Crowdsourcing using Mechanical Turk: Quality Management and Scalability

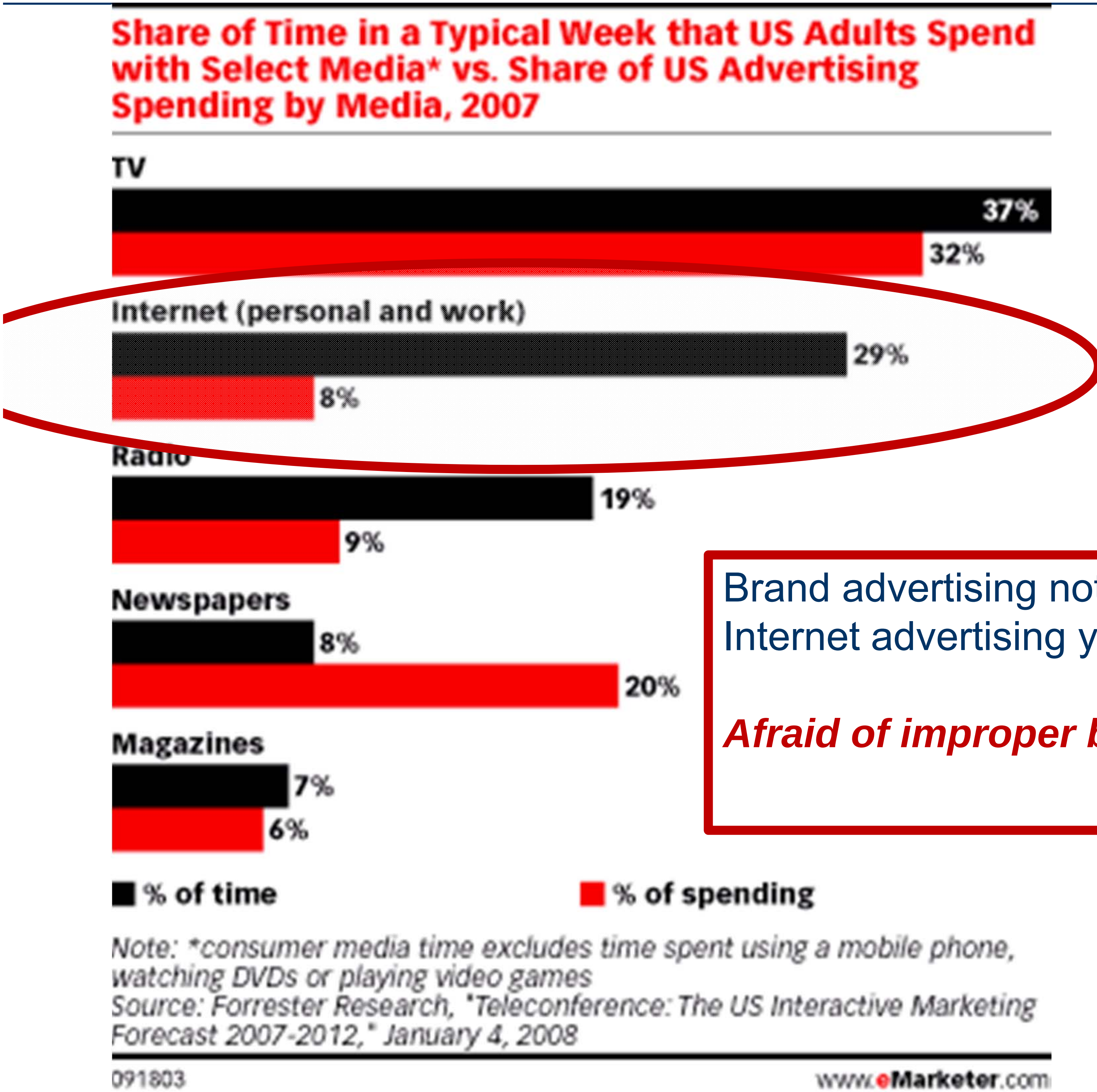
Panos Ipeirotis

New York University & oDesk

Joint work with: Jing Wang, Foster Provost, Josh Attenberg, and Victor Sheng; Special thanks to AdSafe Media

Twitter: @ipeirotis

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



Brand advertising not fully embraced Internet advertising yet…

Afraid of improper brand placement

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

Subscribe to The Times | Welcome, sinan74 | Log Out | Help | TimesPeople

The New York Times

Politics

Search All NYTimes.com

Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

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

Marnta 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

RECOMMEND
TWITTER
E-MAIL
SEND TO PHONE
PRINT
REPRINTS
SHARE

Related

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

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?

Log In With Facebook

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.

