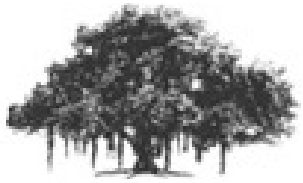


IIIT-H

Web Mining
Lecture 26: Crowdsourcing

Manish Gupta

16th Nov 2013

Slides borrowed (and modified) from
http://www.cs.ucsb.edu/~gangw/MAE-gang.pptx
http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html
WWW 2011: Managing Crowdsourced Human Computation (Panos Ipeirotis)

Recap of Lecture 25: User Understanding by Log Mining

- Personalized search
- User behavior modeling
- Privacy in Web Search Query Logs

Announcements

- Submit your project reports and presentation slides by Nov 19 (midnight)
  - I will bind all presentation slides into a single slideshow

Today's Agenda

- Introduction to Crowdsourcing
- Applications of Crowdsourcing
- Challenges in Crowdsourcing
- Managing Quality for Simple Tasks

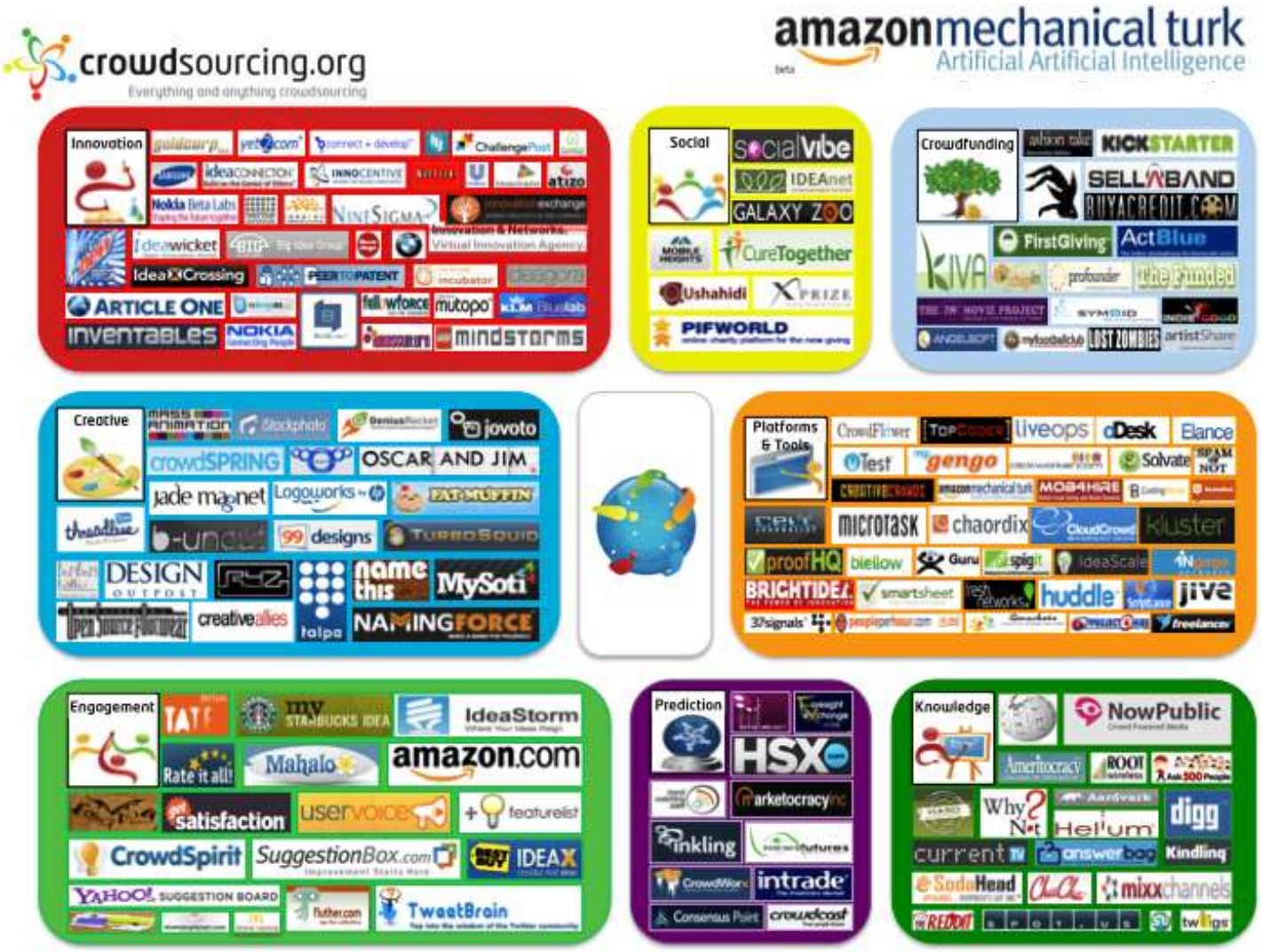
Today's Agenda

- **Introduction to Crowdsourcing**
- Applications of Crowdsourcing
- Challenges in Crowdsourcing
- Managing Quality for Simple Tasks

What is Crowdsourcing?

- June 2006: Jeff Howe created the term for his article in the Wired magazine "The Rise of Crowdsourcing".
- A company uses people outside of their organization to help them solve problems
  - Paid or unpaid
  - Amateur or expert
  - Variety of fields and tasks
- Crowdsourcing is an online, distributed problem-solving and production model

Crowdsourcing Industry Landscape



Tasks/Hitapps

Translate words from Tibetan to English — View a HIT in this group

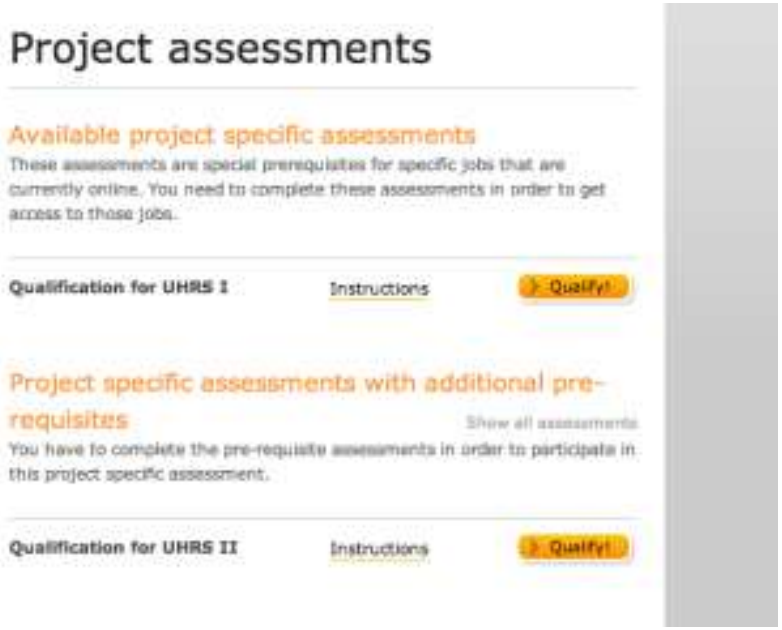
Requester: Chris Callison-Burch **HIT Expiration Date:** Apr 22, 2012 (2 weeks 1 day) **Reward:** $0.15

Time Allotted: 60 minutes **HITs Available:** 1247

Description: Translate 10 words from Tibetan to English

Keywords: translation, vocabulary, dictionary, Tibetan, English, language, research, JHU

Qualifications Required:
HIT approval rate (%) is greater than 85

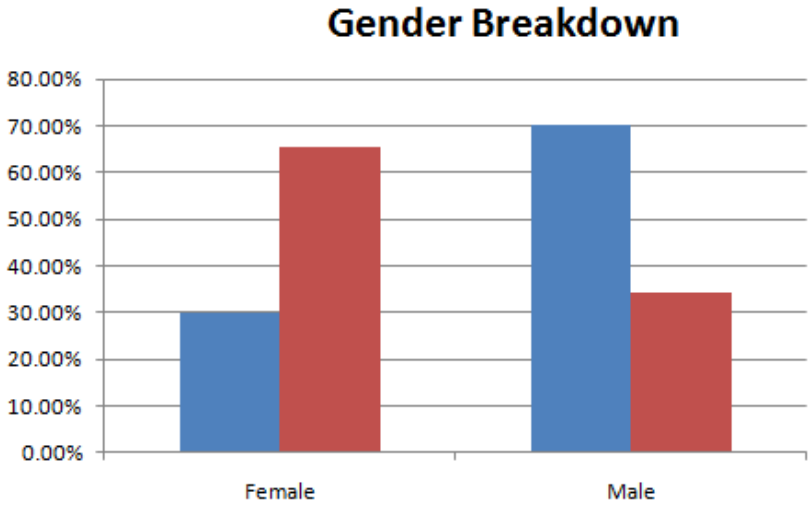


- Mass work, Distributed work, or just tedious work
- Creative work
- Look for specific talent
- Testing
- Support
- To offload peak demands
- Tackle problems that need specific communities or human variety
- Any work that can be done cheaper this way

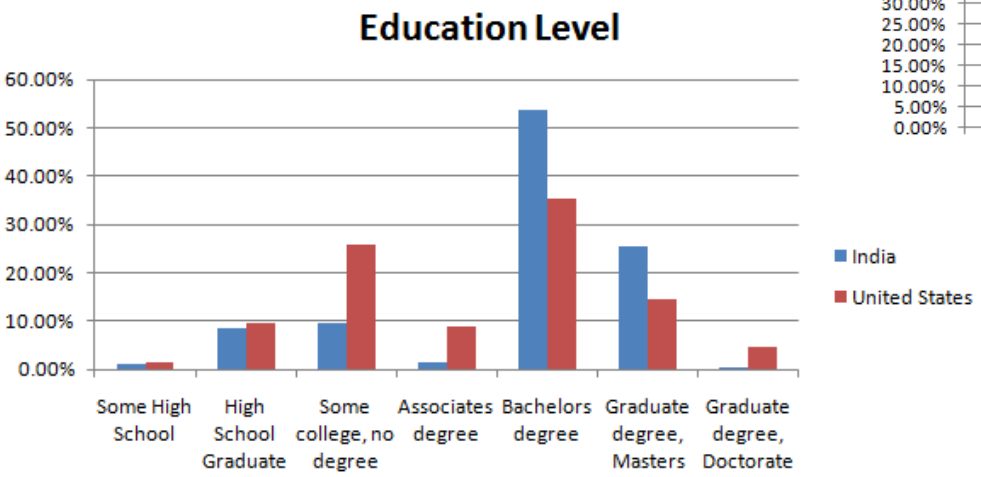
Crowd Motivation

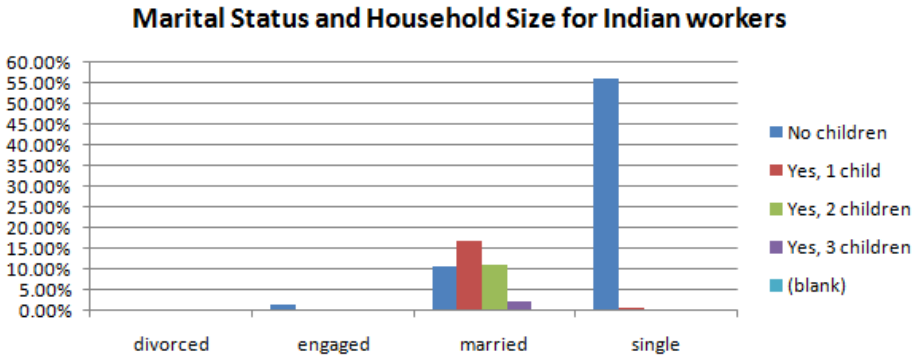
- Money
- Self-serving purpose (learning new skills, get recognition, avoid boredom, enjoyment, create a network with other profesionals)
- Socializing, feeling of belonging to a community, friendship
- Altruism (public good, help others)

Demographics of Mechanical Turk Workers (1)

