Speech Transcription with Crowdsourcing

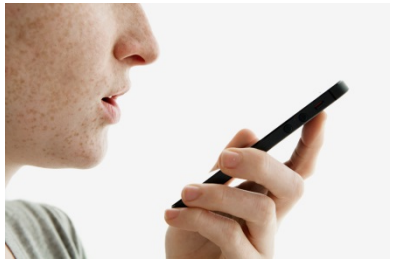
Scott Novotney

11/20/13

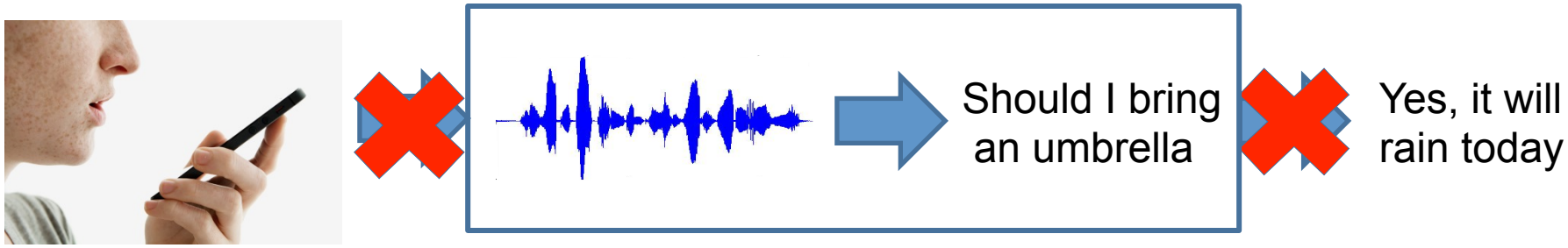
Why I'm here

1. Get more data, not better data
2. Use other Turkers to do QC for you
3. Non-English crowdsourcing is not easy

Siri in Five Minutes



Siri in Five Minutes



Automatic Speech Recognition

Digit Recognition

Digit Recognition

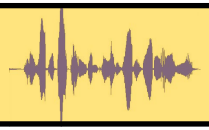
Digit Recognition

P(one| ) =

Digit Recognition

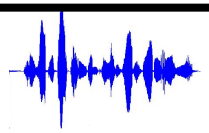
$$P(\text{one}|\ ) = \frac{P(\ |\text{one})\ P(\text{one})}{P(\ )}$$

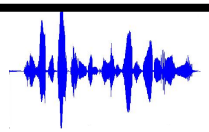
Digit Recognition

P(one| ) = P( |one) P(one)

Acoustic Model **Language Model**

Digit Recognition

$$P(\text{one}|\ \text{  }\ ) = \frac{P(\ \text{  }\ |\text{one})\ P(\text{one})}{P(\ \text{  }\ )}$$

$$P(\text{two}|\ \text{  }\ ) = \frac{P(\ \text{  }\ |\text{two})\ P(\text{two})}{P(\ \text{  }\ )}$$

Digit Recognition

$$P(\text{one}|\ ) = \frac{P(\ |\text{one})\ P(\text{one})}{P(\ )}$$

$$P(\text{two}|\ ) = \frac{P(\ |\text{two})\ P(\text{two})}{P(\ )}$$

.
.
.

$$P(\text{zero}|\ ) = \frac{P(\ |\text{zero})\ P(\text{zero})}{P(\ )}$$

Digit Recognition

$$P(\text{one}|\ ) = \frac{P(\ |\text{one})\ P(\text{one})}{P(\ )}$$

$$P(\text{two}|\ ) = \frac{P(\ |\text{two})\ P(\text{two})}{P(\ )}$$

.
.
.

$$P(\text{zero}|\ ) = \frac{P(\ |\text{zero})\ P(\text{zero})}{P(\ )}$$

Evaluating Performance

Reference THIS IS AN EXAMPLE SENTENCE

Evaluating Performance

Reference	THIS	IS	AN	EXAMPLE	SENTENCE	
Hypothesis	THIS	IS		EXAMPLE	CENT	TENSE

Evaluating Performance

Reference	THIS	IS	AN	EXAMPLE	SENTENCE	
Hypothesis	THIS	IS		EXAMPLE	CENT	TENSE
Score			*Del.*		*Subs.*	*Insert.*

Evaluating Performance

Reference	THIS	IS	AN	EXAMPLE	SENTENCE	
Hypothesis	THIS	IS		EXAMPLE	CENT	TENSE
Score			*Del.*		*Subs.*	*Insert.*

$$WER = \frac{\#sub + \#ins + \#del}{\#ref} = \frac{1+1+1}{5} = 60\%$$

Evaluating Performance

Reference	THIS	IS	AN	EXAMPLE	SENTENCE	
Hypothesis	THIS	IS		EXAMPLE	CENT	TENSE
Score			*Del.*		*Subs.*	*Insert.*

$$WER = \frac{\#sub + \#ins + \#del}{\#ref} = \frac{1+1+1}{5} = 60\%$$

- **Some Examples (lower is better)**
  - Youtube: ~50%
  - Automatic closed captions for news: ~12%
  - Siri/Google voice: ~5%

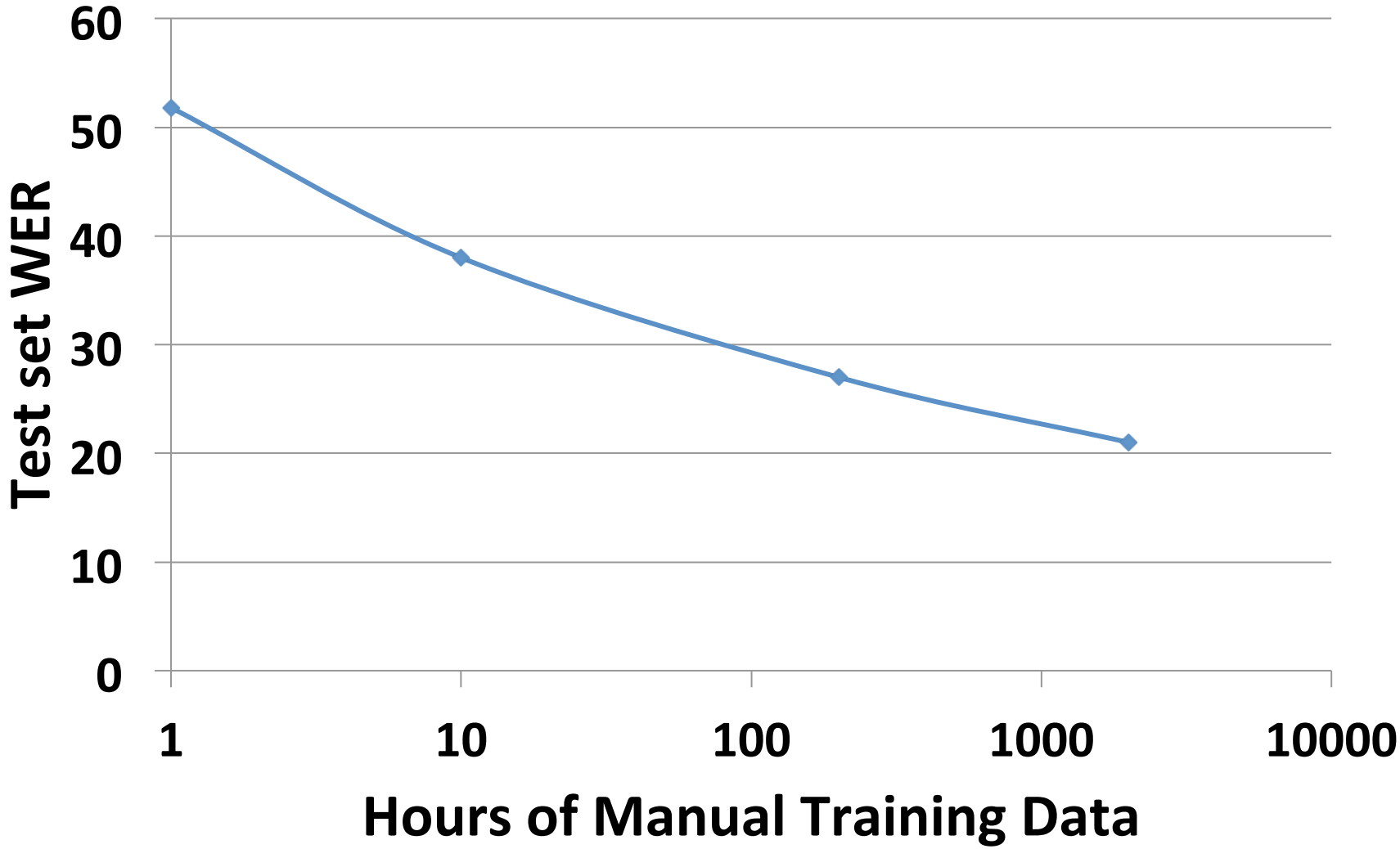
Probabilistic Modeling

$$\arg\max_{\mathbf{W}} \; P(\text{audio}\,|\mathbf{W})\; P(\mathbf{W})$$

Acoustic Model **Language Model**

- **Both models are <u>statistical</u>**
  - I'm going to completely skip over how they work
- **Need training data**
  - Audio of people saying "one three zero four"
  - Matching transcript "one three zero four"

Why do we need data?



