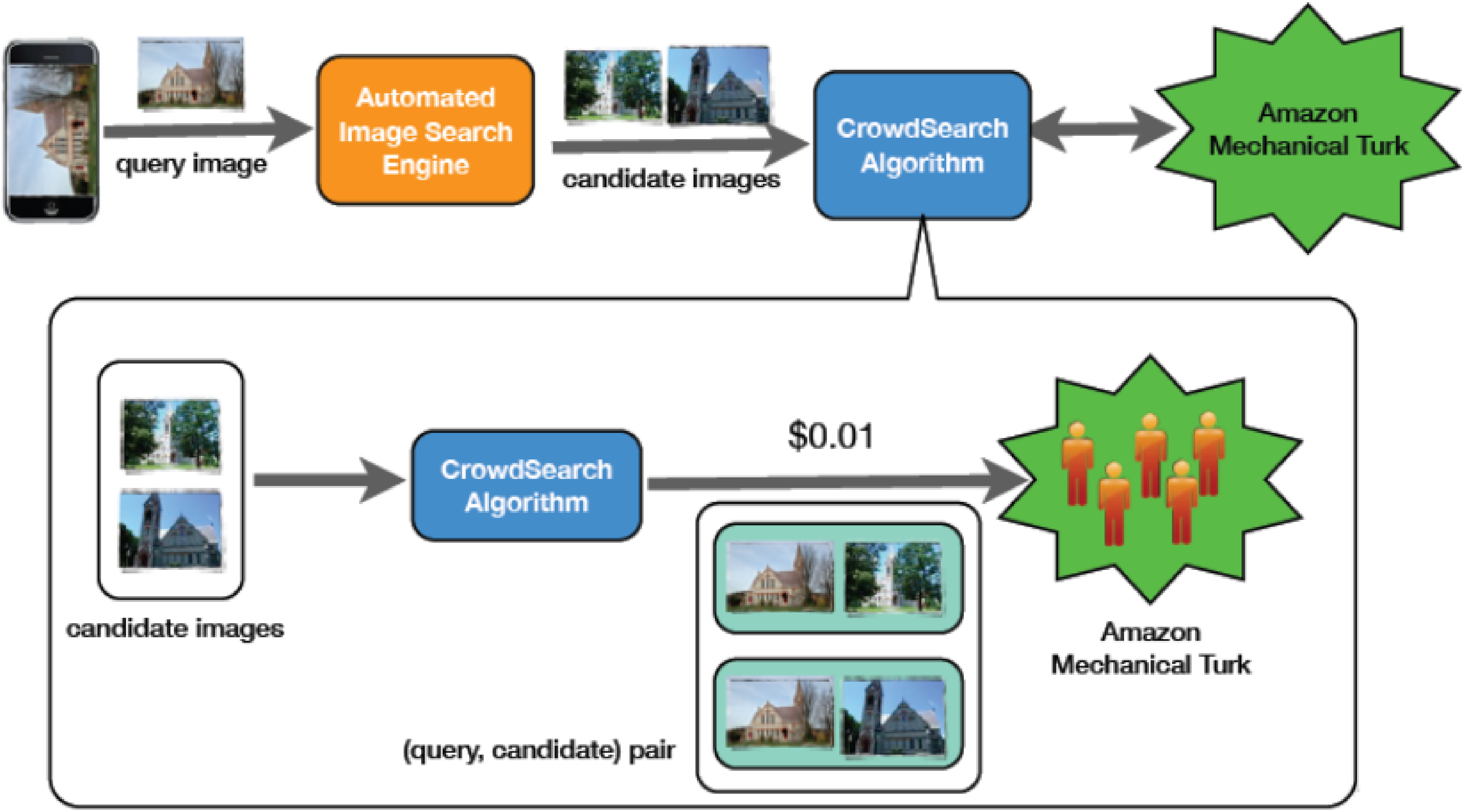
Estimating Consensus from Crowdsourced Annotations

SFIT R & D Lecture Series
Nov 2022

Dr. Santosh Chapaneri

Crowdsourcing



Tingxin Yan, Vikas Kumar, Deepak Ganesan, "*CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones*", MobiSys 2010:77-90

Crowdsourcing

Does this image depict a woman?