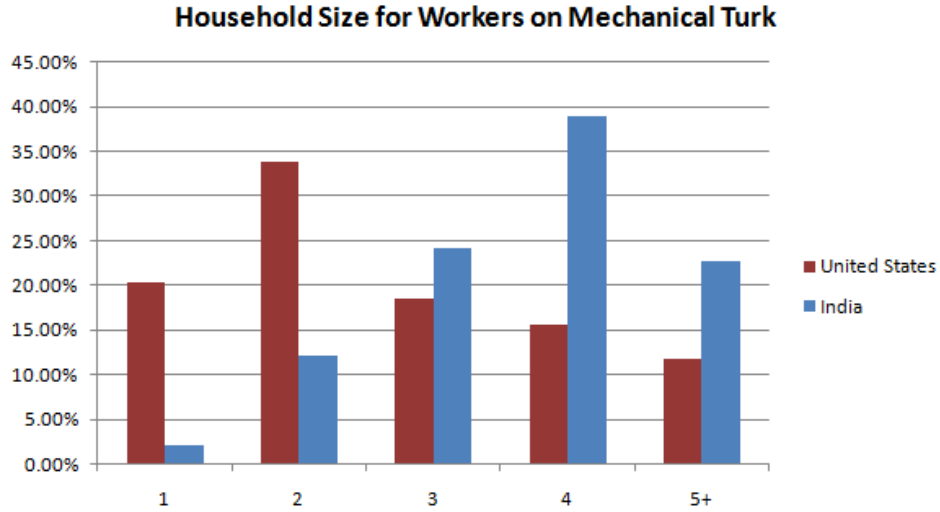


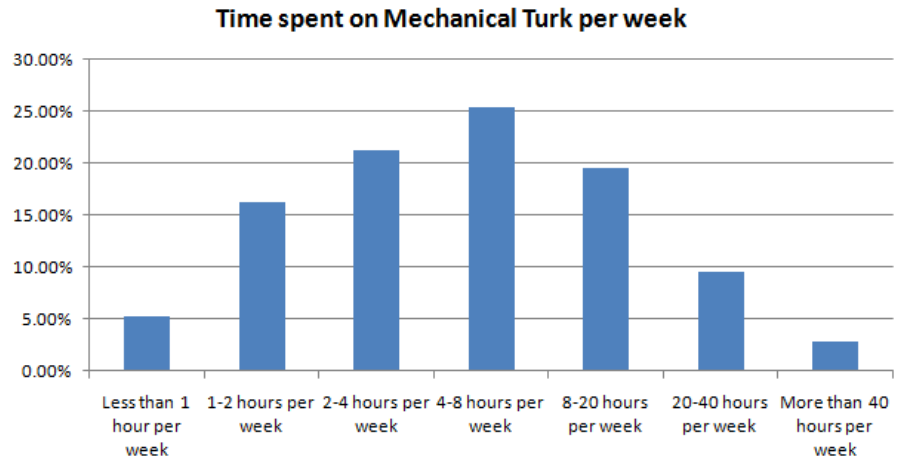
- United States: 46.80%
- India: 34.00%
- Miscellaneous: 19.20%

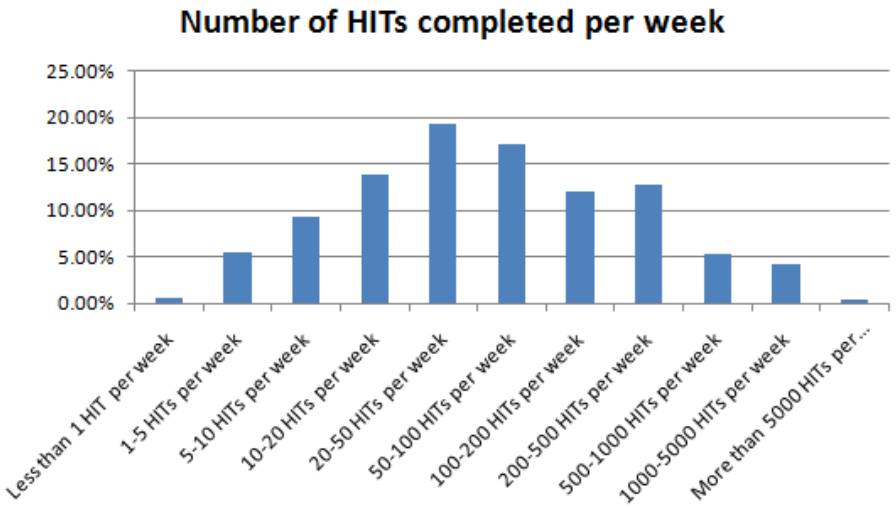


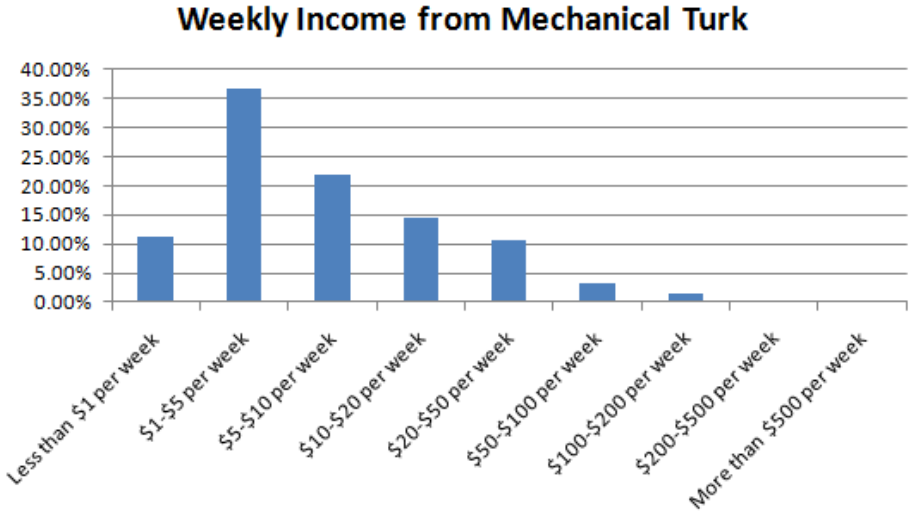


Demographics of Mechanical Turk Workers (2)









Pros of Crowdsourcing

- Quicker: Parallelism reduces time
- Cheap, even free
- Creativity, Innovation
- Quality (depends)
- Availability of scarce resources: Taps on the 'long tail'
- Multiple feedback
- Allows to create a community (followers)
- Business Agility
- Scales up!

Today's Agenda

- Introduction to Crowdsourcing
- **Applications of Crowdsourcing**
- Challenges in Crowdsourcing
- Managing Quality for Simple Tasks

Applications of Crowdsourcing

- R&D, Innovation
  - Inno-Centive, Netflix
- Market Research
  - Affinova
- Forecasting
  - NewsFutures/Lumenogic
- Customer Service
  - User feedback
- Knowledge Management
  - Wikis
- Systems Testing
  - NIST cryptography, Peer-to-Patent
- Crisis Response
  - Cajun Navy
  - Katrina People Finder Project
- Micropayment business model
  - iStockphoto

Case Study: Threadless.com

- It crowdsources the design process by ongoing online compition
- Started by Jake Nickell (20) and Jacob DeHart (19) in 2000
- As of June 2006, Threadless was ‘selling 60,000 T-shirts a month
- Had a profit margin of 35 per cent and was on track to gross $18 million in 2006’, all with ‘fewer than 20 employees’

How Threadless.com Works

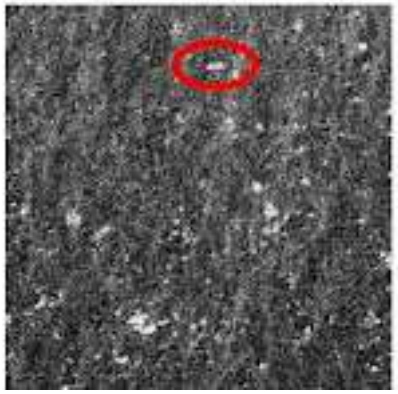
- Free Membership with valid email address
- Member can vote and submit design online
- Submissions are rated from 1 to 5 scale with ‘I’d buy it!’ box
- New designs are available for 2 weeks from the day of submission and the highest scoring designs are selected by Threadless staff to be printed and made available for sale
- Winners receive awards worth of 2,000 USD

Searching for Jim Gray (2007)

- Jim Gray, Turing Award winner
- Missing with his sailboat outside San Francisco Bay, Jan 2007
- No result from searches of coastguard and private planes

Use satellite image to search for Jim Gray's sailboat
Problem: the search cannot be automated by computer
Solution

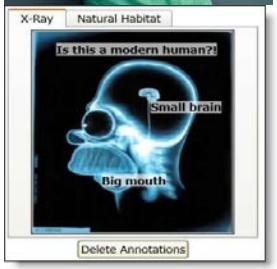
Split the satellite image into many small images
Volunteers look for his boat in each image

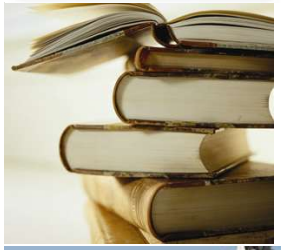


100,000 tasks completed in 2 days

Crowdsourcing Applications

- Natural language processing (NLP)
  - Data labeling [Snow2008] [Callison-Burch2009]
  - Searching results validation [Alonso2008]
  - Database query: CrowdDB [Franklin2011], Qurk [Marcus2011]
- Image processing
  - Image annotation [Ahn2004] [Chen2009]
  - Image search [Yan2010]
- Content generation/knowledge sharing
  - Wikipedia, Quora, Yahoo! Answers, StackOverflow
  - Real-time Q&A: Vizwiz [Bigham2010], Mimir [Hsieh2009]
- Human sensor
  - Google Map traffic monitoring
  - Twitter earthquake report [Sakaki2010]





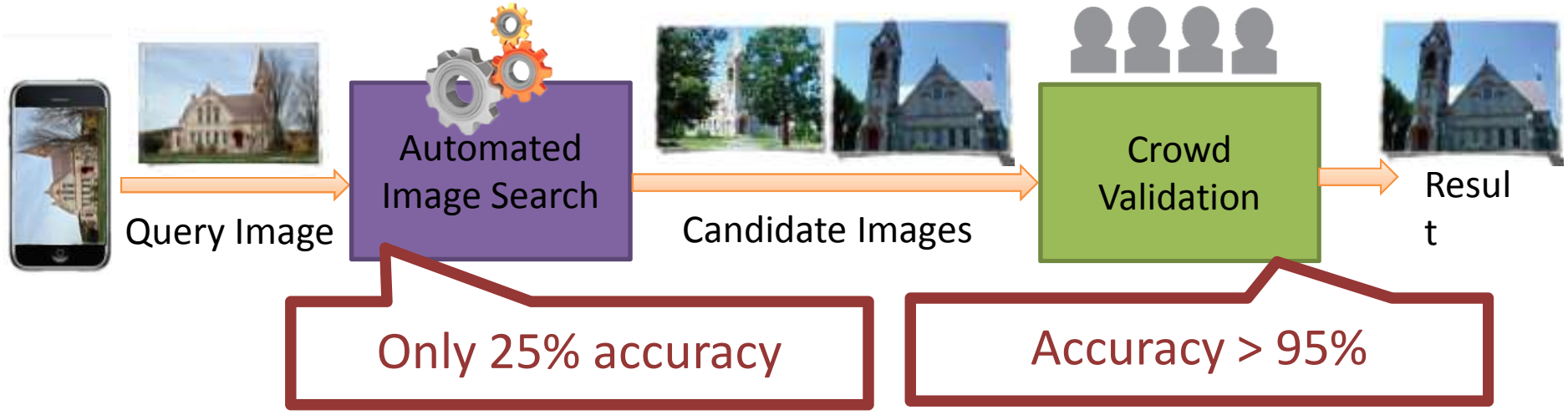
Data Annotation

- Natural Language Processing (NLP) problems
  - Evaluating machine translation quality [Callison-Burch2009]
  - Labeling text content (e.g. emotions) [Snow2008]
- Challenges
  - Difficult for software automation
  - Experts are expensive and slow
- Benefits of using crowdsourcing
  - Non-experts are cheap and fast
  - Non-expert results (processed) are as good as experts

Image Search

CrowdSearch [Yan2010]

- Accurate image searching for mobile devices by combining
  - Automated image searching
  - Human validation of searching results via crowdsourcing



