Quality Assurance Techniques

Irwin King

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Agenda

- Introduction
- Examples
  - Task Matching
  - Drug Design
  - Game Theory Basics
  - Case Studies Based on Game Theory
  - **Quality Assurance Techniques**
  - Conclusion

Quality Assurance Techniques

- Ground Truth Seeding
- Expert Review
- Automatic Check
- Redundancy (Vote-based)
  - Majority vote
  - Quality adjusted vote
  - Gold Testing
- Active Data Collection

Ground Truth Seeding

- Start with a small number of tasks for which ground truth has been provided by a trusted source
- Mix in questions with known answers
- Used in post-processing to estimate the true worker error-rate
  - For instance, we have the true labels of 10 examples
  - Worker A gives 8 accurate labels among these 10 examples
  - P_{ml} *(Worker A's error-rate) = 0.2*

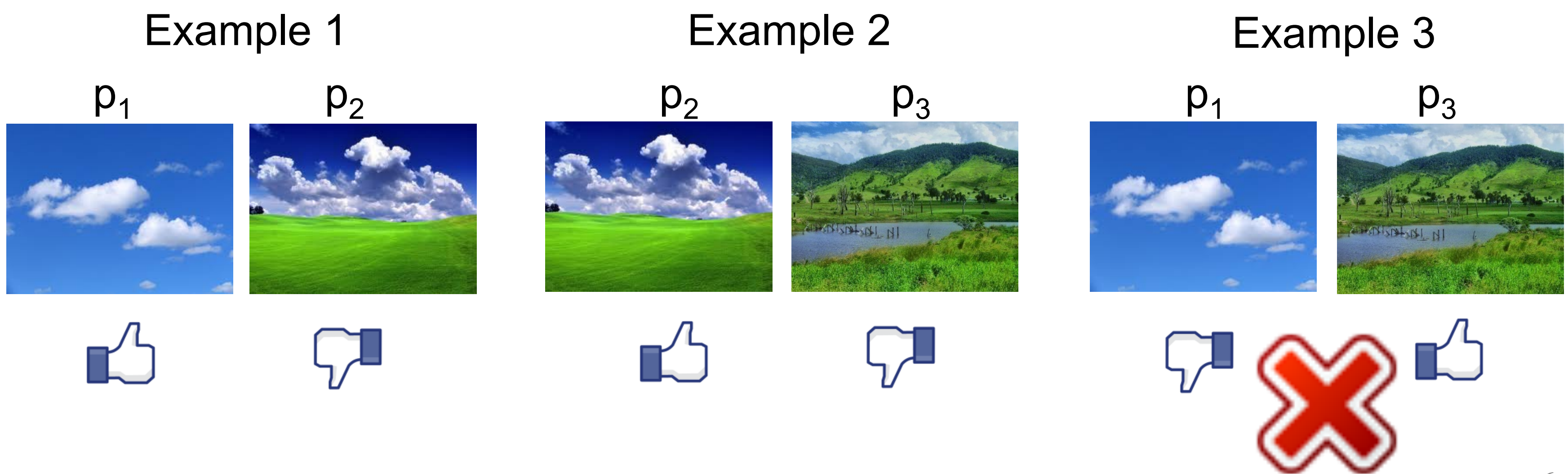
Expert Review

- A trusted expert skims or cross-checks contributions for relevance and apparent accuracy
  – For example, with Mechanical Turk, people who post tasks may review the work and choose whether to pay or not

Automatic Check

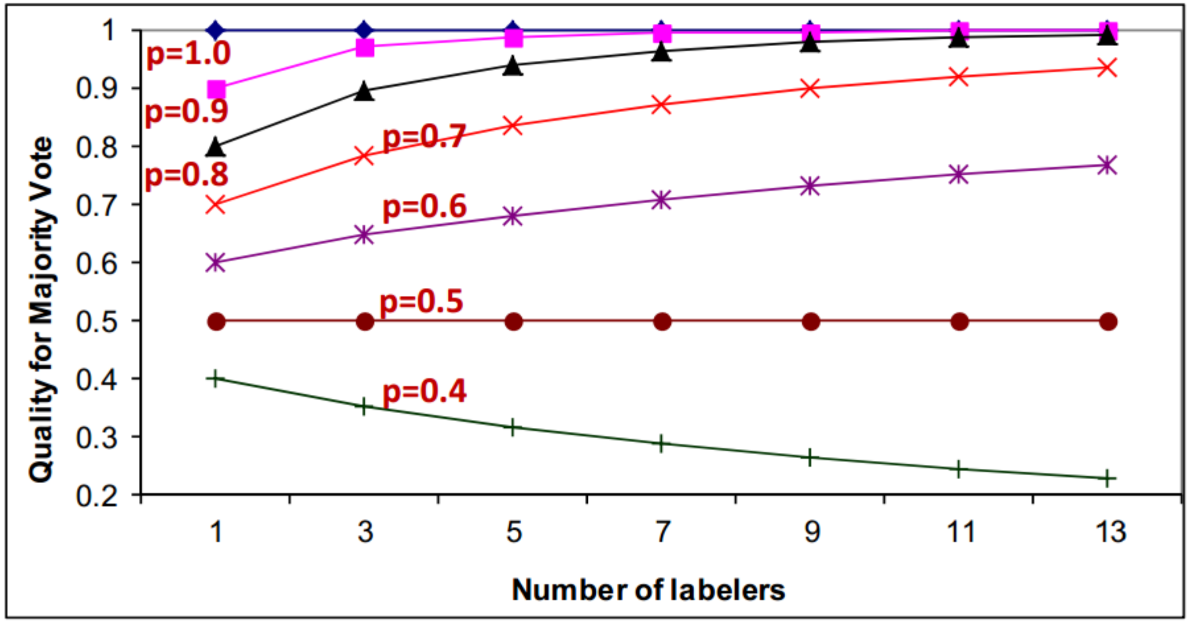
p2

- Reasoning → Paradox Detection
- Example: Image Search
  - Task: Given each search query, select which of the two alternative results (images) is more relevant
  - Query: sky



