Workers must have
intrinsic motivation - which are built through non-pecuniary
incentives.

- Mason and Watts (2010) showed that paying more results in higher *volume* of work but not higher *quality*.

Microlabor Incentives

Traditional economic theory: higher wage \Rightarrow higher performance.
Large body of research in economics and psychology that shows
this isn't the case, and not only on MTurk. Workers must have
intrinsic motivation - which are built through non-pecuniary
incentives.

- Mason and Watts (2010) showed that paying more results in higher *volume* of work but not higher *quality*.
- Many other crowdsourcing platforms e.g. galaxyzoo (classify galaxies in photographs), duolingo (translate and learn) motivate by providing an educational experience.

Microlabor Incentives

- Chandler and Kapelner (2013) found that increasing “meaningfulness” of a task will increase participation and quantity but not quality. To decrease quality, the worker had to be convinced their work would *never* be looked at.

Microlabor Incentives

- Chandler and Kapelner (2013) found that increasing “meaningfulness” of a task will increase participation and quantity but not quality. To decrease quality, the worker had to be convinced their work would *never* be looked at.
- “Gamification” is a new, hot field!

Microlabor Incentives

- Chandler and Kapelner (2013) found that increasing “meaningfulness” of a task will increase participation and quantity but not quality. To decrease quality, the worker had to be convinced their work would *never* be looked at.
- “Gamification” is a new, hot field! Making the task fun via create leaderboards, collections, point systems, earning credit, skill level-ups (like a video game), timed response makes a big difference. Horton and Chilton (2010) found workers to be “target earners” who stop after reaching some predetermined goal (e.g. exactly \$1).

Microlabor Incentives

- Chandler and Kapelner (2013) found that increasing “meaningfulness” of a task will increase participation and quantity but not quality. To decrease quality, the worker had to be convinced their work would *never* be looked at.
- “Gamification” is a new, hot field! Making the task fun via create leaderboards, collections, point systems, earning credit, skill level-ups (like a video game), timed response makes a big difference. Horton and Chilton (2010) found workers to be “target earners” who stop after reaching some predetermined goal (e.g. exactly \$1).

Not so relevant for surveys, *small* labeling tasks, or randomized experiments, but very useful for massive crowdsourced projects.

Ethical Dilemmas

- In our experience, we've measured hourly wages at about \$1.12/hr (Chandler and Kapelner, 2013) and Chilton (2010) measured the reservation wage at \$1.38/hr (both *pretax*).

Ethical Dilemmas

- In our experience, we've measured hourly wages at about \$1.12/hr (Chandler and Kapelner, 2013) and Chilton (2010) measured the reservation wage at \$1.38/hr (both *pretax*). These estimates are substantially under minimum wage and the Turkers receive zero benefits.

Ethical Dilemmas

- In our experience, we've measured hourly wages at about \$1.12/hr (Chandler and Kapelner, 2013) and Chilton (2010) measured the reservation wage at \$1.38/hr (both *pretax*). These estimates are substantially under minimum wage and the Turkers receive zero benefits. Is this a “new digital sweatshop” that Harvard law professor Zittrain claims?

Ethical Dilemmas

- In our experience, we've measured hourly wages at about \$1.12/hr (Chandler and Kapelner, 2013) and Chilton (2010) measured the reservation wage at \$1.38/hr (both *pretax*). These estimates are substantially under minimum wage and the Turkers receive zero benefits. Is this a “new digital sweatshop” that Harvard law professor Zittrain claims?
- Turkers are at the whim of requesters who get to veto their work, at times dishonestly, and they do not get paid.

Ethical Dilemmas

- In our experience, we've measured hourly wages at about \$1.12/hr (Chandler and Kapelner, 2013) and Chilton (2010) measured the reservation wage at \$1.38/hr (both *pretax*). These estimates are substantially under minimum wage and the Turkers receive zero benefits. Is this a “new digital sweatshop” that Harvard law professor Zittrain claims?
- Turkers are at the whim of requesters who get to veto their work, at times dishonestly, and they do not get paid.
- Requesters are under no obligation to inform the Turker of the meaning / purpose of the HITs. Turkers could be *deceived* into contributing to something they find personally unethical without their knowledge. (We will talk about Institutional Review Board stuff later).

Ethical Dilemmas

Ethical Dilemmas

- MTurk is considered the “wild west” by some in the legal community: no governmental regulation to date (age? max hours? unionization rights?)

Ethical Dilemmas

- MTurk is considered the “wild west” by some in the legal community: no governmental regulation to date (age? max hours? unionization rights?)
- Should human mental cycles be an “undifferentiated commodity” without regulation?

Ethical Dilemmas

- MTurk is considered the “wild west” by some in the legal community: no governmental regulation to date (age? max hours? unionization rights?)
- Should human mental cycles be an “undifferentiated commodity” without regulation?

for more information see:

Horton (2011a), Kittur et al (2013), Zittrain (2008), Felstiner (2010)

Checkpoint

We've covered:

- crowdsourcing

Checkpoint

We've covered:

- crowdsourcing
- microlabor on MTurk: what it looks like, how it's created

Checkpoint

We've covered:

- crowdsourcing
- microlabor on MTurk: what it looks like, how it's created
- MTurk labor market: requesters and Turkers

Checkpoint

We've covered:

- crowdsourcing
- microlabor on MTurk: what it looks like, how it's created
- MTurk labor market: requesters and Turkers
- how industry and academia make use of MTurk

Checkpoint

We've covered:

- crowdsourcing
- microlabor on MTurk: what it looks like, how it's created
- MTurk labor market: requesters and Turkers
- how industry and academia make use of MTurk

Goal A Complete

Know what MTurk is used for in academics and industry *and thereby* recognize where crowdsourcing could be useful in your work

Surveys



Surveying on MTurk

Timer: 00:00:00 of 15 minutes

Want to work on this HIT?

Accept HIT

Total Earned: Unavailable

Total HITs Submitted: 0

Short Survey on Web Usage

Requester: Siddharth Suri

Reward: \$0.01 per HIT

HITs Available: 1

Duration: 15 minutes

Qualifications Required: HIT approval rate (%) is not less than 95

Answer a short survey about your web usage

1. Where do you find the pages you bookmark? (select all that apply)

- News sites
- E-Mail
- Social networking sites
- Web search
- Other
- I don't use bookmarks.

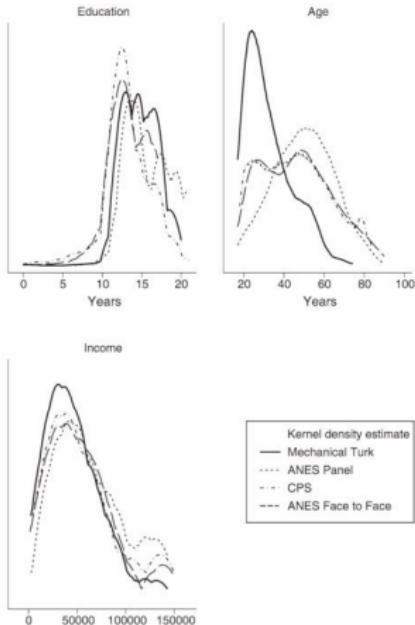
2. Do you use bookmarks more or less now than you did a year ago?

- I use bookmarks more now than a year ago.
- I use bookmarks about the same now as I did a year ago.
- I use bookmarks less now than a year ago.

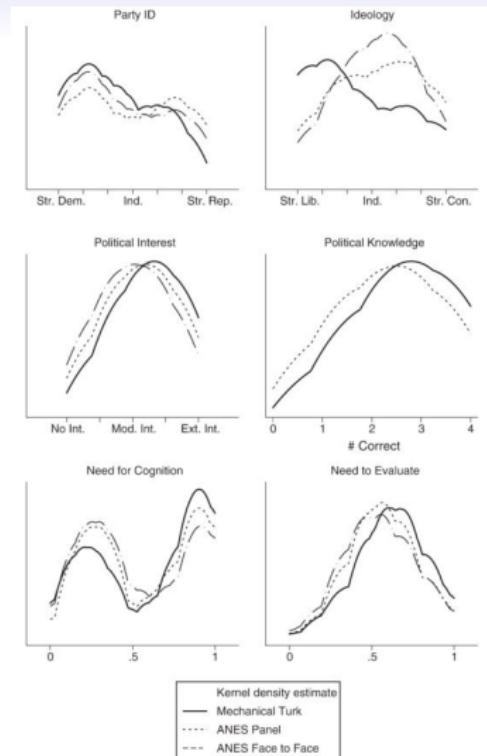
Surveys are one of the most common tasks on MTurk for *both* industry and research.

Survey Validity

The main issue in survey research is whether the survey properly reflects the population of interest. Political scientists Berinsky et al (2012) asks this question for the US population. They find the following:



Survey Validity



Survey Validity

They also studied whether or not Turkers were “chronic survey-takers” and found over 4 months, of ≈ 1600 unique Turkers, 70% took only one survey. Hence there probably is a small effect of *cross-survey stimuli*.

Survey Validity

They also studied whether or not Turkers were “chronic survey-takers” and found over 4 months, of ≈ 1600 unique Turkers, 70% took only one survey. Hence there probably is a small effect of *cross-survey stimuli*.

Conclusions: the Turkers are younger, more politically liberal, slightly less wealthy, slightly more educated, enjoy cognitively-demanding tasks slightly more than gold standard probability samples of Americans.

Survey Validity

They also studied whether or not Turkers were “chronic survey-takers” and found over 4 months, of ≈ 1600 unique Turkers, 70% took only one survey. Hence there probably is a small effect of *cross-survey stimuli*.

Conclusions: the Turkers are younger, more politically liberal, slightly less wealthy, slightly more educated, enjoy cognitively-demanding tasks slightly more than gold standard probability samples of Americans.

Thus, for survey research, you can build stratified samples.

Problems with collecting surveys

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000).

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question,

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information,

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information,

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response.

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

- Response order: more likely to fill out the first or last response on a multiple response question, thus, randomize!

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

- Response order: more likely to fill out the first or last response on a multiple response question, thus, randomize!
- Acquiescence bias: more likely say “agree / true / yes” with the surveyor than “disagree / false / no,” thus, avoid these questions!

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

- Response order: more likely to fill out the first or last response on a multiple response question, thus, randomize!
- Acquiescence bias: more likely say “agree / true / yes” with the surveyor than “disagree / false / no,” thus, avoid these questions!
- No opinion filter: more likely to pick the “no opinion” or “don’t know” option if given the opportunity, don’t give it!

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

- Response order: more likely to fill out the first or last response on a multiple response question, thus, randomize!
- Acquiescence bias: more likely say “agree / true / yes” with the surveyor than “disagree / false / no,” thus, avoid these questions!
- No opinion filter: more likely to pick the “no opinion” or “don’t know” option if given the opportunity, don’t give it!
- Rankings inaccuracy: ranking many items puts undue stress on the survey-taker, thus ratings are recommended

Problems with collecting surveys

Many respondents “cheat” or are lazy when filling out these surveys (see Krosnick, 2000). There are four steps: interpretation of question, retrieval of information, judging retrieved information, and interpreting and selecting a response. Expenditure of the minimum amount of effort to complete the survey may involve skipping some of these steps.

Other survey design problems observed:

- Response order: more likely to fill out the first or last response on a multiple response question, thus, randomize!
- Acquiescence bias: more likely say “agree / true / yes” with the surveyor than “disagree / false / no,” thus, avoid these questions!
- No opinion filter: more likely to pick the “no opinion” or “don’t know” option if given the opportunity, don’t give it!
- Rankings inaccuracy: ranking many items puts undue stress on the survey-taker, thus ratings are recommended

Krosnick recommends **minimizing task difficulty** and **maximizing subject motivation**.

A way to mitigate satisficing

Kapelner and Chandler (2010) recommend fading in each word...

A way to mitigate satisficing

Kapelner and Chandler (2010) recommend fading in each word...

Question: 3 / 32

Sports Participation

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. In order to facilitate our research on decision making we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read the instructions, please ignore the sports items below, as well as the continue button. Instead, simply click on the title at the top of this screen (i.e., "sports participation") to proceed to the next screen. Thank you very much.