Motivation

- **Speech recognition models hunger for data**
  - ASR requires thousands of hours of transcribed audio
  - In-domain data needed to overcome mismatches like language, speaking style, acoustic channel, noise, etc…
- **Conversational telephone speech transcription is difficult**
  - Spontaneous speech between intimates
  - Rapid speech, phonetic reductions and varied speaking style
  - Expensive and time consuming
    - $150 / hour of transcription
    - 50 hours of effort / hour of transcription
- **Deploying to new domains is slow and expensive**

Evaluating Mechanical Turk

- **Prior work judged quality by comparing Turkers to experts**
  - 10 Turkers match expert for many NLP tasks (*Snow et al 2008*)
- **Other Mechanical Turk speech transcription had low WER**
  - Robot Instructions ~3% WER *(Marge 2010)*
  - Street addresses, travel dialogue ~6% WER *(McGraw 2010)*
- **Right metric depends on the data consumer**
  - Humans: *WER on **transcribed** data*
  - Systems: *WER on **test** data decoded with a trained system*

English Speech Corpus

- **English Switchboard corpus**
  - Ten minute conversations about an assigned topic
  - Two existing transcriptions for a twenty hour subset:
    - LDC – high quality, ~50xRT transcription time
    - Fisher 'QuickTrans' effort – 6xRT transcription time

- **Callfriend language-identification corpora**
  - Korean, Hindi,Tamil, Farsi, and Vietnamese
  - Conversations from U.S. to home country between friends
  - Mixture of English and native language
  - Only Korean has existing LDC transcriptions

Transcription Task

Transcribe English phone conversations

-
-
-
-
-
-
-

ally no

rd

Thanks so much for your help!

Pay:

OH WELL I GUESS RETIREMENT THAT KIND OF THING WHICH I DON'T WORRY MUCH ABOUT

UH AND WE HAVE A SOCCER TEAM THAT COMES AND GOES WE DON'T EVEN HAVE THAT PRETTY

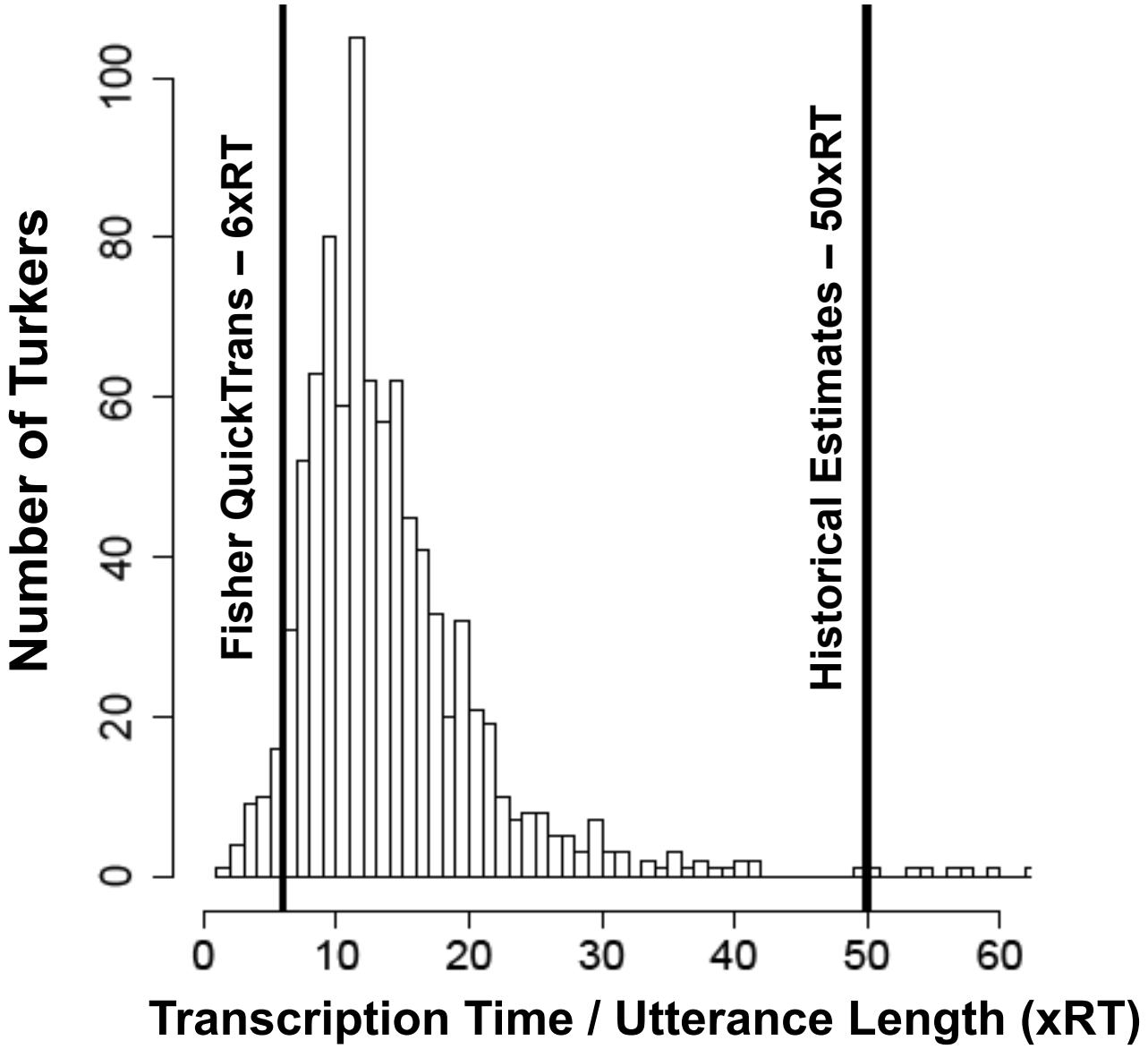
Speech Transcription for $5/hour

- **Paid $300 to transcribe 20 hours of Switchboard three times**
  - $5 per hour of transcription ($0.05 per utterance)
  - 1089 Turkers completed the task in six days
  - 30 utterances transcribed on average (earning 15 cents)
  - 63 Turkers completed more than 100 utterances

- **Some people complained about the cost**
  - *"wow that's a lot of dialogue for $.05"*
  - *"this stuff is really hard. pay per hit should be higher"*

- **Many enjoyed the task and found it interesting**
  - *"Very interesting exercise. would welcome more hits."*
  - *"You don't grow pickles they are cucumbers!!!!"*

Turker Transcription Rate



Dealing with Real World Data

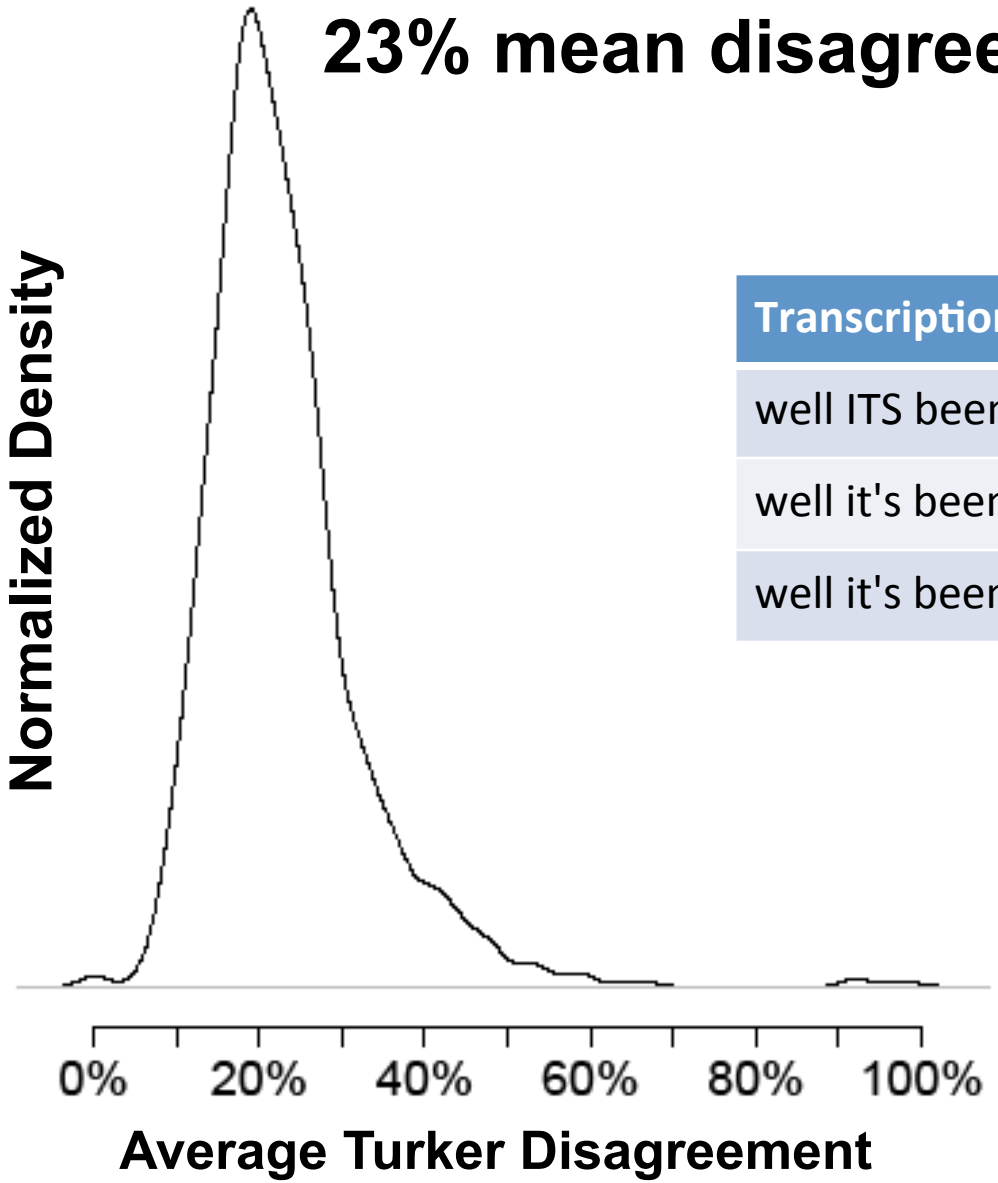
- **Every word in the transcripts needs a pronunciation**
  - Misspellings, new proper name spellings, jeez vs. geez
  - Inconsistent hesitation markings, myriad of ‘uh-huh’ spellings
  - 26% of utterances contained OOVs (10% of the vocabulary)