Majority Voting and Label Quality

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

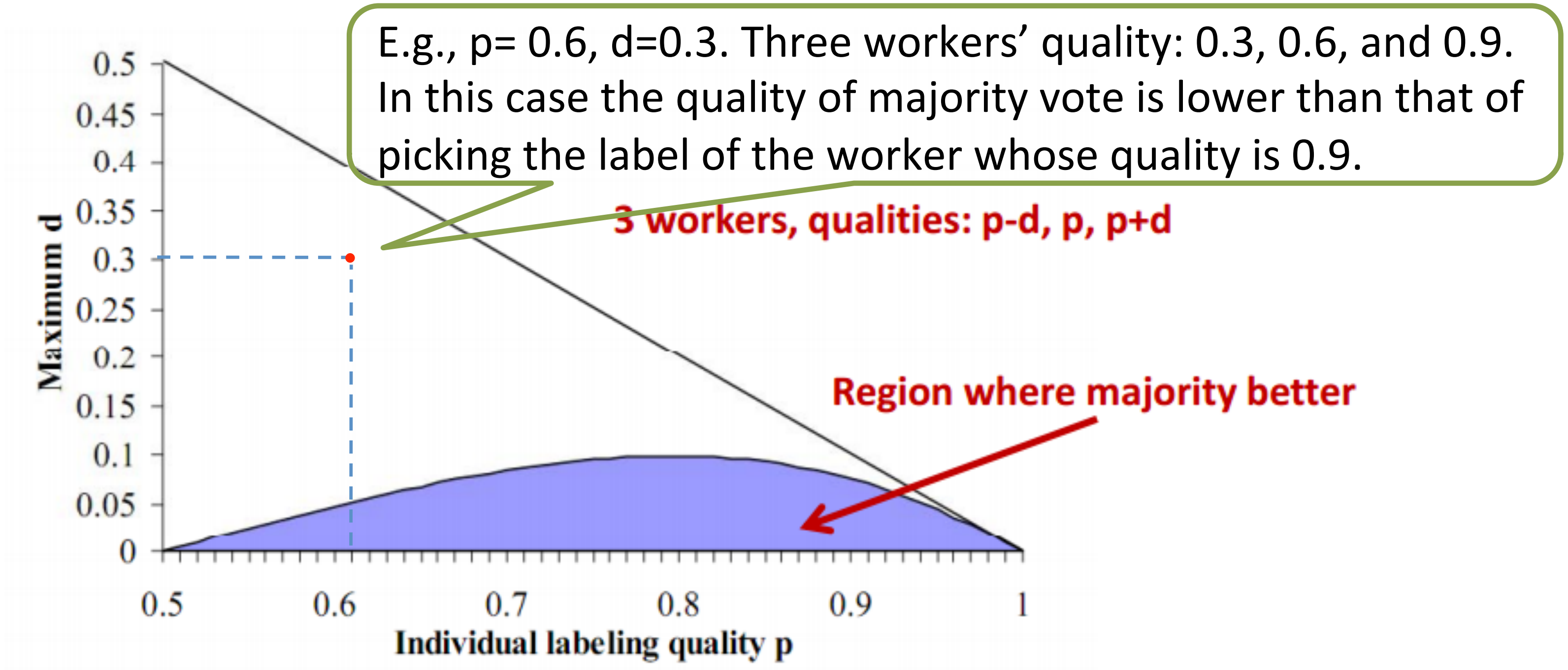


P is probability of individual **labeler** being correct

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

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

What if Qualities of Workers are Different?



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

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

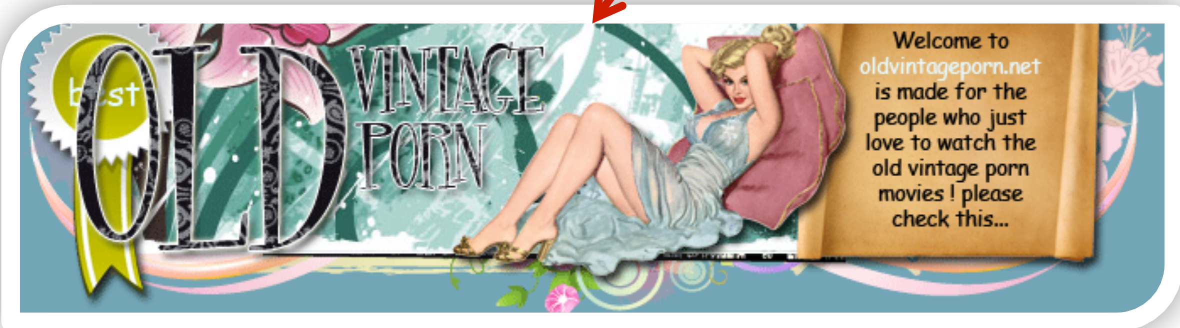
G (general audience) **PG** (parental guidance) **R** (restricted) **X** (porn)

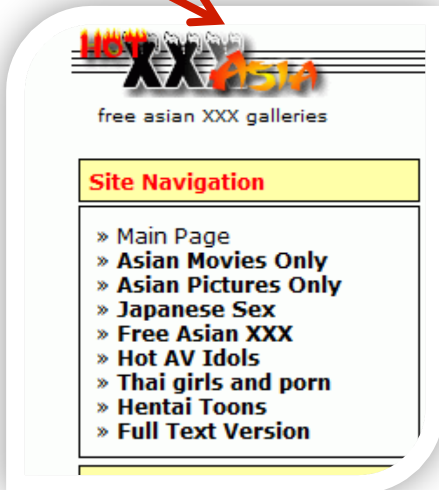
Cost/Speed Statistics:

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

Spammers!