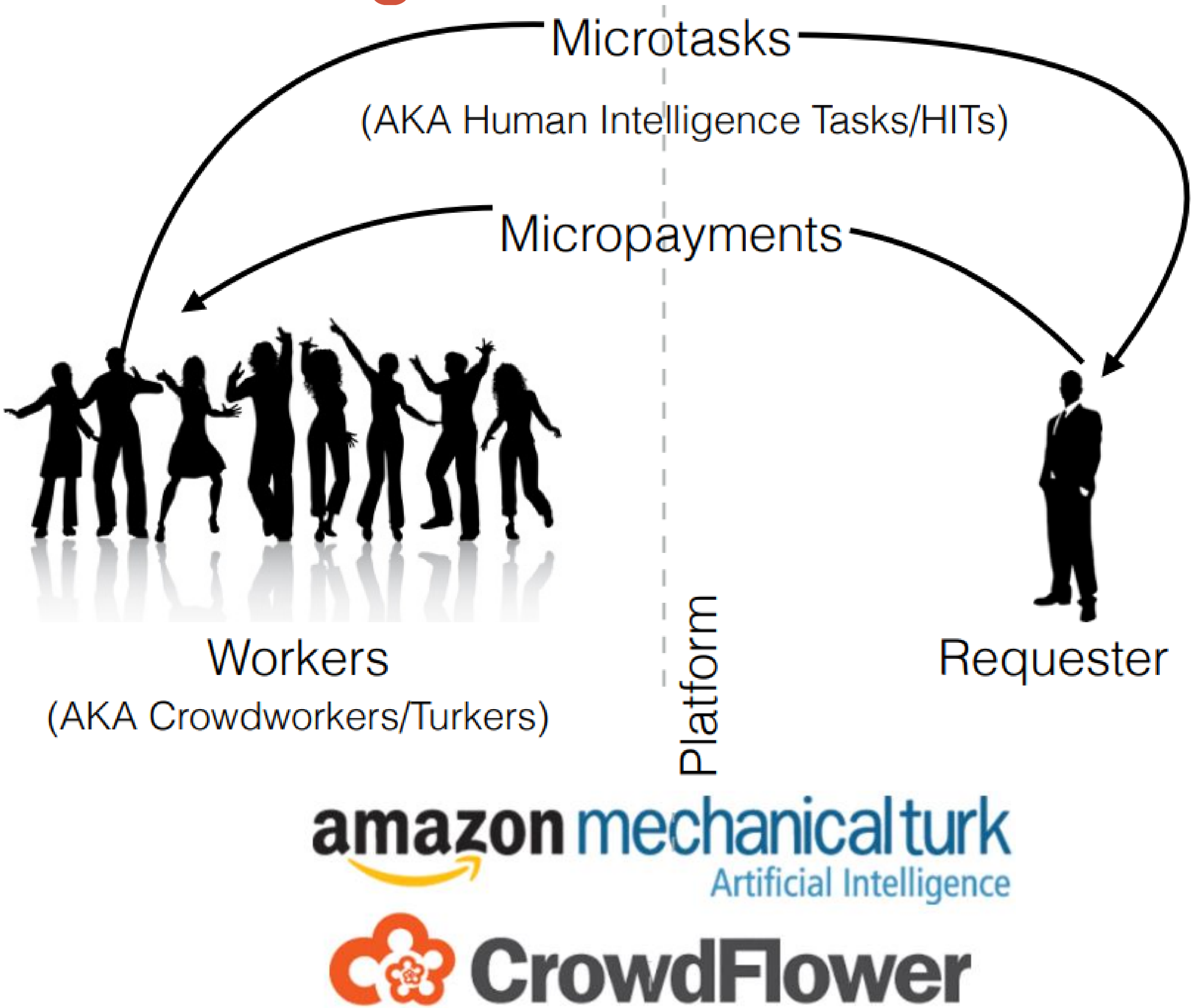
Crowdsourcing

Does this image depict a woman?