Which of these activities do you engage in regularly?
(click on all that apply)

skiing soccer snowboarding running hockey
 football swimming tennis basketball cycling

Continue

The “IMC”

Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk

Fade-in results

$(N = 727)$	without controls b (se)	with controls b (se)
Treatment		
<i>Exhortation</i>	-1.0 (4.4)	-1.0 (4.4)
<i>Timing Control</i>	6.6 (4.2)	7.3 (4.2)
<i>Kapcha</i>	12.8** (4.0)	13.0*** (3.9)
Gender (male)		-7.7* (3.1)
Age (26-35)		11.4** (4.0)
Age (36-45)		16.3*** (4.3)
Age (over 45)		17.1*** (4.6)
Completed college		4.2 (2.9)
Reported motivation		1.9 (1.1)
# words in feedback		0.3*** (0.1)
Need for cognition		3.8 (2.5)
Break for $\geq 2\text{min}$		-13.1 (7.7)
Other covariates		✓
Intercept	77.4*** (3.2)	27.1 (14.3)
R^2	0.020	0.165

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Fade-in results

Table 6: Question A: Increase in willingness to pay for a soda due to subtle word changes involving whether source of soda was a fancy resort or run-down grocery store (with and without other controls)

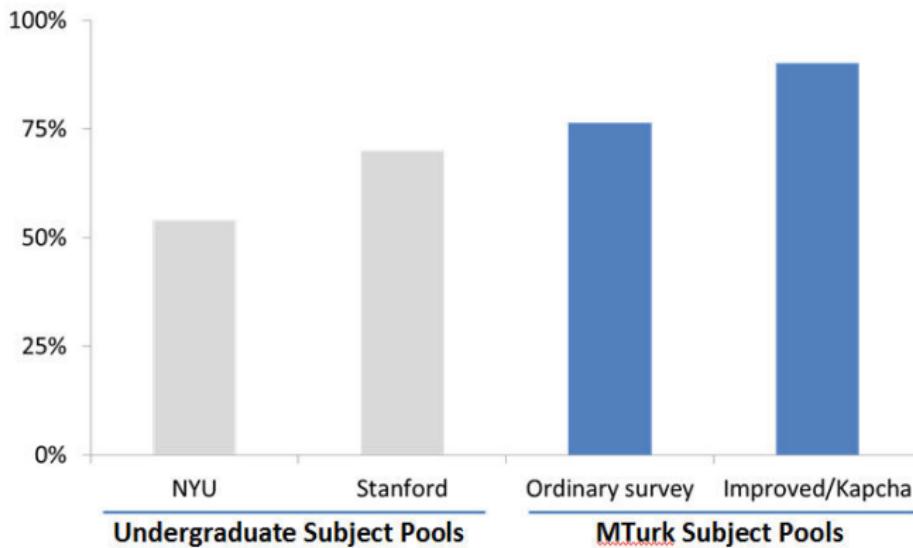
	Overall	Control	Exhortation	Timing control	Kapcha
No controls					
“fancy” b (se)	0.239** (0.086)	0.080 (0.173)	0.072 (0.176)	0.303 (0.176)	0.543*** (0.148)
<i>R</i> ²	0.011	0.001	0.001	0.015	0.077
With controls^a					
“fancy” b (se)	0.249** (0.089)	0.143 (0.180)	-0.021 (0.184)	0.319 (0.180)	0.533** (0.182)
<i>R</i> ²	0.083	0.210	0.185	0.254	0.205
<i>N</i> ^b	714	163	201	190	160

*p < 0.05, **p < 0.01, ***p < 0.001

^a Includes same controls as table 5

^b We excluded 13 prices that were not numbers between \$0 and \$10

MTurkers found to beat Ivy-leaguers



IMC pass rate by subpopulation

MTurk Survey Studies

MTurk Survey Studies

- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)

MTurk Survey Studies

- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)
- What do terms like “unlikely” and “improbable” mean to people mathematically? (Teigen et al., 2013)

MTurk Survey Studies

- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)
- What do terms like “unlikely” and “improbable” mean to people mathematically? (Teigen et al., 2013)
- How “quickly” do people perceive life has changed is changing, and will change for different age groups to investigate the “end of history illusion.” (Quoidbach et al., 2013)

MTurk Survey Studies

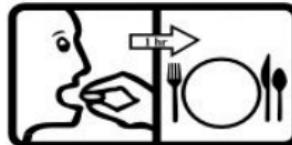
- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)
- What do terms like “unlikely” and “improbable” mean to people mathematically? (Teigen et al., 2013)
- How “quickly” do people perceive life has changed is changing, and will change for different age groups to investigate the “end of history illusion.” (Quoidbach et al., 2013)
- Ascertain levels of narcissism, psychopathy and Machiavellianism to test the validity of psychological models. (Jonason and Luevano, 2013)

MTurk Survey Studies

- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)
- What do terms like “unlikely” and “improbable” mean to people mathematically? (Teigen et al., 2013)
- How “quickly” do people perceive life has changed is changing, and will change for different age groups to investigate the “end of history illusion.” (Quoidbach et al., 2013)
- Ascertain levels of narcissism, psychopathy and Machiavellianism to test the validity of psychological models. (Jonason and Luevano, 2013)
- Measure “intolerance” differences between those leaning politically left and right (Crawford and Pilansky, 2013)

MTurk Survey Studies

- Finding: Childhood socioeconomic status influences subject likelihood to “diversify” and hedge their bets (White et al, 2013)
- What do terms like “unlikely” and “improbable” mean to people mathematically? (Teigen et al., 2013)
- How “quickly” do people perceive life has changed is changing, and will change for different age groups to investigate the “end of history illusion.” (Quoidbach et al., 2013)
- Ascertain levels of narcissism, psychopathy and Machiavellianism to test the validity of psychological models. (Jonason and Luevano, 2013)
- Measure “intolerance” differences between those leaning politically left and right (Crawford and Pilansky, 2013)
- See if medical pictograms (e.g. “take pill with food”) are understandable (Yu et al, 2013).



Checkpoint

We've covered:

Checkpoint

We've covered:

- validity / stratified sampling on MTurk

Checkpoint

We've covered:

- validity / stratified sampling on MTurk
- understanding satisficing issues on MTurk and apply it to design

Checkpoint

We've covered:

- validity / stratified sampling on MTurk
- understanding satisficing issues on MTurk and apply it to design
- a quick review of survey research on MTurk

Checkpoint

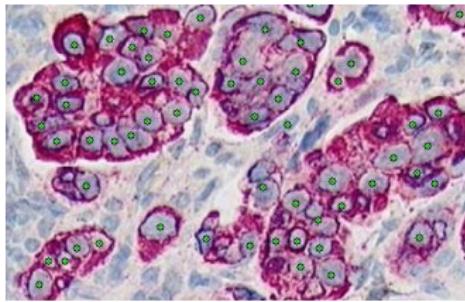
We've covered:

- validity / stratified sampling on MTurk
- understanding satisficing issues on MTurk and apply it to design
- a quick review of survey research on MTurk

Goal B Complete

Use MTurk for survey research

Labeling and Machine Learning



Crowdsourcing for labeling

- The first widely used application
- Often machine vision or language
 - Objects in images (or their features)
 - Word or document sense, sentiment, translation, ...
 - Audio transcription

Collection often goes beyond labels: proposed translation or summary

The labeling process

- Define task and make instructions
 - HITs are usually fast and simple
- Pretest in-house and refine instructions
- Get IRB approval or exemption
- Decide on collection process
 - How to select items to label?
 - How many labels?
 - Batch or adaptive item selection?
- Code M-turk application
- Collect data
- Screen for spammers
- Estimate model

Test HITS in-house

- What is confusing?
- How high is inter-annotator agreement?
- Refine instructions

How positive or negative or negative is the sentiment?

“I just bought a Corolla”

“I like my friends but their driving me crazy”

Get IRB approval

- Institutional Review Board (IRB) approval or exemption
- Takes longer than you think
- Inconsistent across schools
- Big issue: privacy
- Usually expedited due to being exempt under category #2

Code M-turk application

- How hard is it?
 - a couple programmer weeks to modify existing code to a new task
- What code exists to do this already?
- Gotchas?

Screen for spammers

- Not a problem in our experience – but others have different experiences
- Single turker methods
 - speed, variety, accuracy
- Agreement with other annotators

Case Study: Word Sense Disambiguation

"The level of good-faith participation by Turkers is surprisingly high, given the generally small nature of the payment."
(Callison-Burch, 2009)



Kapelner, A, Kaliannan, K, Schwartz, A, Ungar, L, Foster, D, New Insights from Coarse Word Sense Disambiguation in the Crowd In Proceedings of CoLING, 2012, pp539–548

Case Study: Word Sense Disambiguation

"The level of good-faith participation by Turkers is surprisingly high, given the generally small nature of the payment."
(Callison-Burch, 2009)



Kapelner, A., Kaliannan, K., Schwartz, A., Ungar, L., Foster, D., New Insights from Coarse Word Sense Disambiguation in the Crowd In Proceedings of CoLING, 2012, pp539–548

What is Word Sense Disambiguation? (WSD)

Language is ambiguous and many words can denote different "meanings"

- I can hear *bass* sounds.
- They like grilled *bass*.

WSD has been described as "AI-complete" which means it is equivalent to solving central problems in artificial intelligence *a la* the Turing Test.

Our HIT: the WSD labeling task

Word Meaning Task

Read the following snippet which will fade in slowly:

Apple shares fell 75 cents in over-the-counter trading to close at \$48 a share. Fiscal fourth-quarter sales grew about 18% to \$1.38 billion from \$1.17 billion a year earlier. Without the Adobe gain, Apple's full-year operating profit edged up 1.5% to \$406 million, or \$3.16 a **share**, from \$400.3 million, or \$3.08 a share. Including the Adobe gain, full-year net was \$454 million, or \$3.53 a share. Sales for the year rose nearly 30% to \$5.28 billion from \$4.07 billion a year earlier.

Please pick the meaning of the word **share** which best fits the context of the paragraph above:

- capital stock in a corporation
- a tool for tilling soil
- a portion or percentage of a whole

Submit my definition of "share" (and whatever optional feedback I left below)

My feedback:

- A subset of the Penn Treebank Wall Street Journal Corpus
- A collapsed version of Wordnet 2.1
- Nouns and verbs only

This data is from the “OntoNotes” project (2006). They iteratively make the senses more and more “coarse-grained” until they achieve a 90% ITA among the annotators. For example, the word “share:”

One snippet, one word, all coarse senses.
Optional feedback box. Prevent cheating /
satisficing by (1) fading in the words slowly
(2) randomly ordering the senses.

1,000 words, each disambiguated by 10
Turkers each

The Data and the questions we want to answer

595 Turkers, 10,000 HITs at 10¢ each

total time: 51hr — throughput: 4700 labels / day with cost: \$110.

Krippendorf's $\alpha = 0.66$ which is fair interrater reliability.

We tried to answer the following in data analysis:

- Can we combine Turker responses to boost accuracy?
- Is the workers' performance heterogeneous?
- Which features in the target word, the snippet text, and the text of the sense choices affect accuracy?
- Which characteristics in the Turker's engagement of the task affect accuracy?

Today: the first two goals.

Boosting Accuracy

Can we combine Turker responses naively to boost accuracy?

Combine annotations and use plurality vote arbitrating ties randomly:

# of Dis	2	3	4	5	6	7	8	9	10	2.4 (1st pl)
Accuracy	.73	.80	.81	.82	.83	.84	.84	.84	.86	.81

Combining 10 untrained Turkers is competitive with the accuracy of state-of-the-art algorithms.

Ipeirotis, 2011 claims that “majority vote works best when workers have similar quality” and if not:

- Find best worker and use that label
- Model worker quality and combine

Using labels from different-quality workers

Dawid and Skene (1979) proposed an expectation-maximization algorithm:

- 0 Initialize with the aggregate labels being the majority vote
- 1 Estimate confusion matrix for each worker
- 2 Estimate aggregate labels again (if gold data exists, keep it)
- 3 Repeat steps 2-3 until convergence criteria is met

For a bayesian version see Raykar et al (2010).

Using labels from different-quality workers

Ipeirotis et al (2010a) is able to tease out bias and error. For K categories, cost matrix $\{c_{ij}\}$, and soft labels p_1, \dots, p_K for a given worker,

$$\text{Expected Cost}(p_1, \dots, p_K) := \sum_{i=1}^K \sum_{j=1}^K p_i p_j c_{ij}$$

Assigned Label	"Soft" Label	Cost
G	<G: 25%, P: 25%, R: 25%, X: 25%>	0.75
G	<G: 99%, P: 1%, R: 0%, X: 0%>	0.0198

Using labels from different-quality workers

There are high costs when probabilities are diffuse and low costs when less diffuse. Now compute a quality score, where “spammer” is defined as a discrete uniform distribution over labels:

$$\text{Quality Score}(\text{worker}) := 1 - \frac{\text{Expected Cost (worker)}}{\text{Expected Cost (spammer)}}$$

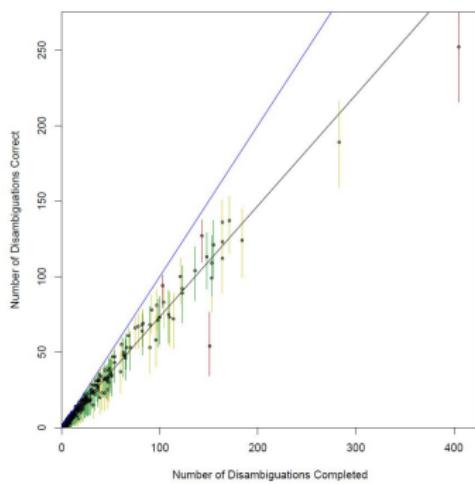
Plug this into the EM algorithm on previous slide.

Conclusions: you may want to have some gold data when benefit of learning quality differences outweighs cost of gold data (Wang et al, 2011), try to weight workers appropriately when combining.

Are all workers the same in our task?