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





Worker **ATAMRO447HWJQ**

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

From Aggregate Labels to Worker Quality

Look at our spammer **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 the error rate for each worker

Algorithm of (Dawid & Skene, 1979)

[and many recent variations on the same theme]

- Iterative process to estimate worker error rate

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)
    - Keep labels for "gold data" unchanged (coming later)
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%**

ATAMRO447HWJQ marked **almost all** sites as **G**. Seems like a spammer…

The probability of labeling X websites as G websites is 99.153%

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?

Challenge 1: Spammers are Lazy and Smart!

Confusion matrix for a spammer

P[X → X]=0% **P[X → G]=100%**
P[G → X]=0% **P[G → G]=100%**

Confusion matrix for a 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
- Total Error rate of the spammer = 0% * 85% + 100% * 15% = 15%
- Total Error rate of the good worker = 85% * 20% + 85% * 20% = 20%

False negatives: Spam workers pass as legitimate

Challenge 2:
Workers 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
- Total Error rate of spammer (all G) = 0% * 85% + 100% * 15% = 15%
- Total Error rate of biased worker = 80% * 85% + 100% * 5% = 73%

False negatives: Spam workers pass as legitimate

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 **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%**

Assume equal priors:

- 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**

The results are highly ambiguous. No information provided!

Expected Misclassification Cost

- High cost when "soft" labels have probability spread across classes
- Low cost when "soft" labels have 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%>	0.01

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

soft label cost = 1 – P(correct labeling)

Quality Score

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

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

- QualityScore is useful for the purpose of blocking bad workers and rewarding good ones

Empirical Results and Observations

- 500 web pages in G, P, R, X, manually labeled
- 100 workers per page
- Lots of noise on MTurk. 100 votes per page:
  - 95% accuracy with majority
  - 99.8% accuracy after modeling worker quality
- Blocking based on error rate: Only 1% of labels dropped
- Blocking based on quality score: 30% of labels dropped

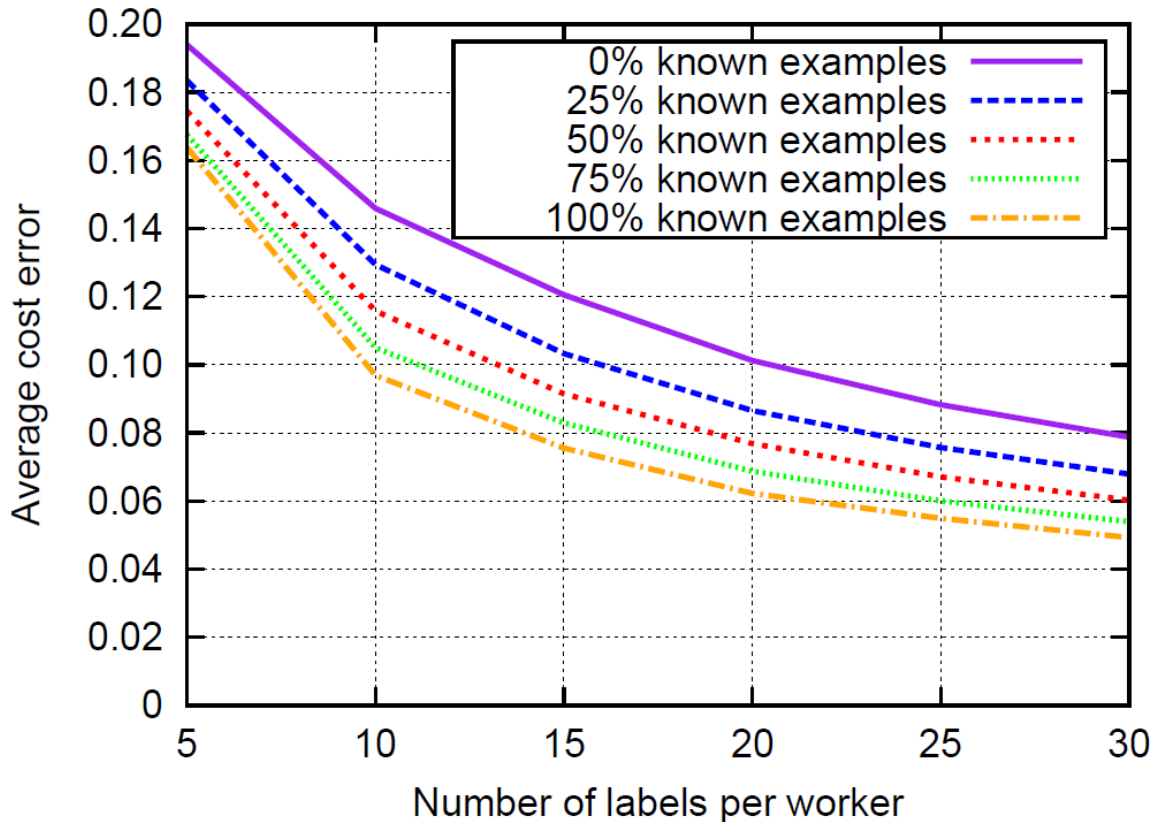
Gold Testing

- Algorithm of (Dawid & Skene, 1979)

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)
    - **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

