Adventures in Crowdsourcing

Panos Ipeirotis

Stern School of Business
New York University

Thanks to: Jing Wang, Marios Kokkodis, Foster Provost, Josh Attenberg, Victor Sheng, Evgeniy Gabrilovich, Chun How Tan, Ari Kobren, Gabrielle Paolaci, Jesse Chandler

Twitter: @ipeirotis

"A Computer Scientist in a Business School"
http://behind-the-enemy-lines.com

Broad Goal

Integrate machine and human intelligence

Create hybrid "intelligence integration" processes

With **paid** users and with **unpaid** participants

Example Application

Detect Inappropriate Ad Placement

Integral
Ad Science

The New York Times

Politics

Arizona Suspect's Online Trail Offers Hints of Alienation

By ERIC LIPTON, CHARLIE SAVAGE and SCOTT SHANE
Published: January 8, 2011

WASHINGTON — His MySpace page included a photograph of a United States history textbook, on top of which he had placed a handgun. He prepared a series of Internet videos in which he posted odd statements about the gold standard, the community college he attended and SWAT teams.

Enlarge This Image

Mamta Popat/Arizona Daily Star, via Associated Press

Jared Lee Loughner, the suspected gunman, at the 2010 Tucson Festival of Books in March.

Jared Lee Loughner, in these few public hints, offered a sense of his alienation from society, confusion, anger as well as foreboding that his life could soon come to an end. Friends talked of how he had become reclusive in recent years, and his public postings raised questions, in retrospect at least, about his mental state.

Still, his comments offered little indication as to why, as police allege, he would go to a Safeway supermarket in northwest Tucson on Saturday morning and begin shooting at a popular Democratic congresswoman and more than a dozen others, killing six and wounding 19.

There was evidence of recent trouble, though. Mr. Loughner, 22, was suspended in late September from Pima Community College, where he had been attending classes, because the school became aware of a disturbing YouTube video. Mr. Loughner subsequently withdrew from the College

Related
2nd Suspect Sought in Arizona Shooting (January 9, 2011)

What's Popular Now
For Law School Graduates, Debts if Not Job Offers
Arizona Orders Tucson to End Mexican-American Studies Program

Politics E-Mail
Keep up with the latest news from Washington with the daily Politics e-mail newsletter. See Sample
sinan_aral@yahoo.com Sign Up
Change E-mail Address | Privacy Policy

MOST POPULAR
E-MAILED | BLOGGED | SEARCHED | VIEWED

Gabrielle Giffords Shooting, Tucson, AZ, Jan 2011

Anatidaephobia - The Fear That You are Being Watched by a Duck

December 08, 2008 by Tammy Duffey

What Is Anatidaephobia?

Anatidaephobia is defined as a pervasive, irrational fear that one is being watched by a duck. The anatidaephobic individual fears that no matter where they are or what they are doing, a duck watches.

Anatidaephobia is derived from the Greek word "anatidae", meaning ducks, geese or swans and "phobos" meaning fear.

What Causes Anatidaephobia?

As with all phobias, the person coping with Anatidaephobia has experienced a real-life trauma. For the anatidaephobic individual, this trauma most likely occurred during childhood.

Perhaps the individual was intensely frightened by some species of water fowl. Geese and swans are relatively well known for their aggressive tendencies and perhaps the anatidaephobic person was actually bitten or flapped at. Of course, the Far Side comics did little to minimize the fear of being watched by a duck.

While we may be tempted to smile at the memory of those comics or at the mental image of being watched by a duck, for the anatidaephobic person, that fear is uncontrollable. Whatever the cause, the anatidaephobic person can experience emotional turmoil and anxiety that is completely disruptive to daily functioning.

Detect Inappropriate content

- Ad hoc topics, with **no existing training data**
  - Hate speech, Violence, Guns & Bombs, Gossip…
- Classification models need to be trained and deployed within *days*
- **Crowdsourcing** allows for fast data collection
  - labor is accessible on demand
  - using Mechanical Turk, oDesk, etc
  - but quality may be lower than experts

Amazon Mechanical Turk

All HITs

1-10 of 1984 Results

Sort by: HITs Available (most first) GO! Show all details | Hide all details 1 2 3 4 5 › Next » Last

Find the email address for the company and website					View a HIT in this group
Requester:	Sam GONZALES	**HIT Expiration Date:**	Dec 13, 2010 (1 week 2 days)	**Reward:**	$0.01
		Time Allotted:	30 minutes	**HITs Available:**	39172
Identify Arabic Dialect in Text					View a HIT in this group
Requester:	Chris Callison-Burch	**HIT Expiration Date:**	Dec 31, 2010 (3 weeks 6 days)	**Reward:**	$0.05
		Time Allotted:	15 minutes	**HITs Available:**	14240
POI Verfication for USA Cities					View a HIT in this group
Requester:	nutella42	**HIT Expiration Date:**	Dec 17, 2010 (2 weeks)	**Reward:**	$0.08
		Time Allotted:	30 minutes	**HITs Available:**	2446
Preference Judgements between Search Engine Results					View a HIT in this group
Requester:	jaime arquello	**HIT Expiration Date:**	Dec 10, 2010 (7 days)	**Reward:**	$0.03
		Time Allotted:	5 minutes	**HITs Available:**	1952
Keyword Category Verification					View a HIT in this group
Requester:	Andy K	**HIT Expiration Date:**	Dec 9, 2010 (6 days 2 hours)	**Reward:**	$0.03
		Time Allotted:	60 minutes	**HITs Available:**	1949

Example: Build an "Adult Content" Classifier

- Need a large number of labeled sites for training
- Get people to look at sites and label them as:

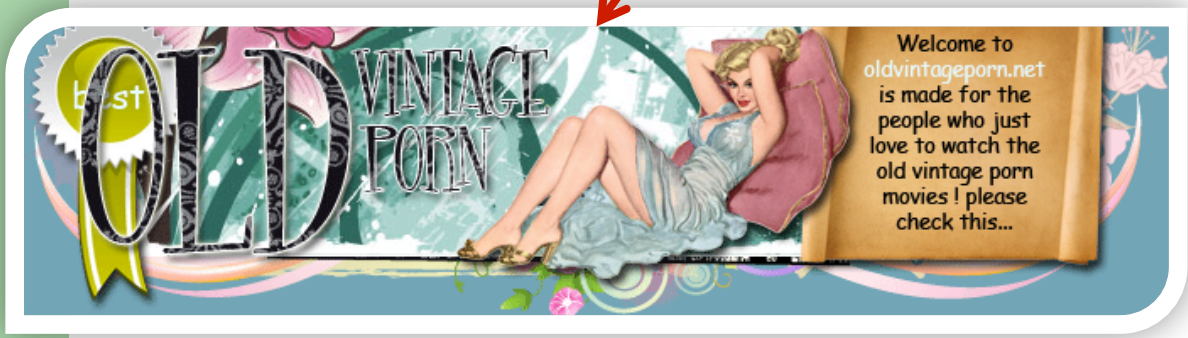
G (general audience) **PG** (parental guidance) **R** (restricted) **X** (porn)

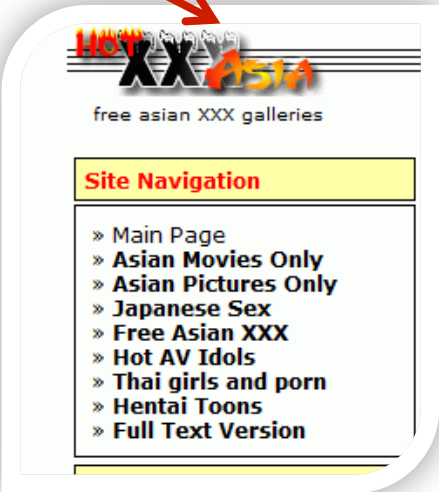
Cost/Speed Statistics

- **Undergrad intern**: 200 websites/hr, cost: $15/hr
- **Mechanical Turk**: 2500 websites/hr, cost: $12/hr

Bad news: Spammers!