Do we have spammers and superstars? Is anyone different from the average, $p_{\text{avg}} = .73$? Let C_w be the number correct for Turker w.

$$H_0 : C_1, C_2, \dots, C_W \stackrel{\text{ind}}{\sim} \text{Binomial}(n_w, p_{\text{avg}})$$

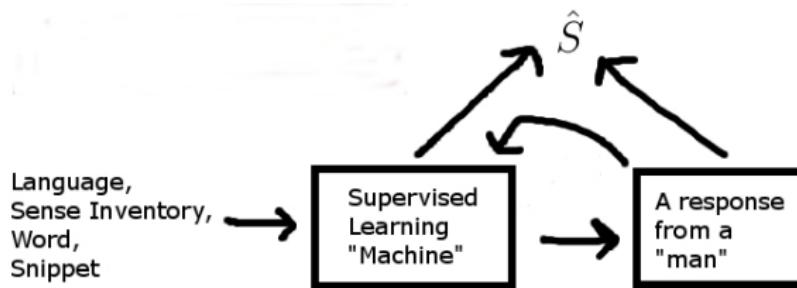


(with Bonferroni-corrected Binomial proportion CI's $\alpha = 5\%$).

We are forced to reject on the account of four workers.
To a first order approximation, all Turkers are the same. Also note: no learning effect was found (unshown).

The Goal: Active Learning Loop

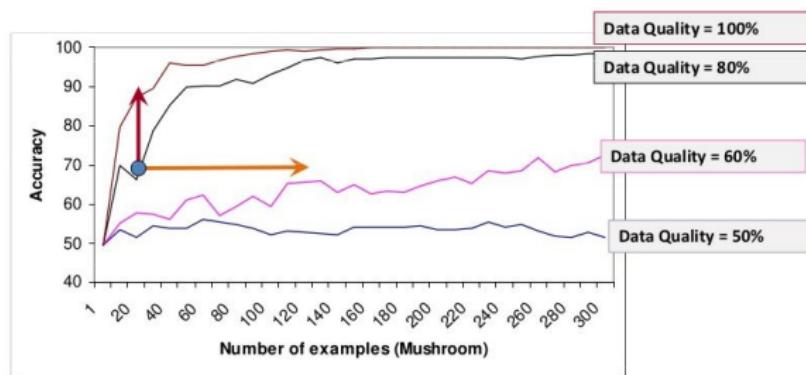
“Crowdsourcing is cheap, but not free” — so use active learning.



Here, \hat{S} is a machine learning classifier. This “active learning” model uses \hat{S} when the machine is “confident” otherwise humans. The human data is then integrated into the model, and \hat{S} improves.

Integrating MTurking with machine learning

Tradeoff between getting *higher quality* data (either via combining multiple labels or getting better Turkers) and getting *more* data.



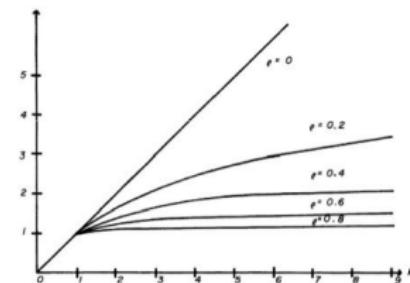
(Ipeirotis, WWW'11)

MTurk labeling data often has block structure

Building \hat{S} requires a machine learning algorithm which usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?



Thus, $n_{\text{eff}} \in [12, 31]$, an inconvenient reality that off-the-shelf ML algorithms do not consider:

$$\mathcal{E} \sim e \left(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{12} \end{bmatrix} \right)$$

Handling the block structure

- Non-parametric, black-box algorithms which model the classification or regression function cannot usually support a correlation structure natively.
- Prior work includes a vast literature in the medical domain.
- Typical simple methods use random-effects linear models or maximum likelihood linear models ($\theta = [\rho, \beta]$). Use interactions built with your favorite model builder (stepwise, etc). For classification, you can use generalized estimating equations (Liang and Zeger, 1986).
- More recent work uses methods like random forests for cluster-correlated data (Karpievitch et al., 2009) or random-effects expectation-maximization trees (Sela and Simonoff, 2011)

Handling the block structure

Our best advice: ignore dependence if:

- you built a sophisticated linear models and test $H_0 : \rho = 0$ (for random effects: $\tau^2 = 0$) and it fails to reject
- your out-of-sample RMSE on gold data for the black boxes is much higher than models estimating intercepts or correlations, you may just want to ignore dependence since you're getting more juice out of the non-parametric ability of the black box.

Group Project

Spend 5 minutes with your neighbors coming up with a labeling task that integrates with a machine learning classifier.



We'll discuss feasibility of projects.

Some literature on MTurk for labeling

- Classic work crowdsourcing natural language processing tasks (Snow et al., 2008)
- Document relevance (Agrawal et al., 2009)
- Machine translation of different languages (Callison-Burch, 2009)
- Personality and “well-being” on Facebook (Schwartz, Kapelner, Ungar et al., 2013)
- Similarity of different melodies (Urbano et al., 2010)
- Transcription (Marge et al., 2010)
- Extract medical terms from patient-written text (Maclean et al., 2013)

Checkpoint

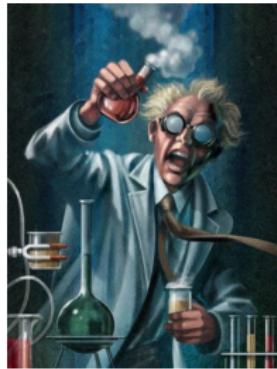
We've covered:

- A natural language processing case study about labeling on MTurk
- Handling noisy / biased labels
- Using data to create a machine learning classifier
- Handling label dependence due to worker portfolios

Goal C Complete

- C For crowdsourced labeling tasks:
- 1 Evaluate if the labeling task is suitable for crowdsourcing
 - 2 Design the labeling task
 - 3 Analyze the resulting data

Experiments



Case Study

“Conduct a wide range of experiments involving potentially large numbers of participants (hundreds or even thousands, but probably not millions) quickly and cheaply” - Mason and Watts (2010)

Case Study

“Conduct a wide range of experiments involving potentially large numbers of participants (hundreds or even thousands, but probably not millions) quickly and cheaply” - Mason and Watts (2010)



Chandler, D. and Kapelner, A. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets, *Journal of Economic Behavior & Organization* 90, 123-133, 2013

Study testing the effect of “meaningfulness” on three outcome measures: labor supply, work quantity and work quality.

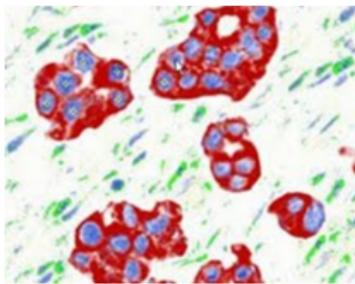
Case Study

“Conduct a wide range of experiments involving potentially large numbers of participants (hundreds or even thousands, but probably not millions) quickly and cheaply” - Mason and Watts (2010)



Chandler, D. and Kapelner, A. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets, *Journal of Economic Behavior & Organization* 90, 123-133, 2013

Study testing the effect of “meaningfulness” on three outcome measures: labor supply, work quantity and work quality.



Directions: click on “tumor cells” / “objects of interest.”

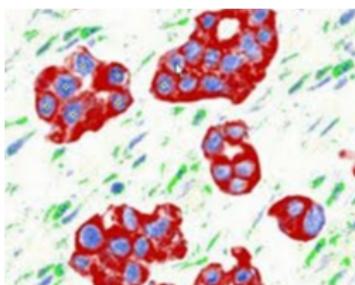
Case Study

“Conduct a wide range of experiments involving potentially large numbers of participants (hundreds or even thousands, but probably not millions) quickly and cheaply” - Mason and Watts (2010)



Chandler, D. and Kapelner, A. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets, *Journal of Economic Behavior & Organization* 90, 123-133, 2013

Study testing the effect of “meaningfulness” on three outcome measures: labor supply, work quantity and work quality.



Directions: click on “tumor cells” / “objects of interest.” Experiment run in US and India for comparison.



Treatments and Hypotheses

Three experimental treatments:

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Hypotheses: Monotonically with *meaningfulness*, more workers will...

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Hypotheses: Monotonically with *meaningfulness*, more workers will...

- “show up” (*i.e.* the labor supply increases)

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Hypotheses: Monotonically with *meaningfulness*, more workers will...

- “show up” (*i.e.* the labor supply increases)
- work longer (*i.e.* on more images)

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Hypotheses: Monotonically with *meaningfulness*, more workers will...

- “show up” (*i.e.* the labor supply increases)
- work longer (*i.e.* on more images)
- work more accurately (*i.e.* clicking closer to the center)

Treatments and Hypotheses

Three experimental treatments:

- *meaningful* treatment: Turkers were told that they were helping cancer researchers mark tumor cells in medical images
- *zero-context* treatment: Turkers were told to mark “objects of interest” in images
- *shredded* treatment: Turkers were told to mark “objects of interest” in images *and* their labelings will be discarded upon submission

Hypotheses: Monotonically with *meaningfulness*, more workers will...

- “show up” (*i.e.* the labor supply increases)
 - work longer (*i.e.* on more images)
 - work more accurately (*i.e.* clicking closer to the center)
- + Minor hypotheses about US vs. India

The HIT Listing

amazon mechanical turk Artificial Artificial Intelligence

Your Account HITs Qualifications 164,523 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Search for **HITs** containing that pay at least \$ 0.00 for which you are qualified Go

HITs

1-10 of 164,523 Results

Sort by: **HIT Creation Date (newest first)** Go! Show all details | Hide all details 1 2 3 4 5 > Next >> Last

Find Objects of Interest in Images (updated) --- \$0.10USD + unlimited bonus!! View a HIT in this group

Requester: [Robert Simmons](#) **HIT Expiration Date:** Mar 23, 2010 (3 hours 55 minutes) **Reward:** \$0.10
Time Allotted: 4 hours **HITs Available:** 1

Description: You will be given an image and "objects of interest" to find. Your task is to click on the "objects of interest". Before beginning, there will be a brief tutorial. After completing the first image, there will be a potential opportunity to complete ***unlimited*** similar HITs. This hit is now updated.

Keywords: [find](#), [objects](#), [images](#), [click](#), [on](#), [areas](#), [of](#), [interest](#), [object](#), [labeling](#), [computer](#), [vision](#)

Qualifications Required: Your Value

Location is US US You meet this qualification requirement [Contact the Requester of this HIT](#)

Note the fake name.

The HIT Listing

The screenshot shows the Amazon Mechanical Turk HIT Listing interface. At the top, there are tabs for 'Your Account', 'HITs' (which is selected), and 'Qualifications'. To the right, it says '164,523 HITs available now'. Below the tabs, there's a search bar with the placeholder 'Search for HITs containing [text] that pay at least \$ [amount] for which you are qualified' and a 'GO' button.

HITs
1-10 of 164,523 Results

Sort by: HIT Creation Date (newest first) Show all details | Hide all details 1 2 3 4 5 > Next >> Last

Find Objects of Interest in Images (updated) --- \$0.10USD + unlimited bonus!!

Requester:	Robert Simmons	HIT Expiration Date:	Mar 23, 2010 (3 hours 55 minutes)	Reward:	\$0.10
		Time Allotted:	4 hours	HITs Available: 1	

Description: You will be given an image and "objects of interest" to find. Your task is to click on the "objects of interest". Before beginning, there will be a brief tutorial. After completing the first image, there will be a potential opportunity to complete ***unlimited*** similar HITs. This hit is now updated.

Keywords: [find](#), [objects](#), [images](#), [click](#), [on](#), [areas](#), [of](#), [interest](#), [object](#), [labeling](#), [computer](#), [vision](#)

Qualifications Required: Your Value

Location is US US You meet this qualification requirement

Note the fake name. Old screenshot: corrections: each HIT lasted one hour only, then it was reposted. We reposted in batches of 1000 so it would appear to the Turkers that many HITs were available.

The HIT Preview

amazon mechanical turk
Artificial Artificial Intelligence

Your Account HITS Qualifications 35,995 HITS available now

Robert Simmons | Account Settings | Sign Out | Help

Search for: HITS containing

Timer: 00:00:00 of 4 hours

All HITS | HITS Available To You | HITS Assigned To You

that pay at least \$ 0.00 for which you are qualified (6)

Want to work on this HIT? Want to see other HITs?

Accept HIT **Skip HIT**

Total Earned: \$0.95
Total HITs Submitted: 13

Find Objects of Interest in Images (updated!) — \$0.10/USD + unlimited bonus!!
Requester: Robert Simmons
Qualifications Required: Location is US

Click to accept

Our box →

Object identification in images

In this HIT, you will be presented with images that have "objects of interest" to click on. After completing your first image, you will be able to work on an **unlimited** number of additional images of comparable difficulty. The first image pays more to compensate for training.

Be careful not to leave or return this HIT, because you will not have a chance to do another one after leaving.

Before doing the first image, we will show you a short instructional video explaining how to identify the objects. You will then be asked to answer a few questions to confirm your understanding.

Additionally, this HIT only works if you have a Windows operating system, working sound, and Firefox.

We would also appreciate any feedback you have. We believe we've worked out all the major technical issues. However, if you notice something strange, we would appreciate any comments you have (rsimmons2010@gmail.com).

Click on "Accept HIT" above to get started.

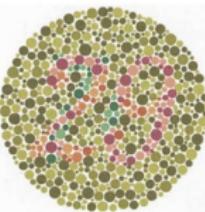
Reward: \$0.10 per HIT HITS Available: 54 Duration: 4 hours

Payment amount = 10 cents + *unlimited bonus*

The pre-HIT

Since the tasks you will perform require you to be able to differentiate color, we have to ask you a few questions that will determine if you may be colorblind.