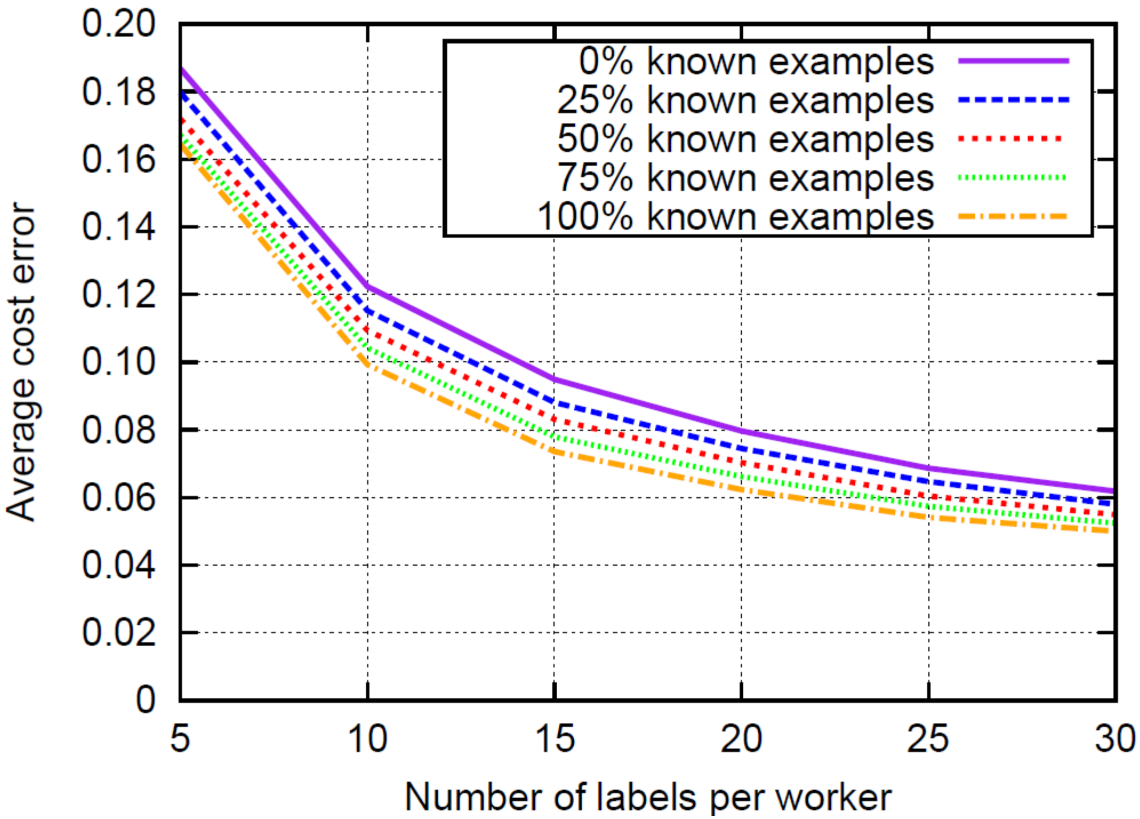


No significant advantage under "good conditions" (balanced datasets, good worker quality)

http://www.behind-the-enemy-lines.com/2010/09/worker-evaluation-in-crowdsourcing-gold.html, crawled on May 15, 2013.

Gold Testing

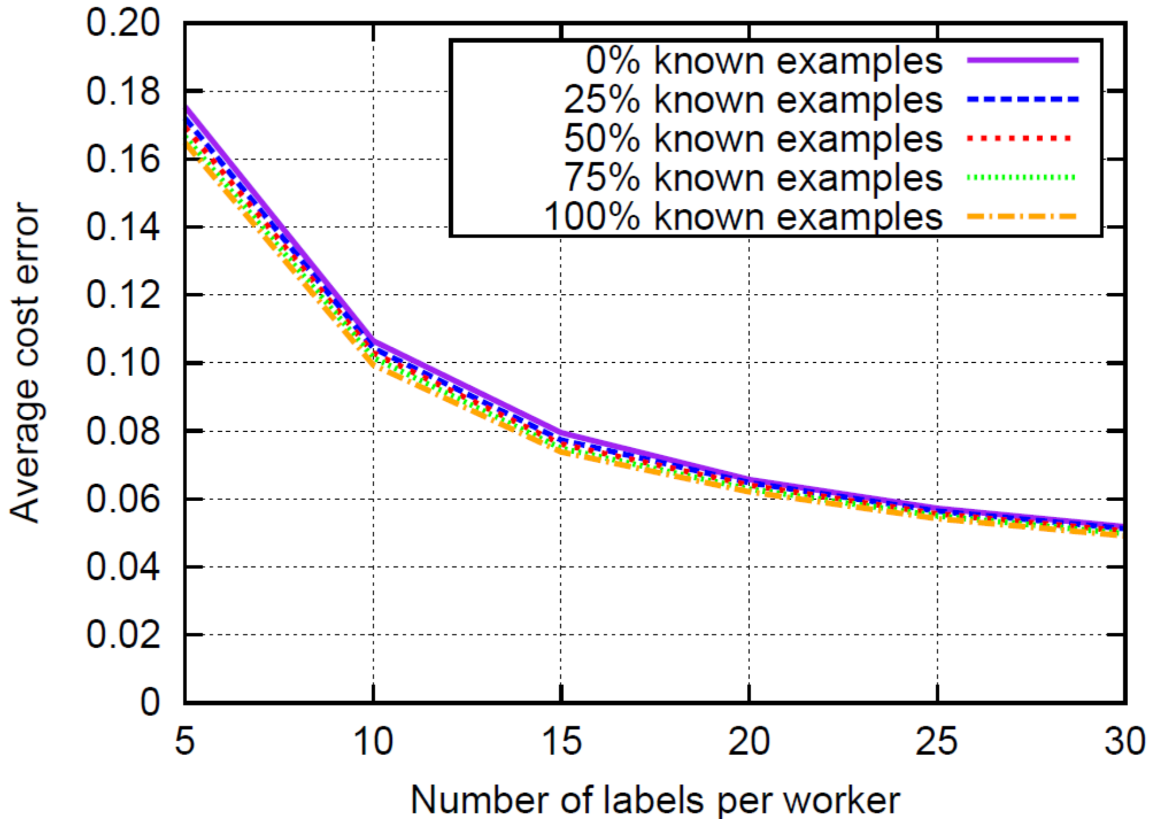
- **5 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:0.05:1.0
- 200 labelers