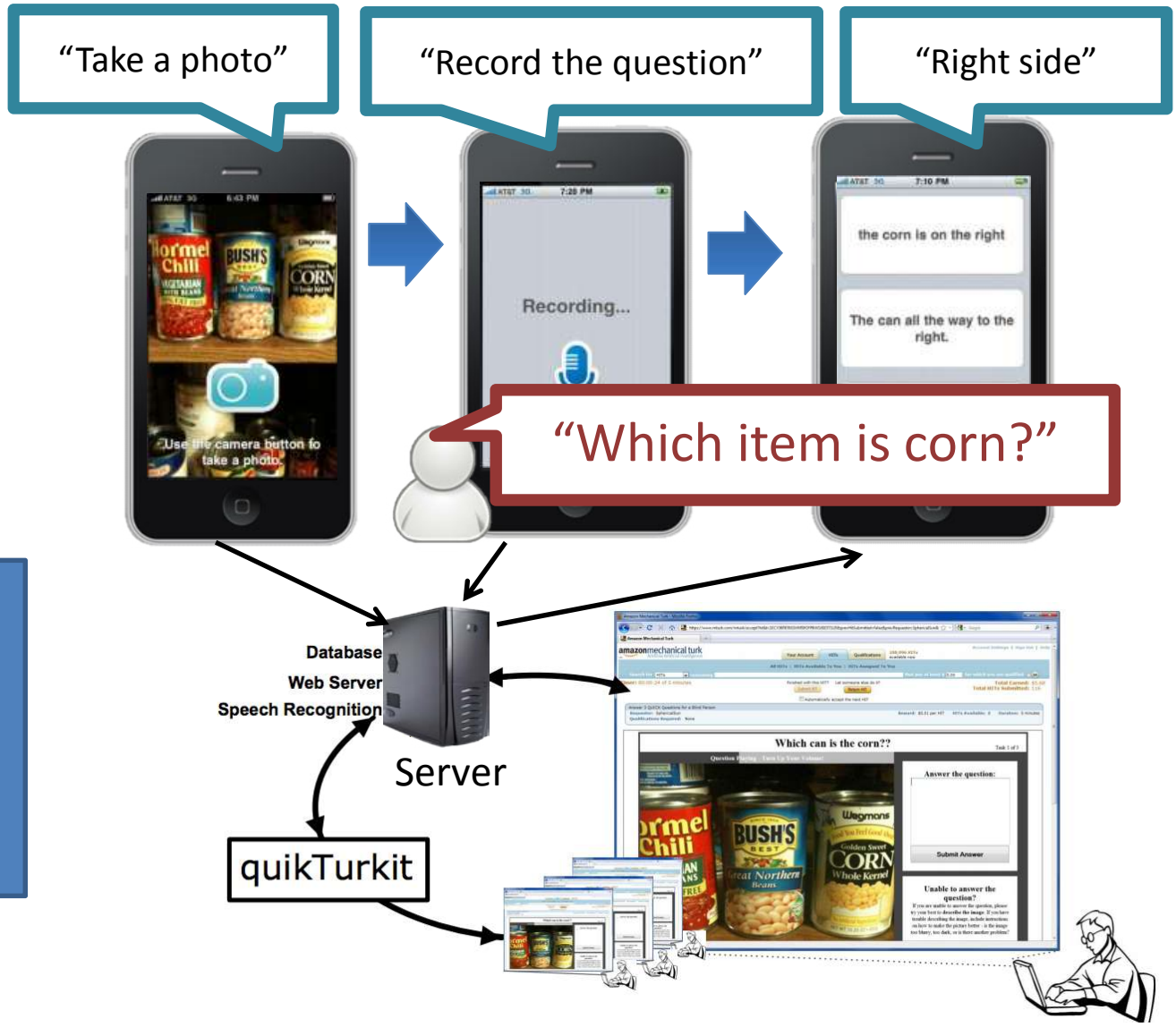
Use human intelligence to improve automation

Question and Answer

VizWiz
[Bigham2010]

- Help blind people
- Answer questions
- Near real-time

- Example
  - Shopping scenario



- Use the crowd to help people in need
- Replace an "expensive" personal assistant

Micro-Crowdsourcing Example: Labeling Images

- Two player online game: ESP
  - Partners don't know each other and can't communicate
  - Object of the game: type the same word
  - The only thing in common is an image

Today's Agenda

- Introduction to Crowdsourcing
- Applications of Crowdsourcing
- **Challenges in Crowdsourcing**
- Managing Quality for Simple Tasks

Cons of Crowdsourcing

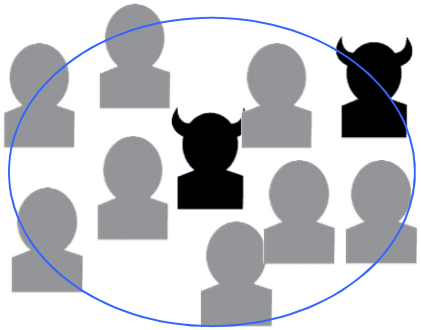
- Lack of professionalism: Unverified quality
- Too many answers
- No standards
- No organisation of answers
- Not always cheap: Added costs to bring a project to conclusion
- Too few participants if task or pay is not attractive
- If worker is not motivated, lower quality of work
- Global language barriers.
- Different laws in each country: adds complexity
- No written contracts, so no possibility of non-disclosure agreements.
- Hard to maintain a long term working relationship with workers.
- Difficulty managing a large-scale, crowdsourced project.
- Can be targeted by malicious work efforts.
- Lack of guaranteed investment, thus hard to convince stakeholders.

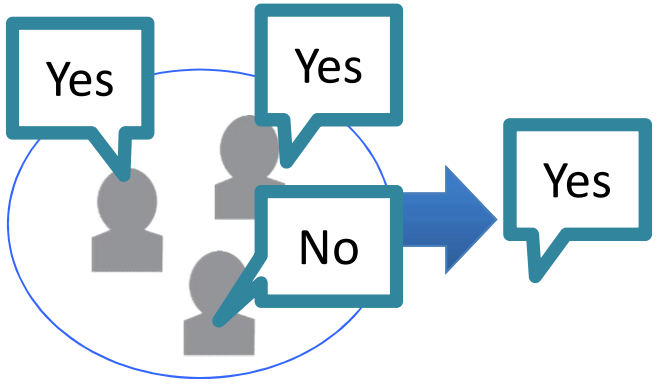
Challenges in Crowdsourcing

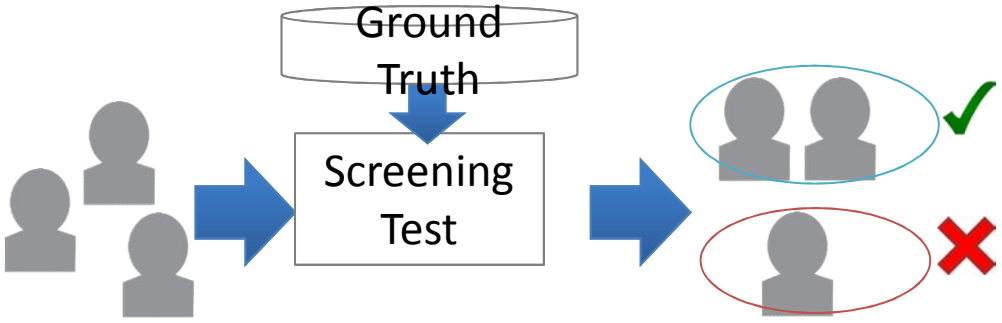
- Quality control
  - High diversity in worker background and expertise
- Incentives
  - Encourage participation
  - Improve work quality
- Task management
  - Perform complex/real-time tasks
  - Coordinate workers and requesters
- Security
  - Spammy/cheating workers, fraud requesters
  - Using crowdsourcing systems for malicious attacks
- Intellectual property

Quality Control

- Fundamental problem: the crowd is not reliable
  - [Oleson2011] [Snow2008] [Callison-Burch2009] [Yan2010] [Franklin2011]
  - Workers make mistakes
  - Workers spam the system
- Existing strategies
  - Majority voting
  - Pre-screening to test workers
  - Statistic models to clear data bias





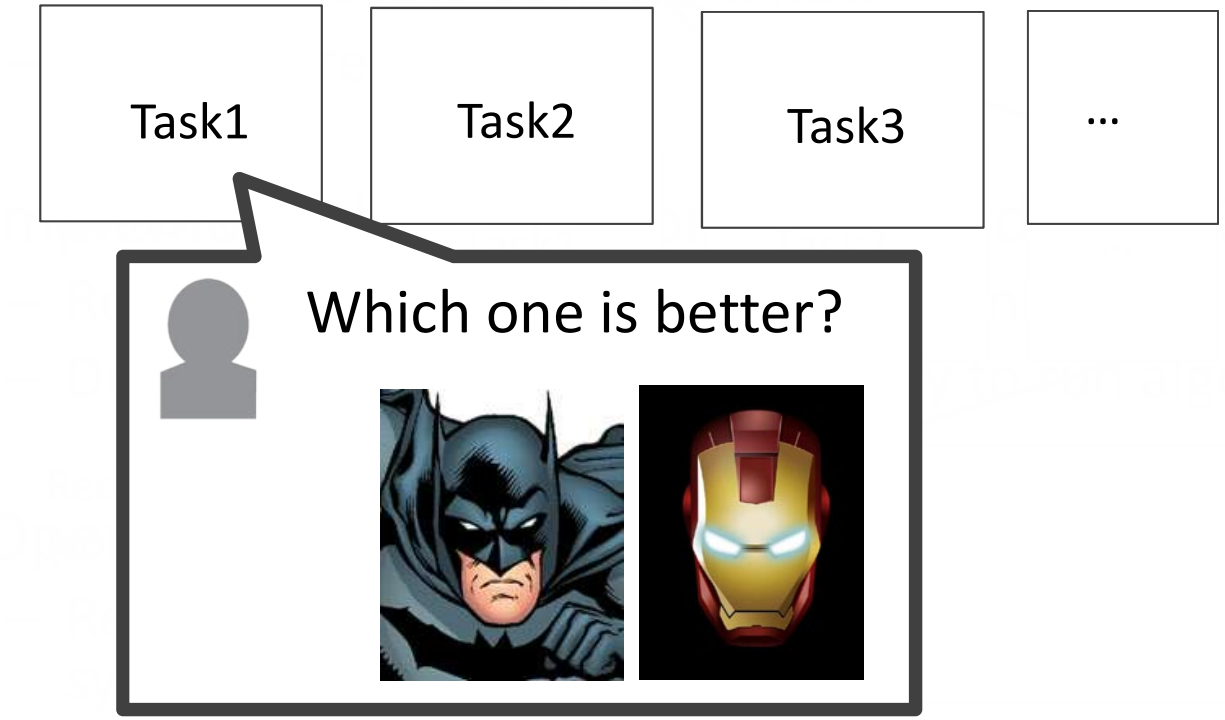


Incentives

- Basic questions: how to set the right price of the tasks?
  - Can you improve work quality by raising payment?
  - Can you attract more workers by raising payment?
- Empirical study on worker incentives [Mason2010] [Hsieh2010]
  - High payment helps to recruit workers faster and increase participation
  - Money does not improve quality
  - Punishment/bonus based quality control
    - Pay the minimum $0.01 for all workers and $0.01 for bonus
- Common problem for all applications

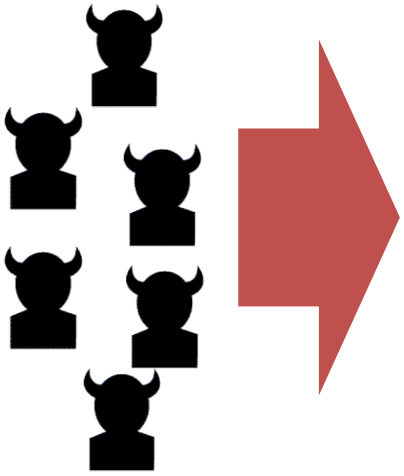
Example: Writing a travel book for New York City

Example: use *bubble sort* algorithm to sort pictures



Security Challenges

- Attacks inside crowdsourcing systems
  - Spammy workers give random/bad answers
  - Dishonest requesters
- Using crowdsourcing system to carry out malicious campaigns
  - Real-user can perform all kinds of malicious tasks
  - Crowdsourcing makes it possible to scale



- Write fake reviews
- Create fake accounts (Sybils)
- Generate social network spam
- Solve CAPTCHA
- Give biased voting
- Build back links (SEO)

The Sybil attack is an attack wherein a reputation system is subverted by forging identities in peer-to-peer networks

freelancer

Post Project | Find Freelancers | Browse Projects | Make Money

Bids	Avg Bid
8	$56 USD

Project ID: 1662641 **Project Type:** Fixed

Budget: $30-$250 USD

Project Description:

Need to have 5-Starts Ratings on a review website. (15 Reviews per Job)
1. Review written in English a
provided. We may ask you to
not relevant.