1. Look at the below image.



Do you see a number? If so, enter it into this box:

2. Are you male or female?

Male Female

3. Have you ever had trouble differentiating between reds and greens?

Yes No

4. Have you ever had trouble differentiating between blues and yellows?

Yes No

5. How old are you?

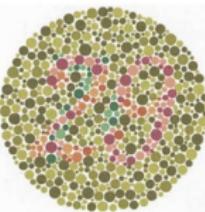
6. Listen to the following sound clip () and enter the word below:

Check demographics,

The pre-HIT

Since the tasks you will perform require you to be able to differentiate color, we have to ask you a few questions that will determine if you may be colorblind.

1. Look at the below image.



Do you see a number? If so, enter it into this box:

2. Are you male or female?

Male Female

3. Have you ever had trouble differentiating between reds and greens?

Yes No

4. Have you ever had trouble differentiating between blues and yellows?

Yes No

5. How old are you?

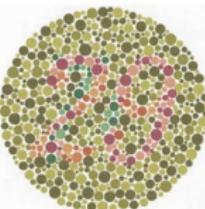
6. Listen to the following sound clip () and enter the word below:

Check demographics, eliminate spam,

The pre-HIT

Since the tasks you will perform require you to be able to differentiate color, we have to ask you a few questions that will determine if you may be colorblind.

1. Look at the below image.



Do you see a number? If so, enter it into this box:

2. Are you male or female?

Male Female

3. Have you ever had trouble differentiating between reds and greens?

Yes No

4. Have you ever had trouble differentiating between blues and yellows?

Yes No

5. How old are you?

6. Listen to the following sound clip () and enter the word below:

Check demographics, eliminate spam, ensure audio works (few Turkers excluded).

Note that this was all custom-built software externally hosted using Ruby-on-Rails. This was *not* built using Amazon's HIT builder tools.

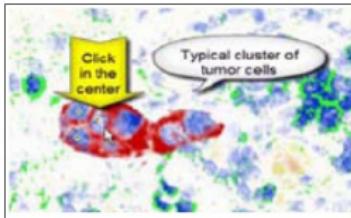
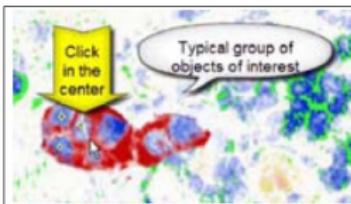
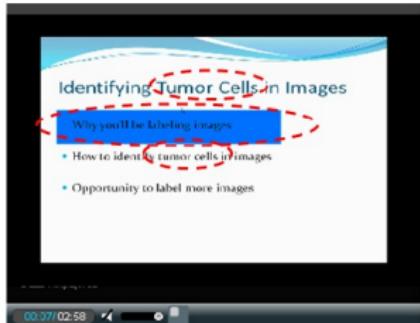
Instructional Video

Now... randomize... to M, Z, S

Instructional Video

Now... randomize... to M, Z, S

How to Identify Tumor Cells



Please pay close attention to this video and watch it in FULL SCREEN. This video will explain how to identify cancerous tumor cells. Afterwards, you will have to take a brief quiz based on the video that will be very easy if you've watched the video.

Instructional Video

Now... randomize... to M, Z, S

How to Identify Tumor Cells

Identifying Tumor Cells in Images

- Who you'll be labeling images
- How to identify tumor cells in images
- Opportunity to label more images

Please pay close attention to this video and watch it in FULL SCREEN. This video will explain how to identify cancerous tumor cells. Afterwards, you will have to take a brief quiz based on the video that will be very easy if you've watched the video.

3min instructional video with 15 treatment cues in the video, e.g

[Your job will be to help identify tumor cells in images and we appreciate your help. /
In this task, you'll look at images and find objects of interest.]

Instructional Video

~~Scientists constantly producing images that need to be analyzed~~

- Scientists are always producing images
- It takes lots of time to analyze the images



The “Quiz”

How to Identify Tumor Cells



Please answer the below questions. Once you answer these questions, you will be qualified to help identify tumor cells.

1 - You should adjust the magnification in order to...

- Make a prettier picture
- Make the tumors exactly 10 pixels across
- Find tumors and make clicks as close to the center as possible

2 - When you are clicking on tumor cells, how many times do you click on the tumor?

- Once
- Twice
- As many dots as you can fit inside the tumor

3 - When you incorrectly click on an area, you should...

- Give up the HIT
- Reload the page
- Use the right mouse button

Continued here

Use the right mouse button

4 - What will happen if you don't accurately click on the tumor?

Scientists who are depending on you to identify tumors will not have accurate results.

- Your HIT may be rejected
- You will not be allowed to do additional HITs with us
- All of the above.

5 - Your HIT will be rejected if...

- You do not click on all the tumors
- You don't click in the center of the tumor
- You click multiple times on the same tumor or on things that are NOT tumors
- All of the above.

6 - As part of this HIT, how many images will you have the chance to identify assuming you correctly label the tumors?

- As many as you want within a 4hr period (as long as you complete each image within 15 minutes)
- You can label up to 4 more images
- You must submit since you can only label one image

Previous Next

The Experimental Task

Label as many as you want at a decreasing wage scale: 10¢, 9¢, 8¢, 7¢, ..., 2¢, 2¢, ...

The Experimental Task

Label as many as you want at a decreasing wage scale: 10¢, 9¢, 8¢, 7¢, ..., 2¢, 2¢, ...

Submit this task and do another task for more pay.

Left click to make a point, right click to delete a point
If you do not see the image below, reduce the magnification, or restart your browser.

Total earnings: \$0.00 Time left: 14:45

Magnification: (0.8x) DO NOT RELOAD THE PAGE - YOU WILL LOSE YOUR POINTS

Find all cancerous tumor cells in the image below:

Tumor

Top Reasons HITs are Rejected

1. Tumors were not clicked on at all.
2. Tumors were marked multiple times. There should only be one unique click per tumor.
3. Tumors were not marked in their centers. They should be marked in the centers.
4. Too many tumors were missed (in the case of the goal being all tumors).

The Experimental Task

Label as many as you want at a decreasing wage scale: 10¢, 9¢, 8¢, 7¢, ..., 2¢, 2¢, ...

Submit this task and do another task for more pay.

Left click to make a point, right click to delete a point
If you do not see the image below, reduce the magnification, or restart your browser.

Total earnings: \$0.00 Time left: 14:45

Magnification: (0.8x) DO NOT RELOAD THE PAGE - YOU WILL LOSE YOUR POINTS

Find all cancerous tumor cells in the image below:

Tumor

Top Reasons HITs are Rejected

1. Tumors were not clicked on at all.
2. Tumors were marked multiple times. There should only be one unique click per tumor.
3. Tumors were not marked in their centers. They should be marked in the centers.
4. Too many tumors were missed (in the case of the goal being all tumors).

This training screen is an adaptation of DistributeEyes, a project funded by Susan Holmes of Stanford University under NIH Grant #R01GM086884-02.

Do Another Task?

Thanks for helping to locate the objects of interest in this image! We really appreciate your help.

Your work will be reviewed in the next 15-30 minutes and you will be paid.

You may do another image of similar difficulty especially for you. [Click on the button below:](#)

Click to do
a HIT for
\$0.09

Note: You do not have to go through training again

If you are finished, you can click below:

No more tasks. I am finished. Pay me \$0.10

Thanks for helping to locate the tumor cells in this image! We really appreciate your help.

Your work will be reviewed in the next 15-30 minutes and you will be paid.

You may do another image of similar difficulty especially for you. [Click on the button below:](#)



Click to do
a HIT for
\$0.09

Note: You do not have to go through training again

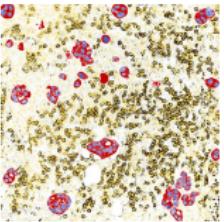
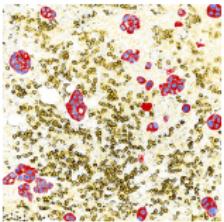
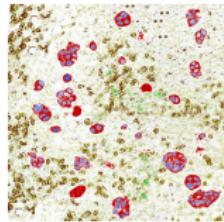
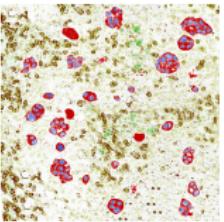
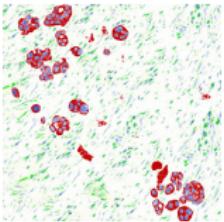
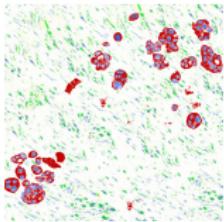
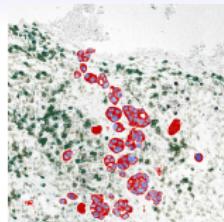
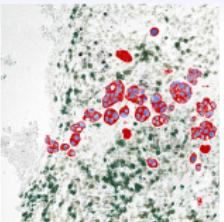
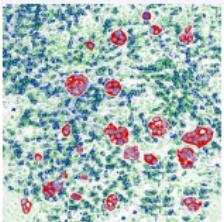
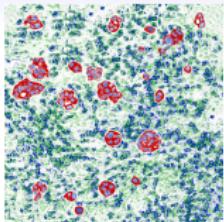
If you are finished, you can click below:

No more tasks. I am finished. Pay me \$0.10

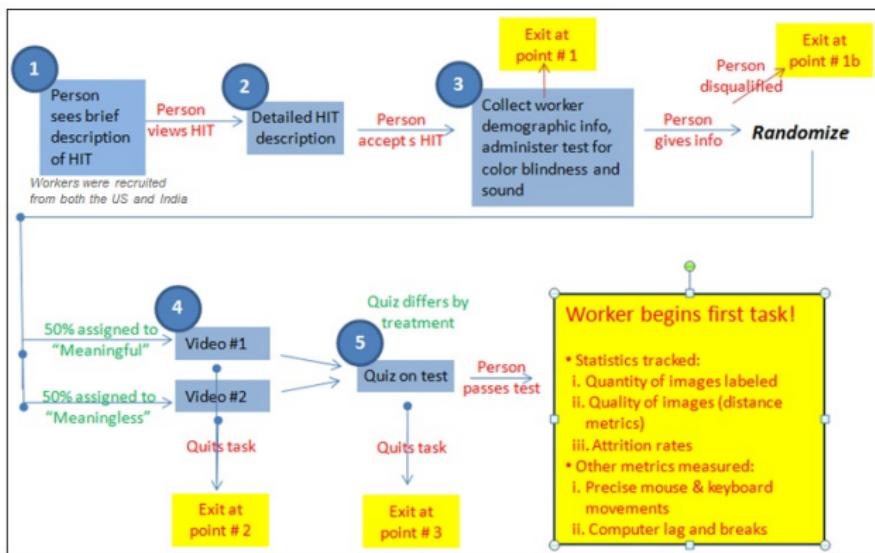
(a) Zero-context / Shredded treatments

(b) Meaningful treatment

Possible Images



The Experimental Flow



Outcome metrics

- Attrition: when did the worker leave?

Outcome metrics

- Attrition: when did the worker leave?
- Number of images labeled

Outcome metrics

- Attrition: when did the worker leave?
- Number of images labeled
- Quality / Accuracy of labeling:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{F-measure} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{average centrality} = \frac{1}{n} \sum_{i=1}^n |\vec{p}_{i_{\text{trained}}} - \vec{p}_{i_{\text{true}}}|$$

fine quality: recall at 2px

Admin Portal: Dashboard

→ Num Workers

→ Num Workers Qualified

→ Induced to do work

→ Avg images done

	US		IN		By Country		By Treatment		Total
	M	L	M	L	US	IN	M	L	
Num Workers	85	84	80	102	169	182	165	186	351
Num Workers Qualified	62	60	27	30	122	57	89	90	179
Induced to do work	87% (67)	76% (68)	67% (36)	56% (48)	81% (135)	61% (84)	80% (103)	68% (116)	74% (219)
Avg images done	4.2	4.1	7.4	5.7	4.1	6.5	5.2	4.6	4.9
Median images done	3	2	3	4	2	3	3	3	3
Avg Quality (F-meas %)	96.9%	97.3%	95.6%	95.7%	97.1%	95.6%	96.3%	96.6%	96.5%
Avg Time to Qualify (min)	6.4	5.5	17.0	18.2	6.0	17.6	9.6	9.8	9.7
1st Time Clicking (min)	4.9	4.8	5.0	5.9	4.9	5.5	4.9	5.2	5.1
Avg Time Clicking (min)	4.2	4.1	3.5	4.2	4.1	3.8	3.9	4.2	4.0
Avg Total Time (min)	29.2	29.3	54.3	54.4	29.2	54.4	36.8	37.8	37.2

→ Percentage Pipeline

Percent Exit	US		IN		By Country		By Treatment		Total
	M	L	M	L	US	IN	M	L	
N	85	84	80	102	169	182	165	186	351
Acceptance	21%	19%	55%	53%	20%	54%	38%	38%	38%
Colorblindness Test	6%	8%	11%	14%	7%	13%	8%	11%	10%
Video	0%	1%	0%	4%	1%	2%	0%	3%	1%
Qualification	5%	10%	4%	3%	7%	3%	4%	6%	5%
Did at least one good image	68%	62%	30%	26%	65%	28%	50%	42%	46%

Controlling for time-of-day effects

```
0 * * * * /usr/local/bin/ruby -C  
..../public/web/current/script/runner  
‘‘HITMaker.create_hit_sets(1000)’’ -e production
```

Create 1000 HITs every hour, on the hour via Linux’s “CRON jobs.” Then just wait as the data rolls in.