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

Gold Testing

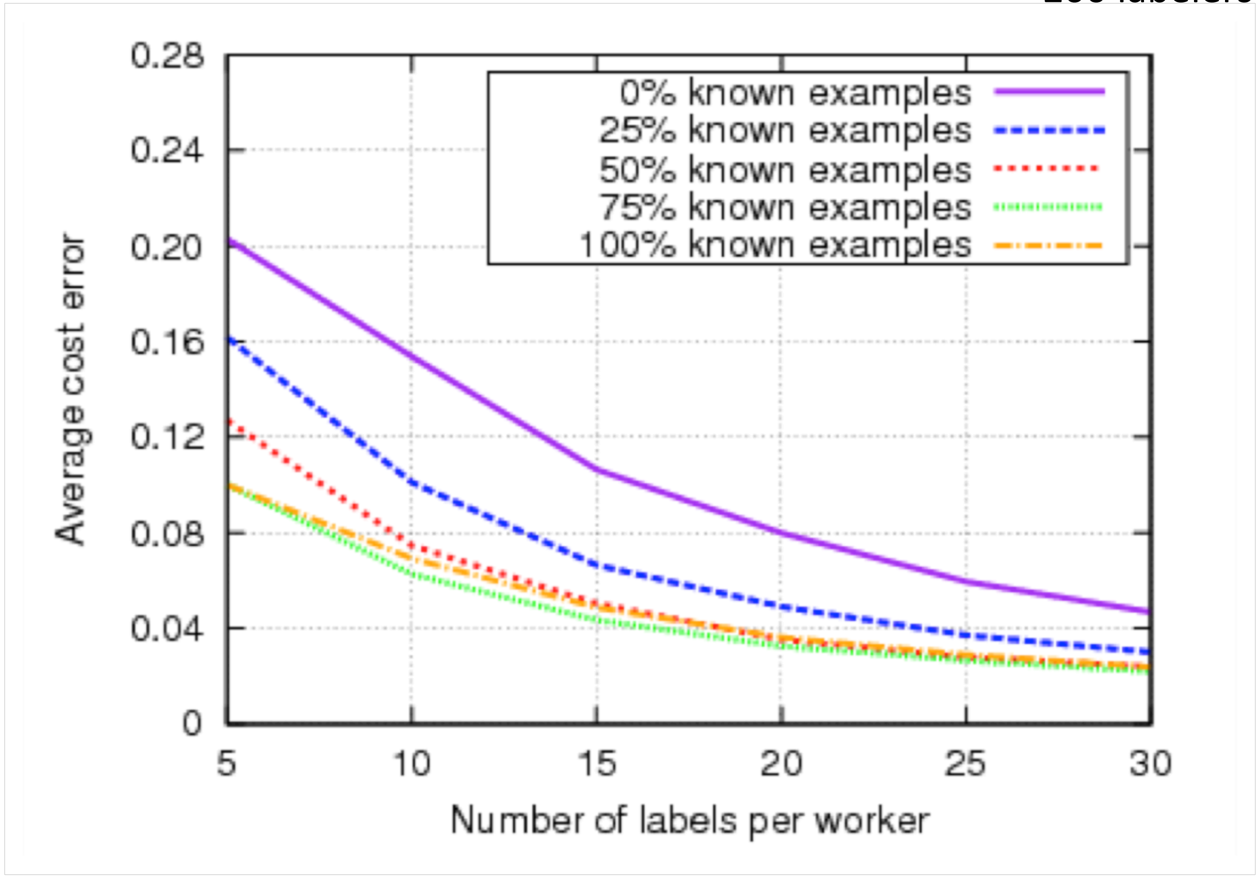
- **10 labels per example**
- 2 categories, 50/50
- Quality range: 0.55:0.05:1.0
- 200 labelers