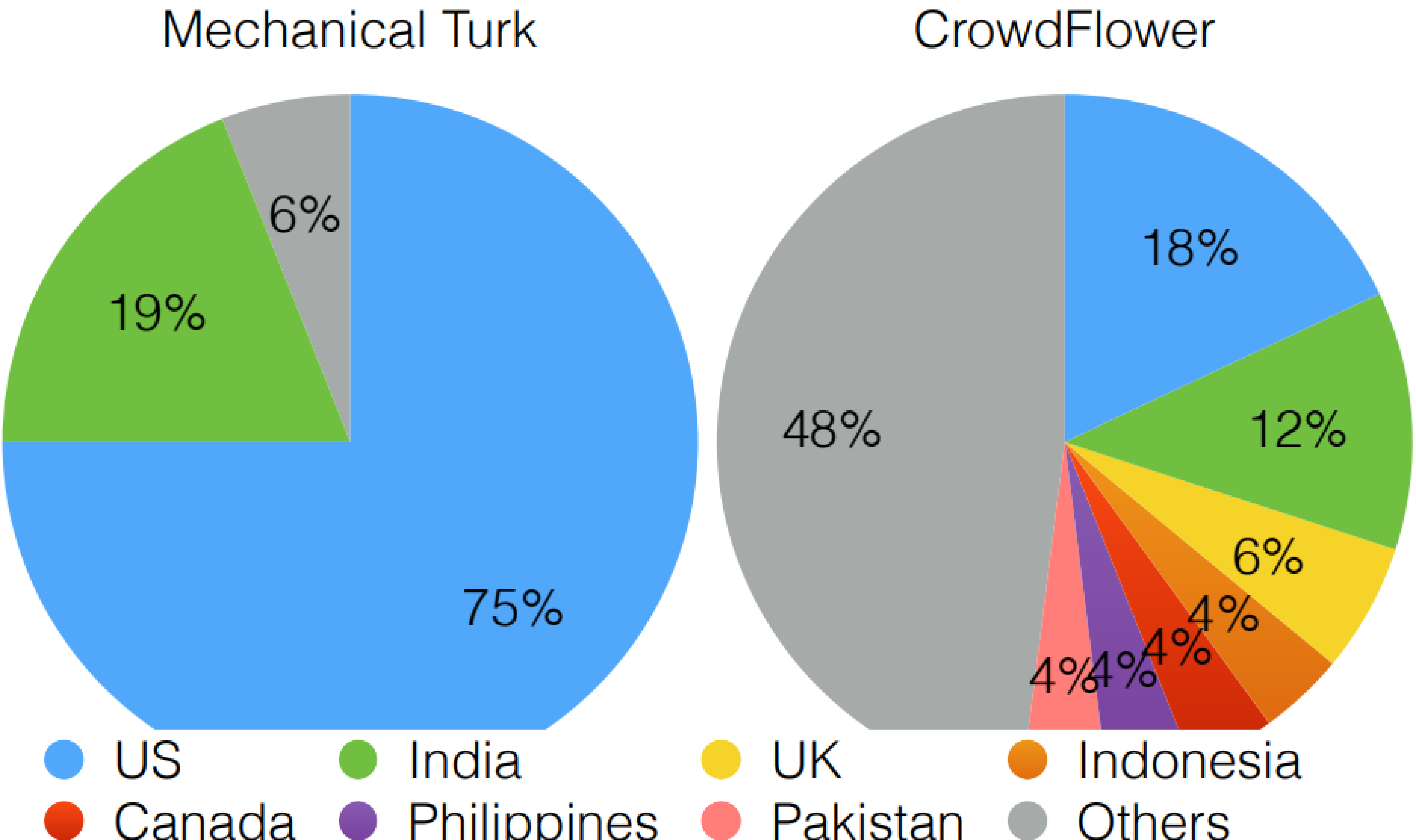
Crowdsourcing

Does this image depict a woman?