- **Lots of elbow grease to prepare phonetic dictionary**

- **Turkers found creative ways not to follow instructions**
  - Comments like “hard to hear” or “did the best I could :)”
  - Enter transcriptions into wrong text box
  - But very few typed in gibberish

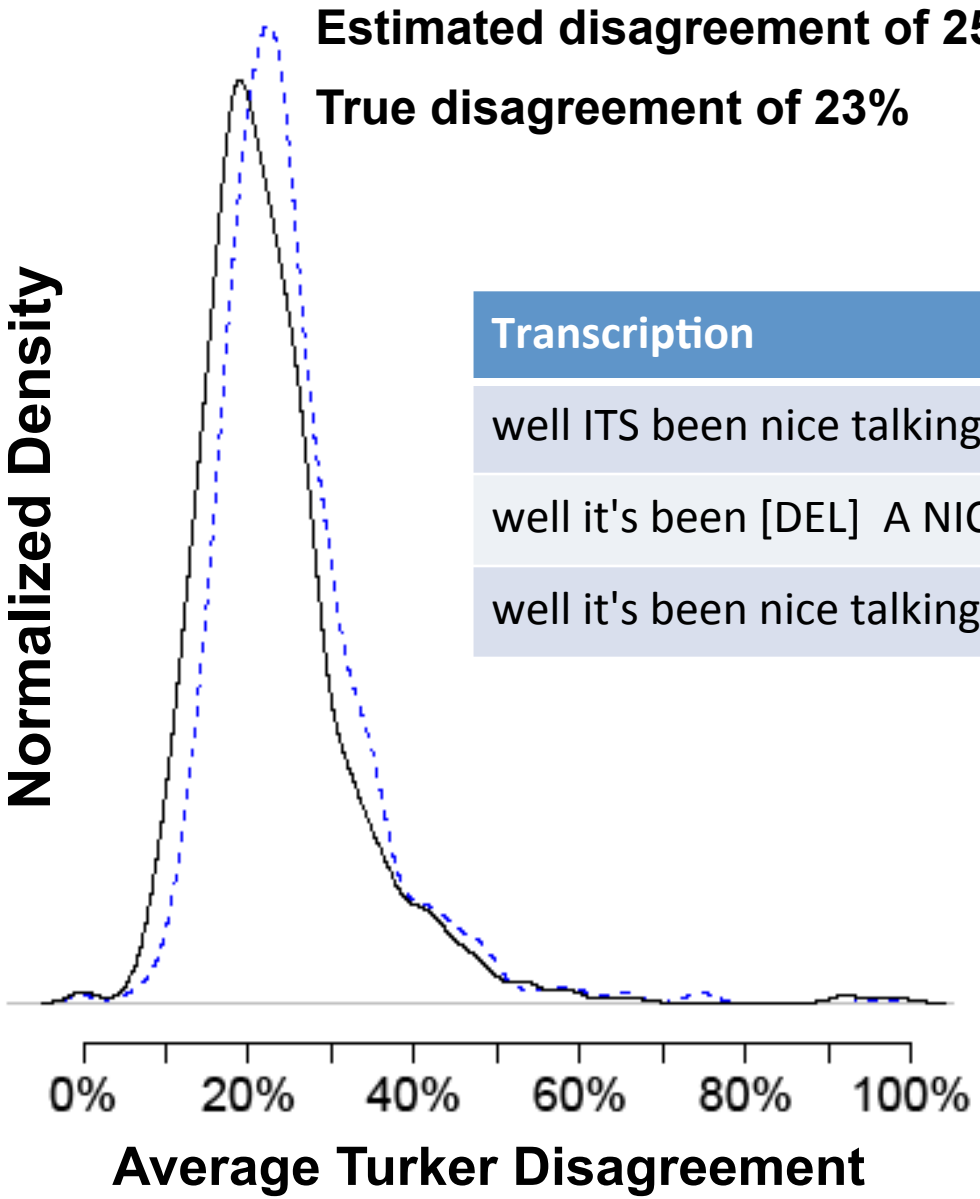
- **We did not explicitly filter comments, etc…**

Disagreement with Experts



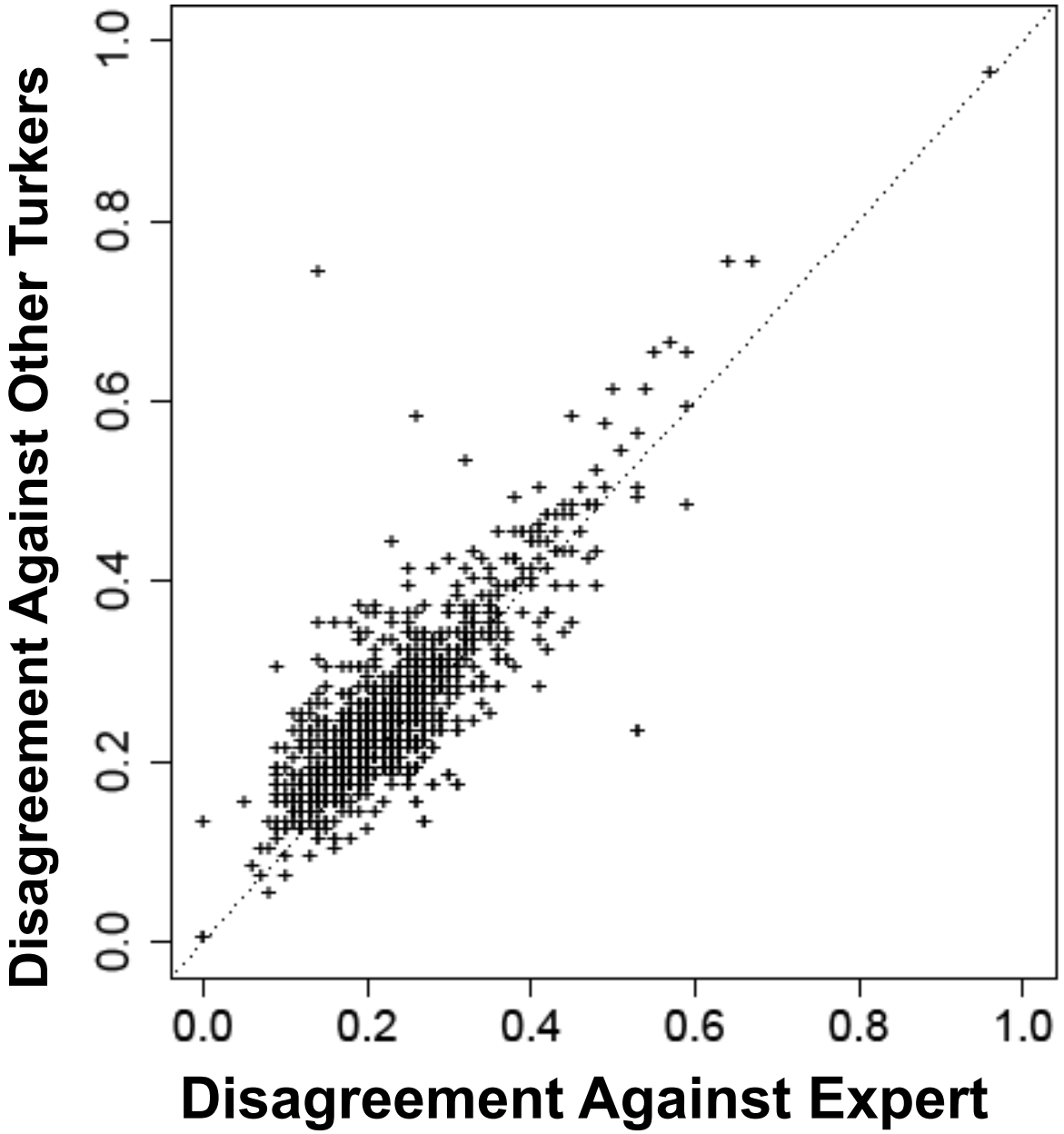
Transcription	WER
well ITS been nice talking to you again	**12%**
well it's been [DEL] A NICE PARTY JENGA	**71%**
well it's been nice talking to you again	**0%**