Controlling for time-of-day effects

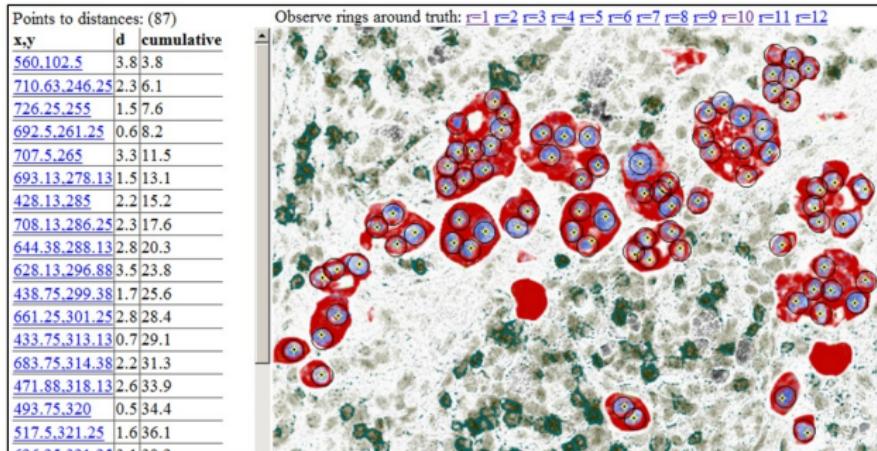
```
0 * * * * /usr/local/bin/ruby -C  
..../public/web/current/script/runner  
‘‘HITMaker.create_hit_sets(1000)’’ -e production
```

Create 1000 HITs every hour, on the hour via Linux’s “CRON jobs.” Then just wait as the data rolls in.



Just crank out the data...

Admin Portal: Subject Analytics



Admin Portal: Labeling Analytics

Images labeled by Worker # 135 (5 total)								Total earnings: \$0.40 Total time training: 19.1min Hourly wage: \$1.26		
ID	Set # / Wage	Prec / Rec / F	Accepted?	Submitted	Duration (min)	Longest Break (s)	Pts Deleted	Num Zooms	U Z	
2077	1 / \$0.10	<u>96.7%</u> / <u>96.7%</u> <u>96.7%</u>	Yes	02/02 09:34 CST	5.7 (5.1)	36	8	13		
2083	2 / \$0.09	<u>97.8%</u> / <u>96.7%</u> <u>97.7%</u>	Yes	02/02 09:39 CST	3.9 (3.8)	9	16	4		
2087	3 / \$0.08	<u>98.9%</u> / <u>96.6%</u> <u>97.7%</u>	Yes	02/02 09:41 CST	2.7 (2.5)	10	4	6		
2088	4 / \$0.07	<u>97.8%</u> / <u>97.8%</u> <u>97.8%</u>	Yes	02/02 09:45 CST	3.4 (3.2)	15	6	11		
2089	5 / \$0.06	<u>96.7%</u> / <u>97.8%</u> <u>97.2%</u>	Yes	02/02 09:49 CST	3.4 (3.2)	9	6	4		
2090				-----	untouched	-----	-----	-----		

Main Results

$n = 2,471$

Main Results

$n = 2,471$ in 36 days, 68 subjects / day

Main Results

$n = 2,471$ in 36 days, 68 subjects / day

Total cost: \$789

Main Results

$n = 2,471$ in 36 days, 68 subjects / day

Total cost: \$789 (31¢ / subject),

Main Results

$n = 2,471$ in 36 days, **68 subjects / day**

Total cost: \$789 (31¢ / subject), total time labored: 701hr

Main Results

$n = 2,471$ in 36 days, **68 subjects / day**

Total cost: \$789 (31¢ / subject), total time labored: 701hr

	Induced to work	Did ≥ 5 labelings	Fine Quality	Average Hourly Wage
Meaningful	↑ 4.6%*	↑ 8.5%***	↑ 0.7%	↓ 4.5%
Meaningful (US)	↑ 5.1%*	↑ 8.9%**	↑ 3.9%	↓ 7.7%
Meaningful (India)	↓ 2.3%	↑ 7.0%*	↓ 3.1%	↑ 0.5%
Shredded	↓ 4.0%	↓ 2.8%	↓ 7.2%***	↑ 5.6%
Shredded (US)	↓ 2.3%	↓ 5.0%	↓ 6.1%*	↑ 9.5%
Shredded (India)	↓ 6.8%	↓ 1.6%	↓ 8.7%**	↓ 1.4%

Baseline for zero context group:

76% / 37% / 65% / \$1.41

25% of Turkers left feedback

25% of Turkers left feedback

Meaningful

"I am hoping you will be offering more HITs like these in the future. It's always nice to have a requester with HITs that do take some thought and mean something to complete"

"nice goal i was a cancer patient i respect the efforts of the researchers"

"Was this really for research, or was this a social psych experiment?"

"I kept thinking of the persons for whom I was confirming the presence of cancer cells, somewhat like I was reading out the death sentence to them. I felt good at completing this task professionally, but there was a profound sense of sadness."

25% of Turkers left feedback

Meaningful

"I am hoping you will be offering more HITs like these in the future. It's always nice to have a requester with HITs that do take some thought and mean something to complete"

"nice goal i was a cancer patient i respect the efforts of the researchers"

"Was this really for research, or was this a social psych experiment?"

"I kept thinking of the persons for whom I was confirming the presence of cancer cells, somewhat like I was reading out the death sentence to them. I felt good at completing this task professionally, but there was a profound sense of sadness."

Zero-Context

"Didn't see the point of doing, only kept doing in hopes of payout increasing"

"Tell people what it is accomplishing to make more interesting."

"I'm wondering if this isn't actually a psychological test seeing how low people would go (monetary wise)"

25% of Turkers left feedback

Meaningful

"I am hoping you will be offering more HITs like these in the future. It's always nice to have a requester with HITs that do take some thought and mean something to complete"

"nice goal i was a cancer patient i respect the efforts of the researchers"

"Was this really for research, or was this a social psych experiment?"

"I kept thinking of the persons for whom I was confirming the presence of cancer cells, somewhat like I was reading out the death sentence to them. I felt good at completing this task professionally, but there was a profound sense of sadness."

Zero-Context

"Didn't see the point of doing, only kept doing in hopes of payout increasing"

"Tell people what it is accomplishing to make more interesting."

"I'm wondering if this isn't actually a psychological test seeing how low people would go (monetary wise)"

Shredded

"That was sort of fun. I have no idea what the point of it was though."

"I felt a little silly since the results weren't going to be used aside from testing."

"Here's a HIT I'd really like to know the purpose of!"

Data Analysis

Simple linear model, test coefficient of treatment indicators:

$$\begin{aligned}Y &= \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \\ \varepsilon &\sim \mathcal{N}_n(0, \sigma^2 I_n)\end{aligned}$$

Data Analysis

Simple linear model, test coefficient of treatment indicators:

$$\begin{aligned}Y &= \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \\ \varepsilon &\sim \mathcal{N}_n(0, \sigma^2 I_n)\end{aligned}$$

For quality, we use quality metrics for each image in a Turker's portfolio.
Either use a random intercept model or cluster the standard errors:

$$\begin{aligned}Y &= \beta_0 + \gamma_w + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \\ \varepsilon &\sim \mathcal{N}_n(0, \sigma^2 I_n) \quad \text{or...}\end{aligned}$$

Data Analysis

Simple linear model, test coefficient of treatment indicators:

$$\begin{aligned} Y &= \beta_0 + \beta_T \mathbf{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \\ \varepsilon &\sim \mathcal{N}_n(0, \sigma^2 I_n) \end{aligned}$$

For quality, we use quality metrics for each image in a Turker's portfolio.
 Either use a random intercept model or cluster the standard errors:

$$Y = \beta_0 + \gamma_w + \beta_T \mathbf{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon,$$

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n) \quad \text{or...}$$

$$Y = \beta_0 + \beta_T \mathbf{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon,$$

$$\varepsilon \sim \mathcal{N}_n\left(0, \sigma^2 \begin{bmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ & & \ddots & \\ 0 & \dots & 0 & D \end{bmatrix}\right), \quad D = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ & & \ddots & \\ \rho & \dots & \rho & 1 \end{bmatrix}$$

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Solutions to this missing data problem?

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Solutions to this missing data problem?

- Listwise deletion? Maybe, but it’s quite possible opting-out is *non-ignorable*.

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Solutions to this missing data problem?

- Listwise deletion? Maybe, but it’s quite possible opting-out is *non-ignorable*.
- Multiple imputation using known covariates? Maybe, unless $\rho(X_j, Y)$ ’s are large, this is as good as using \bar{y} .

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Solutions to this missing data problem?

- Listwise deletion? Maybe, but it’s quite possible opting-out is *non-ignorable*.
- Multiple imputation using known covariates? Maybe, unless $\rho(X_j, Y)$ ’s are large, this is as good as using \bar{y} .
- Create dummy variables? Yes, Angrist (2001) recommends using dummies. We used “Labeled more than 5 images?” — the opt-out’s become 0’s, and you can run a valid regression.

Induced to work \Leftrightarrow Selection Bias

We do not recommend using induced to work. Turkers have a reputation and “returned HITs” are counted against them. This is most likely the reason there wasn’t too much variability.

However, *leaving* the HIT produces problems. The problem is you cannot observe y_i for the subject who jumped ship.

Solutions to this missing data problem?

- Listwise deletion? Maybe, but it’s quite possible opting-out is *non-ignorable*.
- Multiple imputation using known covariates? Maybe, unless $\rho(X_j, Y)$ ’s are large, this is as good as using \bar{y} .
- Create dummy variables? Yes, Angrist (2001) recommends using dummies. We used “Labeled more than 5 images?” — the opt-out’s become 0’s, and you can run a valid regression.
- Conditional-on-positive regression? Yes, as long as you tell your reader. We used “fine quality” *conditional* on being induced-to-work.

Group Project

Spend 5min with your neighbor coming up with an experiment.



We'll discuss feasibility of projects.

Internal and Statistical Validity

Are the causal conclusions in a randomized MTurk experiment warranted?

If anything, it's more internally valid than a lab experiment (as long as differential attrition is taken care of).

Internal and Statistical Validity

Are the causal conclusions in a randomized MTurk experiment warranted?

If anything, it's more internally valid than a lab experiment (as long as differential attrition is taken care of). There's no experimenter bias,

Internal and Statistical Validity

Are the causal conclusions in a randomized MTurk experiment warranted?

If anything, it's more internally valid than a lab experiment (as long as differential attrition is taken care of). There's no experimenter bias, perfect blinding — disinterested randomization,

Internal and Statistical Validity

Are the causal conclusions in a randomized MTurk experiment warranted?

If anything, it's more internally valid than a lab experiment (as long as differential attrition is taken care of). There's no experimenter bias, perfect blinding — disinterested randomization, SUTVA — better than the lab.

Internal and Statistical Validity

Are the causal conclusions in a randomized MTurk experiment warranted?

If anything, it's more internally valid than a lab experiment (as long as differential attrition is taken care of). There's no experimenter bias, perfect blinding — disinterested randomization, SUTVA — better than the lab.

Statistical conclusion validity: we have much higher power versus lab experiments ($n \approx 2,500$ in our example).

External Validity / Generalizability

Berinsky et al (2012) demonstrated that at least in the US, experimental results can be *generalized* to the American population at large unless you suspect heterogeneous effects:

$$\begin{aligned} Y &= \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \underbrace{h(\mathbb{1}_T, X_1, \dots, X_p)}_{\text{large?}} + \epsilon, \\ \epsilon &\sim \mathcal{N}_n(0, \sigma^2 I_n) \end{aligned}$$

External Validity / Generalizability

Berinsky et al (2012) demonstrated that at least in the US, experimental results can be *generalized* to the American population at large unless you suspect heterogeneous effects:

$$Y = \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \underbrace{h(\mathbb{1}_T, X_1, \dots, X_p)}_{\text{large?}} + \mathcal{E},$$
$$\mathcal{E} \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

They gave some of the covariates that could be problematic: age, political ideology, political knowledge, etc. However, there is no guarantee there aren't *other* covariates that they didn't explore. It is better than convenience samples.

External Validity / Generalizability

Berinsky et al (2012) demonstrated that at least in the US, experimental results can be *generalized* to the American population at large unless you suspect heterogeneous effects:

$$Y = \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \underbrace{h(\mathbb{1}_T, X_1, \dots, X_p)}_{\text{large?}} + \mathcal{E},$$
$$\mathcal{E} \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

They gave some of the covariates that could be problematic: age, political ideology, political knowledge, etc. However, there is no guarantee there aren't *other* covariates that they didn't explore. It is better than convenience samples.

External Validity is *never* airtight! But it looks promising for MTurk vis-a-vis most published lab studies.

