

An Analytic Framework for Human Computation

**Crowdsourcing and Human Computation
Lecture 4**

**Instructor: Chris Callison-Burch
TA: Ellie Pavlick**

Website: crowdsourcing-class.org

HW2 discussion

- Did everyone get the code working?
- What people/companies/things did you target for sentiment analysis?
- When creating your own gold standard data, what was the distribution of positive/negative/neutral?

Leftovers from Monday

Guidelines for your own HITs

- Simple instructions are required
- If your task can't be expressed in one paragraph + bullets, then it may need to be broken into simpler sub-tasks

Guidelines for your own HITs

- Quality control is paramount
 - Measuring redundancy doesn't work if people answer incorrectly in systematic ways
 - Embed gold standard data as controls
- Qualification tests v. no qualification test
 - Reduce participation, but usually ensures higher quality

Guidelines for your own HITs

- Controls should be
 - Ones that have a unambiguous answer
 - Scorable automatically or with 2nd pass HIT
 - Not easily detectable

Guidelines for your own HITs

- Remember that MTurk is a reputation system
 - Rejecting Turkers has consequences for them, blocking Turkers is likely to get them expelled from MTurk
 - I reject all work performing near chance, accept all work doing ~80%, reject proportionally to performance on controls for in-between

Guidelines for your own HITS

- Reputation goes both ways
 - Monitor your profile on Turker Nation and TurkOpticon
 - Pay your workers generously
 - Pay quickly, and be responsive to email questions

Guidelines for your own HITs

- Subjective tasks are difficult, but not impossible to control for
 - You should still use embedded controls, but ones which have very clear answers
- If your task is one that requires lots of training, and retaining workers for longer periods:
 - Don't use Mechanical Turk
 - Consider ODesk or other “online temping agencies”

Today's topic



Classification System for Human Computation

- Motivation
- Quality Control
- Aggregation
- Human Skill
- Process Order
- Task-request Cardinality

Motivation

How can we motivate people to participate?

Even with a low barrier to entry (anyone with a computer can contribute) we still need to make a case **why** they should contribute.

Motivation: Pay

- Easiest way to recruit workers.
- Downside: provides incentive to cheat
- Problem might be exacerbated when the crowd workers are anonymous
- MTurk uses micropayments
- Online temping services provide higher wages: LiveOps, ODesk, etc
- CrowdFlower tried non-monetary payments (virtual goods and currencies, SwagBucks)

Motivation: Altruism

- People want to do good
- When Jim Gray went missing, volunteers searched 500k satellite images
- After the Haitian earthquake, diaspora translated 1000 messages per day



<p>Eske lekòl kolej Marie Anne kraze? mes</p>	<p>Was the College Marie Anne school destroyed? Thank you.</p>
<p>Nou pa ka anpeche moustik yo mo`de nou paske yo anpil.</p>	<p>We can't prevent the mosquitoes from biting because there are so many.</p>
<p>tanpri ke`m ap kase mwen pa ka pran nouvel manmanm.</p>	<p>Please heart is breaking because I have no news of my mother.</p>
<p>4636: Opital Medesansan Fwontye `delmas 19 lafe`men. Opital sen lwi gonzag nan delma 33 pran an chaj gratwit- man tout moun ki malad ou blese</p>	<p>4636: The Doctors without Borders Hospital in Delmas 19 is closed. The Saint Louis Gonzaga hospital in Delmas 33 is taking in sick and wounded people for free</p>

Motivation: Reputation

- Sometimes people will contribute in order to build their profile within a community
- Example: stackoverflow

Users

reputation

new us

Type to find users:

**Gordon Linoff**

New York, United States

1,260

sql, mysql, sql-server

**Sotirios Delimanolis****1,020**

java, spring, spring-mvc

**Arun P Johny**

Bangalore, India

960

jquery, javascript, html

**Jon Skeet**

Reading, United Kingdom

915

c#, java, .net

**Hans Passant**

Madison, WI

905

c#, .net, winforms

**Martijn Pieters**

Cambridge, United Kingdom

880

python, python-2.7, list

**alecxe**

Russia

870

python, django, scrapy

**BalusC**

Willemstad, Curaçao

855

java, jsf, jsf-2

**dasblinkenlight**

United States

825

c#, c++, java

**falsetru**

Seoul, South Korea

800

python, regex, ruby

**VonC**

France

780

git, eclipse, java

arshajii

Boston, MA

777

java, python, string

Motivation: Enjoyment

- Games with a purpose is a strategy to try to make a task fun
- In the ESP game two players look at an image and try to guess what words the other is thinking
- In doing so they label images on the web

score
100

ESP Game

Concentrate...

time
2:21

What do you see?

taboo words

peace

lay



guesses

sheeps...

sheep

+ submit

→ pass



Luis Von Ahn
==
Tom Sawyer



Tom Sawyer (Whitewashing the Fence), 1936

Motivation: Implicit Work

- It is sometimes possible to make people do work alongside some other task

The Norwich line steamboat train, from New-London for Boston, this **morning** ran off the track seven miles north of New-London.

morning

morning overtook

Type the two words:

reCAPTCHA™
stop spam.
read books.