61QZ5GG9A12Z548T9AQZ	ATAMRO447HWJQ	http://oldvintageporn.net	G	☐
625ZXHZMQXTMKPMKDZS0	ATAMRO447HWJQ	http://hotxxxasia.com	G	☐





Worker **ATAMRO447HWJQ**

labeled **X (porn)** sites as **G** (general audience)

Challenges

- We do not know the true category for the objects
- We do not know the quality of the workers

- We want to label objects with true categories
- We want (need?) to know the quality of the workers

Redundant votes, infer quality

Look at our lazy friend **ATAMRO447HWJQ** together with other 9 workers

PR7MQ44W2XAZ6FYTYB70	A2VL24C5P7Y3DJ	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	ADU3MDAGZD0UX	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A3LJIDEMXCRZ5R	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A3OHQRF1MDQ99B	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A35GER5TWMH9VP	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A3FN8S0N5JNAL6	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A2JP3HEL3J25AJ	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	A179HLQL4BT5NJ	http://25u.com	G	http://30plus40plus.com	X
PR7MQ44W2XAZ6FYTYB70	ATAMRO447HWJQ	http://25u.com	G	http://30plus40plus.com	G
PR7MQ44W2XAZ6FYTYB70	A2VLOL5DA4M2T1	http://25u.com	G	http://30plus40plus.com	X

- Using redundancy, we can compute error rates for each worker

Expectation Maximization Estimation

Iterative process to estimate worker error rates

1. *Initialize "correct" label for each object (e.g., use majority vote)*
2. Estimate **error rates** for workers (using "correct" labels)
3. Estimate **"correct" labels** (using error rates, weight worker votes according to quality)
4. Go to Step 2 and iterate until convergence

Error rates for ATAMRO447HWJQ

P[G → G]=99.947% **P[G → X]=0.053%**

P[X → G]=99.153% **P[X → X]=0.847%**

The spammer worker marked **almost all** sites as **G**.

Challenge: Humans are biased!

Error rates for the CEO, providing **"expert" labels**

P[G → G]=20.0%	**P[G → P]=80.0%**	P[G → R]=0.0%	P[G → X]=0.0%
P[P → G]=0.0%	P[P → P]=**0.0%**	**P[P → R]=100.0%**	P[P → X]=0.0%
P[R → G]=0.0%	P[R → P]=0.0%	**P[R → R]=100.0%**	P[R → X]=0.0%
P[X → G]=0.0%	P[X → P]=0.0%	P[X → R]=0.0%	**P[X → X]=100.0%**

We have 85% G sites, 5% P sites, 5% R sites, 5% X sites

- Error rate of spammer (all G) = 0% * 85% + 100% * 15% = 15%
- Error rate of biased worker = 80% * 85% + 100% * 5% = 73%

False positives: Legitimate workers appear to be spammers

(important note: bias is not just a matter of "ordered" classes)

Solution: Fix bias first, compute error rate afterwards

Error Rates for CEO

P[G → G]=20.0%	**P[G → P]=80.0%**	P[G → R]=0.0%	P[G → X]=0.0%
P[P → G]=0.0%	P[P → P]=**0.0%**	**P[P → R]=100.0%**	P[P → X]=0.0%
P[R → G]=0.0%	P[R → P]=0.0%	**P[R → R]=100.0%**	P[R → X]=0.0%
P[X → G]=0.0%	P[X → P]=0.0%	P[X → R]=0.0%	**P[X → X]=100.0%**

- When biased worker says G, it is **100% G**
- When biased worker says P, it is **100% G**
- When biased worker says R, it is **50% P, 50% R**
- When biased worker says X, it is **100% X**

Small ambiguity for "R-rated" votes but other than that, fine!

Expected Misclassification Cost

- **High cost:** probability spread across classes
- **Low cost:** probability mass concentrated in one class

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

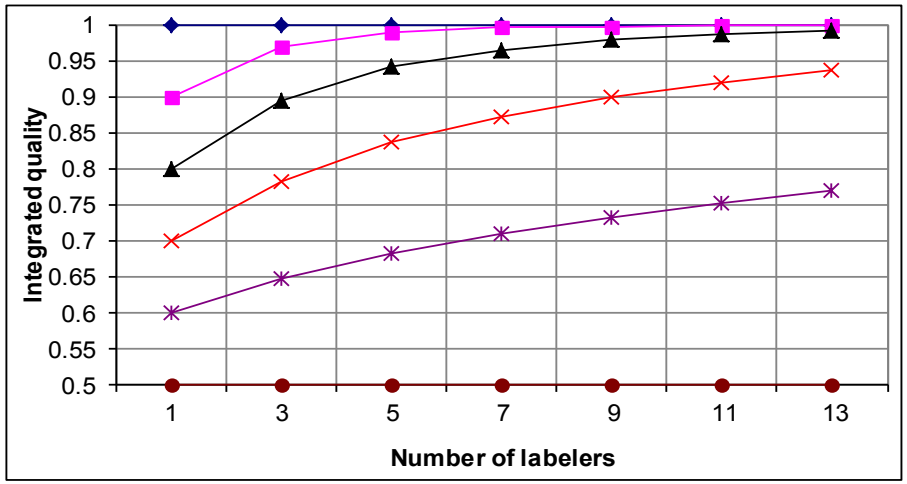
[Assume misclassification cost equal to 1, solution generalizes to arbitrary costs]

Question: How to pay workers?

- **Naïve** solution: Have a quality-score ***threshold***
- **Thresholding rewards gives wrong incentives:**
  - Very different outcomes around the threshold, for similar performance
  - Often uncertain about true performance
  - Decent (but still useful) workers get punished

Quality-sensitive Payment

- Set quality goal and price (e.g., $1 for 90%)
  - For workers above goal: Pay full price
  - For others: Payment divided with redundancy needed to reach goal
    - Need 3 workers with 80% accuracy ➔ Payment = $1/3 = $0.33
    - Need 9 workers with 70% accuracy ➔ Payment = $1/9 = $0.11



How to deal with uncertaintv?