External Validity / Generalizability

Berinsky et al (2012) demonstrated that at least in the US, experimental results can be *generalized* to the American population at large unless you suspect heterogeneous effects:

$$Y = \beta_0 + \beta_T \mathbb{1}_T + \beta_1 X_1 + \dots + \beta_p X_p + \underbrace{h(\mathbb{1}_T, X_1, \dots, X_p)}_{\text{large?}} + \mathcal{E},$$
$$\mathcal{E} \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

They gave some of the covariates that could be problematic: age, political ideology, political knowledge, etc. However, there is no guarantee there aren't *other* covariates that they didn't explore. It is better than convenience samples.

External Validity is *never* airtight! But it looks promising for MTurk vis-a-vis most published lab studies.

Ecological validity? Maybe not... but that always depends!

Why do we care? Natural field Experiments



EXTERNAL VALIDITY PROBLEMS “...lab experiments in isolation are necessarily limited in relevance for predicting field behavior, unless one wants to insist *a priori* that those aspects of economic behavior under study are perfectly general...” -Harrison and List (2004)

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- scrutiny and anonymity:

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched.

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:**

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:**

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:** none!
- **difficulty of replication:**

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:** none!
- **difficulty of replication:** git clone, load up the account with money — perfect, push-button replication!

Possible issues

- **Stakes:**

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:** none!
- **difficulty of replication:** git clone, load up the account with money — perfect, push-button replication!

Possible issues

- **Stakes:** You can argue the wages on MTurk are too low to generalize.

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:** none!
- **difficulty of replication:** git clone, load up the account with money — perfect, push-button replication!

Possible issues

- **Stakes:** You can argue the wages on MTurk are too low to generalize.
- **Selection:**

Natural field Experiments on MTurk

Why are natural field experiments on MTurk *better* than most natural field experiments? Evaluated by the Levitt and List (2007, 2009) rubric:

- **scrutiny and anonymity:** No “Hawthorne effects:” Turkers don’t know they’re being watched. No “Clever Hans effects:” Turkers cannot respond to the researcher’s subtle cues. In fact, no interaction with the experimenter since they don’t know they’re in an experiment!
- **artificial restrictions:** the Turkers are allowed to perform the task however they please (within a time limit).
- **need for cooperation with third parties:** none!
- **difficulty of replication:** git clone, load up the account with money — perfect, push-button replication!

Possible issues

- **Stakes:** You can argue the wages on MTurk are too low to generalize.
- **Selection:** Are Turkers different from the rest of the population. Berinsky et al say “no” but there is room to say “yes.”

Ethical Issues

Research using human subjects has to pass an *institutional review board* (IRB) review process. Simple surveying and labeling tasks on adults are *exempt* from review under “category 2.”

Ethical Issues

Research using human subjects has to pass an *institutional review board* (IRB) review process. Simple surveying and labeling tasks on adults are *exempt* from review under “category 2.” However, you may have noticed we used a fake name and *never* told the Turkers they were in an experiment in order to create a *natural field experiment* which is a form of...

DECEPTION

Ethical Issues

Research using human subjects has to pass an *institutional review board* (IRB) review process. Simple surveying and labeling tasks on adults are *exempt* from review under “category 2.” However, you may have noticed we used a fake name and *never* told the Turkers they were in an experiment in order to create a *natural field experiment* which is a form of...

DECEPTION

Thus, this study had to pass a *full board* IRB meeting and we needed to provide justification.

Ethical Issues

Research using human subjects has to pass an *institutional review board* (IRB) review process. Simple surveying and labeling tasks on adults are *exempt* from review under “category 2.” However, you may have noticed we used a fake name and *never* told the Turkers they were in an experiment in order to create a *natural field experiment* which is a form of...

DECEPTION

Thus, this study had to pass a *full board* IRB meeting and we needed to provide justification. They also asked us to send a debrief message to all subjects post-experiment.

Some literature on MTurk experimentation

- Multi-player MTurk experiment tests altruism and cooperation under various “local public goods” games. (Suri and Watts, 2011)

Some literature on MTurk experimentation

- Multi-player MTurk experiment tests altruism and cooperation under various “local public goods” games. (Suri and Watts, 2011)
- Replicating classical Tversky-Kahneman-type psychological experiments finds agreement with effect sizes. (Horton et al., 2011)

Some literature on MTurk experimentation

- Multi-player MTurk experiment tests altruism and cooperation under various “local public goods” games. (Suri and Watts, 2011)
- Replicating classical Tversky-Kahneman-type psychological experiments finds agreement with effect sizes. (Horton et al., 2011)
- Does “honesty priming” make Turkers more willing to admit to self-destructive alcoholic binges? (Pashler et al., 2013)

Some literature on MTurk experimentation

- Multi-player MTurk experiment tests altruism and cooperation under various “local public goods” games. (Suri and Watts, 2011)
- Replicating classical Tversky-Kahneman-type psychological experiments finds agreement with effect sizes. (Horton et al., 2011)
- Does “honesty priming” make Turkers more willing to admit to self-destructive alcoholic binges? (Pashler et al., 2013)
- Relying on “intuition” over “reasoning” causes about 30% more risk tolerance. (Butler et al., 2013)

Some literature on MTurk experimentation

- Multi-player MTurk experiment tests altruism and cooperation under various “local public goods” games. (Suri and Watts, 2011)
- Replicating classical Tversky-Kahneman-type psychological experiments finds agreement with effect sizes. (Horton et al., 2011)
- Does “honesty priming” make Turkers more willing to admit to self-destructive alcoholic binges? (Pashler et al., 2013)
- Relying on “intuition” over “reasoning” causes about 30% more risk tolerance. (Butler et al., 2013)

Further reading on MTurk as a platform

- “Conducting behavioral research on Amazon’s Mechanical Turk” (Mason and Suri, 2012)
- “The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments” (Rand, 2012)
- “Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research” (Crump et al, 2013)

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail
- a group exercise designing an experiment

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail
- a group exercise designing an experiment
- issues to look out for during analysis

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail
- a group exercise designing an experiment
- issues to look out for during analysis
- internal and external validity when drawing conclusions from the data

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail
- a group exercise designing an experiment
- issues to look out for during analysis
- internal and external validity when drawing conclusions from the data
- MTurk as a platform for high-power *natural field experiments*

Checkpoint

We've covered:

- a case study examining an MTurk experiment in detail
- a group exercise designing an experiment
- issues to look out for during analysis
- internal and external validity when drawing conclusions from the data
- MTurk as a platform for high-power *natural field experiments*

Goal D Complete

D For crowdsourced randomized experiments:

- Evaluate if the experiment is suitable for crowdsourcing
- Design the experiment
- Analyze the resulting data

Further Developments: Some extensions and plugins in the literature

Further Developments: Some extensions and plugins in the literature

- “UserActionTracer” (Stieger and Reips, 2010) collects *paradata* (information about mouse clicks, mouse movements, keystrokes) that we believe can be used in MTurk labeling tasks to inform quality of the label, and in experiments admitting better covariates and increasing efficiency of estimates.

Further Developments: Some extensions and plugins in the literature

- “UserActionTracer” (Stieger and Reips, 2010) collects *paradata* (information about mouse clicks, mouse movements, keystrokes) that we believe can be used in MTurk labeling tasks to inform quality of the label, and in experiments admitting better covariates and increasing efficiency of estimates.
- “Soylent” (Bernstein et al., 2010) is a word processing interface that embeds crowdsourcing labor. It helps users write by being able to have Turkers complete annoying find-fix-verify tasks such as text shortening, spelling and grammar checking, and formating citations and finding figures.

Further Developments: Some extensions and plugins in the literature

- “UserActionTracer” (Stieger and Reips, 2010) collects *paradata* (information about mouse clicks, mouse movements, keystrokes) that we believe can be used in MTurk labeling tasks to inform quality of the label, and in experiments admitting better covariates and increasing efficiency of estimates.
- “Soylent” (Bernstein et al., 2010) is a word processing interface that embeds crowdsourcing labor. It helps users write by being able to have Turkers complete annoying find-fix-verify tasks such as text shortening, spelling and grammar checking, and formating citations and finding figures.
- Little et al (2010) studied tasks broken up iteratively or in parallel for many different types of microtasks. They found that iterative work on a task increases the average accuracy, but many Turkers in parallel increase the maximum accuracy.

Further Developments: Use MTurk to study Statistical Questions

Further Developments: Use MTurk to study Statistical Questions

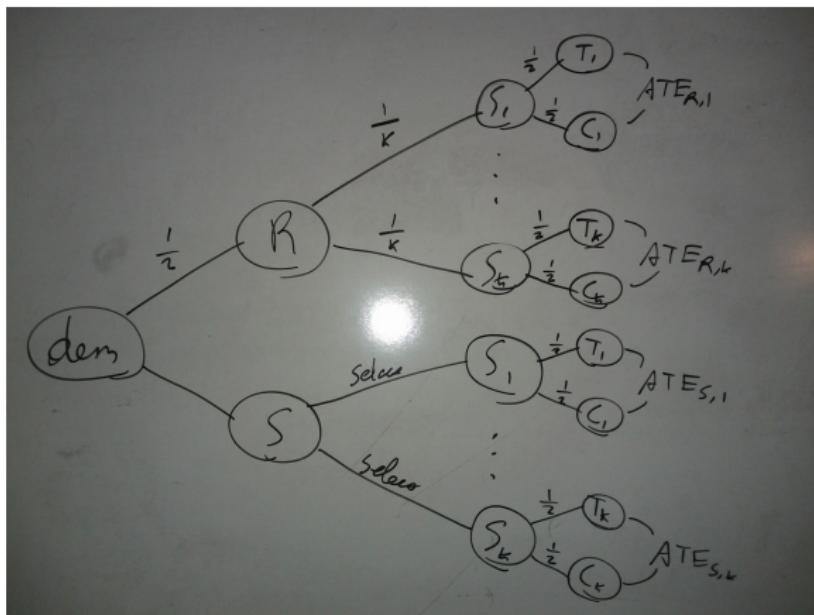
We talked a bit about heterogeneous effects where the conclusions of a study may not be applicable to the population of interest. Many experiments in the real world are run by recruiting subjects who *want* to be in the study.

Further Developments: Use MTurk to study Statistical Questions

We talked a bit about heterogeneous effects where the conclusions of a study may not be applicable to the population of interest. Many experiments in the real world are run by recruiting subjects who *want* to be in the study.

Does *wanting* to be in the study affect the *average treatment effect*? Is there selection bias *into* the randomized trial itself? This can be tested.

An experiment to test a statistical phenomenon



Design flowchart for proposed experiment

Power calculation

$$F^* := F(95\%, 1, 4(r - 1)), \quad W \sim F \left(1, 4(r - 1), \frac{4r\beta_{ST}^2}{\sigma^2} \right),$$

$$\text{POW} := \mathbb{P}(W > F^*), \quad \sigma^2 \approx 12.8 \text{ (from pilot data)}$$

Let r be the number of duplicates, N is total sample size, σ^2 is the variance of the homoskedastic error term, and β_{ST} is the true interaction effect between selecting the study and treatment.

Power calculation

$$F^* := F(95\%, 1, 4(r-1)), \quad W \sim F\left(1, 4(r-1), \frac{4r\beta_{ST}^2}{\sigma^2}\right),$$

$$\text{POW} := \mathbb{P}(W > F^*), \quad \sigma^2 \approx 12.8 \text{ (from pilot data)}$$

Let r be the number of duplicates, N is total sample size, σ^2 is the variance of the homoskedastic error term, and β_{ST} is the true interaction effect between selecting the study and treatment.

N_{study}	r	β_{ST} (\$)							
		0.01	0.05	0.1	0.15	0.2	0.25	0.5	0.75
400	100	0.05	0.06	0.09	0.14	0.22	0.31	0.84	0.99
800	200	0.05	0.07	0.13	0.24	0.39	0.55	0.99	1.00
1200	300	0.05	0.08	0.18	0.34	0.53	0.73	1.00	1.00
1600	400	0.05	0.09	0.22	0.43	0.66	0.84	1.00	1.00
2000	500	0.05	0.10	0.26	0.51	0.75	0.91	1.00	1.00
2400	600	0.05	0.11	0.30	0.58	0.83	0.95	1.00	1.00
2800	700	0.05	0.12	0.35	0.65	0.88	0.97	1.00	1.00

Began thinking: how can I reduce σ^2 ?

Current Research: Matching to Reduce Variance

“Perhaps the best way of reducing the error due to differences between person is to match before random assignments to treatments.”



Cook, Thomas D., and Donald T. Campbell. "Quasi-experimentation: Design and analysis for field setting." MA: Houghton Mifflin (1979).

Current Research: Matching to Reduce Variance

"Perhaps the best way of reducing the error due to differences between person is to match before random assignments to treatments."



Cook, Thomas D., and Donald T. Campbell. "Quasi-experimentation: Design and analysis for field setting." MA: Houghton Mifflin (1979).

$$\begin{aligned} Y &= \beta_T \mathbb{1}_T + X\beta + \mathcal{E}, \quad \text{Var}[\mathcal{E}] = \sigma_e^2 I_n \\ \mathbb{E} [\bar{Y}_T - \bar{Y}_C] &= \beta_T \quad (\text{expectorating over } \mathbb{1}_T) \\ \text{Var} [\bar{Y}_T - \bar{Y}_C] &= \beta^\top \text{Var} [\bar{X}_T - \bar{X}_C] \beta + \underbrace{\text{Var} [\bar{\mathcal{E}}_T - \bar{\mathcal{E}}_C]}_{\approx 4\sigma_e^2/n} \end{aligned}$$



Morgan, K. L., & Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 1263—1282

Current Research: Matching to Reduce Variance

“Perhaps the best way of reducing the error due to differences between person is to match before random assignments to treatments.”



