Brand advertising not fully embraced
Internet advertising yet...

Afraid of improper brand placement
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.

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

December 08, 2006 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.
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

<table>
<thead>
<tr>
<th>Requester</th>
<th>HIT Expiration Date</th>
<th>Reward</th>
<th>Time Allotted</th>
<th>HITs Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam GONZALES</td>
<td>Dec 13, 2010 (1 week 2 days)</td>
<td>$0.01</td>
<td>30 minutes</td>
<td>39172</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>Dec 31, 2010 (3 weeks 6 days)</td>
<td>$0.05</td>
<td>15 minutes</td>
<td>14240</td>
</tr>
<tr>
<td>nutella42</td>
<td>Dec 17, 2010 (2 weeks)</td>
<td>$0.08</td>
<td>30 minutes</td>
<td>2446</td>
</tr>
<tr>
<td>jaime_arquello</td>
<td>Dec 10, 2010 (7 days)</td>
<td>$0.03</td>
<td>5 minutes</td>
<td>1952</td>
</tr>
<tr>
<td>Andy K</td>
<td>Dec 9, 2010 (6 days 2 hours)</td>
<td>$0.03</td>
<td>60 minutes</td>
<td>1949</td>
</tr>
</tbody>
</table>
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!

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

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

| PR7MQ44W2XA76FYTYB70 | A2VL24C5P7Y3DJ | http://25u.com | G | http://30plus40plus.com | X |
| PR7MQ44W2XA76FYTYB70 | ADU3MDAG2D0UX | http://25u.com | G | http://30plus40plus.com | X |
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

- $P[X \rightarrow X] = 0.847\%$
- $P[X \rightarrow G] = 99.153\%$
- $P[G \rightarrow X] = 0.053\%$
- $P[G \rightarrow G] = 99.947\%$

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** = $0\% \times 85\% + 100\% \times 15\% = 15\%$
- Error rate of **good worker** = $85\% \times 20\% + 85\% \times 20\% = 20\%$

**False negatives:** Spam workers pass as legitimate
Challenge 2: Humans are biased!

Error rates for CEO of AdSafe

- $P[G \rightarrow G] = 20.0\%$
- $P[G \rightarrow P] = 80.0\%$
- $P[G \rightarrow R] = 0.0\%$
- $P[G \rightarrow X] = 0.0\%$
- $P[P \rightarrow G] = 0.0\%$
- $P[P \rightarrow P] = 0.0\%$
- $P[P \rightarrow R] = 100.0\%$
- $P[P \rightarrow X] = 0.0\%$
- $P[R \rightarrow G] = 0.0\%$
- $P[R \rightarrow P] = 0.0\%$
- $P[R \rightarrow R] = 100.0\%$
- $P[R \rightarrow X] = 0.0\%$
- $P[X \rightarrow G] = 0.0\%$
- $P[X \rightarrow P] = 0.0\%$
- $P[X \rightarrow R] = 0.0\%$
- $P[X \rightarrow X] = 100.0\%$

- We have 85% G sites, 5% P sites, 5% R sites, 5% X sites
- Error rate of spammer (all G) = $0\% \times 85\% + 100\% \times 15\% = 15\%$
- Error rate of biased worker = $80\% \times 85\% + 100\% \times 5\% = 73\%$

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

- \( P[G \rightarrow G] = 20.0\% \)
- \( P[P \rightarrow G] = 0.0\% \)
- \( P[R \rightarrow G] = 0.0\% \)
- \( P[X \rightarrow G] = 0.0\% \)

- \( P[G \rightarrow P] = 80.0\% \)
- \( P[P \rightarrow P] = 0.0\% \)
- \( P[R \rightarrow P] = 0.0\% \)
- \( P[X \rightarrow P] = 0.0\% \)

- \( P[G \rightarrow R] = 0.0\% \)
- \( P[G \rightarrow X] = 0.0\% \)
- \( P[P \rightarrow R] = 100.0\% \)
- \( P[P \rightarrow X] = 0.0\% \)
- \( P[R \rightarrow R] = 100.0\% \)
- \( P[R \rightarrow X] = 0.0\% \)
- \( P[X \rightarrow R] = 0.0\% \)
- \( P[X \rightarrow X] = 100.0\% \)

- 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!