Instead of blocking: Quality-sensitive Payment

- **Uncertainty hurts:**
  - Small fluctuations in performance may result in drastic payment changes
  - Payment decreases practically equivalent to rejection

- Introduced ***uncertainty "penalty":*** **Pay less** for uncertain estimates (for workers with short working histories)

- **Refund** underpayment when quality estimate more certain

Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

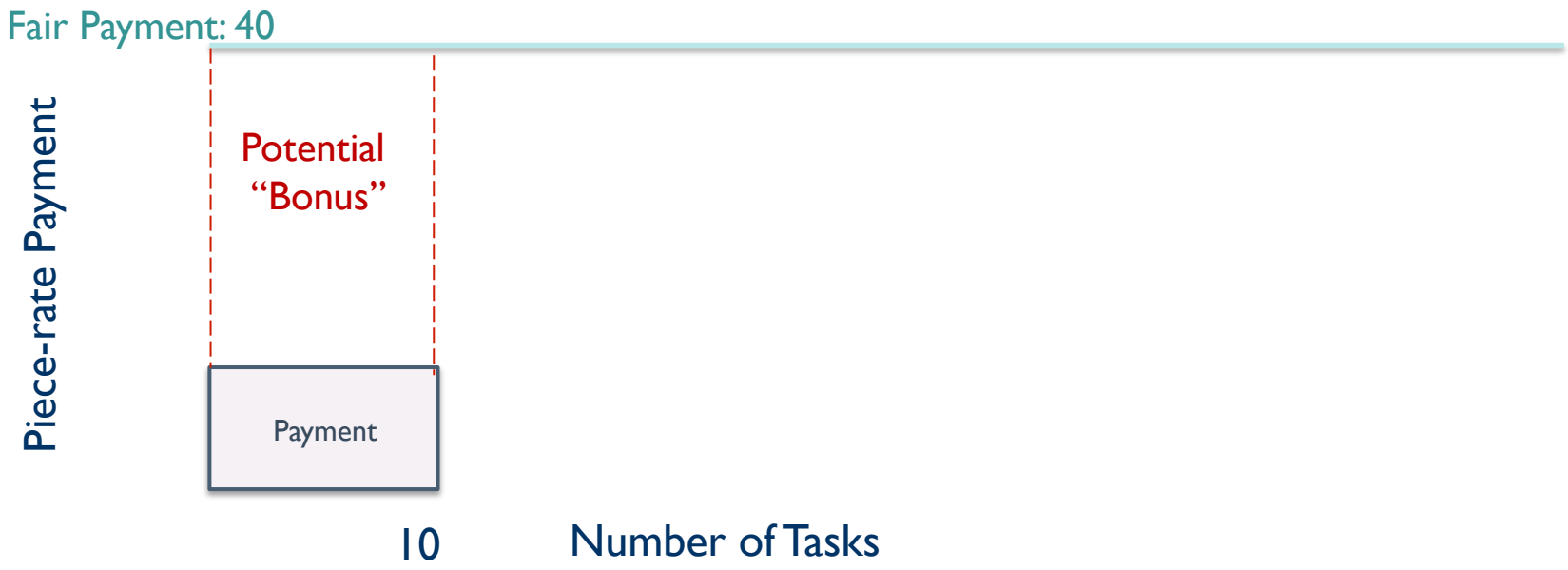
# Tasks	10	20	30	40	Infinity
Piece-rate Payment (cents)	11	18	21	23	40

Fair Payment

Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

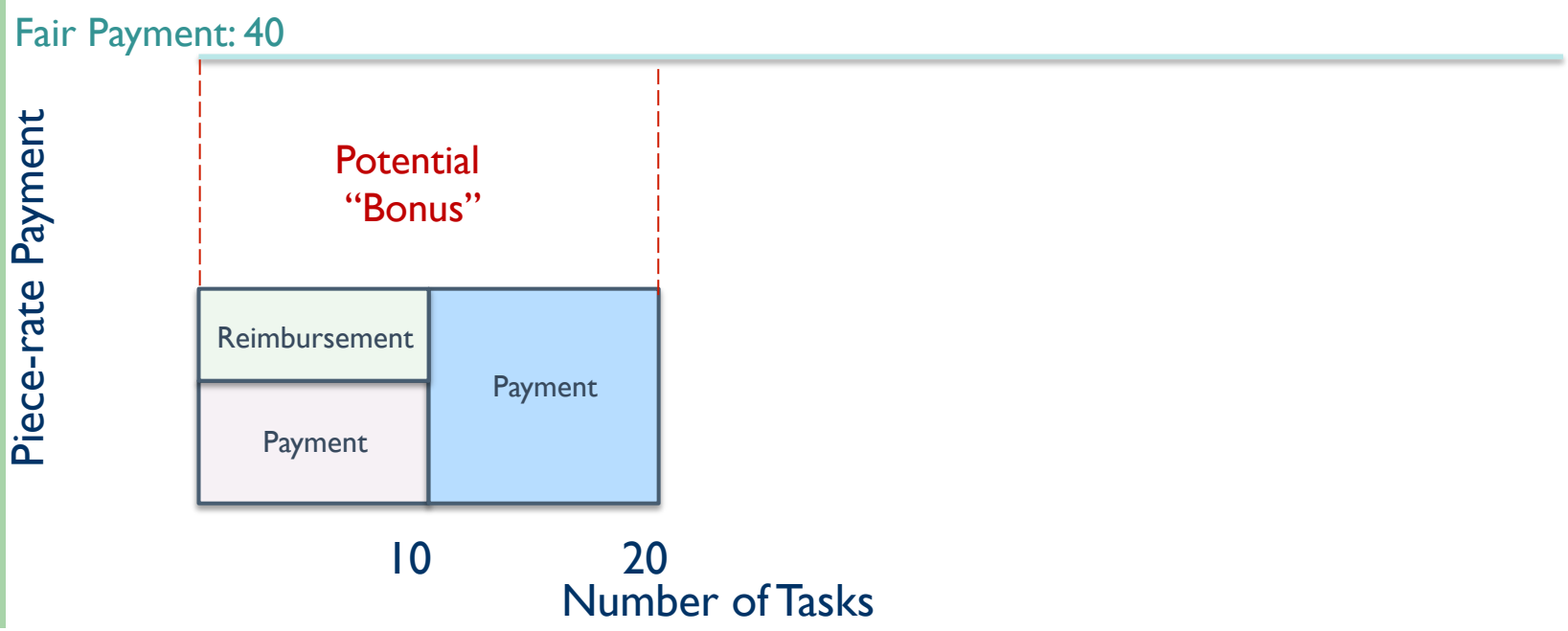
# Tasks	10	20	30	40	Infinity
Piece-rate Payment (cents)	11	18	21	23	40



Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

# Tasks	10	20	30	40	Infinity
Piece-rate Payment (cents)	11	18	21	23	40

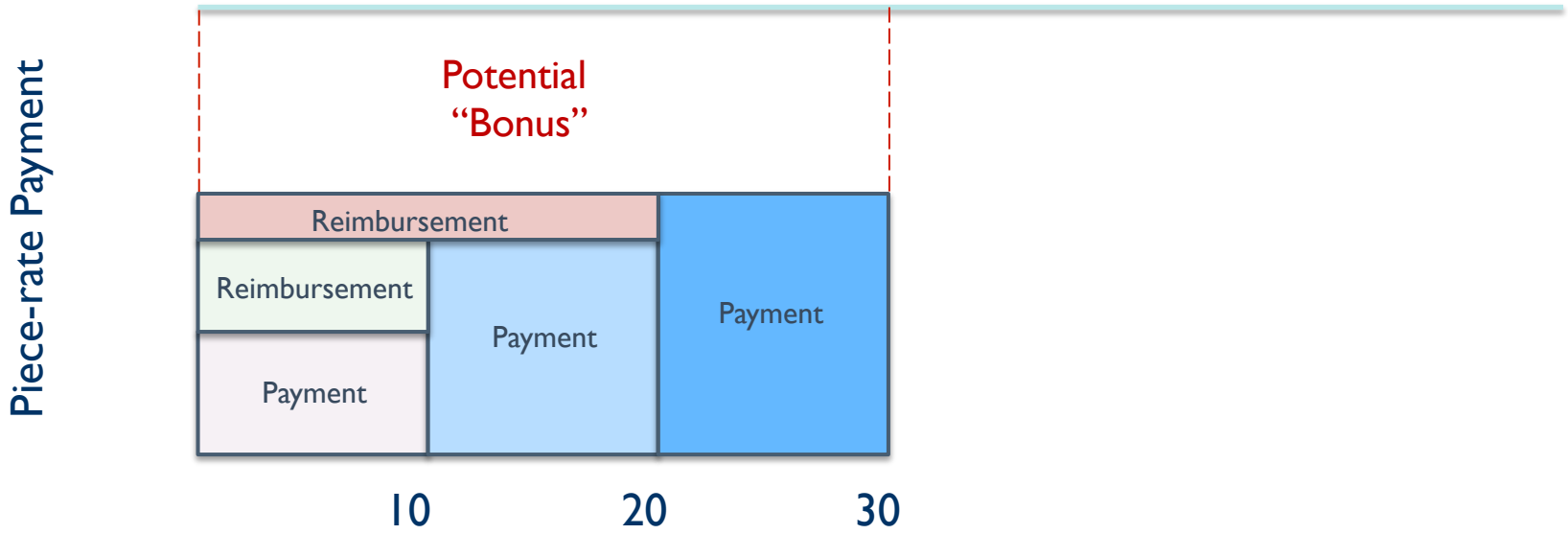


Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

#Tasks	10	20	30	40	Infinity
Piece-rate Payment (cents)	11	18	21	23	40

Fair Payment: 40

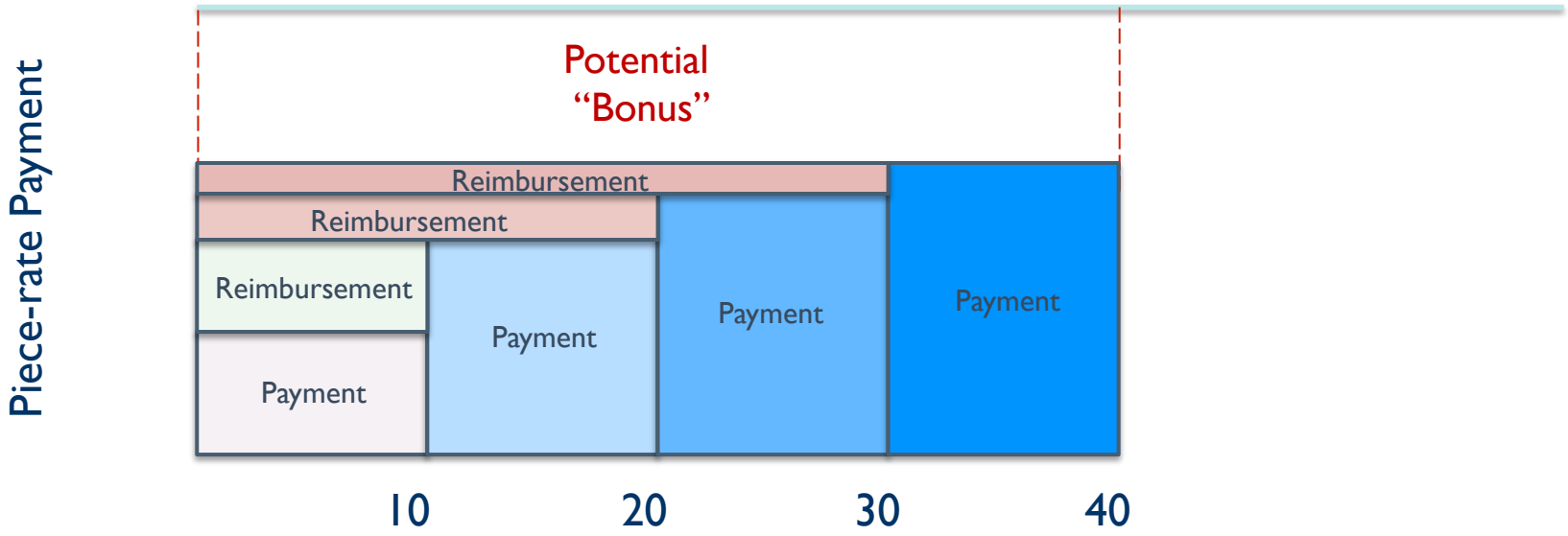


Real-Time Payment and Reimbursement

Example of the piece-rate payment of a worker

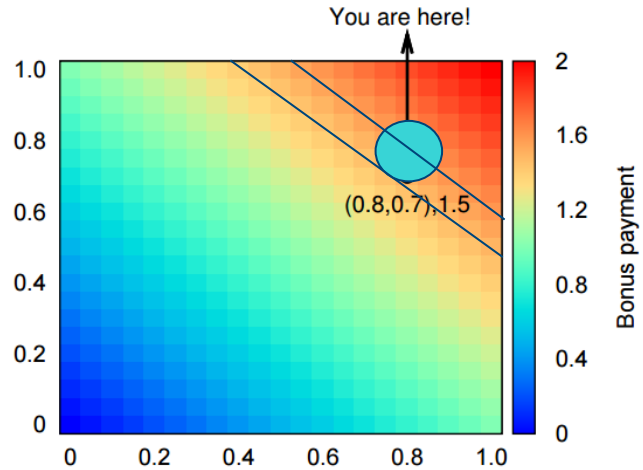
#Tasks	10	20	30	40	Infinity
Piece-rate Payment (cents)	11	18	21	23	40

Fair Payment: 40



Summary of Experimental Results

- Randomized Controlled Trial on oDesk
  - Thresholding,
  - Quality-based payment (QBP)
  - QBP with reimbursements