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

Gold Testing

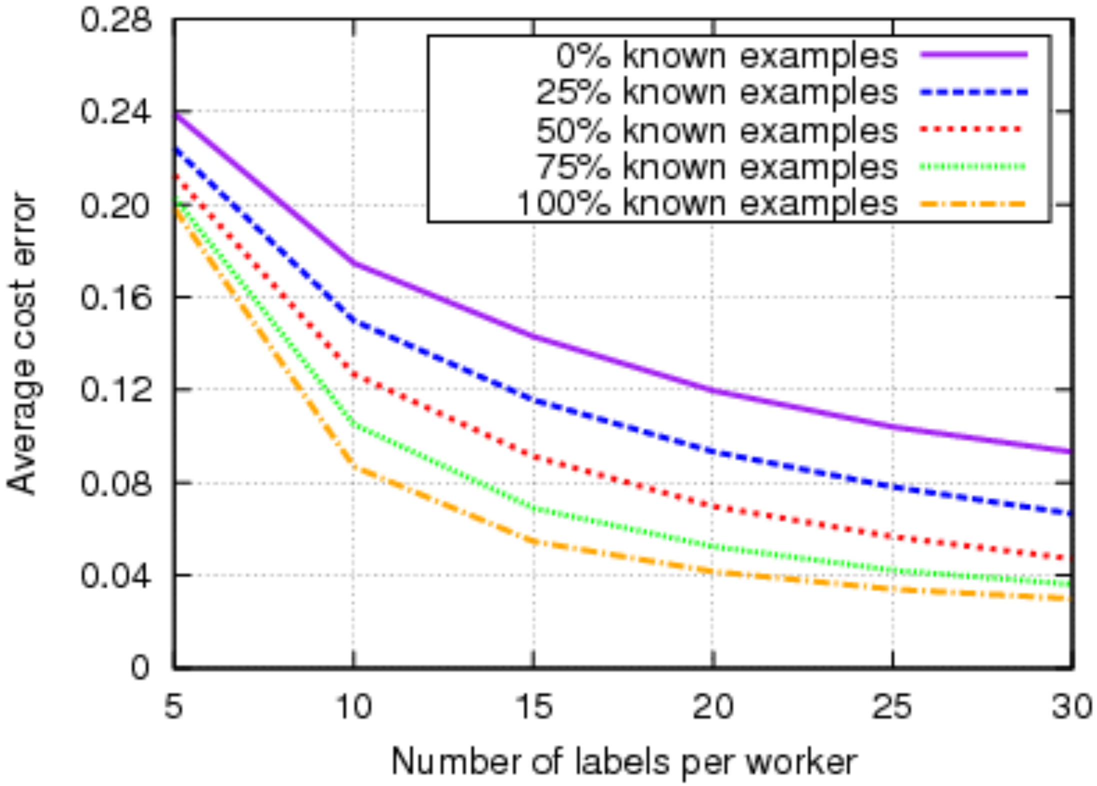
- 10 labels per example
- **2 categories, 90/10**
- Quality range: 0.55:0.05:1.0
- 200 labelers



Advantage under **imbalanced datasets**

Gold Testing

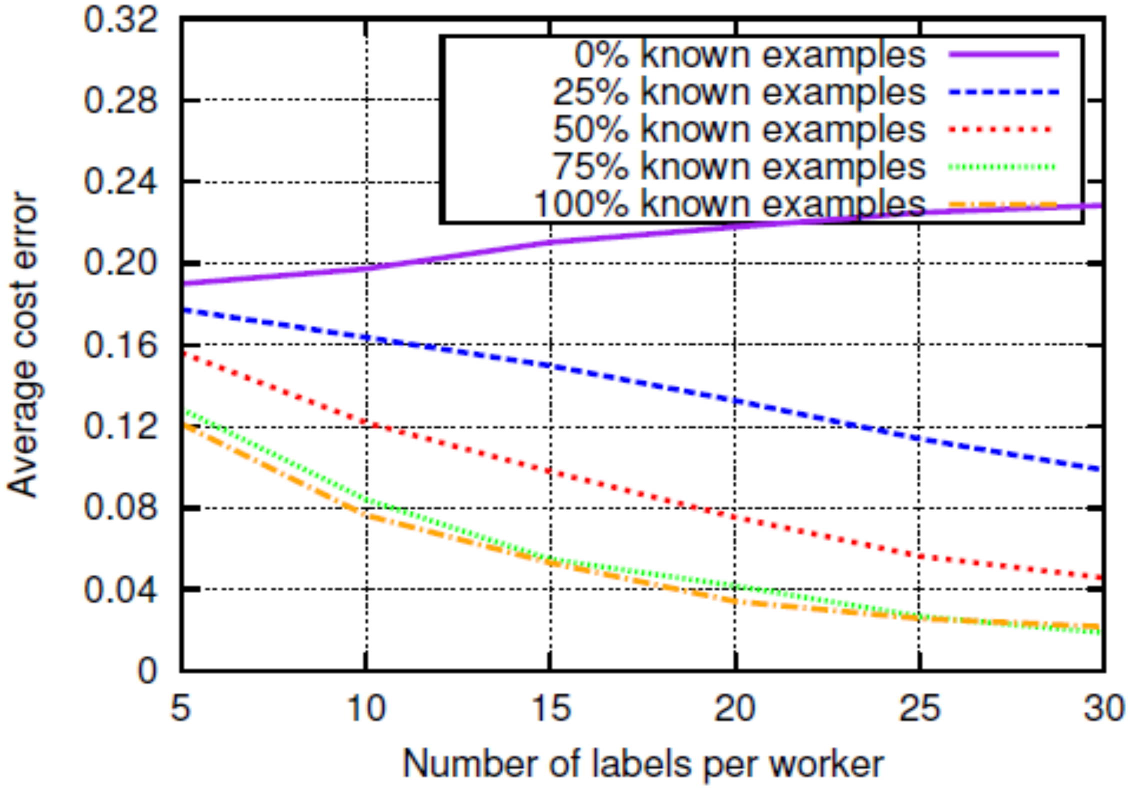
- 10 labels per example
- 2 categories, 90/10
- **Quality range: 0.55:0.65**
- 200 labelers



Advantage under **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)

Active Data Collection

- Intuition: we do not need to label everything same number of times

Rule of Thumb Results:

- With high quality labelers (85% and above): One worker per example (Get more data)
- With low quality labelers (~60-70%): Multiple workers per example (Improve quality)

- Solution: selective repeated-labeling

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
- Example: {+,-,+,-,-,+} vs. {+,+,+,+,-,+}