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

<table>
<thead>
<tr>
<th>Assigned Label</th>
<th>Corresponding “Soft” Label</th>
<th>Expected Label Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spammer: G</td>
<td>&lt;G: 25%, P: 25%, R: 25%, X: 25%&gt;</td>
<td>0.75</td>
</tr>
<tr>
<td>Good worker: P</td>
<td>&lt;G: 100%, P: 0%, R: 0%, X: 0%&gt;</td>
<td>0.0</td>
</tr>
</tbody>
</table>

[***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.

\[
\text{Quality (Worker)} = 1 - \frac{\text{Cost (Worker)}}{\text{Cost (Spammer)}}
\]

• Scalar score, useful for the purpose of ranking workers
• Threshold-ing rewards gives wrong incentives:
  • Decent (but still useful) workers get fired
  • Uncertainty near the decision threshold
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
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
Noisy labels lead to degraded task performance
Labeling quality increases → classification quality increases

Quality and Classification Performance

AUC

Number of examples ("Mushroom" data set)

Quality = 100%
Quality = 80%
Quality = 60%
Quality = 50%

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

![Graph showing accuracy vs. number of examples (Mushroom)]

- Quality = 100%
- Quality = 80%
- Quality = 60%
- Quality = 50%

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
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
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

http://adsafe-beatthemachine.appspot.com/
Error rate for probes significantly higher than error rate on (stratified) random data (10x to 100x higher than base error rate)

<table>
<thead>
<tr>
<th>#</th>
<th>Category</th>
<th>Tasks Running</th>
<th>URL's gathered</th>
<th>Correct URL's gathered</th>
<th>Total Bonus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identify pages that contain hate speech on the web (haf)</td>
<td>206</td>
<td>1023</td>
<td>161</td>
<td>75516</td>
</tr>
<tr>
<td>2</td>
<td>Identify pages related to illegal drug use on the web (drg)</td>
<td>100</td>
<td>500</td>
<td>26</td>
<td>9114</td>
</tr>
<tr>
<td>3</td>
<td>Identify pages that contain reference to alcohol (alc)</td>
<td>100</td>
<td>475</td>
<td>144</td>
<td>55149</td>
</tr>
<tr>
<td>4</td>
<td>Identify adult-related pages (adf)</td>
<td>174</td>
<td>859</td>
<td>132</td>
<td>63523</td>
</tr>
</tbody>
</table>
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 $\rightarrow$ 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*” $\rightarrow$ “*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
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%”
Your workers behave like my mice!

An unexpected connection…
Your workers behave like my mice!

Eh?
Your workers want to use only their motor skills, not their cognitive skills.
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
An unexpected connection at the NAS “Frontiers of Science” conf.

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

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

And here is how I train my mice to behave…
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!

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

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 delights 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)
Rewarding (?) Cognitive Effort

- **Reward good answers** by rewarding the cognitive part of the brain
  - Introduce variety
  - Introduce novelty
  - Give new tasks fast
  - Show score improvements faster (but not the opposite)
  - Show optimistic score estimates
Experiments

- Web page classification
- Image tagging
- Email & URL collection
Experimental Summary (I)

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

- Good workers unaffected
  - No significant effect on participation of workers with good performance
  - Lifetime of participants unaffected
  - Reduction in response time (after removing the “intervention delays”; that was puzzling)
Remember, scheme was for training the mice…

15%-20% of the spammers start submitting good work!
Two key questions

- Why response time was slower for some good workers?
- Why some low quality workers start working well?
System 1: “Automatic” actions

System 2: “Intelligent” actions
System 1 Tasks

- Detect that one object is more distant than another.
- Orient to the source of a sudden sound.
- Complete the phrase “bread and...”
- Make a “disgust face” when shown a horrible picture.
- Detect hostility in a voice.
- Answer to $2 + 2 = ?$
- Read words on large billboards.
- Drive a car on an empty road.
- Find a strong move in chess (if you are a chess master).
- Understand simple sentences.
System 2 Tasks

- Focus attention on the clowns in the circus.
- Look for a woman with white hair.
- Count the occurrences of the letter \( a \) in a page of text.
- Compare two washing machines for overall value.
- Check the validity of a complex logical argument.
Performing Well?

Status: Usage of System 1 ("Automatic")

Performing Well?
Check if System 1 can Handle, **remove** System 2 stimuli

Not Performing Well?
Disrupt and Engage System 2

Not Performing Well?
Hell/Slow ban

Out
Thanks!

Q & A?