Estimation of Turker Skill

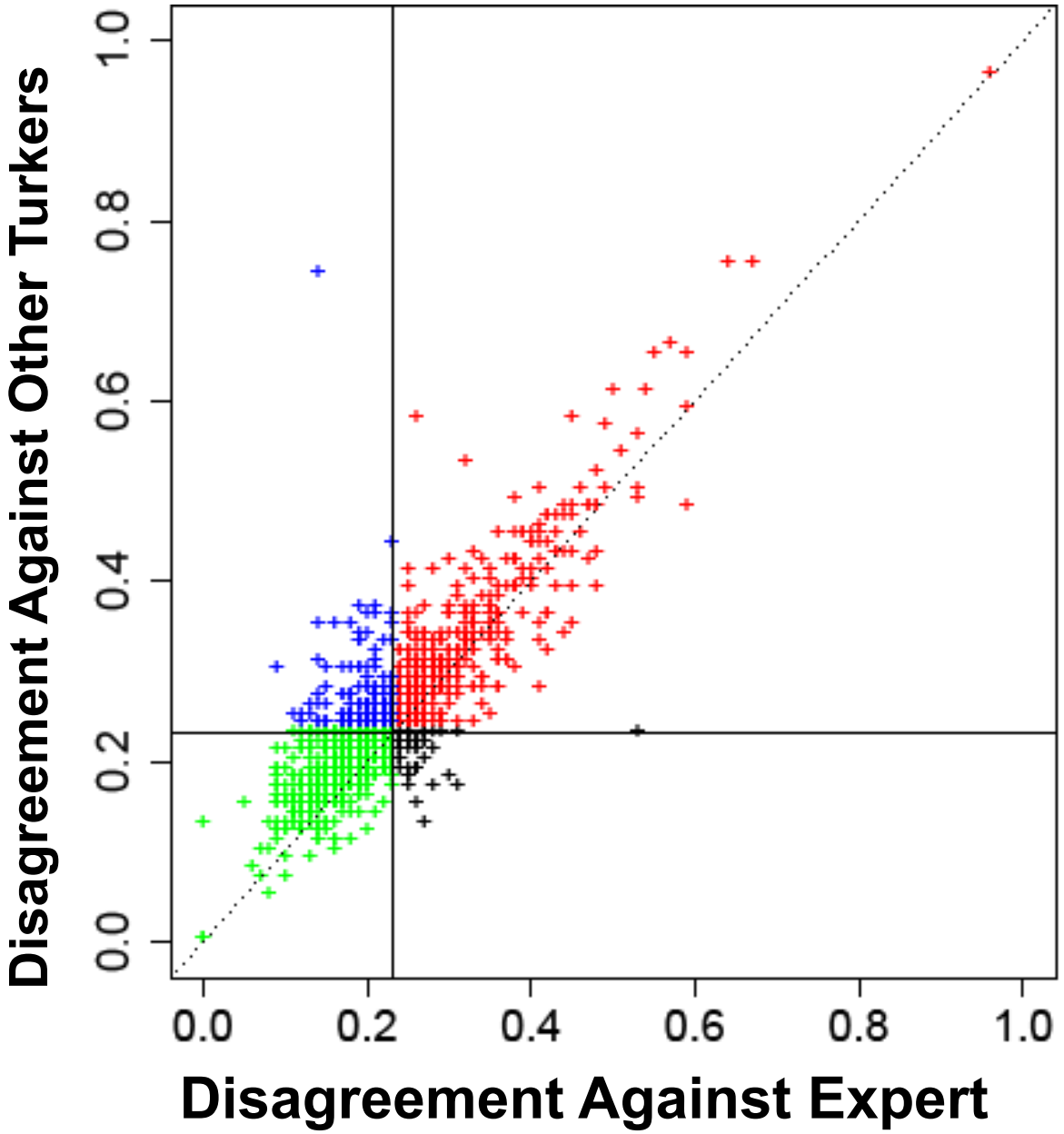


Transcription	WER	Est. WER
well ITS been nice talking to you again	**12%**	**43%**
well it's been [DEL] A NICE PARTY JENGA	**71%**	**78%**
well it's been nice talking to you again	**0%**	**37%**