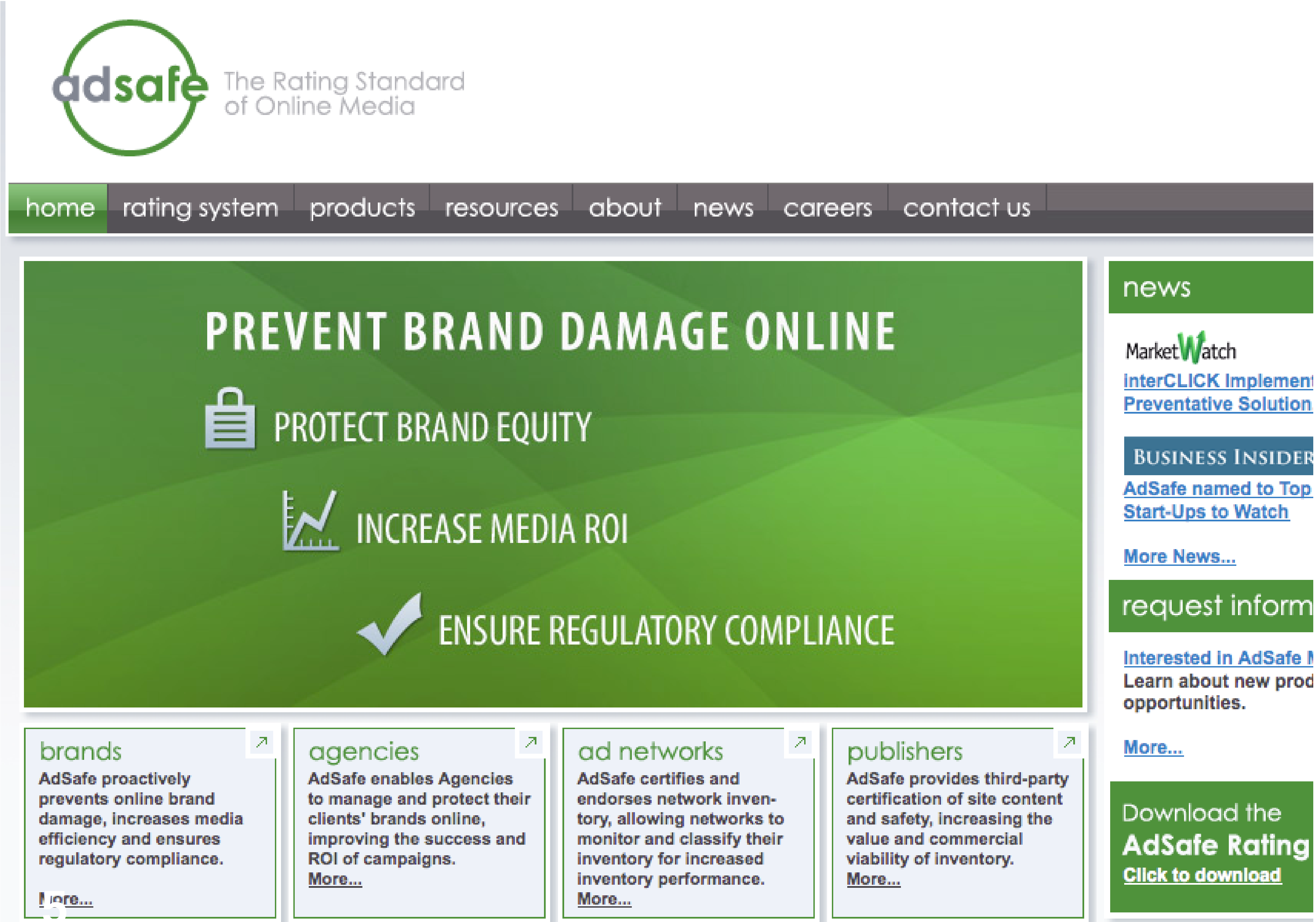
Anatidaephobia is derived from the Greek word "anatidae", meaning ducks, geese or swans and "phobos" meaning fear.

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.


adsafe
The Rating Standard of Online Media
home
rating system
products
resources
about
news
careers
contact us
PREVENT BRAND DAMAGE ONLINE
PROTECT BRAND EQUITY
INCREASE MEDIA ROI
ENSURE REGULATORY COMPLIANCE
news
MarketWatch
interCLICK Implement
Preventative Solution
Business Insider
AdSafe named to Top
Start-Ups to Watch
More News...
request inform
Interested in AdSafe
Learn about new prod
opportunities.
More...
Download the
AdSafe Rating
Click to download
brands
AdSafe proactively prevents online brand damage, increases media efficiency and ensures regulatory compliance.
More...
agencies
AdSafe enables Agencies to manage and protect their clients' brands online, improving the success and ROI of campaigns.
More...
ad networks
AdSafe certifies and endorses network inventory, allowing networks to monitor and classify their inventory for increased inventory performance.
More...
publishers
AdSafe provides third-party certification of site content and safety, increasing the value and commercial viability of inventory.
More...

New Classification Models Needed *within days*

- Pharmaceutical firm does not want ads to appear:
  - In pages that discuss **swine flu** (FDA prohibited pharmaceutical company to display drug ad in pages about swine flu)
- Big fast-food chain does not want ads to appear:
  - In pages that discuss the brand (99% negative sentiment)
  - In pages discussing obesity, diabetes, cholesterol, etc
- Airline company does not want ads to appear:
  - In pages with crashes, accidents, …
  - In pages with discussions of terrorist plots against airlines

Need to build models fast

- **Traditionally**, modeling teams have invested substantial internal resources in data collection, extraction, cleaning, and other preprocessing

No time for such things…

- However, now, we can outsource preprocessing tasks, such as labeling, feature extraction, verifying information extraction, etc.
  - using Mechanical Turk, oDesk, etc.
  - quality may be lower than expert labeling (much?)
  - but low costs can allow massive scale

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 arguello	**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 Web Site" Classifier

- Need a large number of hand-labeled sites
- Get people to look at sites and classify them as:

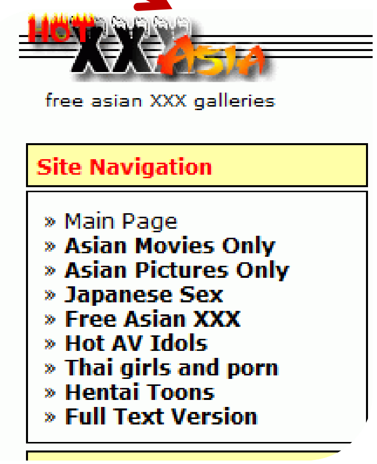
G (general audience) **PG** (parental guidance) **R** (restricted) **X** (porn)

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)

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

Algorithm of (Dawid & Skene, 1979)

[and many recent variations on the same theme]

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%**

Our friend ATAMRO447HWJQ marked **almost all** sites as **G**. Clickety clickey click…

Challenge: From Confusion Matrixes to Quality Scores

Confusion Matrix for ATAMRO447HWJQ

- P[X → X]=0.847%　　P[X → G]=99.153%
- P[G → X]=0.053%　　P[G → G]=99.947%

How to check if a worker is a spammer using the confusion matrix?

(hint: error rate not enough)

Challenge 1:
Spammers are lazy and smart!

Confusion matrix for spammer
- $P[X \rightarrow X]=0\%$ $P[X \rightarrow G]=100\%$
- $P[G \rightarrow X]=0\%$ $P[G \rightarrow G]=100\%$

Confusion matrix for good worker
- $P[X \rightarrow X]=80\%$ $P[X \rightarrow G]=20\%$
- $P[G \rightarrow X]=20\%$ $P[G \rightarrow G]=80\%$

- Spammers figure out how to fly under the radar…
- In reality, we have 85% G sites and 15% X sites
- Error rate of spammer = 0% * 85% + 100% * 15% = 15%
- Error rate of good worker = 85% * 20% + 85% * 20% = 20%

False negatives: Spam workers pass as legitimate

Challenge 2: Humans are biased!