Natural Candidate: Entropy

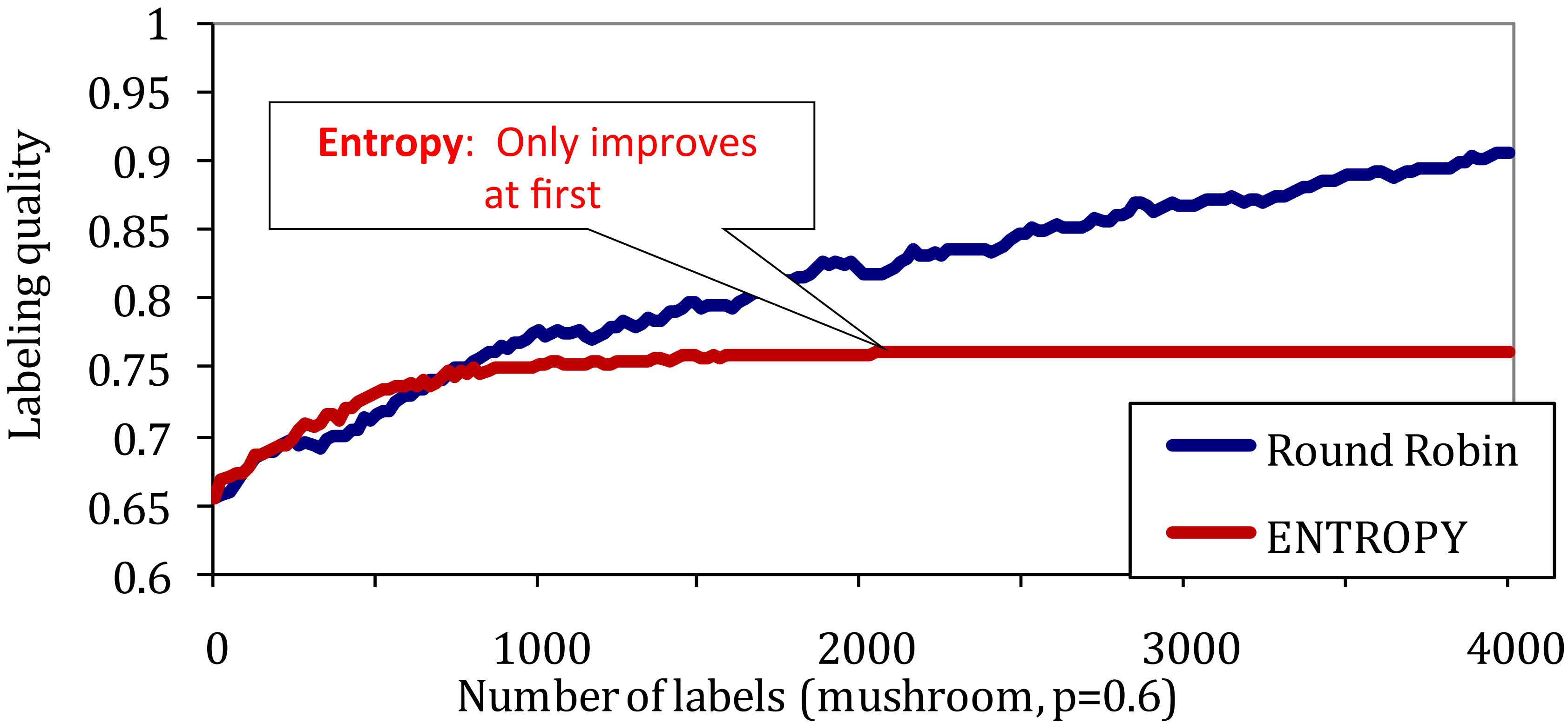
- Entropy is a natural measure of label uncertainty:

$$E(S) = -\frac{|S^+|}{|S|}\log_2\frac{|S^+|}{|S|} - \frac{|S^-|}{|S|}\log_2\frac{|S^-|}{|S|}$$

$$|S^+|: positive - |S^-|: negative$$

- E({+,+,+,+,-,+})=0.65
- E({+,-, +,-, -,+ })=1
- **Strategy**: *Get more labels for high-entropy label multisets*

What Not to Do: Use Entropy



Why?

- In the presence of noise, entropy will be high even with many labels
- Entropy is scale invariant
  - {3+, 2-} has same entropy as {600+ , 400-}, *i.e., 0.97*
- Entropy measures the level of noise