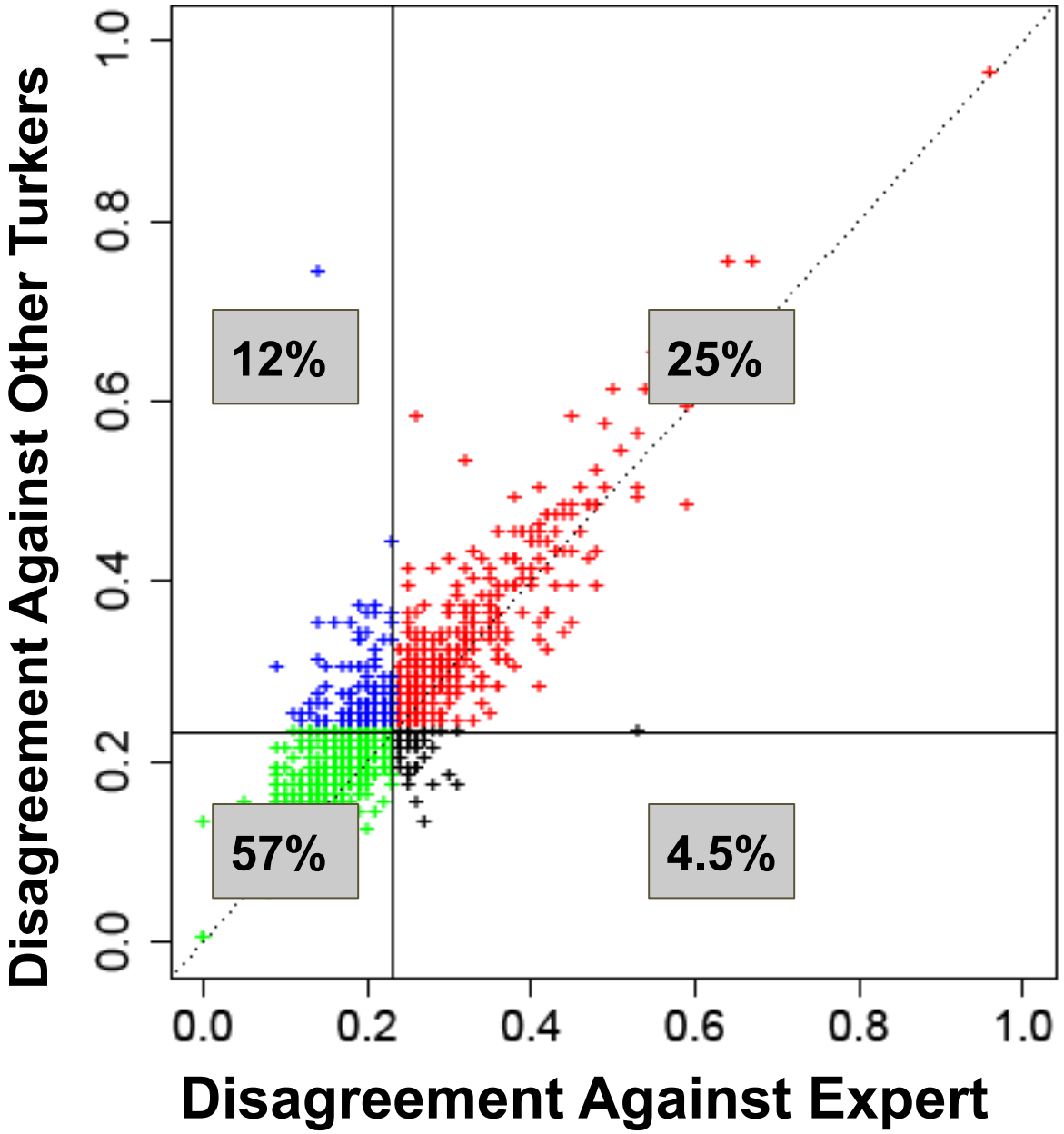
Rating Turkers: Expert vs. Non-Expert



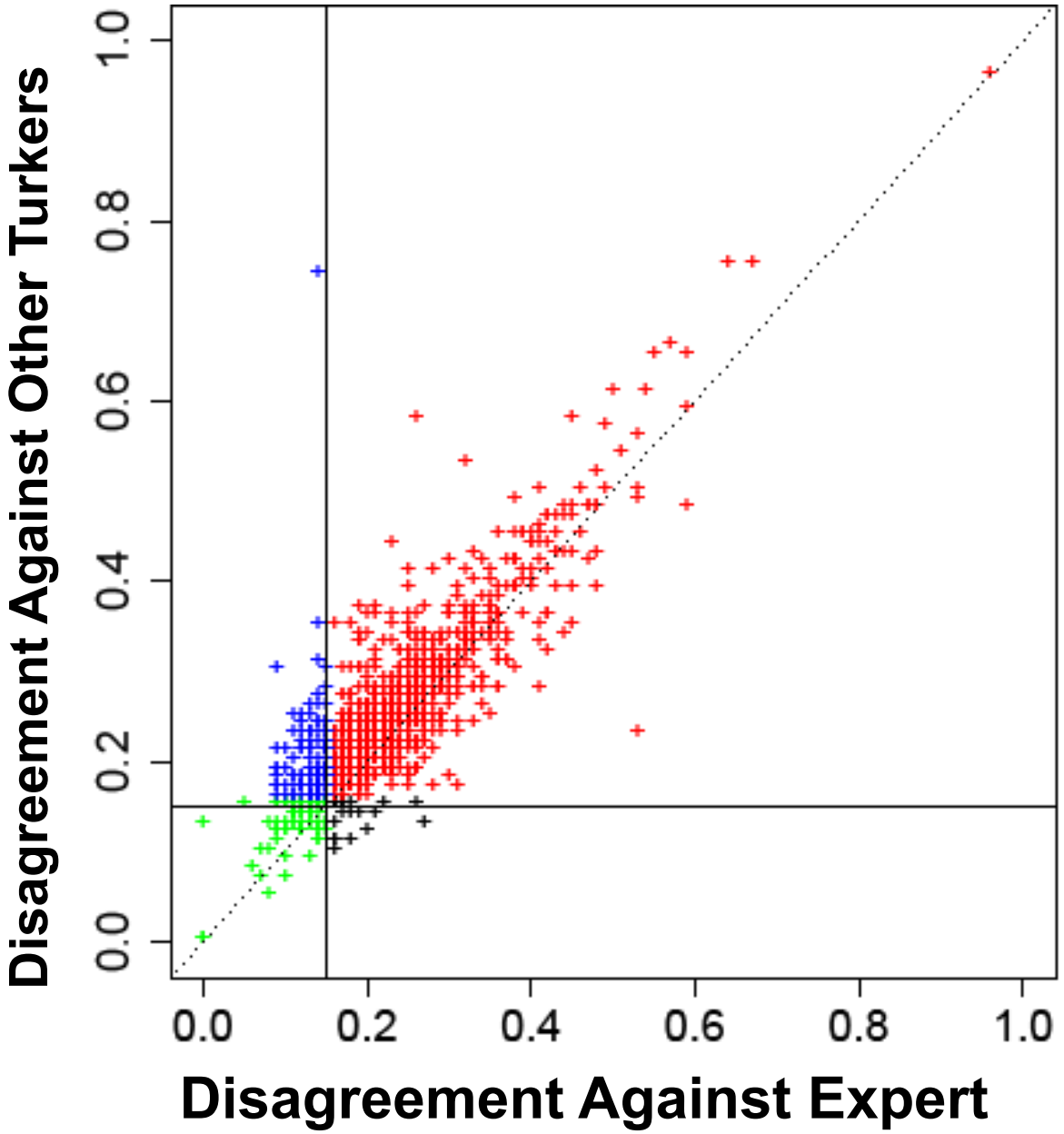
Selecting Turkers by Estimated Skill



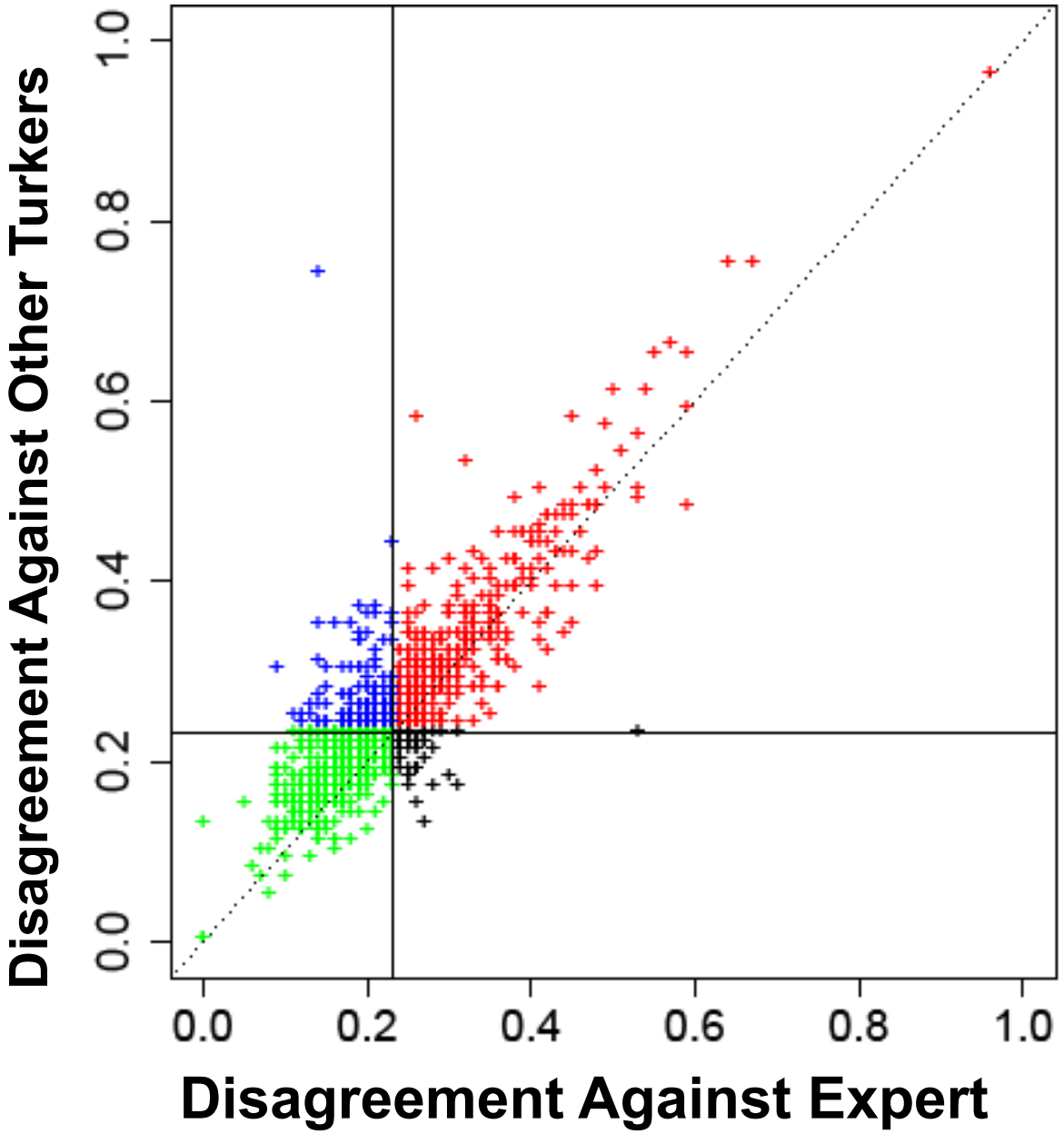
Selecting Turkers by Estimated Skill



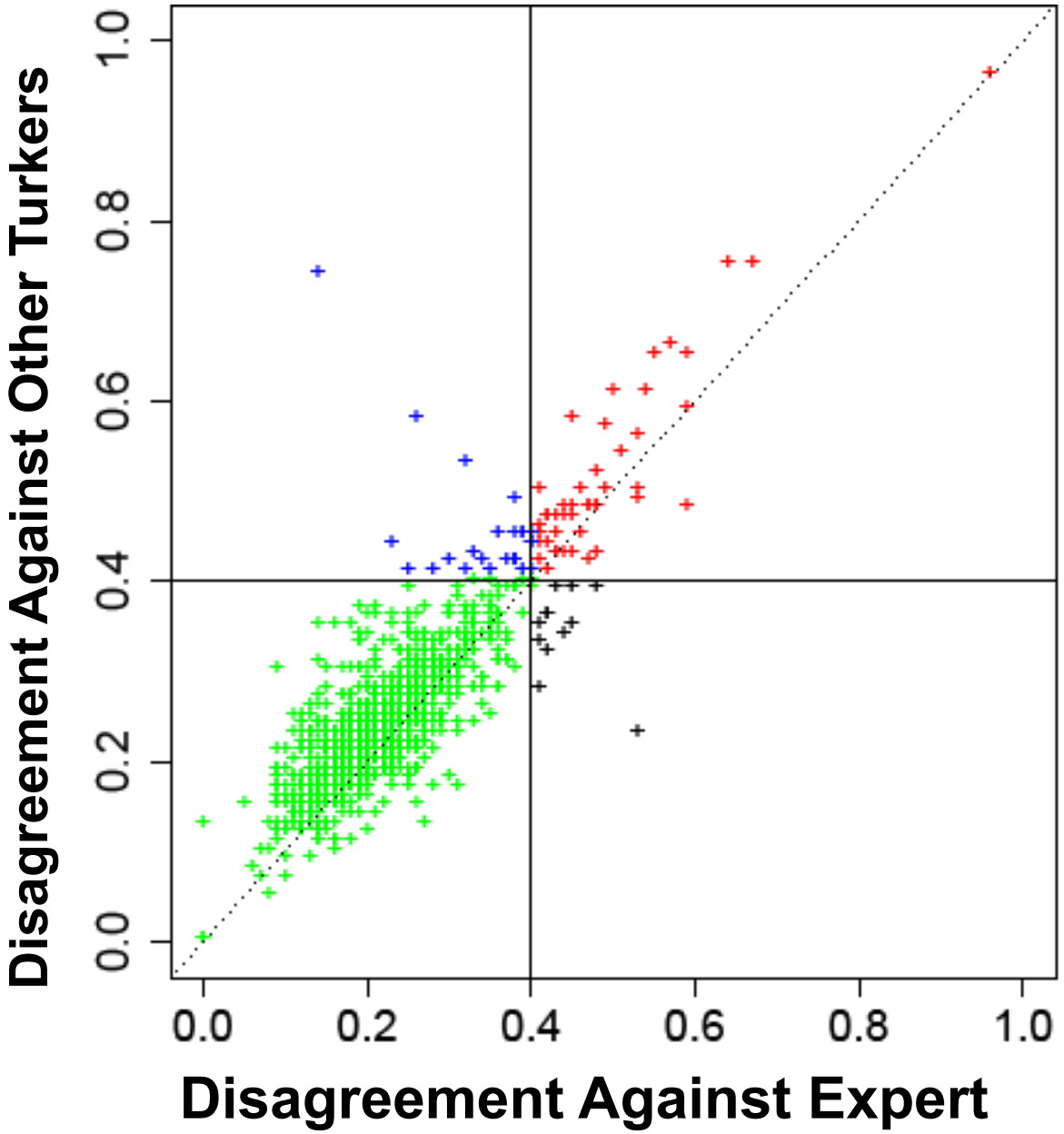
Selecting Turkers by Estimated Skill



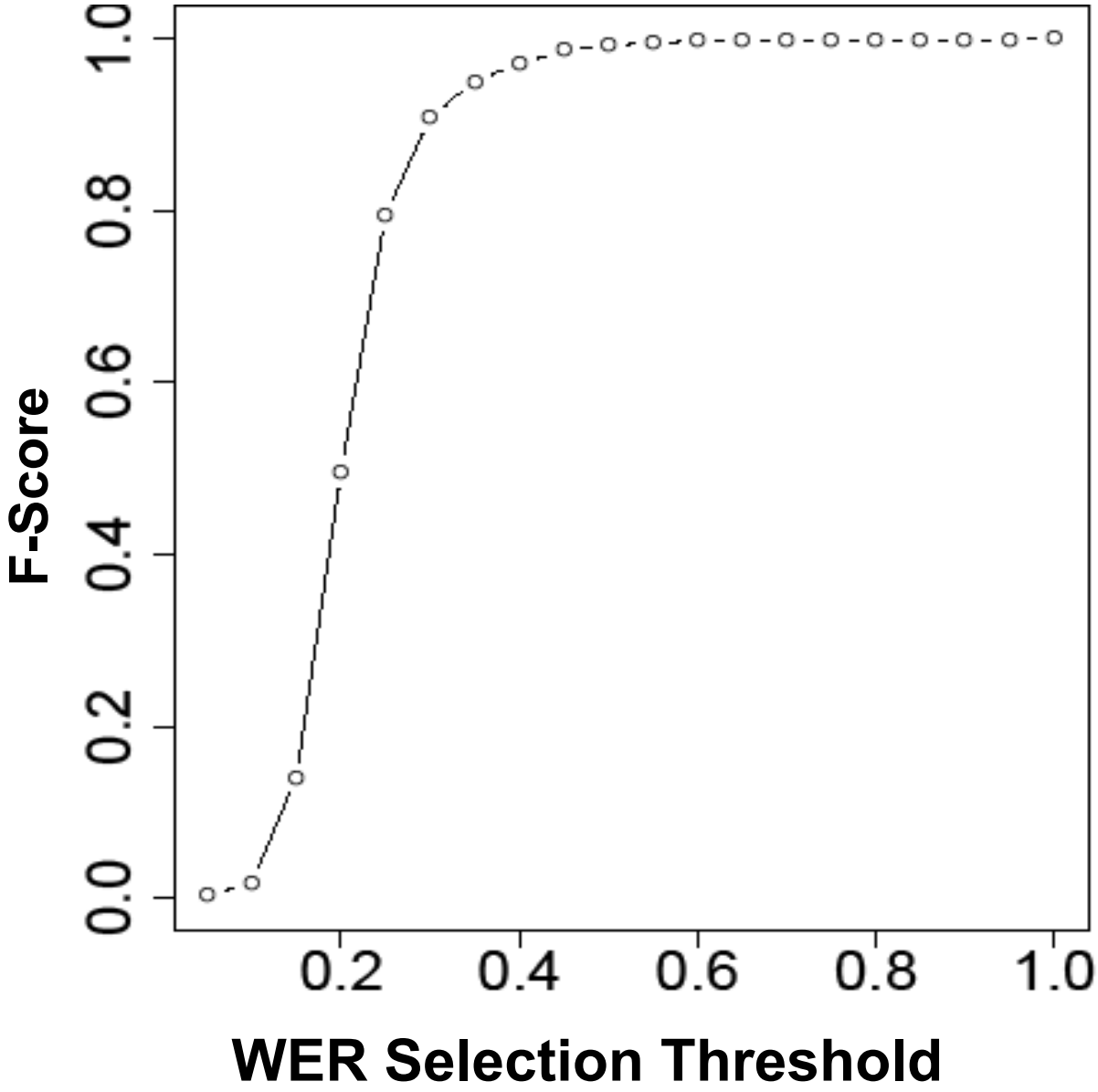
Selecting Turkers by Estimated Skill



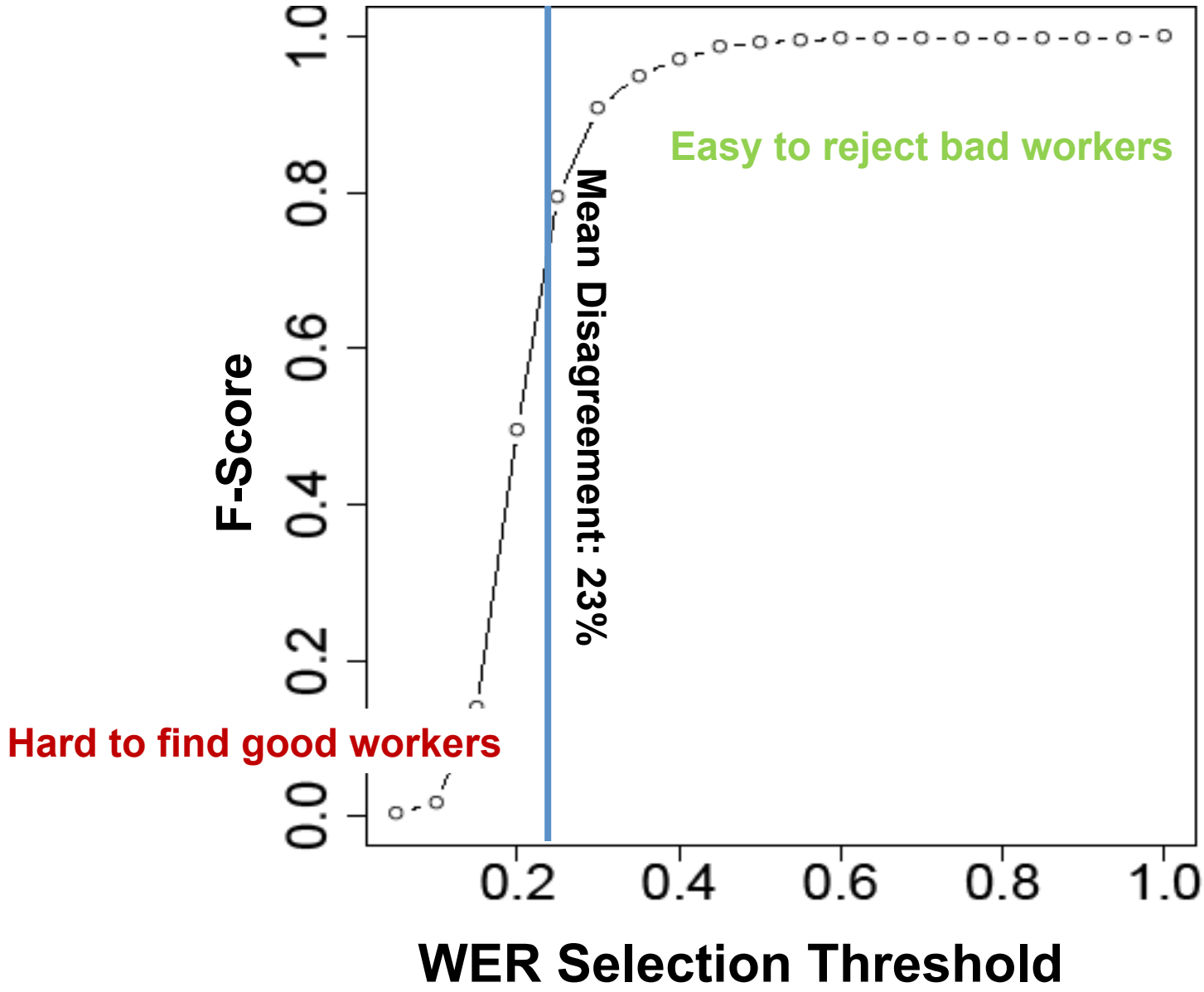
Selecting Turkers by Estimated Skill



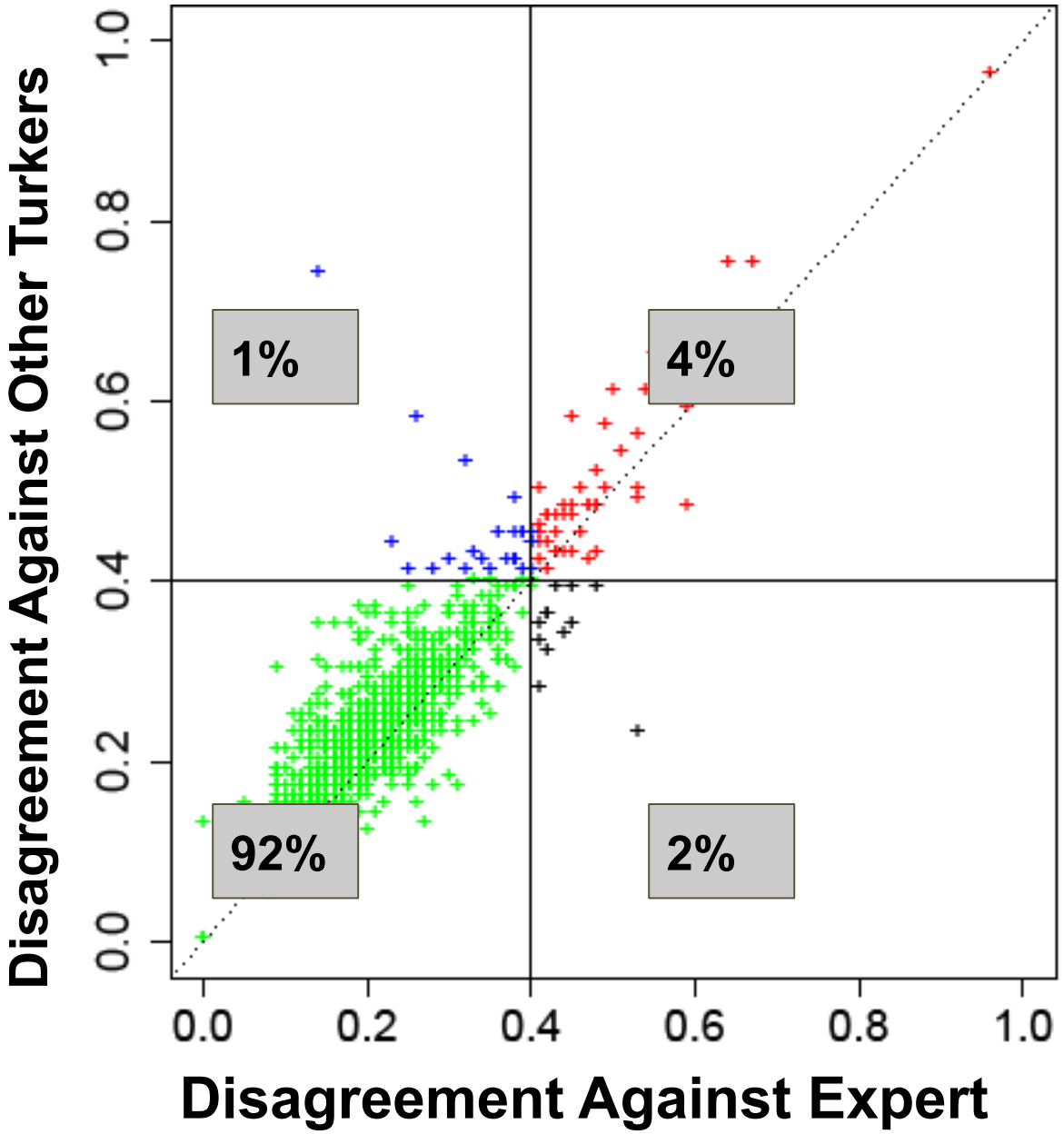
Finding the Right Turkers



Finding the Right Turkers



Selecting Turkers by Estimated Skill



Reducing Disagreement

Selection	LDC Disagreement
None	23%
System Combination	21%
Estimated Best Turker	20%
Oracle Best Turker	18%
Oracle Best Utterance	13%

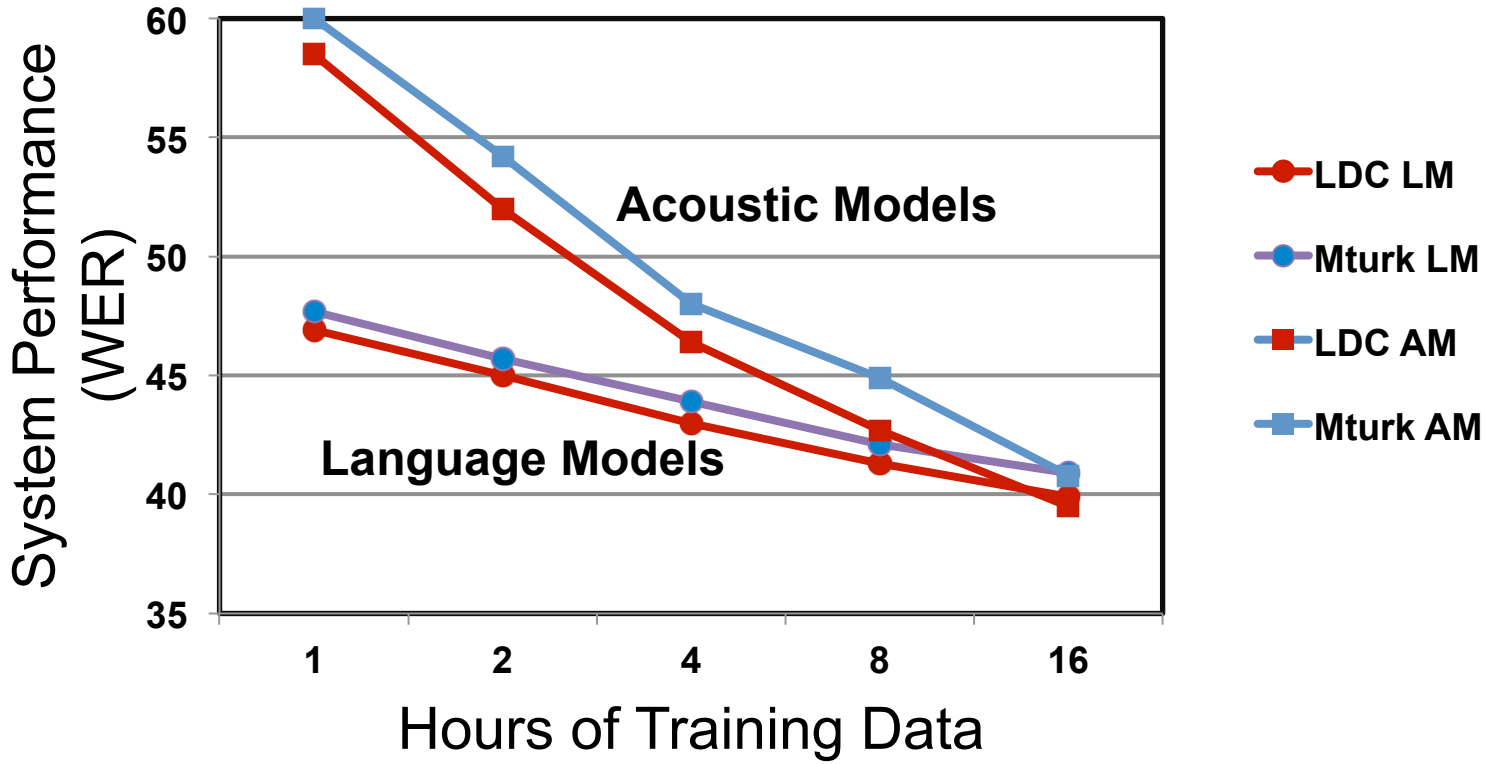
Mechanical Turk for ASR Training