2. Reviews should be spread out
eral weeks (no more than 1-2 reviews per week)

3. Reviews must be from different UK,
Zealand IP addresses

4. Reviews should get passed by review website filter to count for successful 1.

5 star rating and positive review

Use different IP addresses

Bypass existing spam filter

Measuring Malicious Crowdsourcing

- Crowdsourcing malicious tasks
  - Generate Spam, solve CAPTCHA, create fake accounts, Greyhat SEO
- Scale and economics
  - Zhubajie (China): malicious jobs 10K/month, with $1M/month [Wang2012b]
  - FreeLancer (US): malicious jobs 140K/7 years [Motoyama2011]
- Emerging threat
  - International work force
  - Growing exponentially

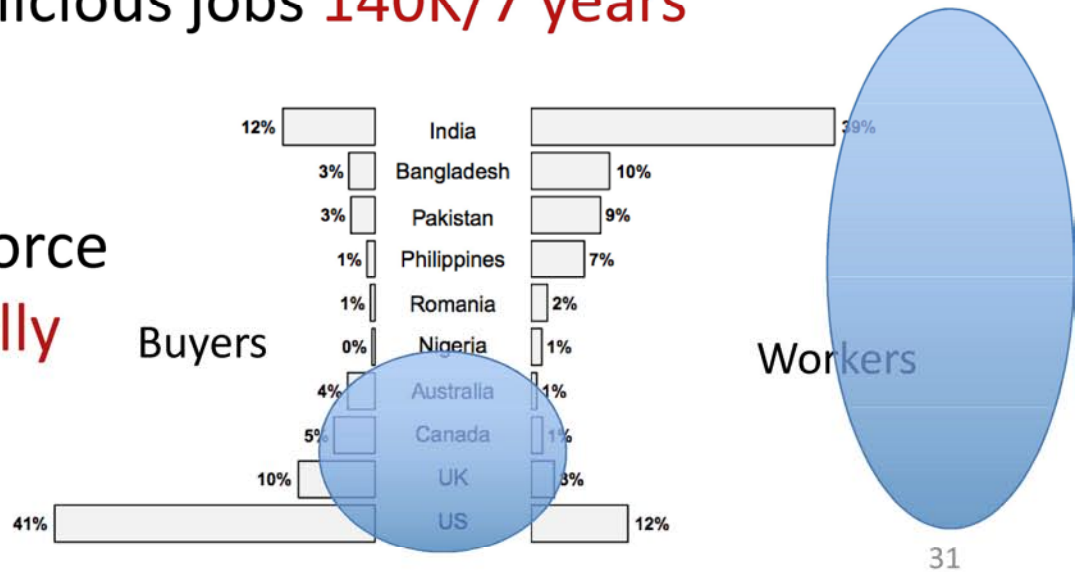


Figure from [Motoyama2011]

Online Reviews: Why Important?

- **80%** of people will check online reviews before purchasing products/travel online.[1]
- Independent restaurants: a one-star increase in Yelp rating leads to a **9%** increase in revenue.[2]

[1] http://www.coneinc.com/negative-reviews-online-reverse-purchase-decisions
[2] Michael Luca. Reviews, reputation, and revenue: The Case of Yelp.com. Harvard Business School Working Paper, 2011

Detecting Fake Reviews

- Detecting review spam [Jindal 2008]
  - Duplicated/Near-duplicated reviews

Dataset	Reviewer	Products	Reviews	Spam Reviews
Amazon	2,146,048	1,195,133	5,838,032	55,319

- Detecting review spammers
  - Classify rating/review behaviors [Lim 2010]
  - Detect synchronized reviews in groups [Mukherjee2012]
- Deception models [Ott2011]
  - Content classification using trained data
  - Psycholinguistic deception detection

Human accuracy 60%
Classifier accuracy 90%

Challenges to Detect Fake Reviews

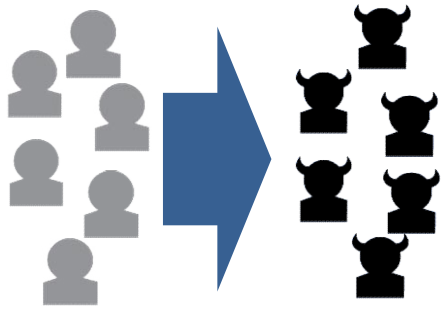
- Current review spam detection solution is limited
  - Assume a few attackers control many accounts
  - Crowdsourcing can break these assumptions
- Deception model (NLP approach) has limitations
  - Domain specific
  - Content analysis still has high false positive (10%)
- Detecting fake reviews is an open problem
  - Low false positive
  - Real-time

Social Network Sybils

- Sybils in Online Social Networks (OSNs)
  - Cheating in social games
  - Spreading spam/malware
  - [Thomas2011], [Gao2010], [Nazir2010]
- Challenges to detect Sybils in the wild
  - Various/adaptive Sybil behavior patterns/attack strategies
  - Increasingly sophisticated/realistic Sybil account profiles
  - Automated mechanisms losing effectiveness
- Use crowdsourcing for Sybil detection

Crowdsourced Sybil Detection

- Basic idea: build a crowdsourced Sybil detector
  - Resilient to changing attacker strategies
- Question: *Can human identify Sybil profiles?* (answer: user study)
  - Ground truth datasets of full user profiles
    - 200 real + 180 fake accounts (Renren, Facebook, Facebook-India)
  - Segmented user groups
    - Renren users (Chinese), Facebook (US), Facebook (Indian)
    - Experts (conscientious, motivated), Turkers (paid per profile, $-driven)
- High level results
  - Experts are accurate; both experts and turkers have near-zero false positives
  - Quality control can improve turker accuracy ~ experts
  - Accurate, scalable, cost-effective



Today's Agenda

- Introduction to Crowdsourcing
- Applications of Crowdsourcing
- Challenges in Crowdsourcing
- **Managing Quality for Simple Tasks**

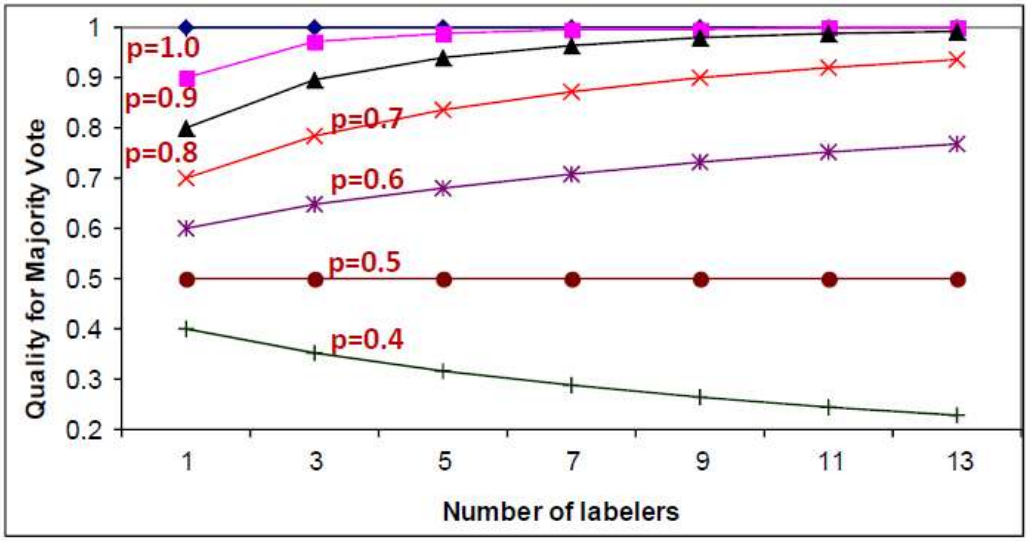
Managing Quality for Simple Tasks

- Quality through redundancy: Combining votes
  - Majority vote
  - Quality-adjusted vote
  - Managing dependencies
- Quality through gold data
- Estimating worker quality (Redundancy + Gold)
- Joint estimation of worker quality and difficulty
- Active data collection

Majority Voting and Label Quality

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

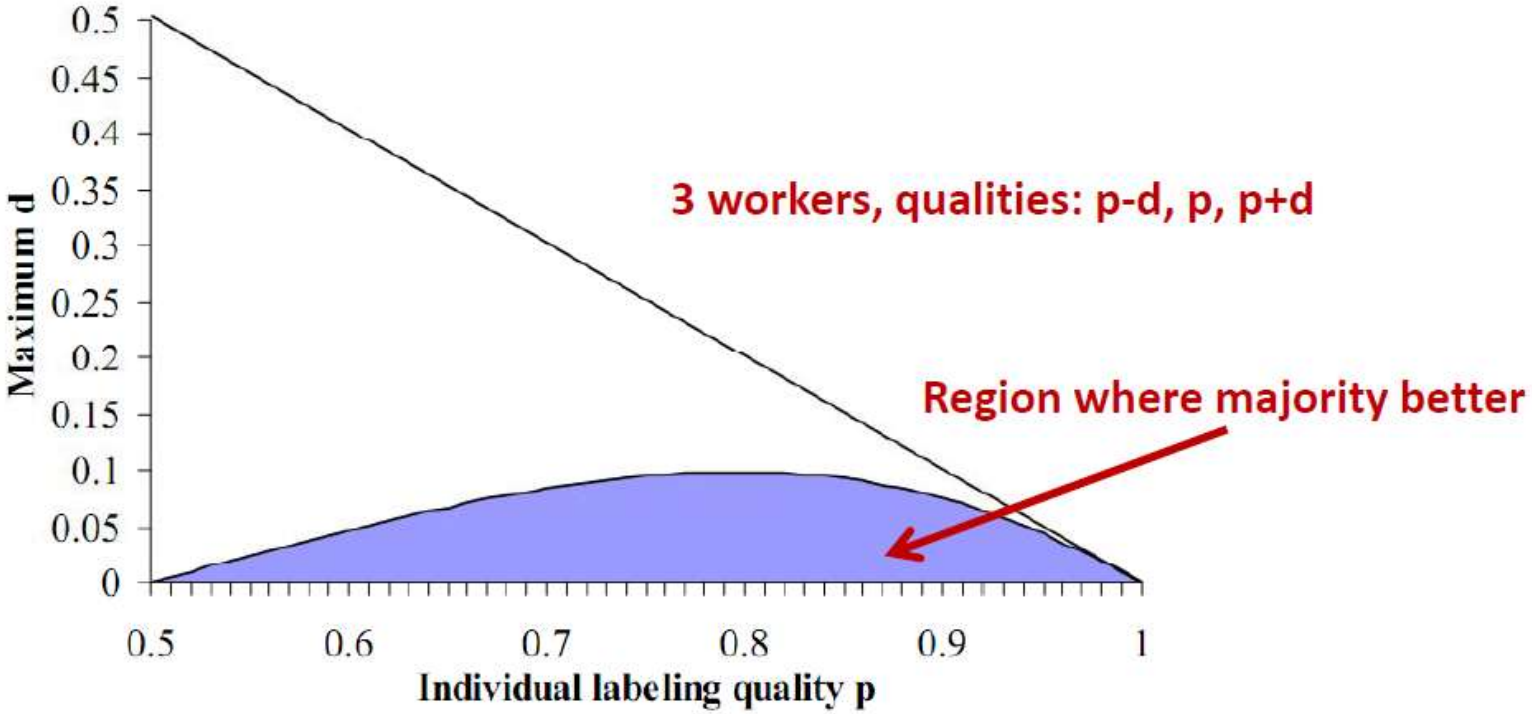
p is probability of individual labeler being correct



Binary classification

$$P_{maj} = \sum_{m=0}^{\lfloor L/2 \rfloor} \binom{L}{m} p^{L-m} (1-p)^{m}$$

What if Qualities of Workers are Different?



- Majority vote works best when workers have similar quality
- Otherwise better to just pick j p the vote of the best worker
- …or model worker qualities and combine

Combining Votes with Different Quality

- Assume prior p_0 for "yes" and each worker w_i voting "yes" with confidence $p_i \in [0,1]$
  - If they say "yes", $p_i = \Pr(yes|w_i\ said\ yes)$
  - If they say "no", $p_i = \Pr(yes|w_i\ said\ no)$
  - Assuming (conditional) independence
  - Overall probability p^* of "yes"
    - $\log\left(\frac{p^*}{1-p^*}\right) = \sum_{i=0}^{N} \log\left(\frac{p_i}{1-p_i}\right)$
  - Same principle for multiple classes
  - [Clemen and Winkler, 1990]

What Happens if we have Dependencies?