Motivation

- Pay
- Altruism
- Reputation
- Enjoyment
- Implicit Work
- **Can you think of others?**

Quality Control

Even if people are motivated to participate, how do we know that they are doing work conscientiously? Can we trust them not to cheat or sabotage the system? Even if they are acting in good faith, how do we know that they're doing things right?

Quality Control: Reputation Check

- Mechanical Turk uses a reputation system
- When a Turker submits poor work, Requesters reject it
- The Turker's approval rate is displayed to all other Requesters

QC: Agreement and Redundancy

- The ESP game uses the labels that two players independently agree on
- Similar technique is often used in MTurk, when each item is done independently
- Redundancy allows a voting on ambiguous answers / opinions
- It is also helpful for identifying workers who are consistently divergent

QC: Gold Standard

- In MTurk we commonly mix in questions with a known answer alongside new questions
- This is similar to agreement, but now we check agreement against experts or trusted workers
- For multiple choice questions, gold standard allows for automatic grading

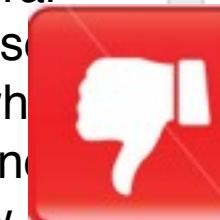
QC: Second-pass review

- Do second-pass grading when gold standard don't allow automatic grading
- Often times the second-pass HIT can be automatically gradable
- This makes the whole pipeline fully automated and ensures high quality

Heather Locklear Arrested for driving under the influence of drugs



The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from parking in Montecito, as revealed to People magazine by a spokesman for the California Highway Police. The witness stated that around 4.30pm M Locklear "hit the accelerator violently, making excessive noise while trying to take her car out from the parking with abrupt and forth maneuvers. While reversing, she passed several times in front of his sunglasses. Shortly after, the witness, who a first time, apparently had not recognized the actress, saw me.



Why was Heather Locklear arrested?

- She was arrested on suspicion of driving under the influence of drugs.

Driving under the influence

Driving while medicated

DUI

Driving while using drugs

Medikamentes

Heather Locklear

Photo by: Santa Barbara County Sheriff's Department

SPONSORED LINKS

QC: Defensive task design

- Try to design tasks so that they are nearly as hard to cheat as they are to complete
- For my translation HITs, people frequently would paste text into Google Translate
- I converted the text into images, then people had to transcribe it.

QC: Statistical models

- Sometimes it is possible to have prior knowledge about what the range of expected answers should be
- Use your statistical knowledge to throw out outliers

QC: Economic incentives

- When money is the motivator, it may be possible to use different incentive structures to illicit better results
- Pay people more when they reach a certain level of mastery, or when their output passes second pass reviews
- CastingWords uses bonuses to do this

[All HITS](#) | [HITS Available To You](#) | [HITS Assigned To You](#)Find **HITS**containing **castingwords**that pay at least \$ **0.00**☐ for which you are qualified☐ require Master Qualification

GO

HITS containing 'castingwords'

1-10 of 47 Results

Sort by: **Reward Amount (most first)** **GO!**[Show all details](#) | [Hide all details](#)1 [2](#) [3](#) [4](#) [5](#) > [Next](#) >> [Last](#)**Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwrdbns: \$7.60) [13:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (56 minutes 15 seconds) **Reward:** \$3.80
Time Allotted: 4 hours 30 minutes **HITS Available:** 1**Express Transcription: Easy #159160 (avg rwrdbns: \$6.3) [21:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (3 hours 58 minutes) **Reward:** \$3.15
Time Allotted: 5 hours 30 minutes **HITS Available:** 1**Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwrdbns: \$4.68) [08:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (56 minutes 13 seconds) **Reward:** \$2.34
Time Allotted: 4 hours **HITS Available:** 1**Express Transcription: Easy #159160 (avg rwrdbns: \$3.9) [13:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (3 hours 58 minutes) **Reward:** \$1.95
Time Allotted: 4 hours 30 minutes **HITS Available:** 1**Express Transcription: Easy #159160 (avg rwrdbns: \$2.4) [08:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (3 hours 58 minutes) **Reward:** \$1.20
Time Allotted: 4 hours **HITS Available:** 1**Express Transcription: Easy #159160 (avg rwrdbns: \$1.5) [05:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (3 hours 58 minutes) **Reward:** \$0.75
Time Allotted: 4 hours **HITS Available:** 1**Express Transcription: Frank King X-ray #159143 (195% premium) (avg rwrdbns: \$1.16) [02:00 mmss]** Not Qualified to work on this HIT ([Why?](#)) | [View a HIT in this group](#)**Requester:** [CastingWords](#) **HIT Expiration Date:** Sep 11, 2013 (56 minutes 9 seconds) **Reward:** \$0.58
Time Allotted: 3 hours 30 minutes **HITS Available:** 1