- **Ultimate test is system performance**
  - Build acoustic and language models
  - Decode test set and compute WER
  - Compare to systems trained on equivalent expert transcription

- **23% professional disagreement might seem worrying**
  - How does it effect system performance?
  - Do reductions in disagreement transfer to system gains?
  - What are best practices for improving ASR performance?

Breaking Down The Degradation

- **Measured test WER degradation from 1 to 16 hours**
  - 3% relative degradation for acoustic model
  - 2% relative degradation for language model
  - 5% relative degradation for both
  - *Despite 23% transcription disagreement with LDC*



Value of Repeated Transcription

- **Each utterance was transcribed three times**

- **What is the value of this duplicate effort?**
  - Instead of dreaming up a better combination method, use oracle error rate as upper bound on system combination

Transcription	LDC Disagreement	ASR WER
Random	**23%**	**42.0%**
Oracle	**13%**	**40.9%**
LDC	**-**	**39.5%**

- **Cutting disagreement in half reduced degradation by half**

- **System combination has at most 2.5% WER to recover**

How to Best Spend Resources?

- **Given a fixed transcription budget, either:**
  - Transcribe as much audio as possible
  - Improve quality by redundantly transcribing

Transcription	Hours	Cost	ASR WER
Mturk	**20**	**$100**	**42.0%**
Oracle Mturk	**20**	**$300**	**40.9%**
MTurk	**60**	**$300**	**37.6%**
LDC	**20**		**39.5%**

- **Get more data, not better data**
  - Compare 37.6% WER versus 40.9% WER
- **Even expert data is outperformed by more lower quality data**
  - Compare 39.5% WER to 37.6% WER

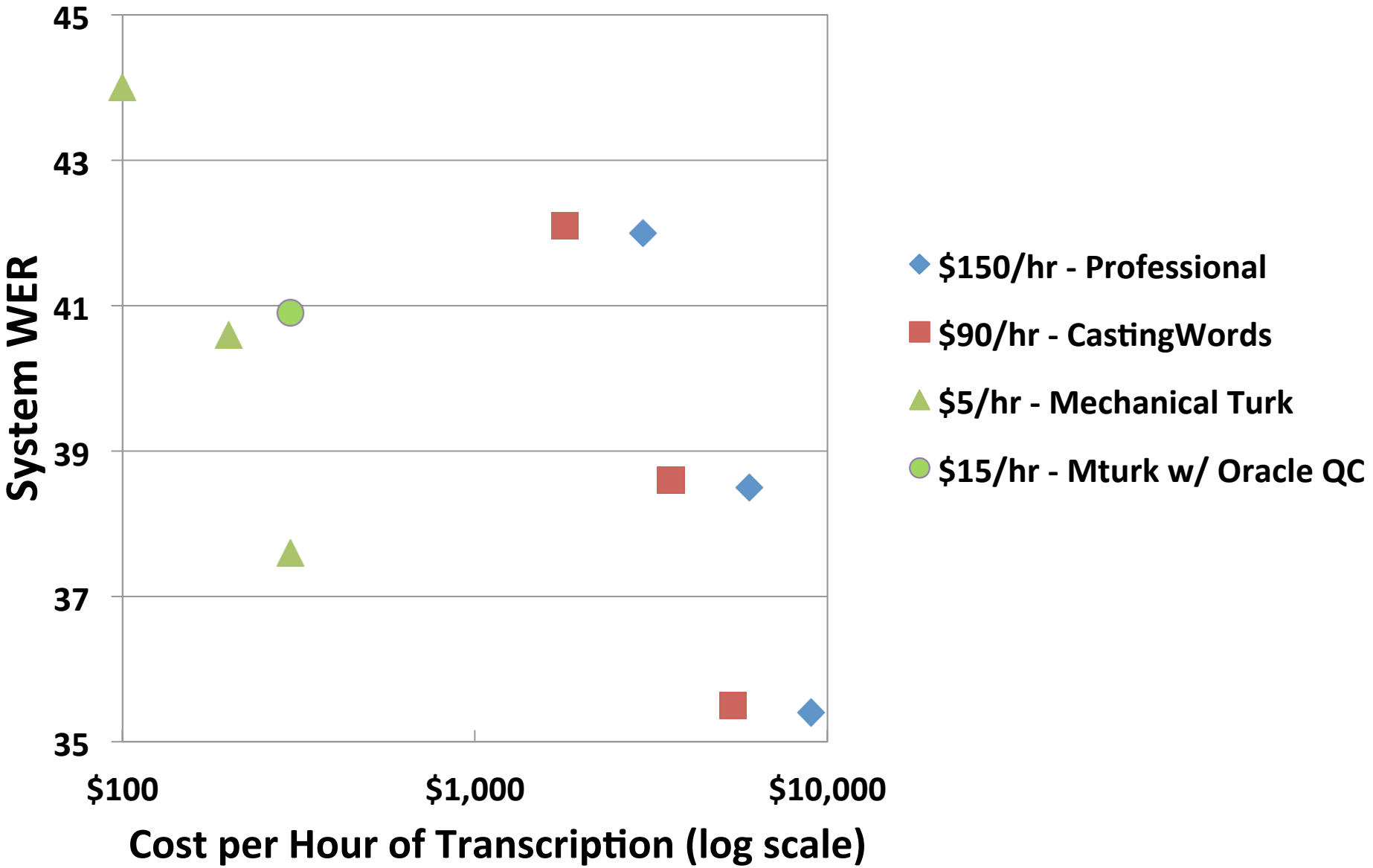
How to Best Spend Resources?