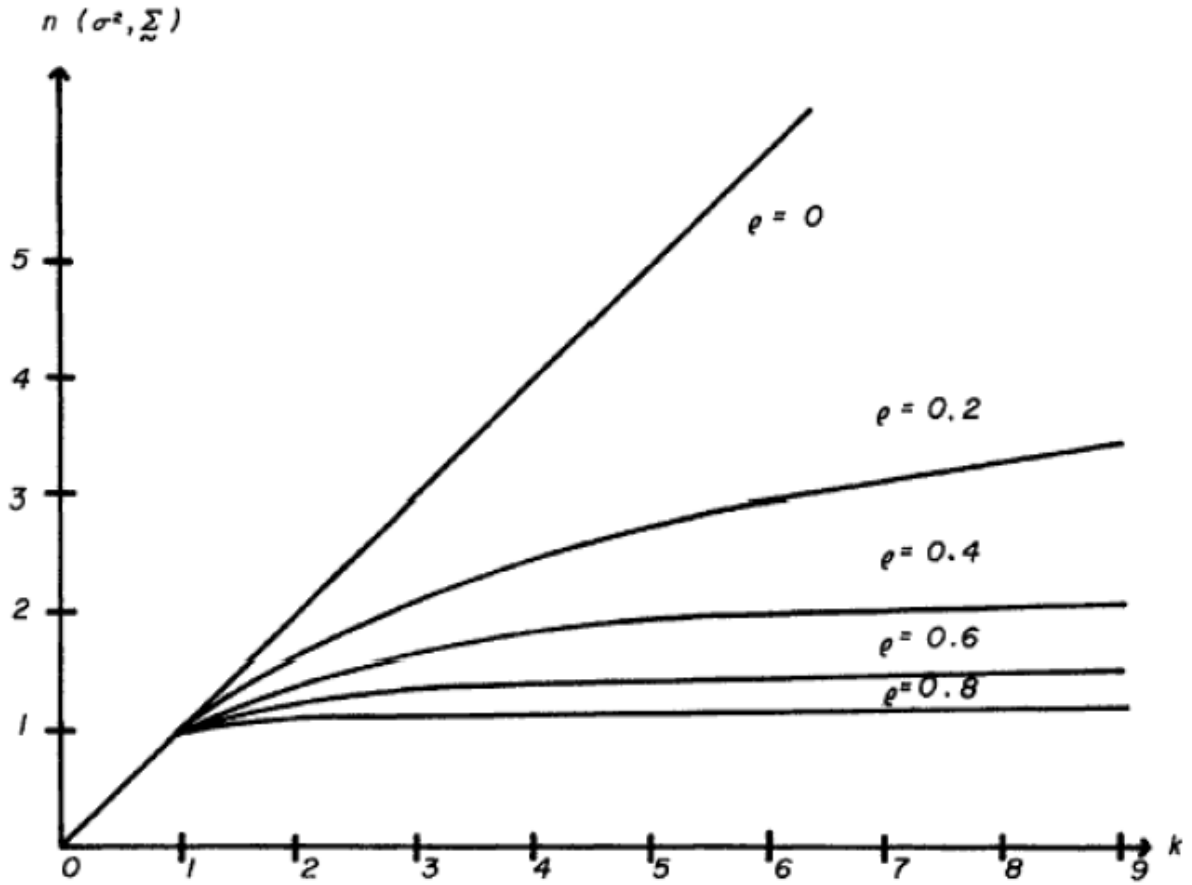


Figure 1. The equivalent number of independent experts as a function of k, the actual number of experts, for selected values of ρ.

Positive dependencies decrease the number of effective labelers

What Happens if we have Dependencies?

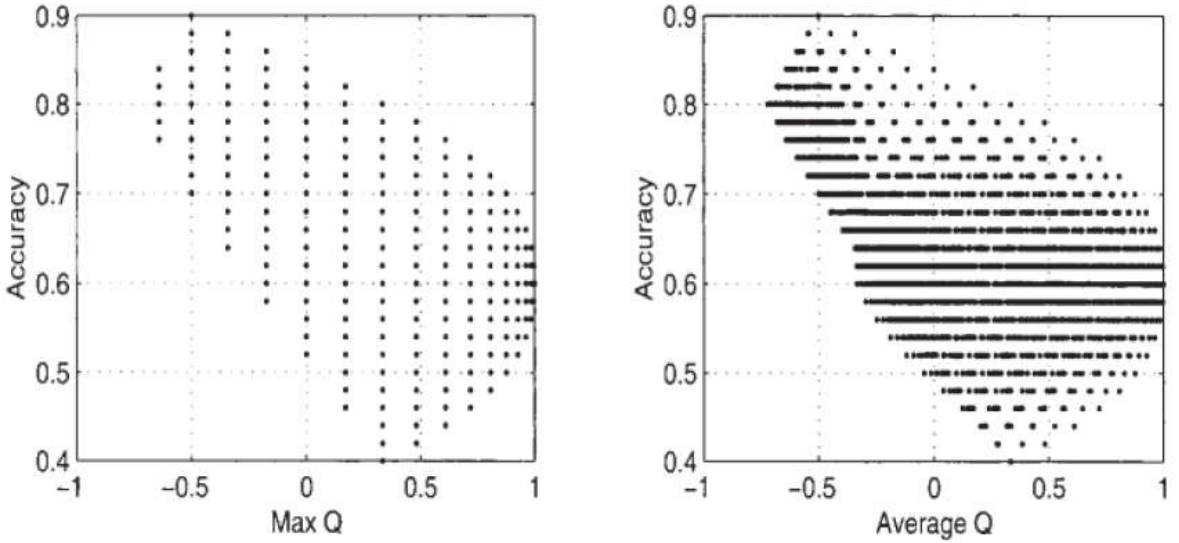


Fig. 2 Plot of P_{maj} against Q_{max} and Q_{avr} for all possible combinations of 3 votes for $N = 50$ objects

a	*b*
c	*d*

$$Q = \frac{ad - bc}{ad + bc}$$

Yule's Q
measure of correlation

Kuncheva et al., PA&A, 2003

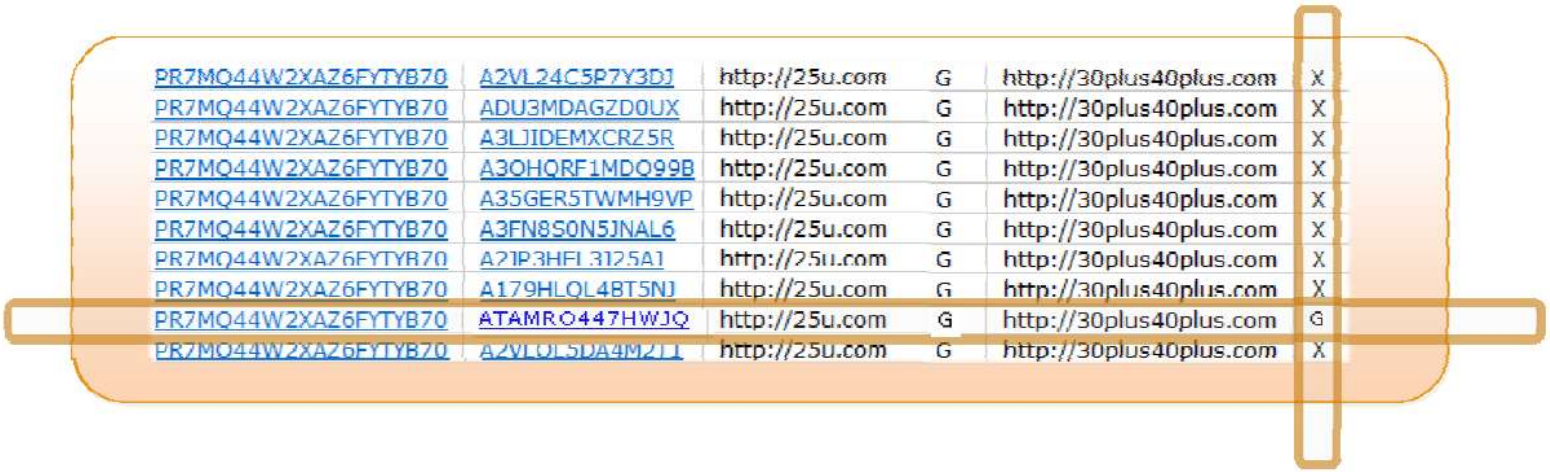
- Positive dependencies decrease the number of effective labelers
- Negative dependencies can improve results (unlikely both workers to be wrong at the same time)

Vote Combination: Meta-studies

- Simple averages tend to work well
- Complex models slightly better but less robust
- [Clemen and Winkler, 1999, Ariely et al. 2000]

From Aggregate Labels to Worker Quality

- Look at our spammer 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	A3QHQRF1MDQ99B	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	A21P3HFL3125A1	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	A2VLQL5DA4M2T1	http://25u.com	G	http://30plus40plus.com	X

- After aggregation, we compute confusion matrix for each worker
- After majority vote, confusion matrix for ATAMRO447HWJQ
  - P[G → G]=100% P[G → X]=0%
  - P[X → G]=100% P[X → X]=0%

Algorithm of Dawid & Skene, 1979

- Iterative process to estimate worker error rates
- 1. Initialize by aggregating labels for each object (e.g., use majority vote)
- 2. Estimate confusion matrix for workers (using aggregate labels)
- 3. Estimate aggregate labels (using confusion matrix)
  - Keep labels for "gold data" unchanged
- 4. Go to Step 2 and iterate until convergence
- Confusion **matrix 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**. Seems like a spammer…

And many Variations…

- van der Linden et al, 1997: Item-Response Theory
- Uebersax, Biostatistics 1993: Ordered categories
- Uebersax, JASA 1993: Ordered categories, with worker expertise and bias, item difficulty
- Carpenter, 2008: Hierarchical Bayesian versions
- And more recently at NIPS
  - Whitehill et al., 2009: Adding item difficulty
  - Welinder et al., 2010: Adding worker expertise

Challenge: From Confusion Matrixes to Quality Scores

- All the algorithms will generate "confusion matrixes" for workers
- 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 → X]=0% P[X → G]=100%
  - P[G → X]=0% P[G → G]=100%
- Confusion matrix for good worker
  - P[X → X]=80% P[X→ G]=20%
  - P[G → X]=20% P[G → G]=80%
- Spammers figure out how to fly under the radar…
- In reality, we have 85% G sites and 15% X sites
- Errors 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%
- In reality, we have 85% G sites, 5% P sites, 5% R sites, 5% X sites
- Errors of spammer (all in G) = 0% * 85% + 100% * 15% = 15%
- Error rate of biased worker = 80% * 85% + 100% * 5% = 73%
- False positives: Legitimate workers appear to be spammers

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!

Quality Score

- $ExpCost(p) = \sum_{i=1}^{L} \sum_{j=1}^{L} p_i . p_j . c_{ij}$
- High cost when "soft" labels have probability spread across classes
- Low cost when "soft" labels have probability mass concentrated in one class
- [Assume equal misclassification costs]

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

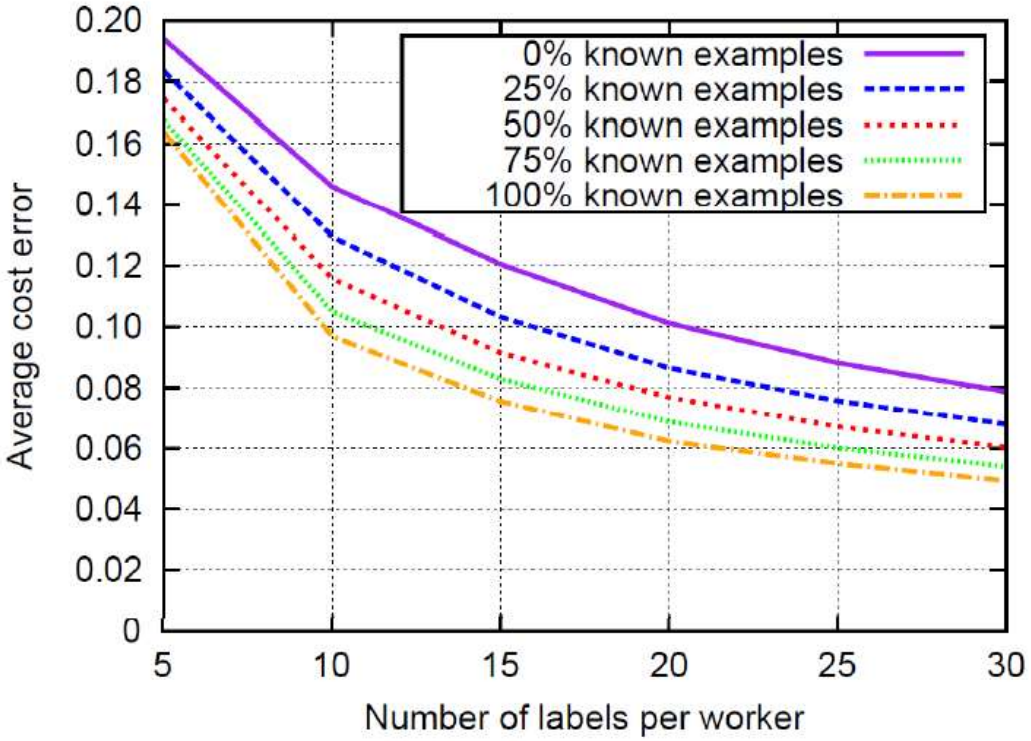
Quality Score

- A *spammer* is a worker who always assigns labels randomly, regardless of what the true class is.
- QualityScore = 1 - ExpCost(Worker)/ExpCost(Spammer)
- QualityScore is useful for the purpose of blocking bad workers and rewarding good ones
- Essentially a multi-class, cost-sensitive AUC metric
  - AUC = area under the ROC curve

What about Gold Testing?