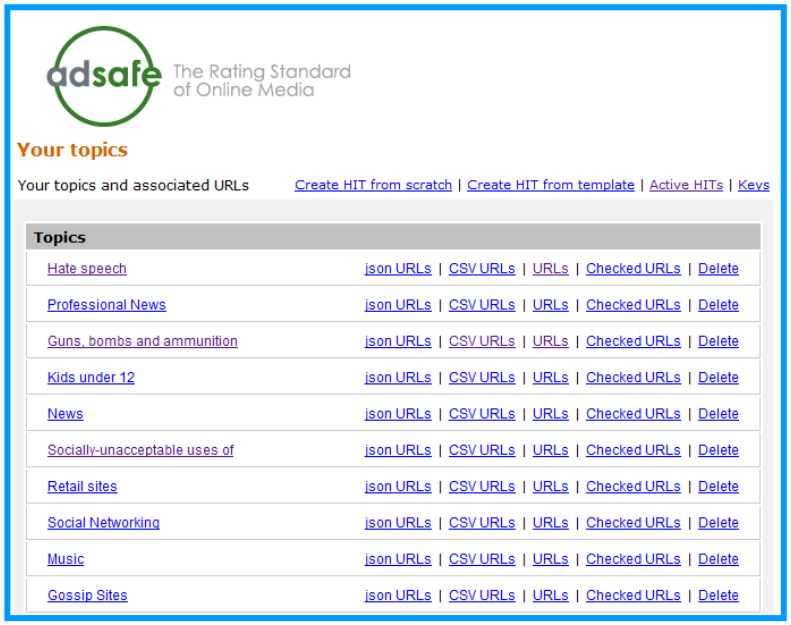
- **Retention**: ~150-300% up over thresholding/QBP
  - No significant differences between QBP/thresholding
  - Decrease in pay, same effect as rejection
- **Cost:** 50%-70% reduction, as we pay for performance
- **Work quality**: Stable

Humans Improving Machine Learning

- With just labeling, workers are **passively** labeling the data that we give them

- Asking instead the workers to search and **find training data**

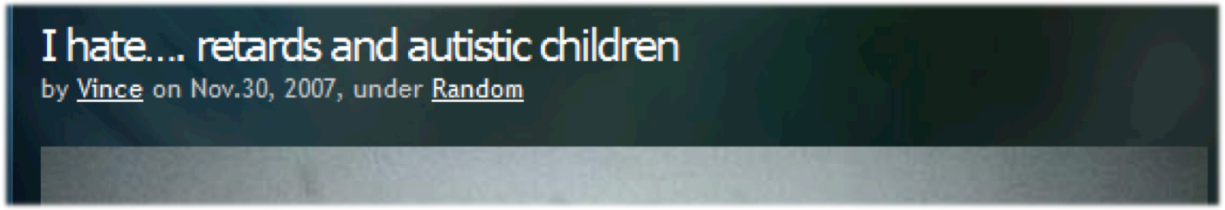
- **Vanilla solution: Use data and build model**



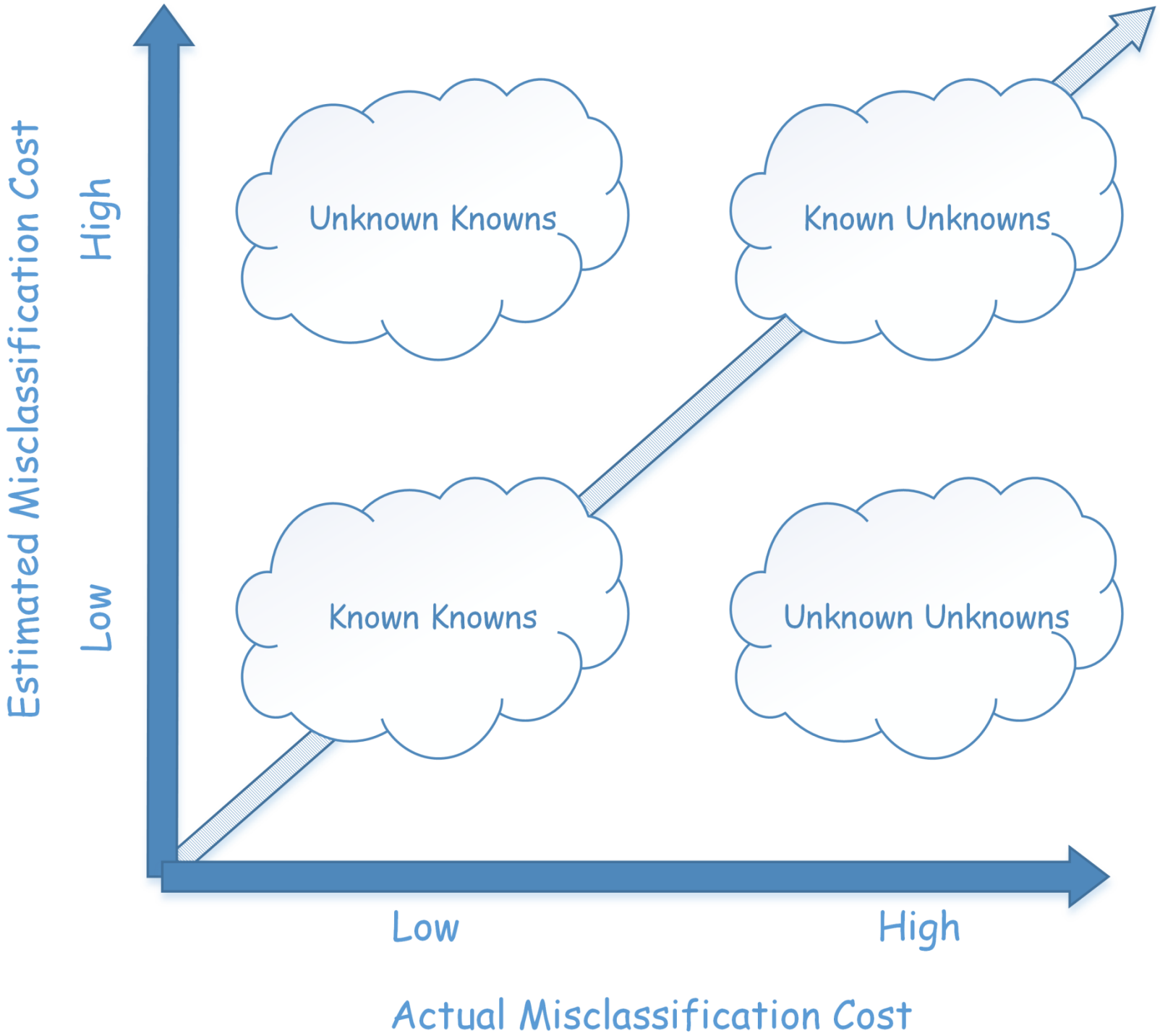
The result? Blissful ignorance…

- Classifier ***seems*** great: Cross-validation tests show excellent performance