Cook, Thomas D., and Donald T. Campbell. "Quasi-experimentation: Design and analysis for field setting." MA: Houghton Mifflin (1979).

$$\begin{aligned} Y &= \beta_T \mathbb{1}_T + X\beta + \mathcal{E}, \quad \text{Var}[\mathcal{E}] = \sigma_e^2 I_n \\ \mathbb{E} [\bar{Y}_T - \bar{Y}_C] &= \beta_T \quad (\text{expectorating over } \mathbb{1}_T) \\ \text{Var} [\bar{Y}_T - \bar{Y}_C] &= \beta^\top \text{Var} [\bar{X}_T - \bar{X}_C] \beta + \underbrace{\text{Var} [\bar{\mathcal{E}}_T - \bar{\mathcal{E}}_C]}_{\approx 4\sigma_e^2/n} \end{aligned}$$



Morgan, K. L., & Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 1263–1282

If the sample averages of the x's in each group are balanced, the variance drops.

Matching to Reduce Variance

Most real world data is “nonlinear.” This is our focus.

Matching to Reduce Variance

Most real world data is “nonlinear.” This is our focus.

$$\begin{aligned} Y &= \beta_T \mathbb{1}_T + z + \mathcal{E}, \quad z_i := f(x_{1i}, \dots, x_{pi}) \\ \mathbb{E} [\bar{Y}_T - \bar{Y}_C] &= \beta_T \quad (\text{expectorating over } \mathbb{1}_T) \\ \mathbb{V}\text{ar} [\bar{Y}_T - \bar{Y}_C] &= \mathbb{V}\text{ar} [\bar{Z}_T - \bar{Z}_C] + \underbrace{\mathbb{V}\text{ar} [\bar{\mathcal{E}}_T - \bar{\mathcal{E}}_C]}_{\approx 4\sigma_e^2/n} \end{aligned}$$

Same mechanism in a non-linear model.

Matching to Reduce Variance

Most real world data is “nonlinear.” This is our focus.

$$\begin{aligned} Y &= \beta_T \mathbb{1}_T + z + \mathcal{E}, \quad z_i := f(x_{1i}, \dots, x_{pi}) \\ \mathbb{E} [\bar{Y}_T - \bar{Y}_C] &= \beta_T \quad (\text{expectorating over } \mathbb{1}_T) \\ \mathbb{V}\text{ar} [\bar{Y}_T - \bar{Y}_C] &= \mathbb{V}\text{ar} [\bar{Z}_T - \bar{Z}_C] + \underbrace{\mathbb{V}\text{ar} [\bar{\mathcal{E}}_T - \bar{\mathcal{E}}_C]}_{\approx 4\sigma_e^2/n} \end{aligned}$$

Same mechanism in a non-linear model.

- if for all subjects $f(x_{T,i}) \approx f(x_{C,i})$ (balance between treatment and control), we achieve a more efficient estimator on average over many experiments

Matching to Reduce Variance

Most real world data is “nonlinear.” This is our focus.

$$\begin{aligned} Y &= \beta_T \mathbb{1}_T + z + \mathcal{E}, \quad z_i := f(x_{1i}, \dots, x_{pi}) \\ \mathbb{E} [\bar{Y}_T - \bar{Y}_C] &= \beta_T \quad (\text{expectorating over } \mathbb{1}_T) \\ \mathbb{V}\text{ar} [\bar{Y}_T - \bar{Y}_C] &= \mathbb{V}\text{ar} [\bar{Z}_T - \bar{Z}_C] + \underbrace{\mathbb{V}\text{ar} [\bar{\mathcal{E}}_T - \bar{\mathcal{E}}_C]}_{\approx 4\sigma_e^2/n} \end{aligned}$$

Same mechanism in a non-linear model.

- if for all subjects $f(x_{T,i}) \approx f(x_{C,i})$ (balance between treatment and control), we achieve a more efficient estimator on average over many experiments
- if $n_T \approx n_C$ we also achieve a more efficient estimator

Student's Identical-Twins Idea (1931)

For among 20,000 children there will be numerous pairs of twins; exactly how many it is not easy to say owing to the differential death rate, but, since there is about one pair of twins in 90 births, one might hope to get at least 160 pairs in 20,000 children. But as a matter of fact the 20,000 children were not all the Lanarkshire schools population, and I feel pretty certain that some 200—300 pairs of twins would be available for the purpose of the experiment.

Of 200 pairs some 50 would be "identicals" and of course of the same sex, while half the remainder would be non-identical twins of the same sex.

Now identical twins are probably better experimental material than is available for feeding experiments carried out on any other mammals, and the error of the comparison between them may be relied upon to be so small that 50 pairs of these would give more reliable results than the 20,000 with which we have been dealing.

The proposal is then to experiment on all pairs of twins of the same sex available, noting whether each pair is so similar that they are probably "identicals" or whether they are dissimilar.

"Feed" one of each pair on raw and the other on pasteurised milk, deciding in each case which is to take raw milk by the toss of a coin.

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control.

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly?”

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	1_T
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	C

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	C
4		46	F	N	Y	Y	

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	C
4		46	F	N	Y	Y	T

Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	C
4		46	F	N	Y	Y	T
5		48	M	Y	Y	N	

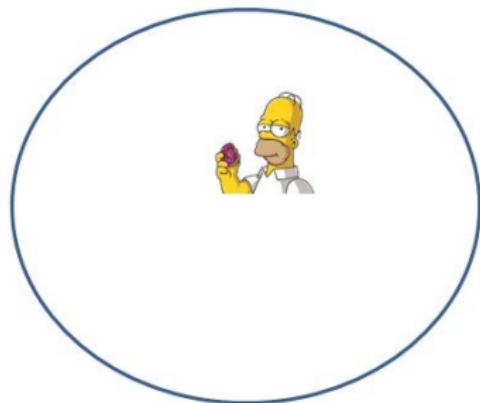
Back to MTurk: Sequential Experiment

Subjects enter sequentially and must immediately be assigned treatment or control. What if we match “on-the-fly”?

t	Subject	Age	Gender	Likes donuts?	Smokes?	Smart?	$\mathbb{1}_T$
1		40	M	Y	N	N	T
2		46	F	N	Y	Y	C
3		46	M	Y	N	Y	C
4		46	F	N	Y	Y	T
5		48	M	Y	Y	N	C

The “Reservoir” and the “Matches”

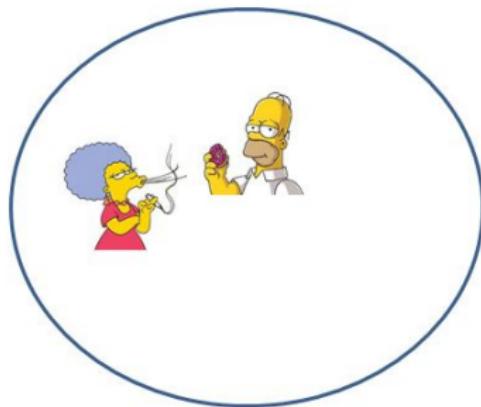
$t = 1$



Reservoir

The “Reservoir” and the “Matches”

$t = 2$

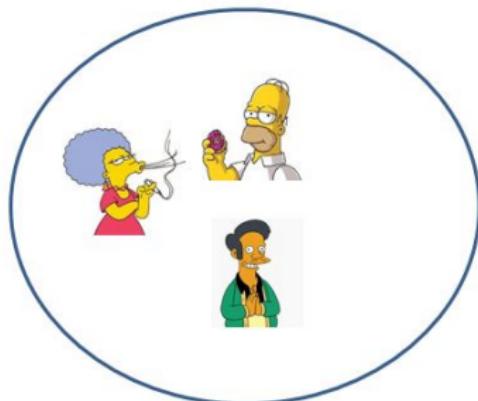


Reservoir

Ending stats $n_R = 1$, $m = 2$.

The “Reservoir” and the “Matches”

$t = 3$

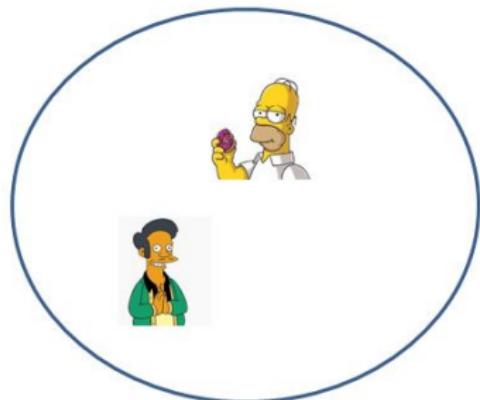


Reservoir

Ending stats $n_R = 1$, $m = 2$.

The “Reservoir” and the “Matches”

$t = 4$



Reservoir

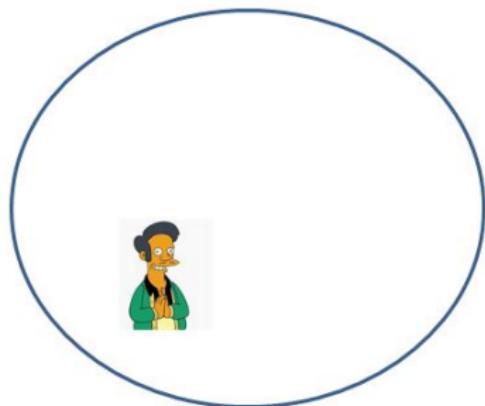


Match 1

Ending stats $n_R = 1$, $m = 2$.

The “Reservoir” and the “Matches”

$t = 5$



Reservoir



Match 1



Match 2

Ending stats $n_R = 1$, $m = 2$.

How to match?

Matching on Mahalanobis distance. Assume covariates are normal enough to use the scaled F distribution.



Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. Optimal multivariate matching before randomization. *Biostatistics* (2004), 5(2), 263—75

How to match?

Matching on Mahalanobis distance. Assume covariates are normal enough to use the scaled F distribution.



Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. Optimal multivariate matching before randomization. *Biostatistics* (2004), 5(2), 263–75

$$D_M^2 := \frac{1}{2}(X_{\text{new}} - X_{\text{old}})^\top S^{-1}(X_{\text{new}} - X_{\text{old}}), \quad \frac{n-p}{p(n-1)} D_M^2 \sim F_{p,n-p}$$

Choose a p_{val} hyperparameter, $\lambda \in (0, 1)$, to represent the cutoff probability to match: the higher λ is, the easier to match.

The Sequential Match Classic Estimator

Consider $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$.

The Sequential Match Classic Estimator

Consider $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$.

Since $S_D^2 \xrightarrow{P} \sigma_D^2$ and $S_R^2 \xrightarrow{P} \sigma_R^2$, we now have a Z-like testing estimator:

The Sequential Match Classic Estimator

Consider $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$.

Since $S_{\bar{D}}^2 \xrightarrow{P} \sigma_{\bar{D}}^2$ and $S_R^2 \xrightarrow{P} \sigma_R^2$, we now have a Z-like testing estimator:

$$\frac{B_T - \beta_0}{\text{SE}[B_T]} \approx \frac{\frac{S_R^2 \bar{D} + S_{\bar{D}}^2 (\bar{Y}_{R,T} - \bar{Y}_{R,C})}{S_R^2 + S_{\bar{D}}^2} - \beta_0}{\sqrt{\frac{S_R^2 S_{\bar{D}}^2}{S_R^2 + S_{\bar{D}}^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

The Sequential Match Classic Estimator

Consider $H_0 : \beta_T = \beta_0$ versus $H_a : \beta_T \neq \beta_0$.

Since $S_{\bar{D}}^2 \xrightarrow{P} \sigma_{\bar{D}}^2$ and $S_R^2 \xrightarrow{P} \sigma_R^2$, we now have a Z-like testing estimator:

$$\frac{B_T - \beta_0}{\text{SE}[B_T]} \approx \frac{\frac{S_R^2 \bar{D} + S_{\bar{D}}^2 (\bar{Y}_{R,T} - \bar{Y}_{R,C})}{S_R^2 + S_{\bar{D}}^2} - \beta_0}{\sqrt{\frac{S_R^2 S_{\bar{D}}^2}{S_R^2 + S_{\bar{D}}^2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Note that in the case where there are no matched pairs, we default to the classic estimator and in the case where there are less than two treatments or controls in the reservoir, we use only the matched pairs component.

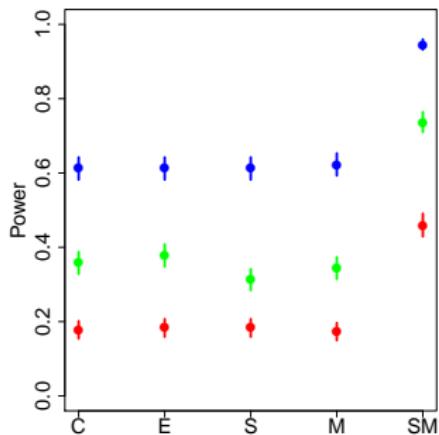
Relative Power

Simulate on a non-linear model:

$$\beta_T \mathbf{1}_{T,i} + x_{1,i} + x_{2,i} + x_{1,i}^2 + x_{2,i}^2 + x_{1,i}x_{2,i} + \mathcal{E}_i$$

$$X_{1,i} \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad X_{2,i} \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad \mathcal{E}_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

$$\beta_T = 1, \quad \sigma_e^2 = 3, \quad N_{\text{sim}} = 1000$$



$$(n = 50, n = 100, n = 200)$$

Results on a Clinical Trial

On 200 simulations and using all covariates...