Error rates for **CEO of AdSafe**

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: Reverse errors first, compute error rate afterwards

Error Rates for CEO of AdSafe

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!

Solution: Reverse errors first, compute error rate afterwards

Error Rates for spammer: ATAMRO447HWJQ

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

- When spammer says G, it is **25% G, 25% P, 25% R, 25% X**
- When spammer says P, it is **25% G, 25% P, 25% R, 25% X**
- When spammer says R, it is **25% G, 25% P, 25% R, 25% X**
- When spammer says X, it is **25% G, 25% P, 25% R, 25% X**

[note: assume equal priors]

The results are highly ambiguous. No information provided!

Expected Misclassification Cost

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

Assigned Label	Corresponding “Soft” Label	Expected Label Cost
Spammer: G	<G: 25%, P: 25%, R: 25%, X: 25%>	0.75
Good worker: P	<G: 100%, P: 0%, R: 0%, X: 0%>	0.0

[***Assume misclassification cost equal to 1, solution generalizes]

Quality Score: A scalar measure of quality

- A *spammer* is a worker who always assigns labels randomly, regardless of what the true class is.

$$QualityScore(\text{Worker}) = 1 - \frac{ExpCost(\text{Worker})}{ExpCost(\text{Spammer})}$$

- Scalar score, useful for the purpose of ranking workers

Instead of blocking: Quality-sensitive Payment

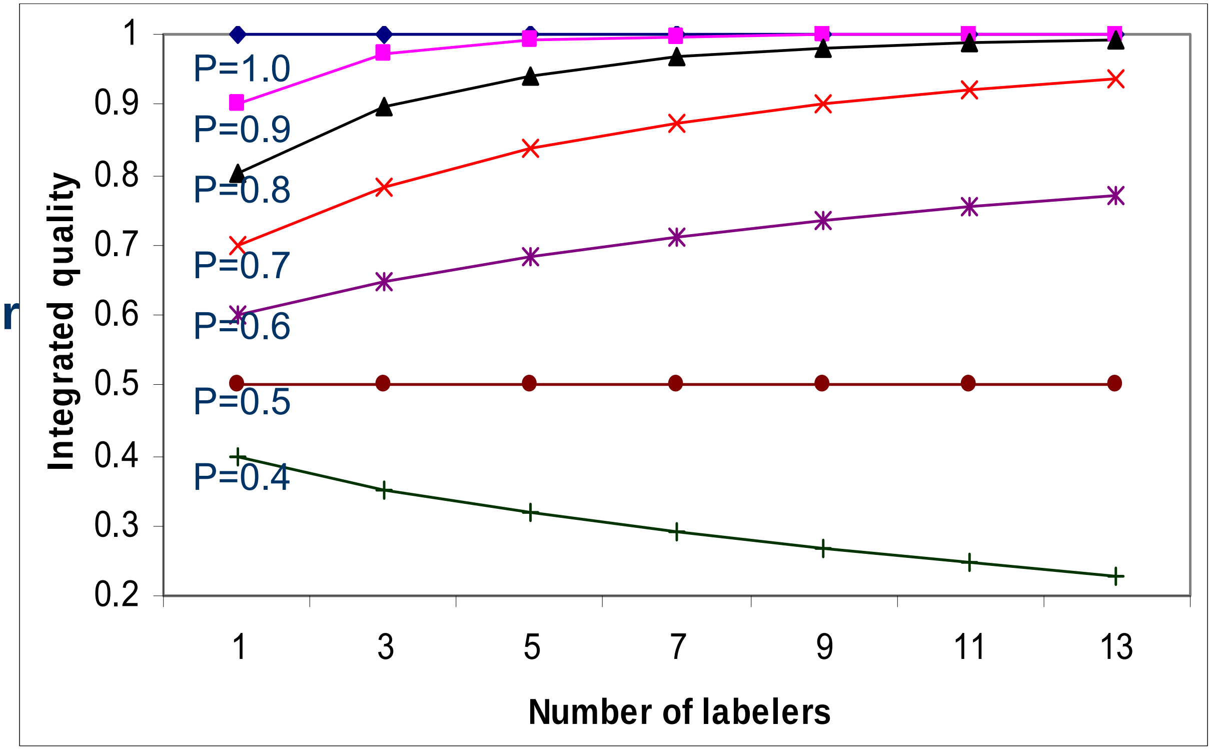
- **Threshold-ing rewards gives wrong incentives:**
  - Decent (but still useful) workers get fired
  - Uncertainty near the decision threshold
- **Instead: Estimate payment level based on quality**
  - Set acceptable quality (e.g., 99% accuracy)
  - For workers above quality specs: Pay full price
  - For others: Estimate level of redundancy to reach acceptable quality (e.g., Need 5 workers with 90% accuracy or 13 workers with 80% accuracy to reach 99% accuracy;)
  - Pay full price divided by level of redundancy

Simple example: Redundancy and Quality

- Ask multiple labelers, keep majority label as “true” label
- Quality is probability of being correct

P is probability of individual **labeler** being correct

P=1.0: perfect
P=0.5: random
P=0.4: adversarial



Implementation

Open source implementation available at:

http://code.google.com/p/get-another-label/

and demo at http://qmturk.appspot.com/

- Input:
  - *Labels from Mechanical Turk*
  - *[Optional] Some "gold" labels from trusted labelers*
  - *Cost of incorrect classifications (e.g., X→G costlier than G→X)*
- Output:
  - *Corrected labels*
  - *Worker error rates*
  - *Ranking of workers according to their quality*
  - *[Coming soon] Quality-sensitive payment*
  - *[Coming soon] Risk-adjusted quality-sensitive payment*

adsafe The Rating Standard of Online Media

Example: Build an "Adult Web Site" Classifier

- Get people to look at sites and classify them as:

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

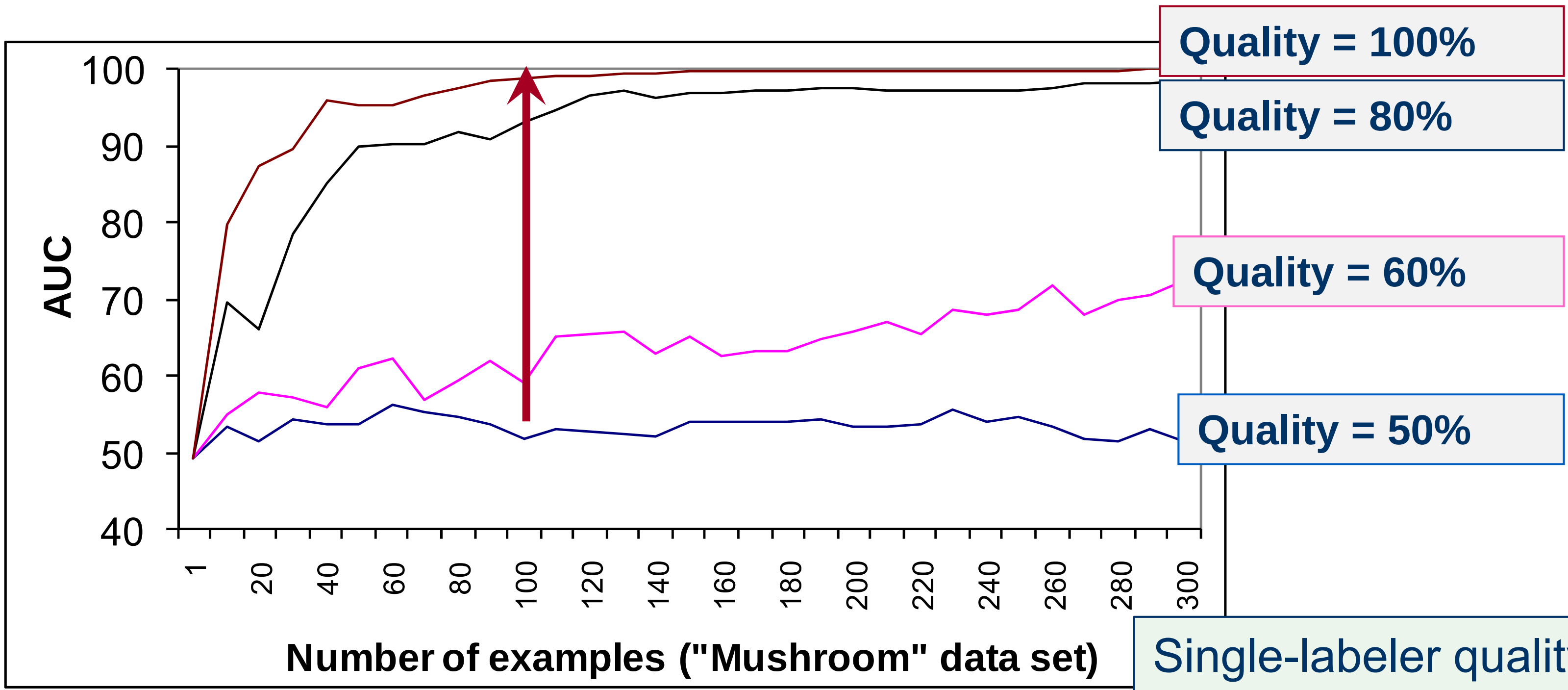
But we are not going to label the whole Internet…

- **Expensive**
- **Slow**

Quality and Classification Performance

Noisy labels lead to degraded task performance

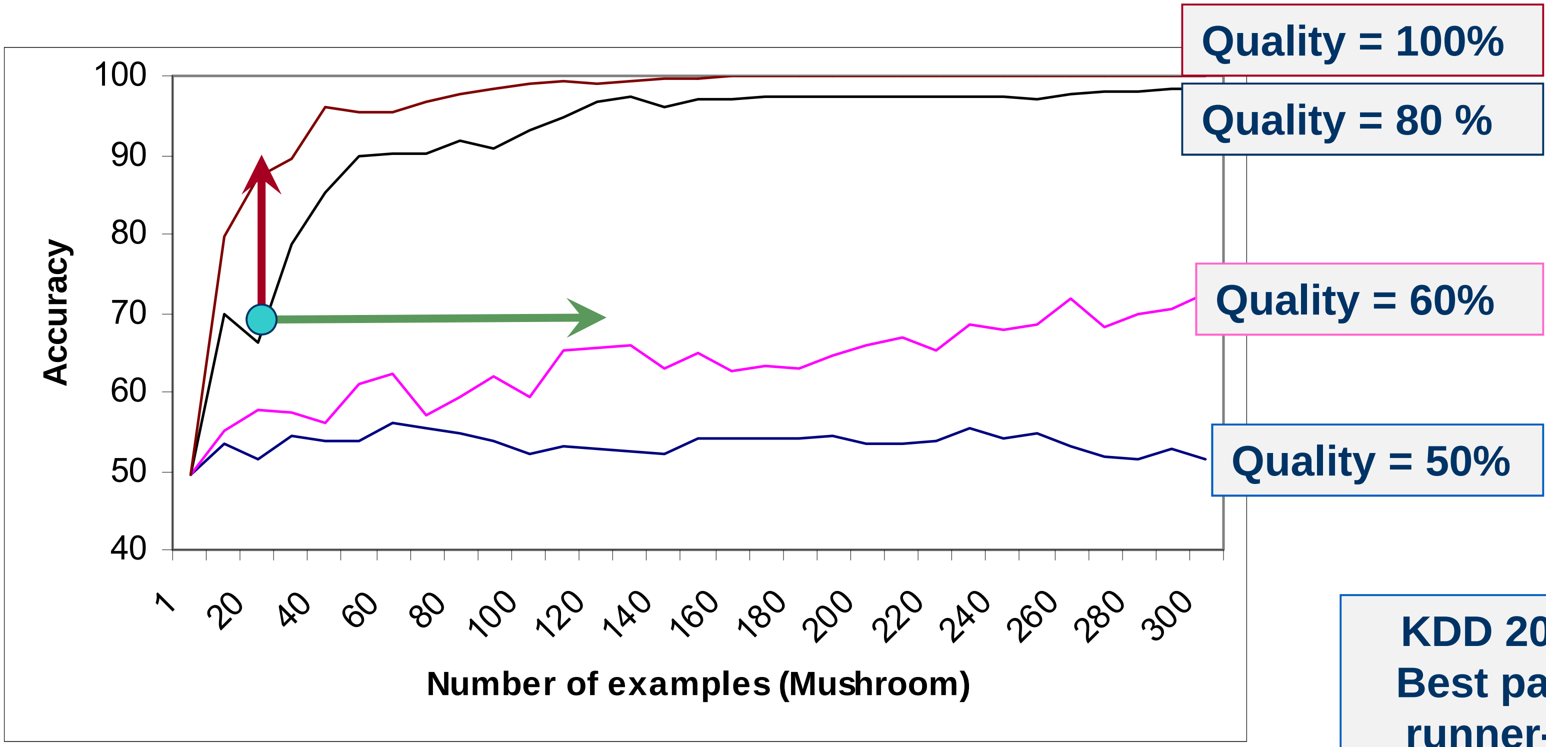
Labeling quality increases → classification quality increases