- Naturally integrated into the latent class model
- *1. Initialize by aggregating labels for each object (e.g., use majority vote)*
- 2. Estimate **error rates** for workers (using aggregate labels)
- 3. Estimate **aggregate labels** (using error rates, weight worker votes according to quality)
  - **Keep labels for "gold data" unchanged**
- 4. Go to Step 2 and iterate until convergence

Gold Testing

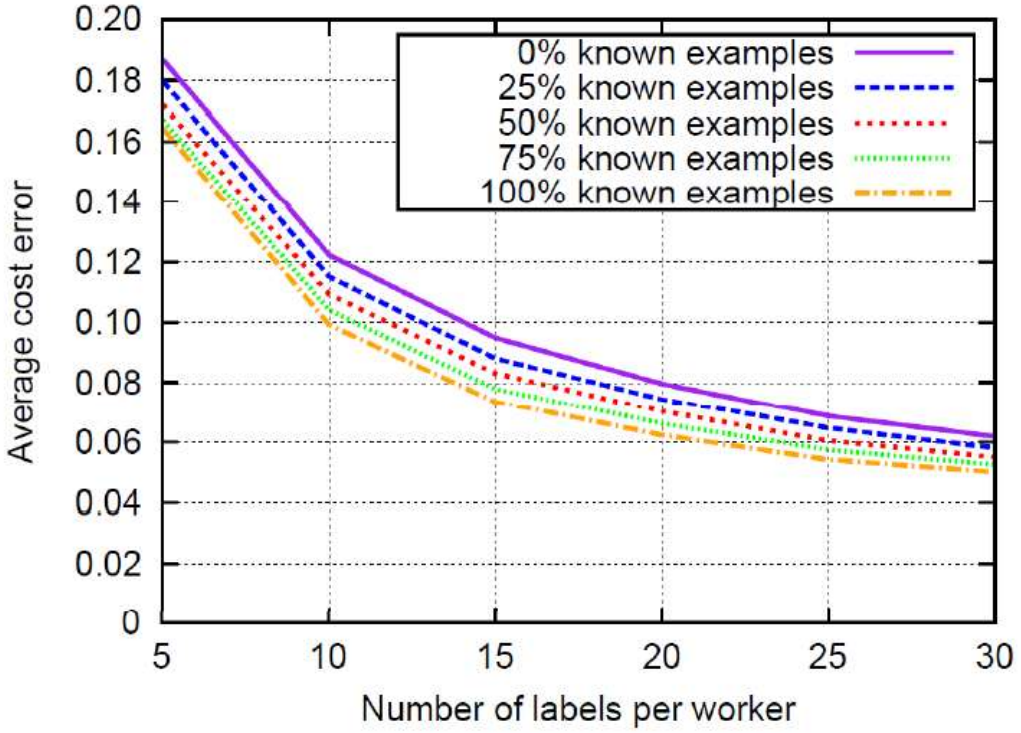


- **3 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:0.05:1.0
- 200 labelers

No significant advantage under "good conditions" (balanced datasets, good worker quality)

http://bit.ly/gold-or-repeated
Wang, Ipeirotis, Provost, WCBI 2011

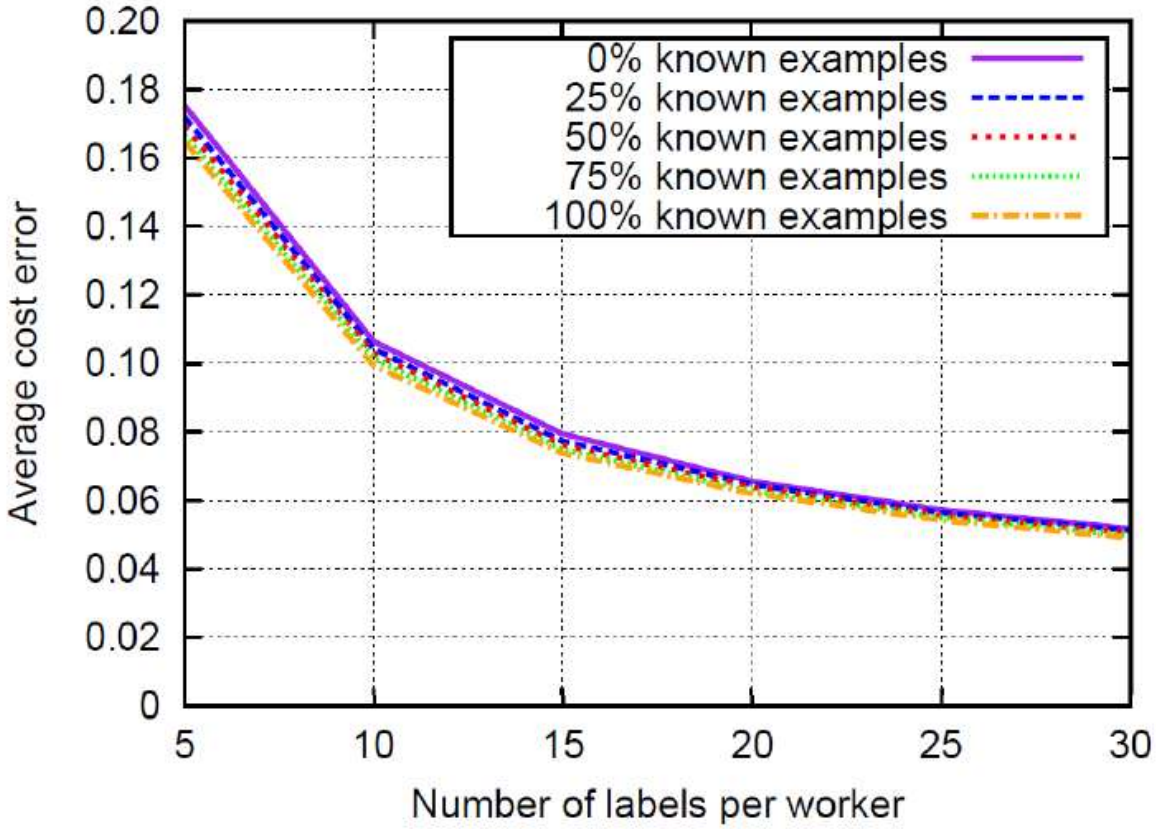
Gold Testing



- **5 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:1.0
- 200 labelers

No significant advantage under "good conditions" (balanced datasets, good worker quality)

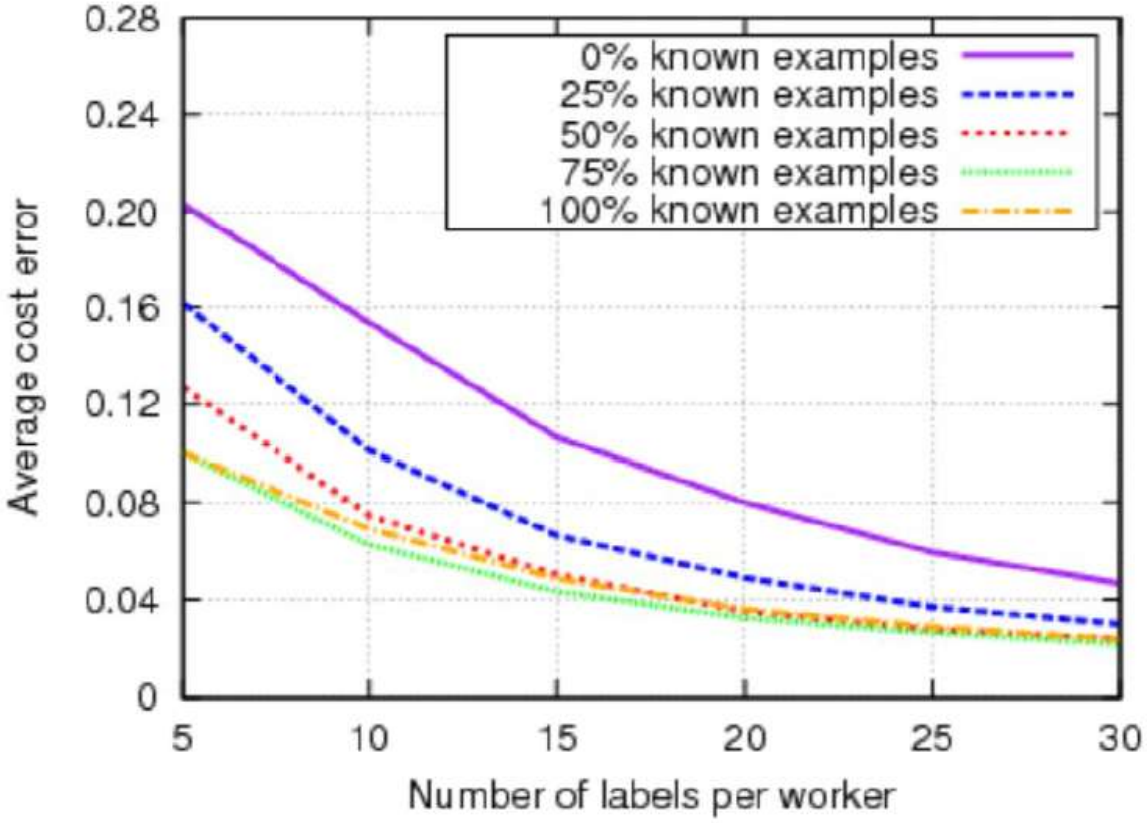
Gold Testing



- **10 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:1.0
- 200 labelers

No significant advantage under "good conditions" (balanced datasets, good worker quality)

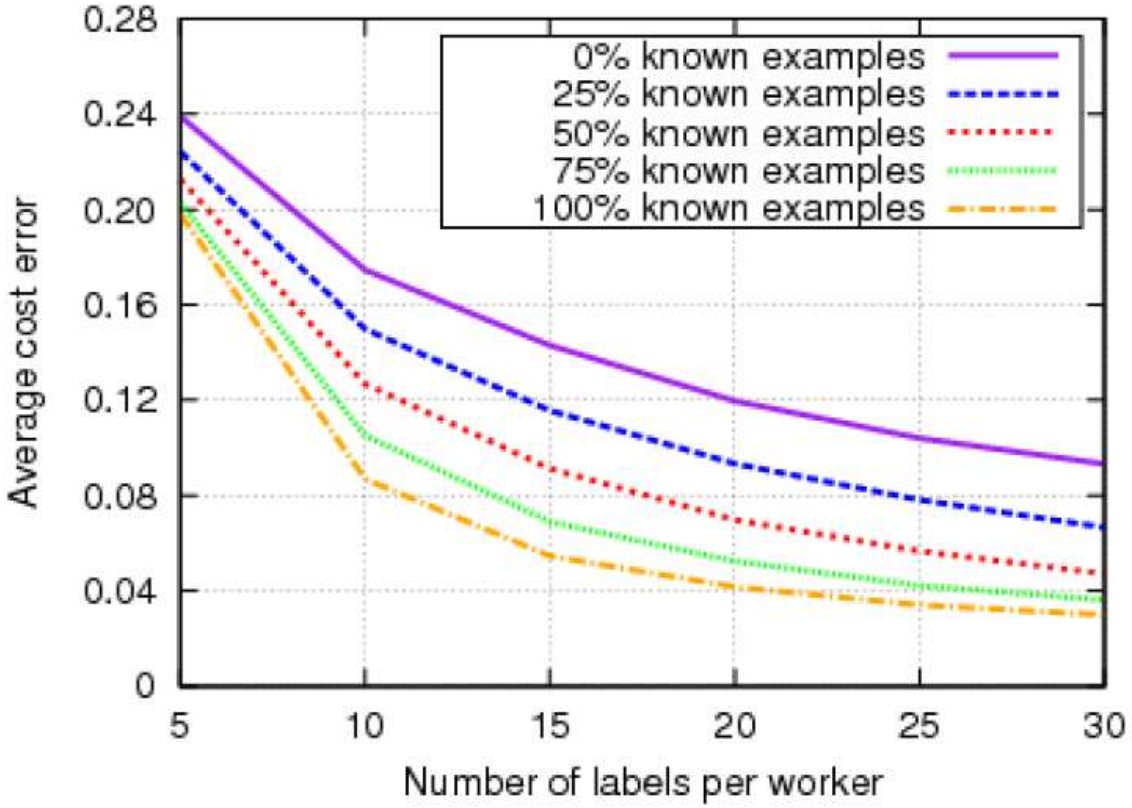
Gold Testing



- **10 labels per example**
- 2 categories, 90/10
- Quality range: 0.55:1.0
- 200 labelers