- Alas, classifier fails: The "*unknown unknowns*" ™



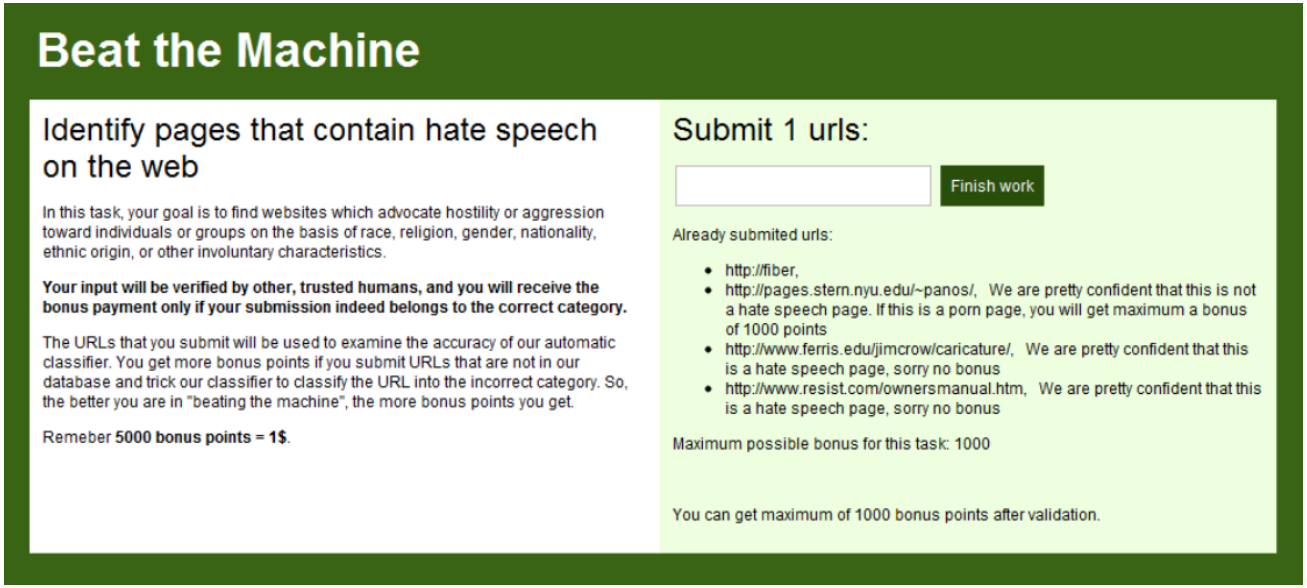
"*Unknown unknowns*" = classifier fails with high confidence


Estimated Misclassification Cost
High
Low
Unknown Knowns
Known Unknowns
Known Knowns
Unknown Unknowns
Low
High
Actual Misclassification Cost

Beat the Machine!

Ask humans to find URLs that

- ***the classifier will classify incorrectly***
- ***another human will classify correctly***

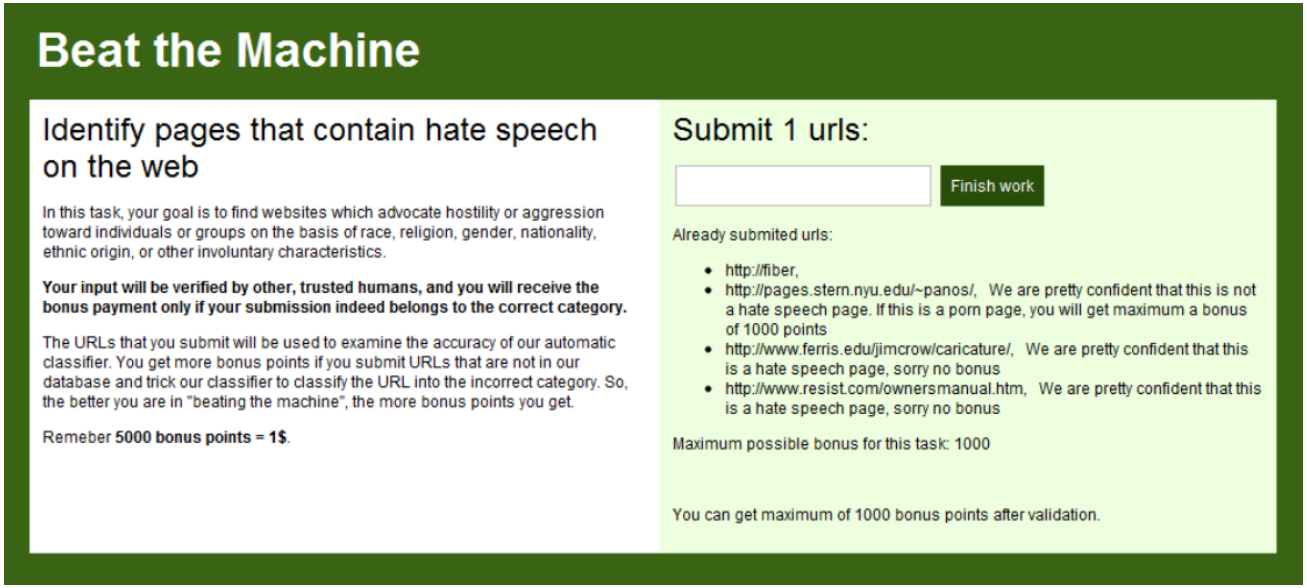


Example:
Find hate speech pages that the machine will classify as benign

Beat the Machine!

Incentive structure:

- ***$1 if you "beat the machine"***
- ***$0.001 if the machine already knows***



Beat the Machine

Identify pages that contain hate speech on the web

In this task, your goal is to find websites which advocate hostility or aggression toward individuals or groups on the basis of race, religion, gender, nationality, ethnic origin, or other involuntary characteristics.

Your input will be verified by other, trusted humans, and you will receive the bonus payment only if your submission indeed belongs to the correct category.

The URLs that you submit will be used to examine the accuracy of our automatic classifier. You get more bonus points if you submit URLs that are not in our database and trick our classifier to classify the URL into the incorrect category. So, the better you are in "beating the machine", the more bonus points you get.

Remeber **5000 bonus points = 1$**.

Submit 1 urls:

Finish work

Already submited urls:

- http://fiber,
- http://pages.stern.nyu.edu/~panos/, We are pretty confident that this is not a hate speech page. If this is a porn page, you will get maximum a bonus of 1000 points
- http://www.ferris.edu/jimcrow/caricature/, We are pretty confident that this is a hate speech page, sorry no bonus
- http://www.resist.com/ownersmanual.htm, We are pretty confident that this is a hate speech page, sorry no bonus

Maximum possible bonus for this task: 1000

You can get maximum of 1000 bonus points after validation.

Example:
Find hate speech pages that the machine will classify as benign

#	Category	Tasks Running	URL's gathered	Correct URL's gathered	Total Bonus
1	Identify pages that contain hate speech on the web (*hat*)	206	1023	161	75516
2	Identify pages related to illegal drug use on the web (*drg*)	100	500	26	9114
3	Identify pages that contain reference to alcohol (*alc*)	100	475	144	55149
4	Identify adult-related pages (*adt*)	174	859	132	63523

Probes Successes

Error rate for probes significantly higher than error rate on (stratified) random data (10x to 100x higher than base error rate)

Conclusion: Humans are good in discovering problematic cases for model testing

Finding People to Beat the Machine

Question: Can we find humans that can and are willing to "beat the machine"?