Single-labeler quality (probability of assigning correctly a binary label)

Tradeoffs: More data or better data?

- Get more examples → Improve classification
- Get more labels → Improve label quality → Improve classification



KDD 2008, Best paper runner-up

Summary of Basic Results

We want to follow the direction that has the highest "learning gradient"

- Estimate improvement with more data (cross-validation)
- Estimate sensitivity to data quality (introduce noise and measure degradation in quality)

Rule-of-thumb results:

With high quality labelers (85% and above):
Get more data (One worker per example)

With low quality labelers (~60-70%):
Improve quality (Multiple workers per example)

Selective Repeated-Labeling

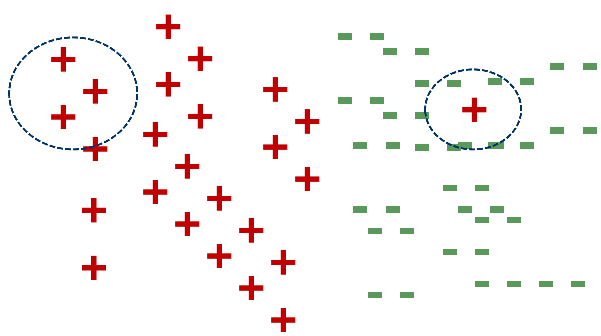
- We do not need to label everything the same way

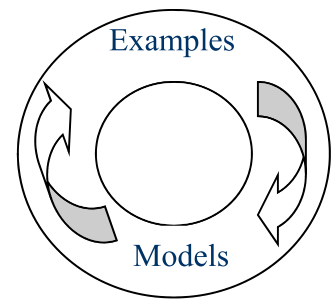
- Key observation: we have additional information to guide selection of data for repeated labeling
  - → the current multiset of labels
  - → the current model built from the data

- Example: {+,-,+,-,-,+} vs. {+,+,+,+,+,+}
  - Will skip details in the talk, see "Repeated Labeling" paper, for targeting using item difficulty, and other techniques

Selective labeling strategy:
Model Uncertainty (MU)

- Learning models of the data additional source of information about label certainty
- **Model uncertainty**: get more labels for instances that cause model uncertainty **in training data (i.e., irregularities!)**

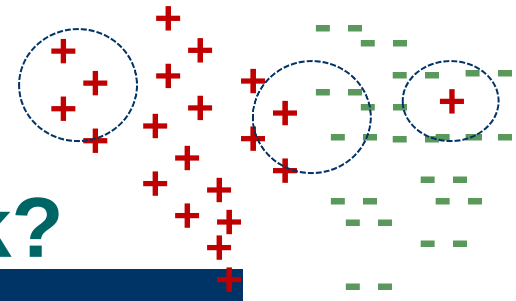




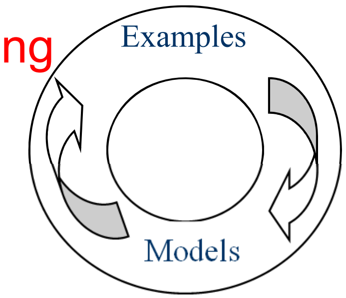
Self-healing process examines irregularities in training data

This is NOT active learning

Why does Model Uncertainty (MU) work?