Quality Control

- Reputation check
- Agreement and redundancy
- Gold standard + automatic reviewing
- Multi-level review
- Defensive task design
- Statistical filtering
- Economic incentives
- **Others?**

Aggregation

Part of the process of human computation is to combine all contributions to solve a global problem. The class of problem may determine what strategy is best.

Aggregation: Statistical Data Processing

In *Wisdom of Crowds*, Surowiecki argues that aggregating answers from a decentralized, disorganized group of people, all thinking independently yields more accurate answers than from individuals.

Individual errors need to be uniformly distributed, so individual judgments must be made independently.

Aggregation: Collection

- Voting
- Prediction markets

Aggregation: Collection

- Sometimes aggregation collects discrete facts in a knowledge base
- A contribution may
 - add a new fact
 - improve quality by correcting, refuting, or confirming existing facts

Status

I have 12 answers so far.

Score: 0

Questions Answered: 0

I am thinking of a sentence

True

False

Don't know

Doesn't make sense

Engines are typically located in storage sheds.

Because I think it might be true.

Aggregation: Search

- Some human computation paradigms are simply asking people to find something in a large number of images
- Stardust@home project - people looked through images of aerogel to find dust from comet's tail gathered by spacecraft



GET STARTED

STARDUST@HOME

ABOUT

NEWS

GET STARTED

COMMUNITY

CLASSROOM

HELP

1. Finding Stardust

2. Tutorial Session

▶ 3. Test & Register

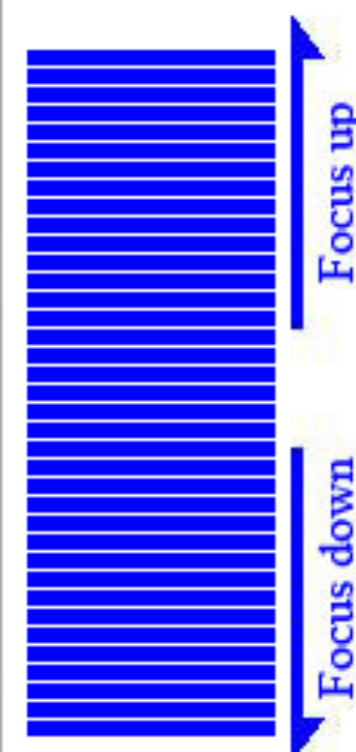
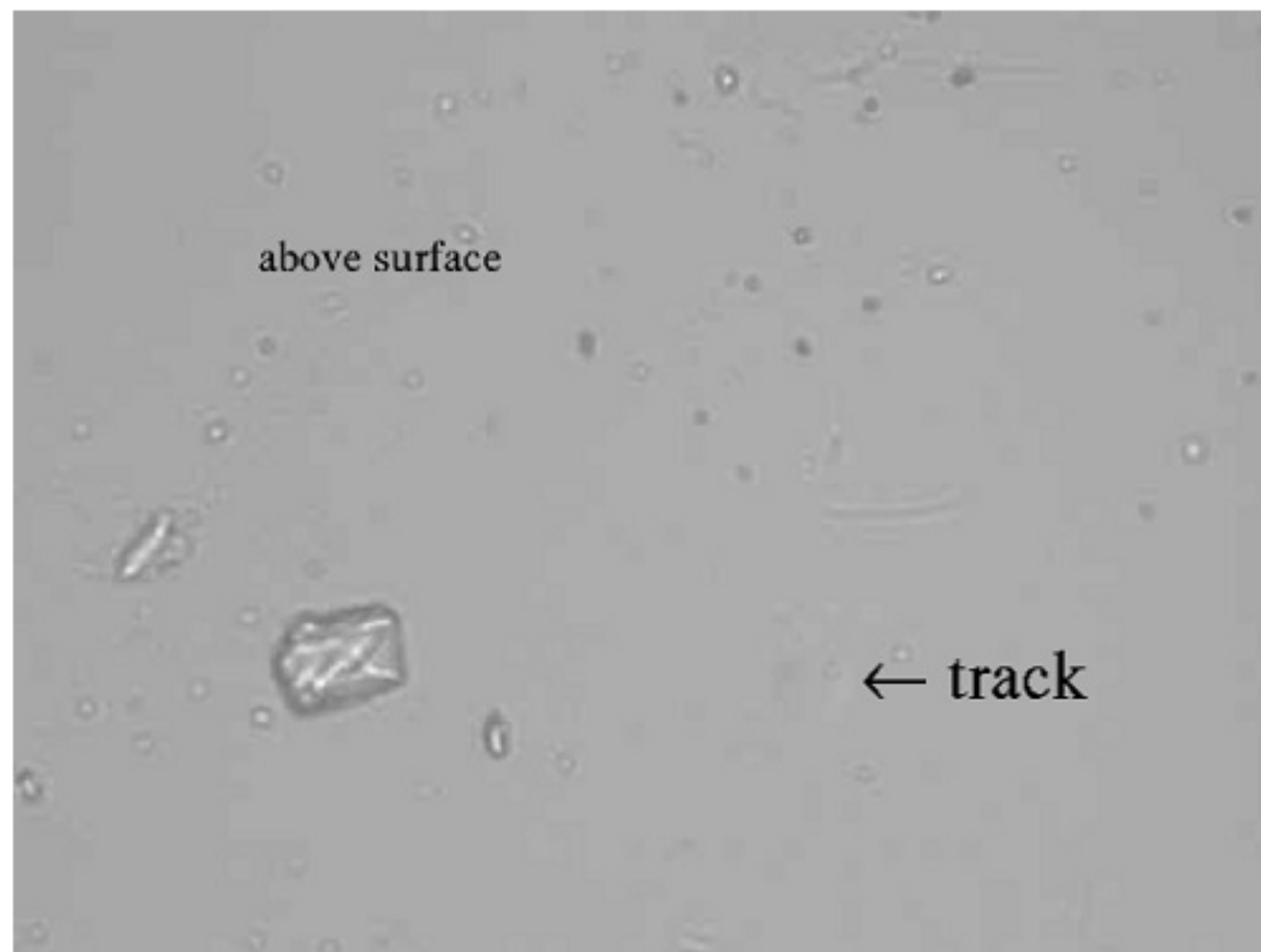
4. Log In & Search