Results on a Clinical Trial

On 200 simulations and using all covariates...

purported <i>n</i>	actual <i>n</i> (average)	average efficiency	approximate sample size reduction
100	75.2	1.10	9.2%
150	111.3	1.05	4.9%
224 (all)	165.5	1.07	6.7%

Results on a Clinical Trial

On 200 simulations and using all covariates...

purported <i>n</i>	actual <i>n</i> (average)	average efficiency	approximate sample size reduction
100	75.2	1.10	9.2%
150	111.3	1.05	4.9%
224 (all)	165.5	1.07	6.7%

Snooping and using the top 4 covariates...

Results on a Clinical Trial

On 200 simulations and using all covariates...

purposed <i>n</i>	actual <i>n</i> (average)	average efficiency	approximate sample size reduction
100	75.2	1.10	9.2%
150	111.3	1.05	4.9%
224 (all)	165.5	1.07	6.7%

Snooping and using the top 4 covariates...

purposed <i>n</i>	actual <i>n</i> (average)	average efficiency	approximate sample size reduction
100	72.9	1.23	18.8%
150	108.6	1.15	13.2%
224 (all)	160.3	1.13	11.3%

How do we pretend to use our sequential matching procedure?

```

nsim: 1 .....|x.x|.x.|xx.x|||...|x|x|..|xx|x..|xx.x|.xx..xxx.|x|||xx.|.|.|..| |x|..|x|.xx|.|x|.xxx.|x|
nsim: 2 .....|x|xx|.x|.x.xxx|||x.xx|xx|.x.x|.x|.|...|xx|xx|.x|.|xxx|||xxxxx.x.xx|||xxxxxx.x|.xx
nsim: 3 .....|x.|.|.xx|||xxxx|.xx|x.|x.|x|x|||...xx|.|.x|..xx|.|||..x|||xx.|.|..|..|..|.|x.|.|xxxxxxxxx|x.x.
nsim: 4 .....|xxx|.|x|.|xx|.|xxx|.|||x.|x|x|x..x.|x|x|.xx|x|.|x.|..|.|.x|x|.x.x.|xx|.xxx|.|xx|.x|...|..|.|
nsim: 5 .....|.xx|.x.|xxx|x|.x|x|x.|.|..x|||..x|xx|||..x|x|.|x|.|x|x|.x|xx|x.|x|xxx|.|.|||..|x.|xxx|.|..x|.|
nsim: 6 .....|x|x|.||.xx.x.|x|.|..x|.|x|x|||..|..x|xx|||..|..x|xx|.x|.|x|...x|xxx|.x|x|xxx|.|x|.|..x|.|.|..x|.|.
nsim: 7 .....|..|xxxxx|||..|x|.xx|x|x|.|xxx|x|.|..|..x|.|..|..x|.|..|..x|.|..|..x|.|..|..x|.|..|..x|.|..|..x|.|..|..x|.|.
nsim: 8 .....|x|xxx|xx|.|x|.|.xxx|.|..|x|.|..|x|x|x|x|x|.|xxx|xxx|.x|x|xxx|.x|.|..|..xxx|.|.x|.|xxx|.|
nsim: 9 .....|x|..x|.|..|..|..|..|x|.|..|xxx|x|x|x|x|x|.|..|.|.|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 10 .....|x|.|x|.|xx|x|.|xx|..|x|.|..|xxx|.|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 11 .....|x|..xx|.||..x|.|x|x|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 12 .....|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 13 .....|x|.|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 14 .....|x|.|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 15 .....|x|.x|xx|.||.xxx|.|||..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|
nsim: 16 .....|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|..|.|

```

Note that sometimes S_t^{-1} cannot be inverted, so we use the Moore-Penrose generalized inverse.

Remember the MTurk data block structure

Machine Learning usually expects:

Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

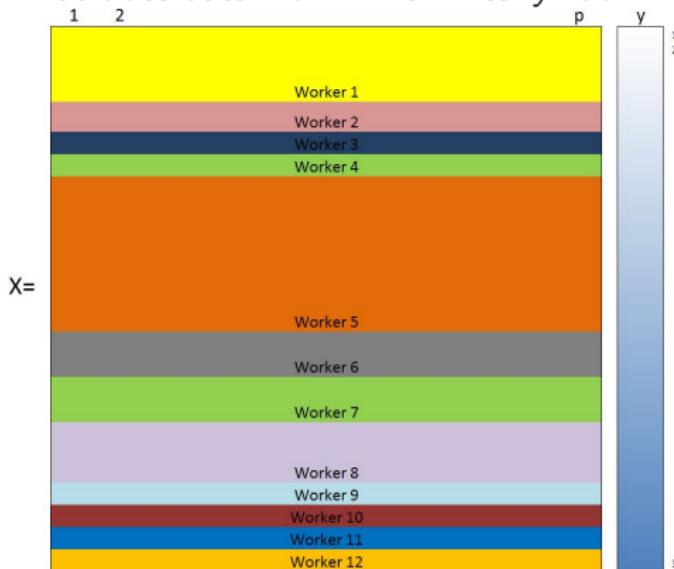
What does data from MTurk *really* look like?

Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?

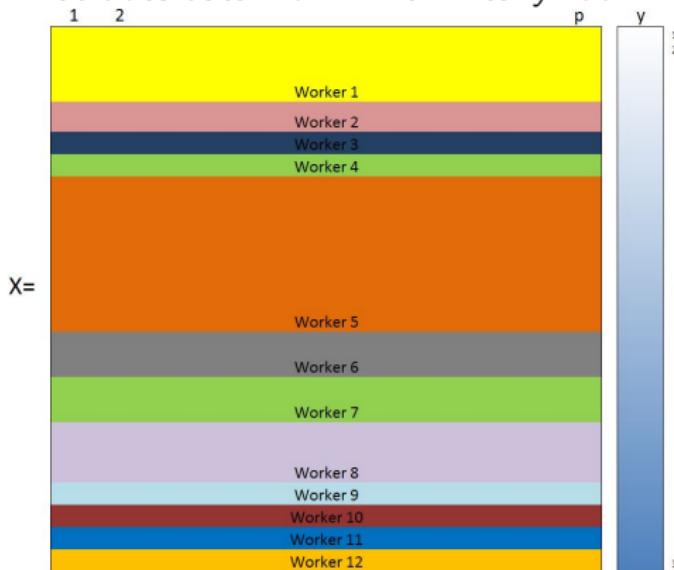


Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?



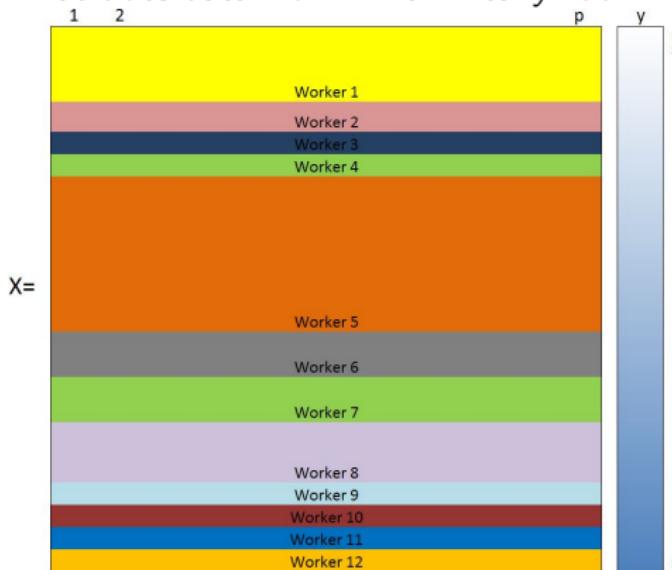
Thus, $n_{\text{eff}} \in [12, 31]$,

Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?



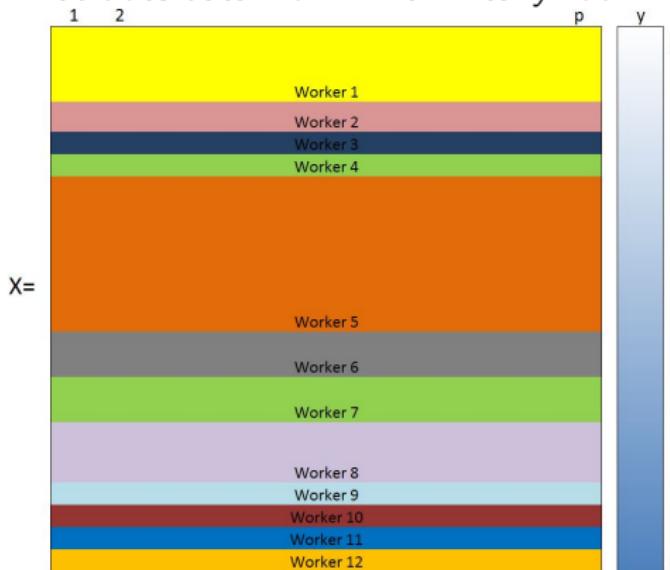
Thus, $n_{\text{eff}} \in [12, 31]$, an inconvenient reality that off-the-shelf ML algorithms do not consider:

Remember the MTurk data block structure

Machine Learning usually expects:

$$Y_i = f(x_{1i}, \dots, x_{pi}) + \mathcal{E}_i, \quad \mathcal{E}_1, \dots, \mathcal{E}_n \stackrel{iid}{\sim} e(0, \sigma^2)$$

What does data from MTurk *really* look like?



Thus, $n_{\text{eff}} \in [12, 31]$, an inconvenient reality that off-the-shelf ML algorithms do not consider:

$$\mathcal{E} \sim e \left(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{12} \end{bmatrix} \right)$$

A new machine learning method to use block structure

“Bayesian Additive Regression Trees” (Chipman et al, 2010) is a machine learning method that builds a sum-of-trees model using a normal likelihood for node data and various priors to control regularization:

$$Y = \underbrace{\psi_1(X) + \psi_2(X) + \dots + \psi_m(X)}_f + \epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

A new machine learning method to use block structure

“Bayesian Additive Regression Trees” (Chipman et al, 2010) is a machine learning method that builds a sum-of-trees model using a normal likelihood for node data and various priors to control regularization:

$$Y = \underbrace{\varphi_1(X) + \varphi_2(X) + \dots + \varphi_m(X)}_f + \epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

and it can be extended to handle

$$Y = \varphi_1(X) + \varphi_2(X) + \dots + \varphi_m(X) + \epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix})$$

A new machine learning method to use block structure

“Bayesian Additive Regression Trees” (Chipman et al, 2010) is a machine learning method that builds a sum-of-trees model using a normal likelihood for node data and various priors to control regularization:

$$Y = \underbrace{\varphi_1(X) + \varphi_2(X) + \dots + \varphi_m(X)}_f + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

and it can be extended to handle

$$Y = \varphi_1(X) + \varphi_2(X) + \dots + \varphi_m(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix}\right)$$

where the block structure could be compound symmetric:

A new machine learning method to use block structure

“Bayesian Additive Regression Trees” (Chipman et al, 2010) is a machine learning method that builds a sum-of-trees model using a normal likelihood for node data and various priors to control regularization:

$$Y = \underbrace{\psi_1(X) + \psi_2(X) + \dots + \psi_m(X)}_f + \epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

and it can be extended to handle

$$Y = \psi_1(X) + \psi_2(X) + \dots + \psi_m(X) + \epsilon, \quad \epsilon \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix}\right)$$

where the block structure could be compound symmetric:

$$D = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \quad \rho \sim \mathbb{P}(\cdot)$$

A new machine learning method to use block structure

“Bayesian Additive Regression Trees” (Chipman et al, 2010) is a machine learning method that builds a sum-of-trees model using a normal likelihood for node data and various priors to control regularization:

$$Y = \underbrace{\psi_1(X) + \psi_2(X) + \dots + \psi_m(X)}_f + \epsilon, \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

and it can be extended to handle

$$Y = \psi_1(X) + \psi_2(X) + \dots + \psi_m(X) + \epsilon, \quad \epsilon \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_k \end{bmatrix}\right)$$

where the block structure could be compound symmetric:

$$D = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \quad \rho \sim \mathbb{P}(\cdot)$$

ρ can be estimated for entire data set or inside of each Turker’s work portfolio.

Crowdsourcing for Statisticians, Designing Applications and Analyzing Data on MTurk

What we covered

- Crowdsourcing and microlabor - a booming field
- Using MTurk for surveying, collecting labeled data, and experimenting
- Some issues and research opportunities in collecting and using Mturk-generated data

We enjoyed having you all in our tutorial! Questions?

Acknowledgements

We would like to thank Larry Brown, Dana Chandler, Dean Foster, John Horton, Panos Ipeirotis, Abba Krieger, Mary Putt, Abraham Wyner. Adam acknowledges the NSF for his graduate research fellowship that made this research possible.