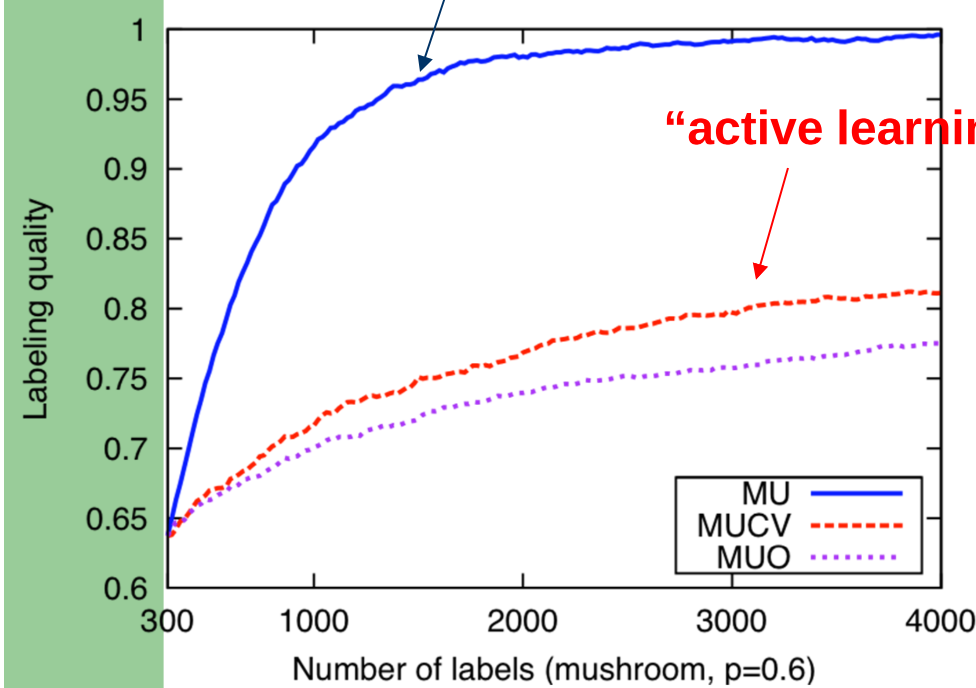


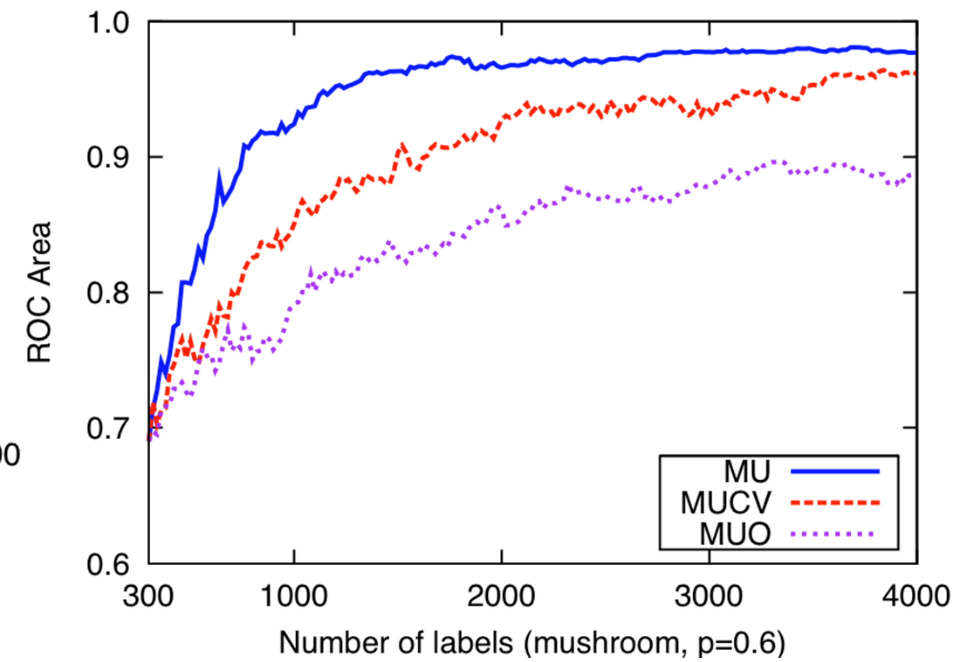
Self-healing process



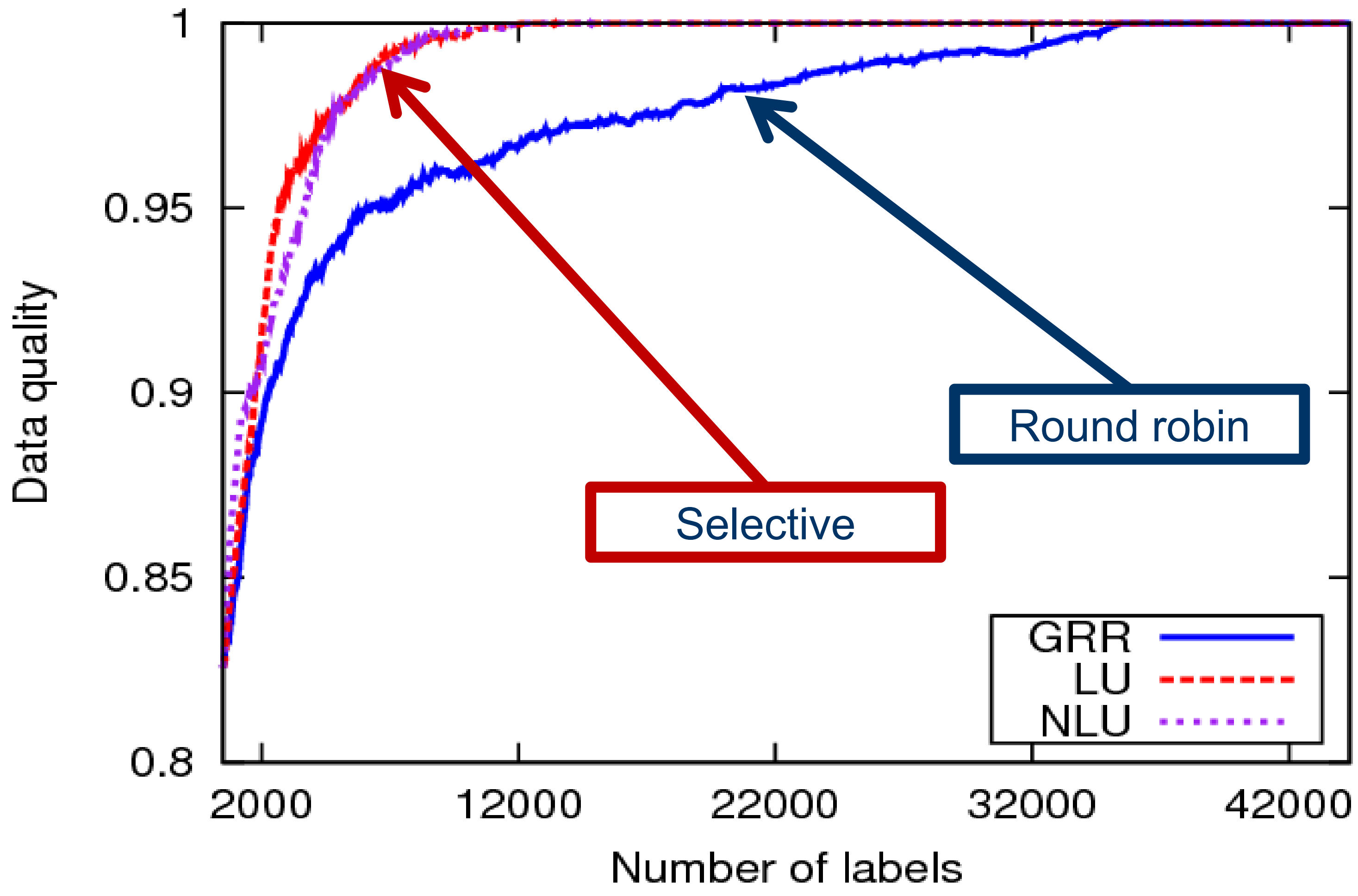
Self-healing MU

“active learning” MU





Adult content classification



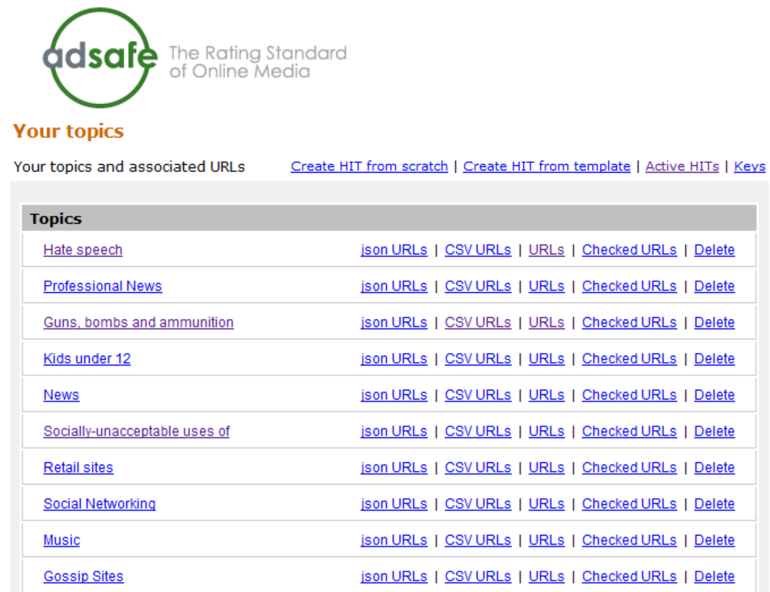
Improving worker participation

- With just labeling, workers are **passively** labeling the data that we give them
- But this can be wasteful when positive cases are sparse
- Why not asking the workers to search themselves and **find training data**

Guided Learning

Ask workers to ***find*** example web pages (great for "sparse" content)

After collecting enough examples, easy to build and test web page classifier