Tech FAQ

Stardust Search

Percent of images loaded:
Images loaded for next movie:

[Continue to the next movie](#) →



Examples

Aggregation: Iterative Refinement

- One person's output is shown to the next person, who is asked to improve upon it
- What would Surowiecki say? (WWSS?)

- Please transcribe as many words as you can.
- Put a * in front of words you are unsure about.

It is important to be sure you , but I am

If a *festival . *two *me . *but *is

you to . I am to you

If . *two If

, but I am to you .

*festival . *festival

Submit

Aggregation

- Wisdom of Crowds
- Collection
- Search
- Iterative improvement
- **Genetic algorithm**
- **None?**
- **Other?**

Human Skill

“Human Computation is a paradigm for utilizing human processing power to solve problems that computers cannot yet solve.”

–Luis von Ahn. Doctoral Thesis, 2005.

Human skill

- What human skills have we seen so far?
- What others might be used for human computation tasks?

Process order

- Three roles in Human Computation: Requester, Worker, Computer
- **Requester** is the end user who benefits from the computation
- **Worker** is the person performing the task
- **Computer** only comes into play when it plays a role in solving the problem (not just aggregating results or being the information channel)

Process Order: CWR

- Computer → Worker → Requester
- In reCAPTCHA:
 - Computer first tries to perform OCR
 - Workers are presented with words that it fails to recognize
 - Their transcriptions are aggregated for the Requester (reader / library)

Process Order: WRC

- Worker → Requester → Computer
- In image labeling games:
 - Players (workers) provide labels
 - Web users (requesters) perform an image search
 - Computer searches the database of labels and presents matches

Process Order: CWRC

- Computer→Worker→Requester→Computer
 - Cyc system has inferred a lot of facts by analyzing text
 - Sends its guesses to FACTory where Workers confirm/correct facts
 - When a user (requester) queries Cyc
 - The Cyc system performs AI inference

Process Order: RW

- Requester → Worker
- Some tasks require no (non-human) computation
- Audio transcription or text dictation
- For small jobs, might not need any sophisticated computer-mediated QC

Task-Request Cardinality

When a service is powered by human computation,
many human workers may produce the result.
Sometimes, just one or a few workers may suffice. The
structure of the problem dictates the cardinality.

Task-Request Cardinality

- One-to-one: ChaCha question answering
- Many-to-many: Image labeling / search
- Many-to-one: Search for Jim Gray
- Few-to-one: VizWiz has a few people respond to each blind person's query

Opportunities for Growth in HComp

- Dimensions: Motivation, Quality Control, Aggregation, Human Skill, Process Order, Task-Request Cardinality
- Consider new pairings of dimensions
- Invent new values for dimensions
- Classify new work and consider variations

HW3: Design a HIT.
Sign up for a company
to profile and a
presentation date.