Example Application: Improving Automatic Essay Scoring

Audience Discovery?

- How can we automate the process of **discovering good users** for arbitrary crowdsourcing applications?

Google Knowledge Graph

Kyrgyzstan

Country

Kyrgyzstan, officially the Kyrgyz Republic, is a country located in Central Asia. Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Wikipedia

Capital: Bishkek

Currency: Kyrgyzstani som

President: Almazbek Atambayev

National anthem: National Anthem of the Kyrgyz Republic

Official languages: Kyrgyz language, Russian Language

Government: Presidential system, Parliamentary republic, Republic

"Things not Strings"

Still incomplete…

- “Date of birth of Bayes” (…uncertain…)
- “Symptom of strep throat”
- “Side effects of treximet”
- “Who is Cristiano Ronaldo dating”
- “When is Jay Z playing in New York”
- “What is the customer service number for Google”
- …

Key Challenge

*"Crowdsource in a **predictable** manner,*
***with knowledgeable** users,*
***without** introducing **monetary rewards**"*

www.quizz.us

Correct Answers: 33/67 Correct (%): 49%

What is a symptom of Morgellons

- Red eye
- Choreoathetosis
- Skin lesion
- Insomnia
- I don't know

Question 1 out of 10

Calibration vs. Collection

- **Calibration** questions (known answer): Evaluating user competence on topic at hand
- **Collection** questions (unknown answer): Asking questions for things we do not know
- *Trust more answers coming from competent users*

Challenges

- Why would **anyone** come and play this game?
- Why would **knowledgeable** users come?
- Wouldn’t it be simpler to **just pay**?

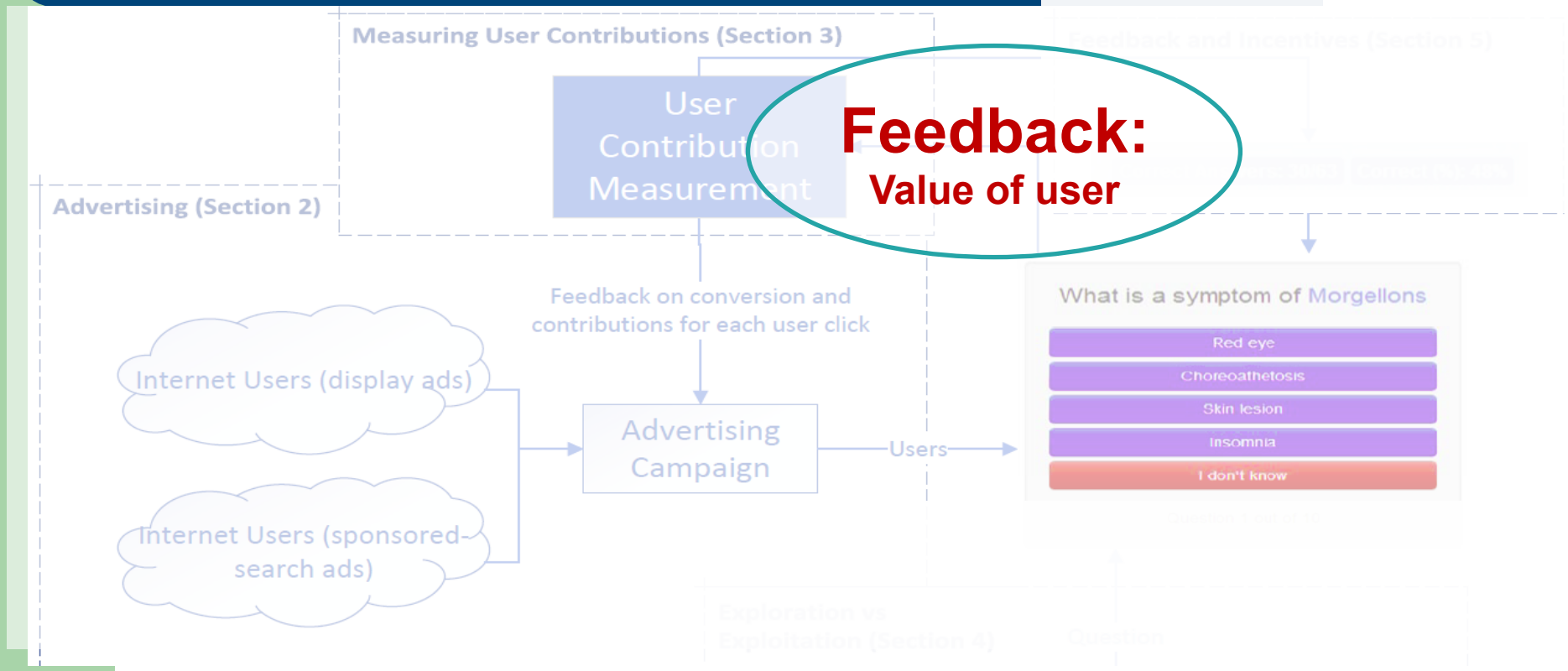
Attracting Visitors: Ad Campaigns

Quiz on disease symptoms

Test how well you can recognize various disease symptoms

www.quizz.us

Treat Quizz as eCommerce Site



- Value of user: ***total*** information gain contributed
- Information gain is additive: **#questions x info/question**

Example of Targeting: Medical Quizzes

- Our initial goal was to use medical topics as a evidence that some topics are ***not*** crowdsourcable
- Our hypothesis failed: They were the best performing quizzes…
- Users coming from sites such as Mayo Clinic, WebMD, … (i.e., “pronsumers”, not professionals)

Immediate feedback helps

Treatment	Effect
Show if user answer correct	**+2.4%**
Show the correct answer	**+20.4%**
Score: % of correct answers	+2.3%
Score: # of correct answers	-2.2%
Score: Information gain	+4.0%
Show statistics for performance of other users	+9.8%
Leaderboard based on percent correct	-4.8%
Leaderboard based on total correct answers	-1.5%

- Knowing the correct answer 10x more important than knowing whether given answer was correct
- Conjecture: Users also want to learn

Showing score moderately helpful

Treatment	Effect
Show if user answer correct	+2.4%
Show the correct answer	+20.4%
Score: % of correct answers	**+2.3%**
Score: # of correct answers	**-2.2%**
Score: Information gain	**+4.0%**
Show statistics for performance of other users	+9.8%
Leaderboard based on percent correct	-4.8%
Leaderboard based on total correct answers	-1.5%

- Be careful what you incentivize ☺
- “Total Correct” incentivizes quantity, not quality