http://url-collector.appspot.com/allTopics.jsp

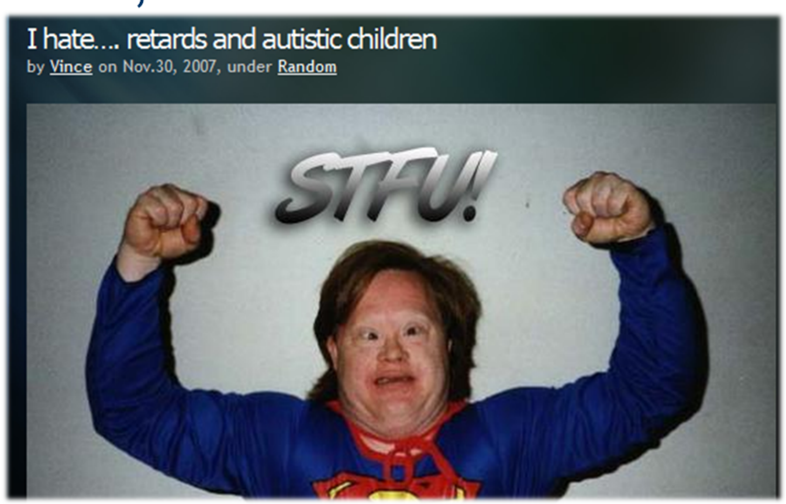
Limits of Guided Learning

- No incentives for workers to find “new” content
- After a while, submitted web pages similar to already submitted ones
- No improvement for classifier

The result? Blissful ignorance…

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

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



No similar training data in training set

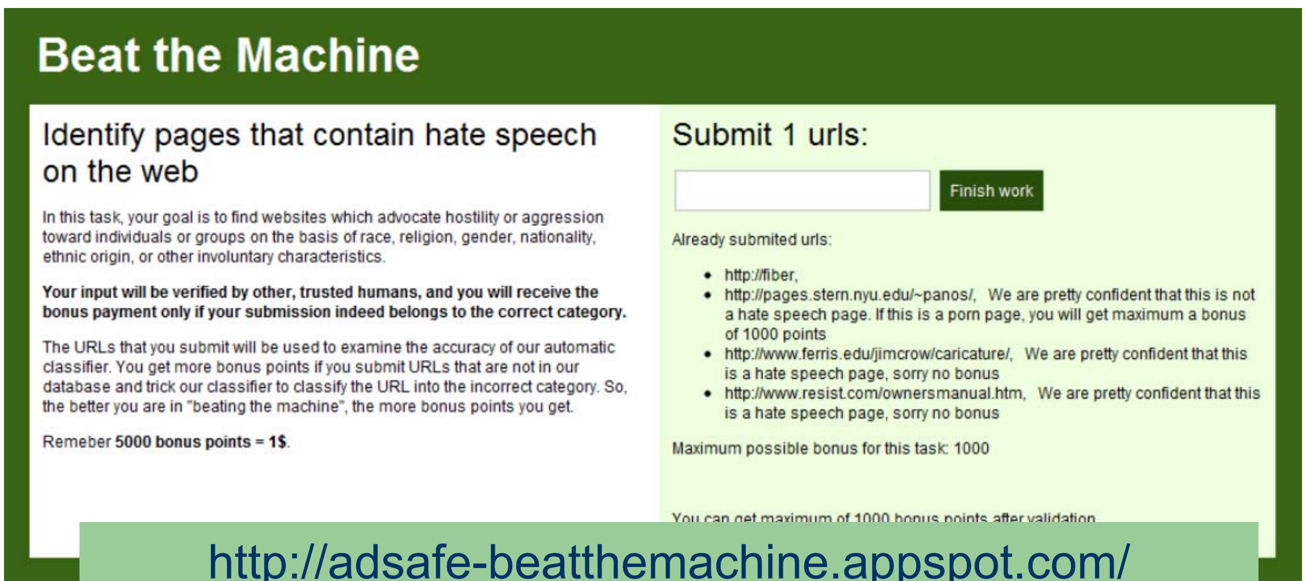
"*Unknown unknowns*" = classifier fails with high confidence

The Child Day Preschools
Children 12mth to 6yrs Active Learning, Low Ratio
www.tcdschools.com
AdChoices

Beat the Machine!