Advantage under imbalanced datasets

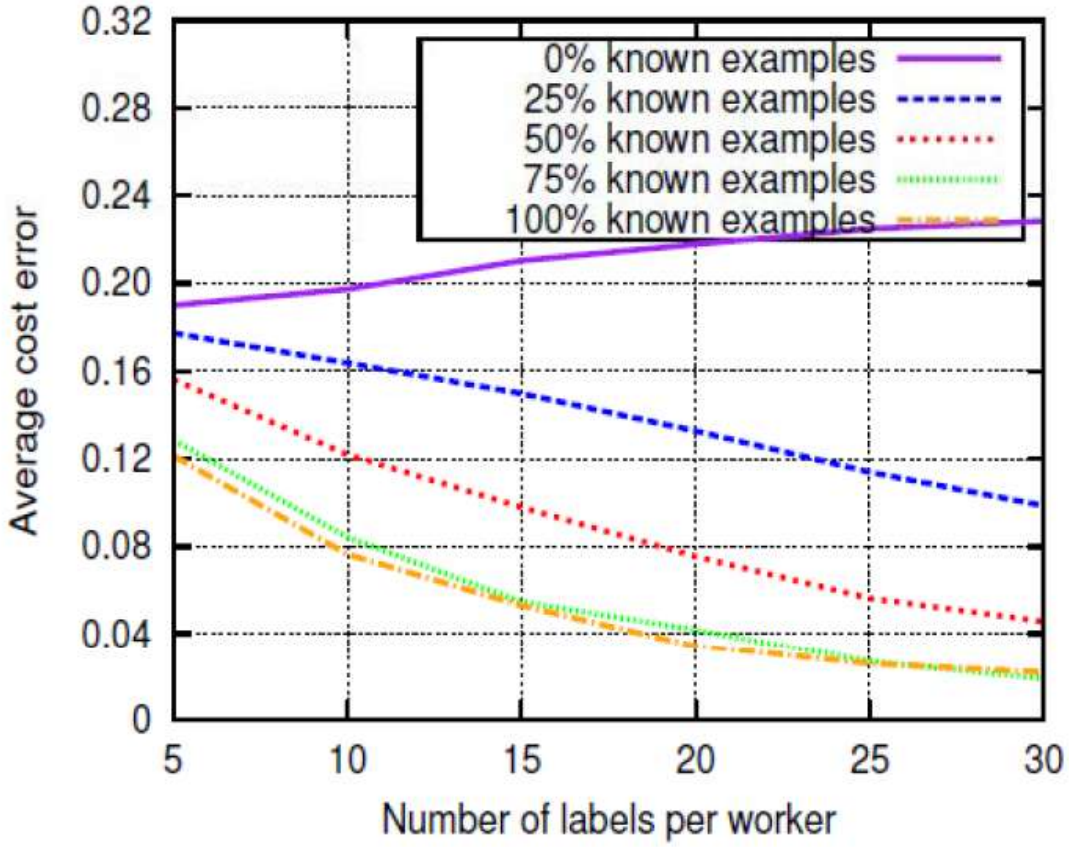
Gold Testing



- **5 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:0.65
- 200 labelers

Advantage with bad worker quality

Gold Testing



- **10 labels per example**
- 2 categories, 90/10
- Quality range: 0.55:0.65
- 200 labelers

Significant advantage under "bad conditions" (imbalanced datasets, bad worker quality)

Testing Workers

- An **exploration-exploitation** scheme
  - **Explore**: Learn about the quality of the workers
  - **Exploit**: Label new examples using the quality

Testing Workers

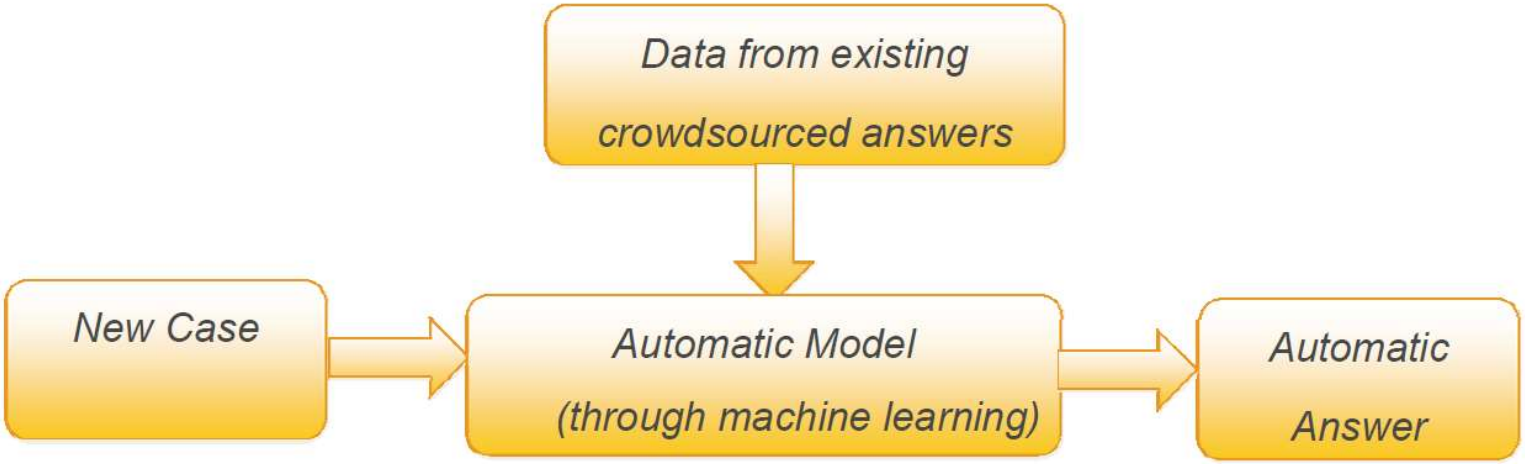
- An exploration-exploitation scheme
  - Assign gold labels when benefit in learning better quality of worker outweighs the loss for labeling a gold (known label) example [Wang et al, WCBI 2011]
  - Assign an already labeled example (by other workers) and see if it agrees with majority [Donmez et al., KDD 2009]
  - If worker quality changes over time, assume accuracy given by HMM and $\phi(\tau) = \phi(\tau\text{-}1) + \Delta$ [Donmez et al., SDM 2010]

Integrating with Machine Learning

- Crowdsourcing **is cheap but not free**
  - Cannot scale to web without help
- Solution: Build automatic classification models using crowdsourced data