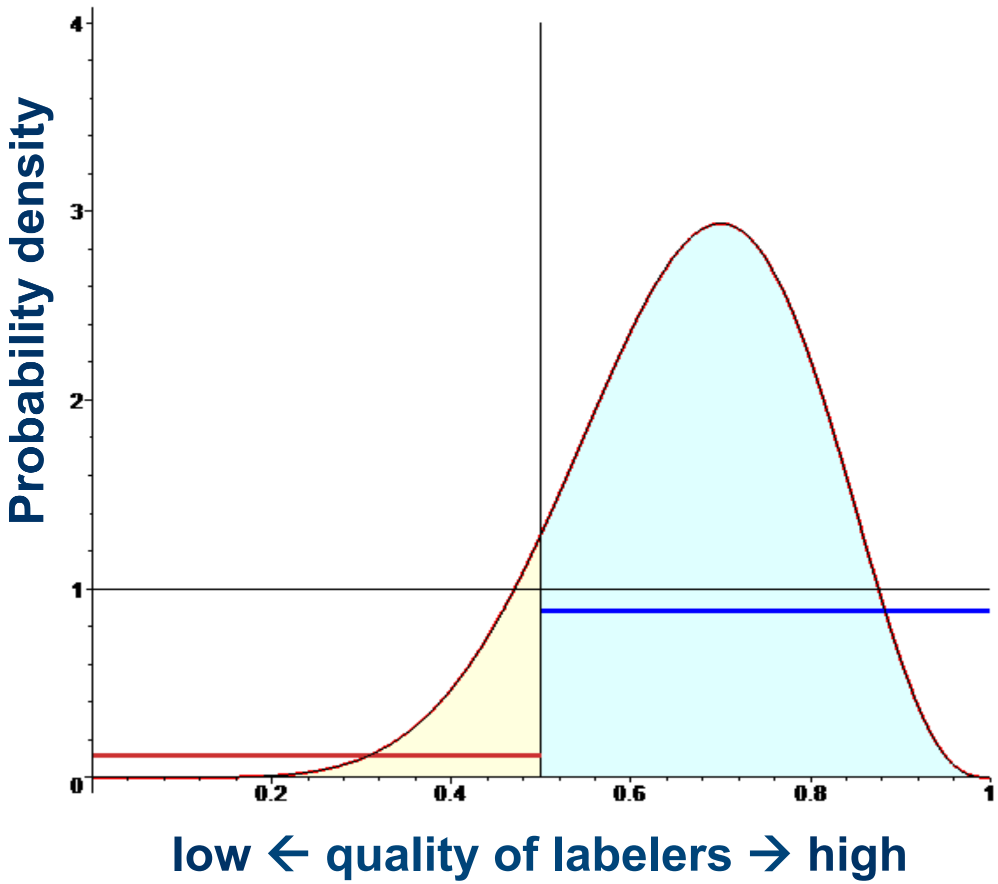
Uncertainty, Not Entropy

- *If* we knew worker quality q, we could estimate class probabilities

$$Pr(+|p,n) = \frac{Pr(p,n|+) \cdot Pr(+)}{Pr(p,n)} = q^p \cdot (1-q)^n \frac{Pr(+)}{Pr(p,n)}$$

- **But we do not know (exact) worker quality q!**
- Estimate first worker quality q for each example

Bayesian Estimate of Labeler Quality

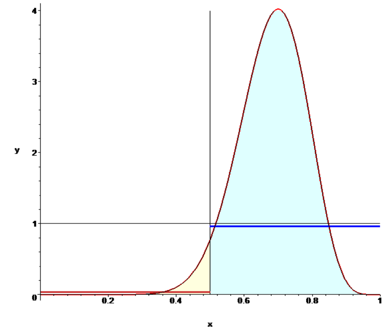


Quality=0.7 (unknown)

- 10 labels (8+, 2-)

- Observe assigned labels and estimate what is the noise level
- Find the **distribution** of possible **quality of labelers for the example**

Label Uncertainty



- Estimate the (distribution of) quality of the workers
- For each quality level, estimate the correct class
- **Integrate across all qualities**
- **Label examples with Pr(class| votes) close to 0.5**

Package for Quality Assurance

Open source implementation available at:

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

- Input:
  - *Labels from Mechanical Turk*
  - *[Optional] Some "gold" labels from trusted labelers*
  - *Cost of incorrect labelings (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*

Questions

- What the advantages and disadvantages of the above techniques?
- When would you choose majority voting in crowdsourcing/human computation?
- Why a spammer's error rate can be lower than normal workers in some cases? Please give an example.

Summary

Technique	Advantage	Disadvantage
Ground Truth Seeding	Easy to control the quality of workers	Sometimes it is not easy to get the ground truth
Expert Review	Easy to control the quality of workers	Sometimes it is not easy to get the labelers from experts
Automatic Check	Avoid manual labels	Not applicable everywhere
Redundancy	Works well when there are enough labels; Avoid manual labels	Need more labels when worker quality is low
Active Data Collection	Adjust the number of workers based on worker quality	Quality estimation may not be so accurate when the number of labels is small

Agenda

- Introduction
- Examples
  - Task Matching
  - Drug Design
  - Game Theory Basics
  - Case Studies Based on Game Theory
  - Quality Assurance Techniques
  - **Conclusion**

References for Quality Assurance

- Dawid, Alexander Philip, and Allan M. Skene. "Maximum likelihood estimation of observer error-rates using the EM algorithm." *Applied Statistics*(1979): 20-28.
- Quinn, Alexander J., and Benjamin B. Bederson. "Human computation: a survey and taxonomy of a growing field." *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011.
- Duguid, Paul. "Limits of self-organization: peer production and laws of quality." *First Monday* 11.10, 2006.
- Little, Greg, and Y. Sun. "Human OCR: Insights from a complex human computation process." *Workshop on Crowdsourcing and Human Computation, Services, Studies and Platforms, ACM CHI*. 2011.
- Le, John, et al. "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution." *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 2010.
- Snow, Rion, et al. "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks." *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008.

References for Quality Assurance

- Ipeirotis, Panagiotis G., Foster Provost, and Jing Wang. "Quality management on amazon mechanical turk." *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2010.
- Ipeirotis, Panos. "Crowdsourcing using Mechanical Turk: quality management and scalability." *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011*. ACM, 2011.
- Ipeirotis, Panagiotis G., and Praveen K. Paritosh. "Managing crowdsourced human computation: a tutorial." *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011.
- Kazai, Gabriella, and Natasa Milic-Frayling. "On the evaluation of the quality of relevance assessments collected through crowdsourcing." *SIGIR 2009 Workshop on the Future of IR Evaluation*. 2009.

The grass is greener on the other side...

Be inspired!

Stories and more stories...

Be informed!

The devil is in the details...

Be challenged!

Q&A


QUALITY ASSURANCE
MONITORING
© Original Artist / Search ID: jfa0411
Rights Available from CartoonStock.com
CARTOONSTOCK
.COM
JOSEPH
FARRIS


ANN
'72
MARI
HARDIN
CARTOONSTOCK
.COM
"It all began when my call
was monitored for Quality
assurance purposes."
© Original Artist / Search ID: pha1071
Rights Available from CartoonStock.com

- Auction (拍卖)
- Quality assurance (质量保证)
- Social computing (社会计算)
- Aggregate labels (总的标签)
- Ground truth (地面真相)
- Paradox detection (悖论检测)
- Error rate (错误率)
- Misclassification (误判)
- Entropy (熵)
- Spammer (垃圾邮件发送者)