Ask humans to find URLs that

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



http://adsafe-beatthemachine.appspot.com/

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

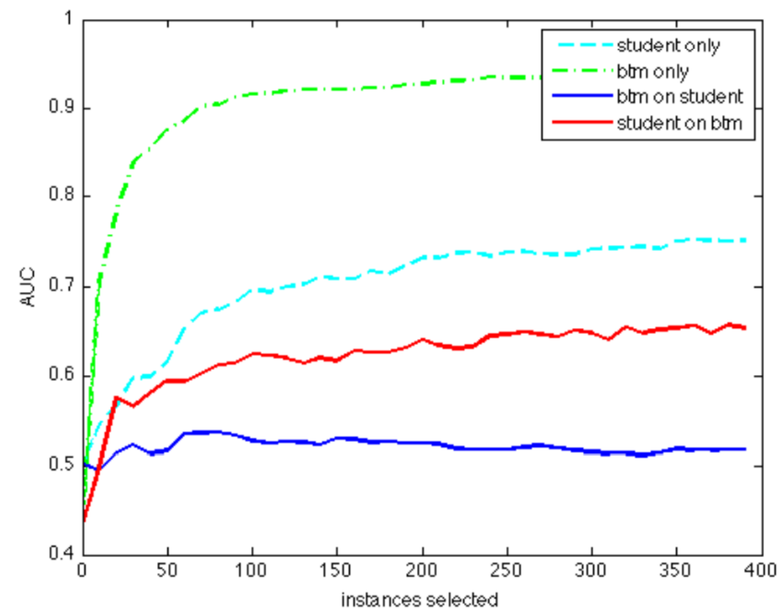
Error rate for probes significantly higher than error rate on (stratified) random data (10x to 100x higher than base error rate)

Structure of Successful Probes

- Now, we identify errors much faster (and proactively)

- Errors not random outliers: **We can "learn" the errors**

- *Could not, however, incorporate errors into existing classifier without degrading performance*



Unknown unknowns → Known unknowns

- Once humans find the holes, they keep probing *(e.g., multilingual porn ☺)*
- However, we **can learn** what we do not know *("unknown unknowns" → "known unknowns")*
- We now know the areas where we are likely to be wrong

Reward Structure for Humans

- High reward higher when:
  - Classifier confident (but wrong) and
  - We ***do not know*** it will be an error
- Medium reward when:
  - Classifier confident (but wrong) and
  - We ***do know*** it will be an error
- Low reward when:
  - Classifier ***already uncertain*** about outcome

Current Directions

- Learn how to best incorporate knowledge to improve classifier
- Measure prevalence of newly identified errors on the web (“query by document”)
  - Increase rewards for errors prevalent in the “generalized” case

Workers reacting to bad rewards/scores

Score-based feedback leads to strange interactions:

The "*angry, has-been-burnt-too-many-times*" worker:

- *"F*** YOU! I am doing everything correctly and you know it! Stop trying to reject me with your stupid 'scores'!"*

The *overachiever* worker:

- *"What am I doing wrong?? My score is 92% and I want to have 100%"*

An unexpected connection at the NAS “Frontiers of Science” conf.

Your bad workers behave like my mice!

An unexpected connection at the NAS “Frontiers of Science” conf.

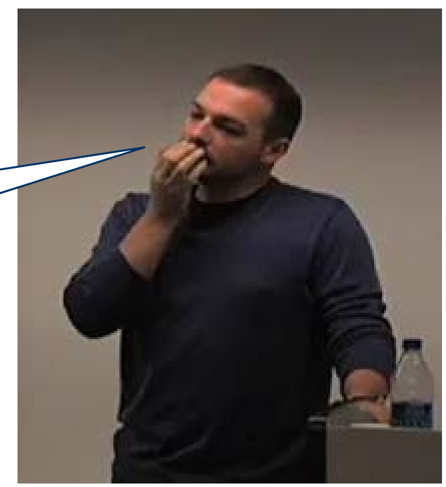
Your bad workers behave like my mice!

Eh?

An unexpected connection at the NAS "Frontiers of Science" conf.

Your bad workers want to engage their brain only for **motor skills,** **not** for **cognitive skills**

Yeah, makes sense…



An unexpected connection at the NAS “Frontiers of Science” conf.

And here is how
I train my mice
to behave…

An unexpected connection at the NAS “Frontiers of Science” conf.

Confuse motor skills!
Reward cognition!

I should try this the moment that I get back to my room

Implicit Feedback using Frustration

- **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"
- **Reward good answers** by rewarding the cognitive part of the brain (e.g, introduce variety/novelty, return results fast)

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

Misery

View Version control

Posted by danielb on *June 22, 2009 at 10:10am*

Misery is a module designed to make life difficult for certain users.

It can be used:

- As an alternative to banning or deleting users from a community.
- As a means by which to punish members of your website.
- To delight in the suffering of others.

Currently you can force users (via permissions/roles, editing their user account, or using Troll IP blacklists) to endure the following misery:

- **Delay:** Create a random-length delay, giving the appearance of a slow connection. (by default this happens 40% of the time)
- **White screen:** Present the user with a white-screen. (by default this happens 10% of the time)
- **Wrong page:** Redirect to a random URL in a predefined list. (by default this happens 0% of the time)
- **Random node:** Redirect to a random node accessible by the user. (by default this happens 10% of the time)
- **403 Access Denied:** Present the user with an "Access Denied" error. (by default this happens 10% of the time)
- **404 Not Found:** Present the user with a "Not Found" error. (by default this happens 10% of the time)

First result

- Spammer workers quickly abandon
- Good workers keep labeling

- Bad: Spammer ***bots*** unaffected
- How to frustrate a bot?
  – Give it a CAPTHCA ☺

Second result (more impressive)

- Remember, scheme was for ***training*** the mice…

- 15% of the spammers start submitting good work!
- *Putting cognitive effort is more beneficial (?)*

- Key trick: Learn to test workers on-the-fly and estimate their quality over streaming data (code and paper coming soon…)

Thanks!

Q & A?