Competitiveness helps

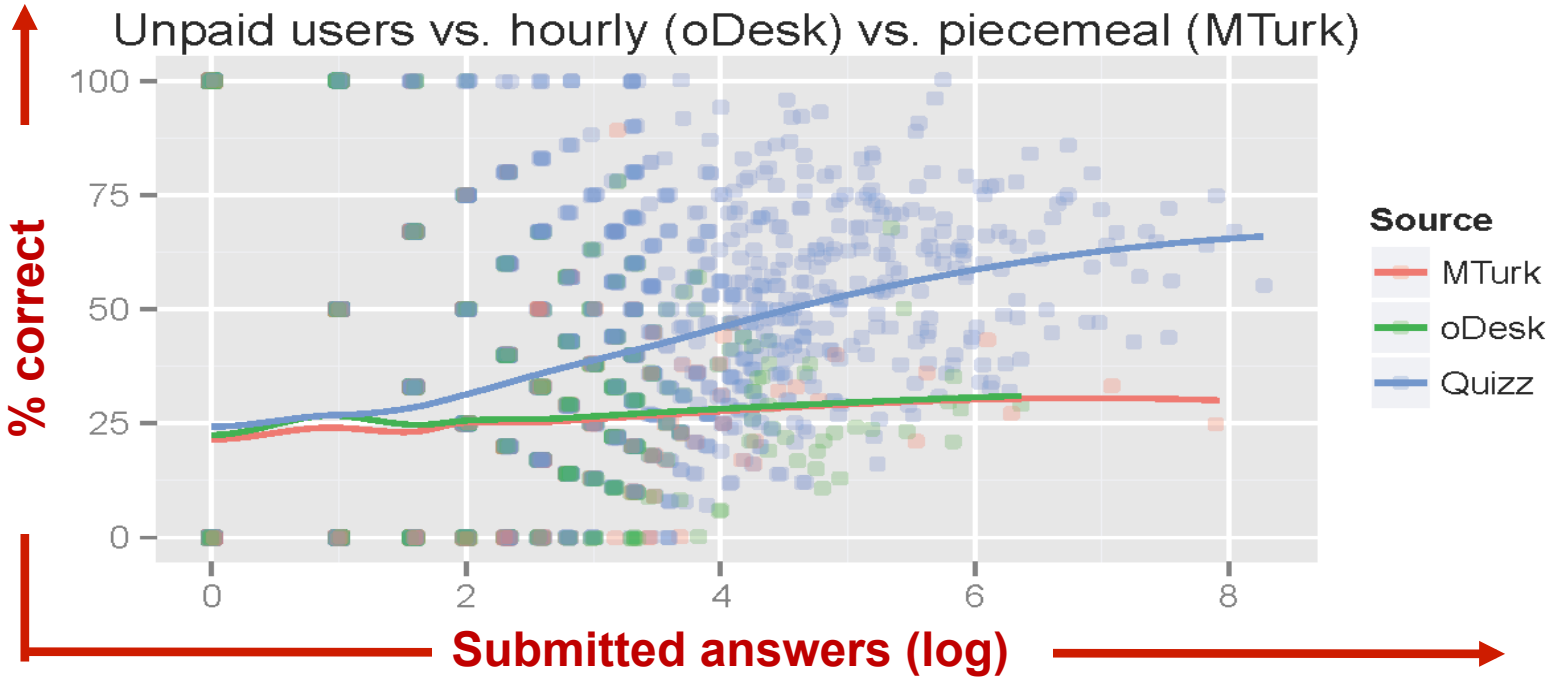
Treatment	Effect
Show if user answer correct	+2.4%
Show the correct answer	+20.4%
Score: % of correct answers	+2.3%
Score: # of correct answers	-2.2%
Score: Information gain	+4.0%
Show statistics for performance of other users	**+9.8%**
Leaderboard based on percent correct	-4.8%
Leaderboard based on total correct answers	-1.5%

Leaderboards are tricky!

Treatment	Effect
Show if user answer correct	+2.4%
Show the correct answer	+20.4%
Score: % of correct answers	+2.3%
Score: # of correct answers	-2.2%
Score: Information gain	+4.0%
Show statistics for performance of other users	+9.8%
Leaderboard based on percent correct	**-4.8%**
Leaderboard based on total correct answers	**-1.5%**

- Initially, strong positive effect
- Over time, effect became strongly negative
- All-time leaderboards considered harmful

Comparison with paid crowdsourcing



Citizen Science Applications

- Google gives **$10K/month** to nonprofits in ad budget
- Climate CoLab experiment
  - Doubled traffic with only $20/day
  - Targets political activist groups (not only climate)
- Additional experiments:
  - Identify users with particular psychological characteristics
  - Engage users with an interest in speech therapy

How can I get rid of users?

National Academy of Sciences "Frontiers of Science" conference

Your workers behave like my mice!

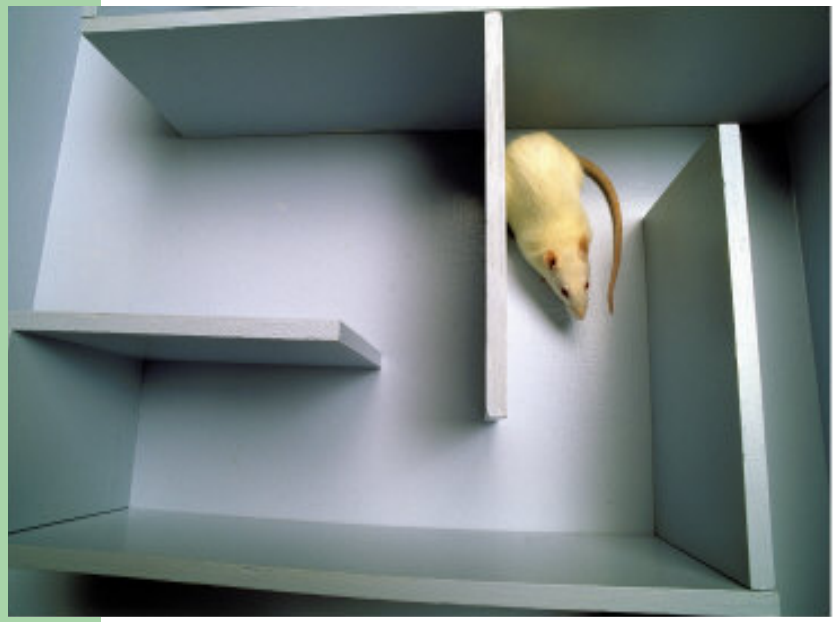
An unexpected connection…

Your workers want to use only their **motor skills**, **not** their **cognitive skills**

The Biology Fundamentals

- Brain functions are biologically expensive (20% of total energy consumption in humans)
- Motor skills are more energy efficient than cognitive skills (e.g., walking)
- Brain tends to delegate easy tasks to part of the neural system that handles motor skills

The Mice Experiment



Cognitive

Solve maze
Find pellet

Motor

Push lever three times
Pellet drops

How to Train the Mice?

Confuse motor skills!
Reward cognition!

Punishing Worker's Motor Skills

- **Punish bad answers** with frustration of motor skills (e.g., add delays between tasks)
  - "Loading image, please wait…"
  - "Image did not load, press here to reload"
  - "404 error. Return the HIT and accept again"

→Make this **probabilistic** to keep feedback implicit

Experimental Summary

- Spammer workers quickly abandon
- No need to display scores, or ban
- Low quality submissions from ~60% to ~3%
- Half-life of low-quality users from 100+ tasks to less than 5

Thanks!

Q & A?