- **Given a fixed transcription budget, either:**
  - Transcribe as much audio as possible
  - Improve quality by redundantly transcribing

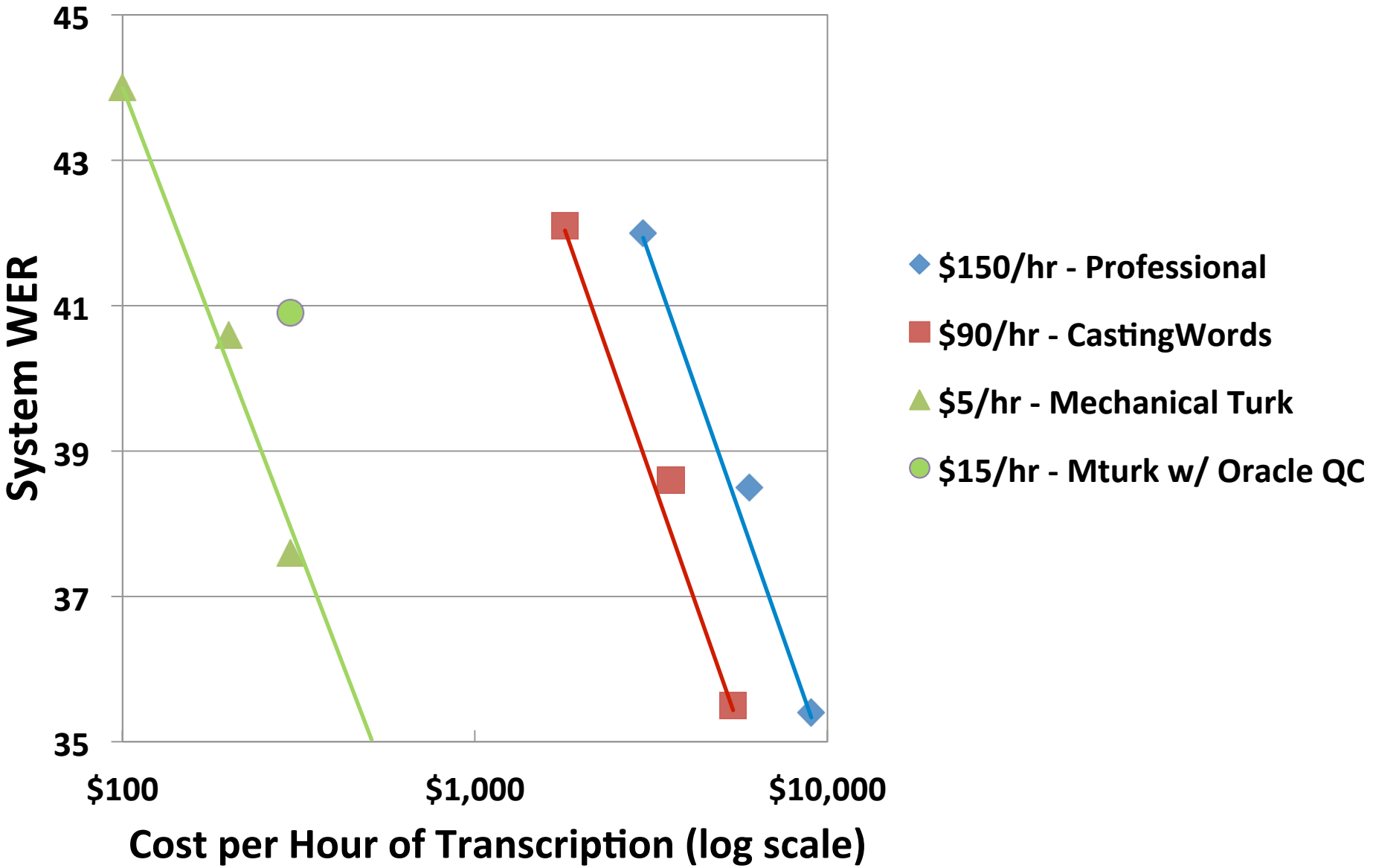
Transcription	Hours	Cost	ASR WER
Mturk	**20**	**$100**	**42.0%**
Oracle Mturk	**20**	**$300**	**40.9%**
MTurk	**60**	**$300**	**37.6%**
LDC	**20**	**~$3000**	**39.5%**

- **Get more data, not better data**
  - Compare 37.6% WER versus 40.9% WER
- **Even expert data is outperformed by more lower quality data**
  - Compare 39.5% WER to 37.6% WER

Comparing Cost of Reducing WER



Comparing Cost of Reducing WER



Korean

- **Tiny labor pool *(initially two Turkers versus 1089 for English)***

- **Posted separate 'Pyramid Scheme' HIT**
  - Paid referrer 25% of what referred earns transcribing
  - Transcription costs $25/hour instead of $20/hour
  - 80% of transcriptions came from referrals

- **Transcribed three hours in five weeks**
  - Paid 8 Turkers $113 at a transcription rate of 10xRT

- **Despite 17% CER, test CER only goes down by 1.5% relative**
  - from 51.3% CER to 52.1% CER
  - Reinforces English conclusions about the usefulness of noisy data for training an ASR system

Tamil and Hindi

- **Collected one hour of transcripts**
  - Much larger labor pool – how many?
  - Paid $20/hour, finished in 8 days
  - Difficult to accurately convey instructions
    - Many *translated* Hindi audio to English

- **No clear conclusions**
  - A private contractor provided transcriptions
  - Very high disagreement (80%+) for both languages
    - Reference transcripts inaccurate
    - Colloquial speech, poor audio quality
    - English speech irregularly transliterated into Devanagari
    - Lax gender agreement both for speaking *and* transcribing
  - Hindi ASR might be a hard task

English Conclusions

- **Mechanical Turk can quickly and cheaply transcribe difficult audio like English CTS**
  - 10 hours a day for $5 / hour
- **Can reasonably predict Turker skill w/out gold standard data**
  - But this turns out not to be as important as we thought
  - Oracle selection still only cuts disagreement in half
- **Trained models show little degradation despite 23% professional disagreement**
  - Even perfect expert agreement has small impact on system performance (2.5% reduction in WER)
  - Resources better spent getting *more* data than *better* data

Foreign Language Conclusions

- **Non-English Turkers are on Mechanical Turk**
  - But not a field of dreams
    - "If you post it, ~~they will come~~"
- **Korean results reinforce English conclusions**
  - 0.8% system degradation despite 17% disagreement
  - $20/hour (still very cheap)
- **Small amounts of errorful data is useful**
  - Poor models can still produce useable systems
    - 90% topic classification accuracy possible despite 80%+ WER
  - Semi-supervised methods can bootstrap initial models
    - 51% WER reduced to 27% with a one hour acoustic model
- **Noisy data is much more useful than you think**

Swahili and Amharic (Gelas, 2011)

- **Two under-resourced African languages**
  - 17M speak Amharic in Ethiopia
  - 50M speak Swahili in East Africa (Kenya, Congo, etc…)
- **Not many workers on Mturk**
  - 12 Amharic, 3 Swahili
- **And they generated data very slowly**
  - 0.75hrs after 73 days, 1.5hrs after 12 days
- **But despite being worse than professionals**
  - 16% WER, 27.7% WER
- **ASR systems performed as well as professionals**

- **At the end of the day, researchers paid grad students at $103/hr of transcription to get 12 hours vs. $37/hr on MTurk**

Other Speech Tasks

- **Use MTurk to elicit speech for the target domain**
  - Data collected on microphone, so point them to an app instead
- **Use Turkers to perform verification and correction**
  - Listen to <audio, transcript> pairs and verify right or wrong
  - Correct automatic speech output
- **Speech Science**
  - How sensitive are humans to noise?
  - Can they detect accent, fluency, etc…
- **System Evaluation**
  - Synthesized Speech (but again non-English was tough)
  - Spoken Dialog Systems *a.k.a.* Siri

If You're Curious

- **Praat - http://www.fon.hum.uva.nl/praat/**
  - Speech analysis
- **Kaldi - Open Source State of the Art Recognizer**
  - http://kaldi.sourceforge.net/
- **Linguistic Data Consortium**
  - Based right here at Upenn!
  - Creates almost all of the speech corpora used in research

BACKUP

Cheaply Estimating Turker Skill