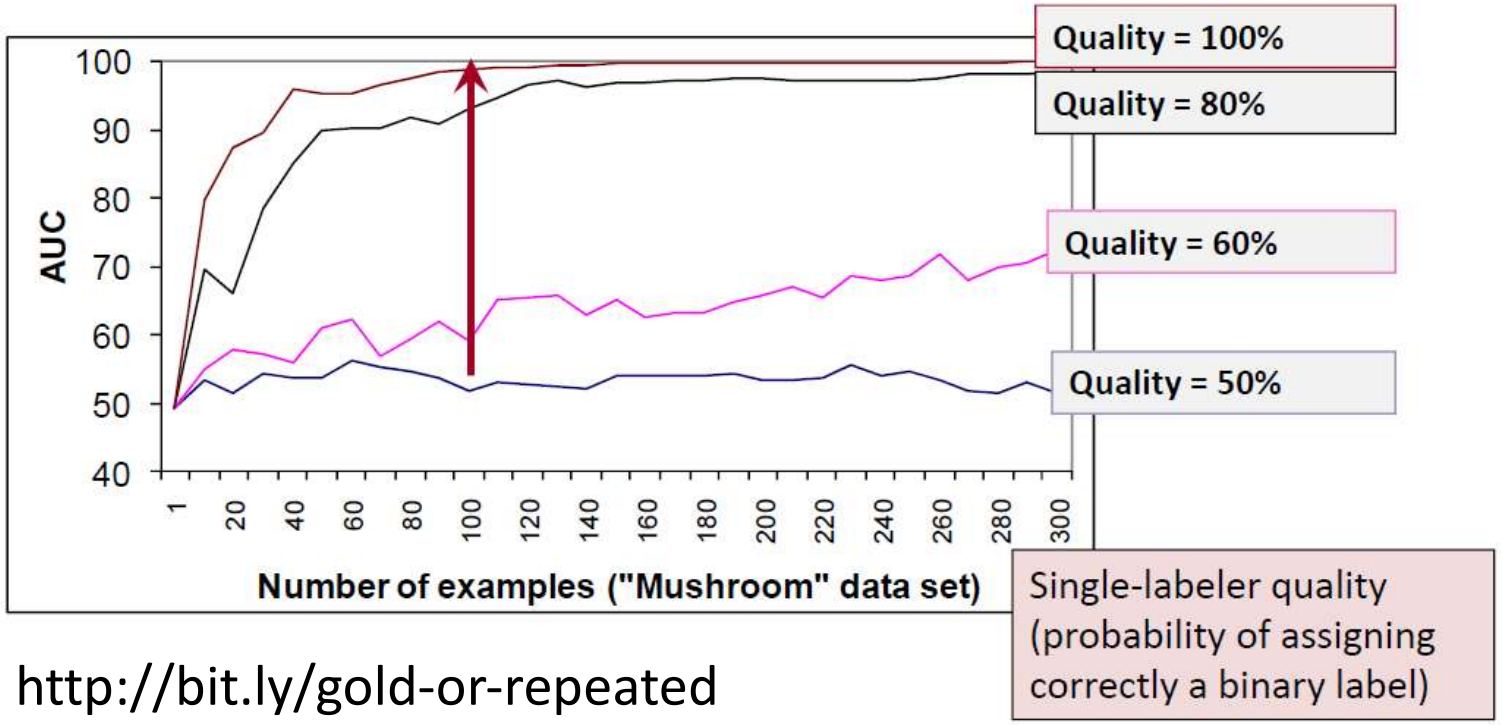
Simple Solution

- Humans label training data
- Use training data to build model



Quality and Classification Performance

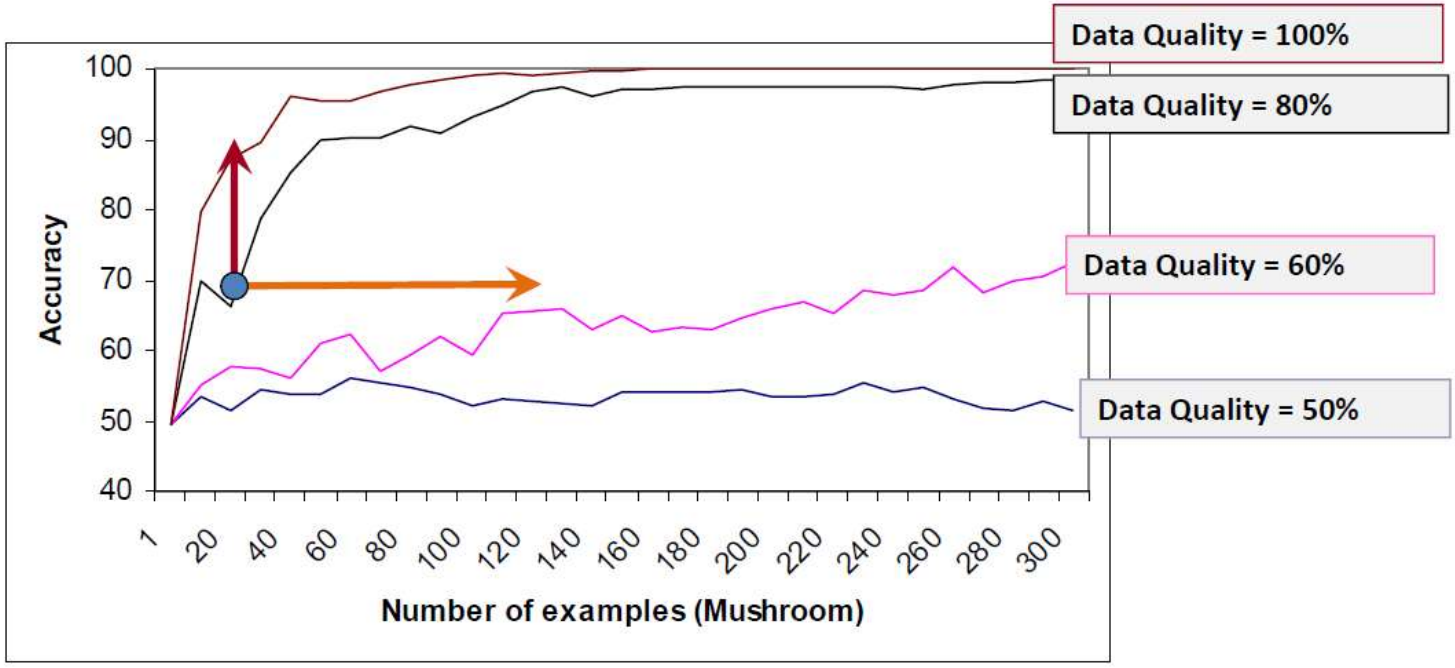
- *Noisy labels lead to degraded task performance*
- Labeling quality increases → classification quality increases



- http://bit.ly/gold-or-repeated
- Sheng, Provost, Ipeirotis, KDD 2008

Tradeoffs for Machine Learning Models

- Get more data → Improve model accuracy
- Improve data quality → Improve classification

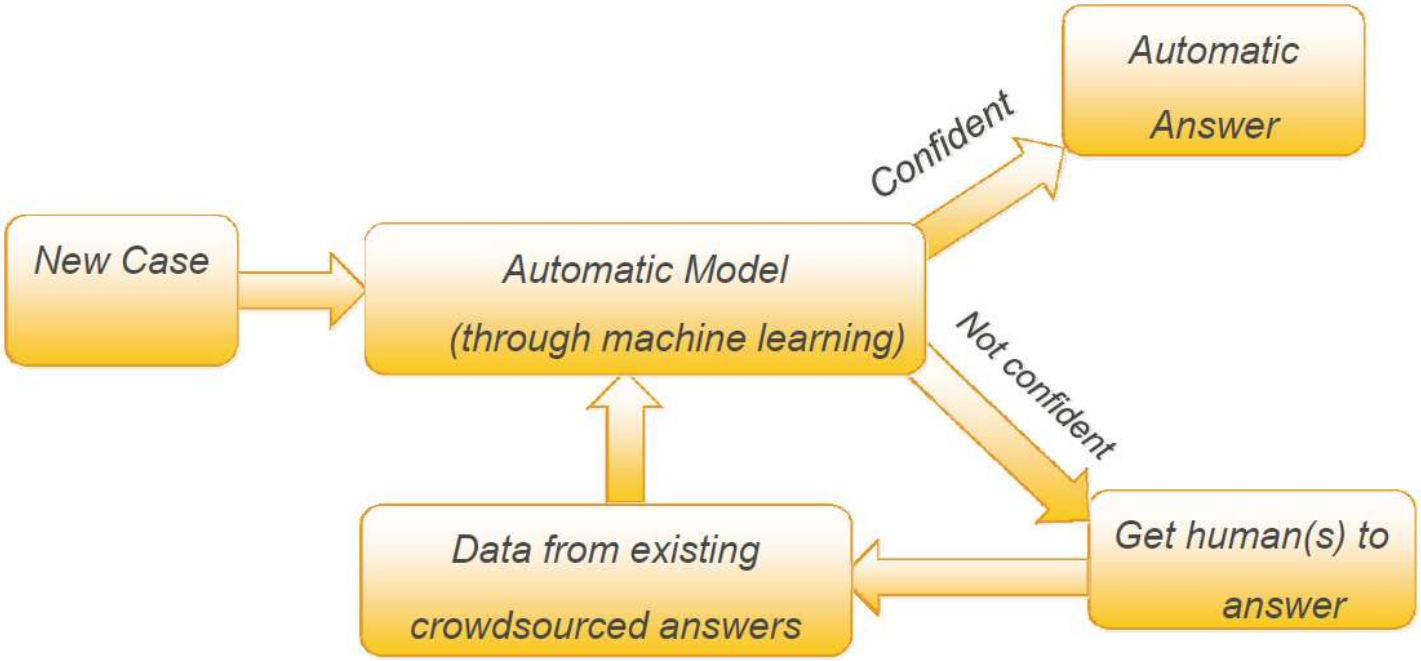


Tradeoffs for Machine Learning Models

- **Get more data**: Active Learning, select which unlabeled example to label [Settles, http://active-learning.net/]
- **Improve data quality**
  - Repeated Labeling, label again an already labeled example [Sheng et al. 2008, Ipeirotis et al, 2010]

Scaling Crowdsourcing: Iterative training

- Use model when confident, humans otherwise
- Retrain with new human input → Improve model → reduce need for humans



Rule of Thumb Results

- With high quality labelers (80% and above): One worker per case (more data better)
- With low quality labelers (~60%): Multiple workers per case (to improve quality)
- [Sheng et al, KDD 2008; Kumar and Lease, CSDM 2011]

Dawid & Skene meets a Classifier

- [Raykar et al. JMLR 2010]: Use the Dawid&Skene scheme but add a classifier as an additional worker
- Classifier in each iteration learns from the consensus labeling

Selective Repeated-Labeling

- We do not need to label everything same number of times
- **Key observation**: we have additional information to guide selection of data for repeated labeling
  - the current multiset of labels
- Example: {+,-,+,-,-,+} vs. {+,+,+,+,+,+}

Label Uncertainty: Focus on Uncertainty

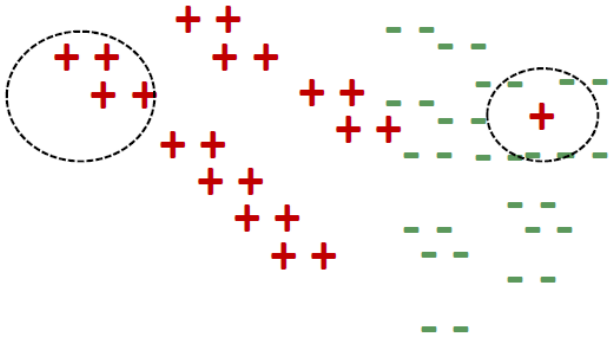
- If we know worker qualities, we q , can estimate log-odds for each example:

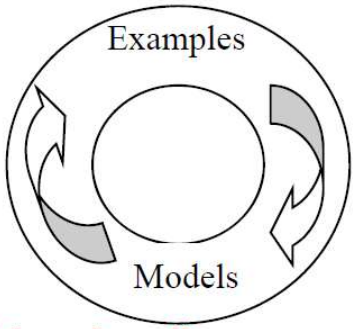
$$-\log\left(\frac{p^*}{1-p^*}\right) = \sum_{i=0}^{N} \log\left(\frac{p_i}{1-p_i}\right)$$

- Assign labels first to examples that are **most uncertain** (logodds close to 0 for binary case)

Model Uncertainty (MU)

- Learning models of the data provides an alternative source of information about label certainty
- **Model uncertainty**: get more labels for instances that cause model uncertainty
- Intuition?
  - **for modeling**: why improve training data quality if model already is certain there?
  - **for data quality**, low-certainty "regions" may be due to incorrect labeling of corresponding instances

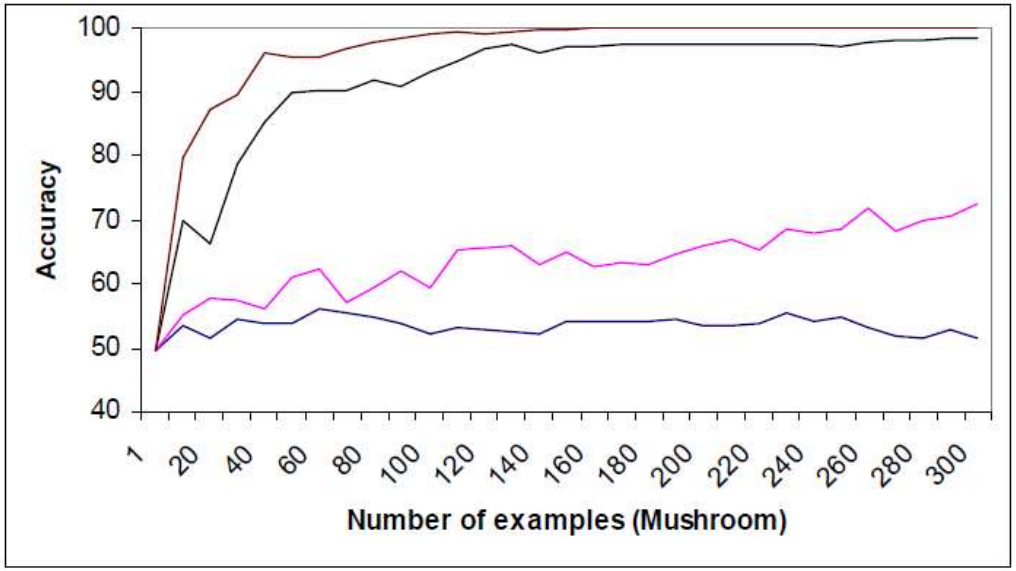




"Self-healing" process
[Brodley et al, JAIR 1999]
[Ipeirotis et al, NYU 2010]

Learning from Imperfect Data

- With inherently noisy data, good to have learning algorithms that are robust to noise.
- Or use techniques designed to handle explicitly noisy data
- [Lugosi 1992; Smyth, 1995, 1996]



Take-away Messages

- Humans are still better at doing many tasks compared to machines
- Human labor is cheap in third world countries
- Crowdsourcing is very useful for many applications
- However, there are many challenges that need to be dealt with
- We saw some mechanisms to deal with data quality from crowdsourcing platforms

Further Reading

- AAAI 2011 (w HCOMP 2011): Human Computation: Core Research Questions and State of the Art (Edith Law and Luis von Ahn)
- WSDM 2011: Crowdsourcing 101: Putting the WSDM of Crowds to Work for You (Omar Alonso and Matthew Lease) http://ir.ischool.utexas.edu/wsdm2011_tutorial.pdf
- LREC 2010 Tutorial: Statistical Models of the Annotation Process (Bob Carpenter and Massimo Poesio) http://lingpipe-blog.com/2010/05/17/lrec-2010-tutorial-modelingdata-annotation/
- ECIR 2010: Crowdsourcing for Relevance Evaluation. (Omar Alonso http://wwwcsif.cs.ucdavis.edu/~alonsoom/crowdsourcing.html
- CVPR 2010: Mechanical Turk for Computer Vision. (Alex Sorokin and Fei-Fei Li) http://sites.google.com/site/turkforvision/
- CIKM 2008: Crowdsourcing for Relevance Evaluation (Daniel Rose) http://videolectures.net/cikm08_rose_cfre/
- WWW 2011: Managing Crowdsourced Human Computation (Panos Ipeirotis)

Preview of Lecture 27: Course Project Presentations

- Five minute presentations
  - Schedule has been published
  - Slide 1: Name of the project, problem definition, one motivating example explaining the problem.
  - Slide 2: Summary of Related previous work (This requires you to read at least one paper related to this project. Even if there are no papers on the same problem as your project, you should discuss at least one paper that you think is the most related to your work.
  - Slide 3: Your methodology in short. This slide is important because it should tell everyone about the "intelligent sparkling idea" that you introduced in your project to solve the problem.
  - Slide 4: Results and summary

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!