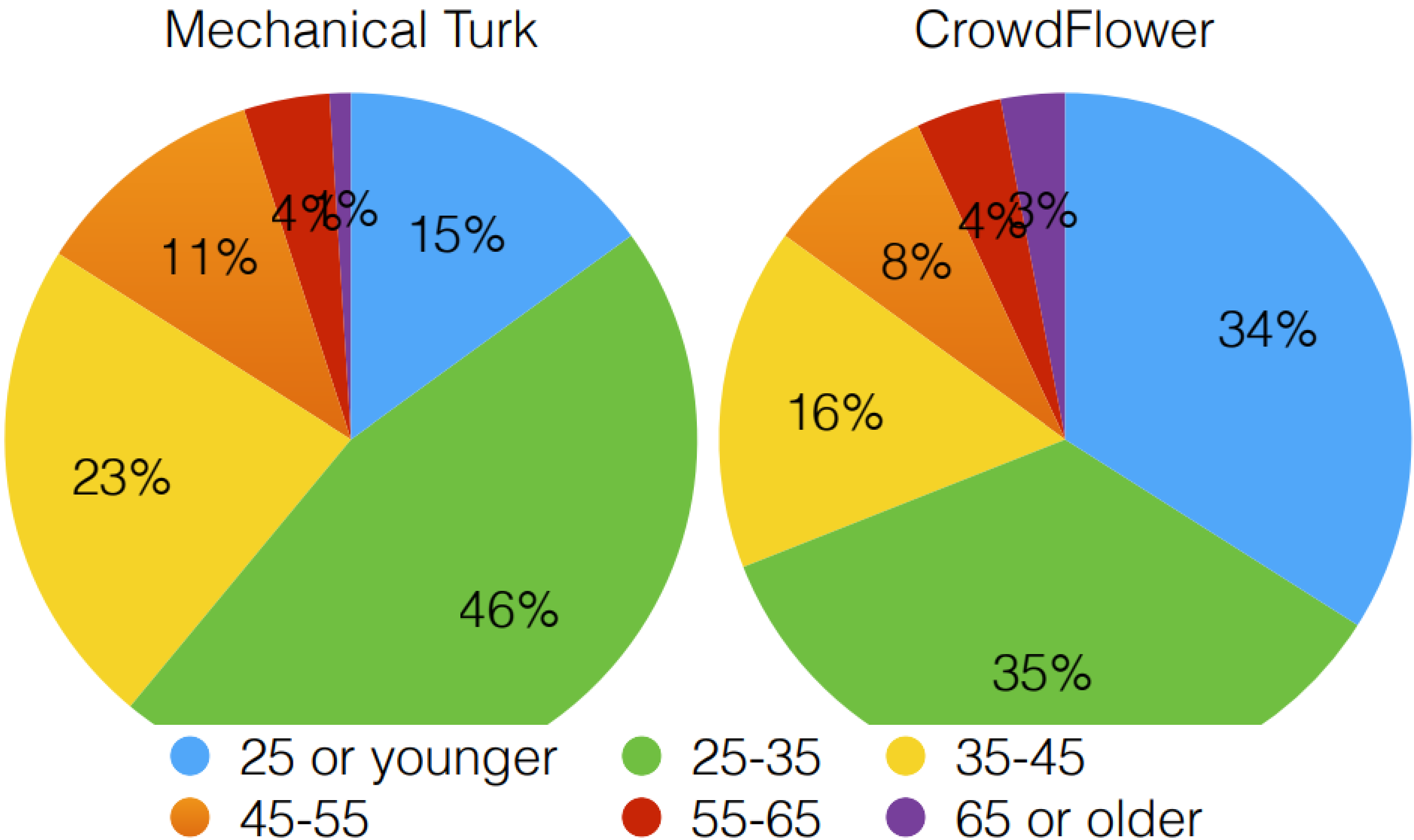
Crowdsourcing



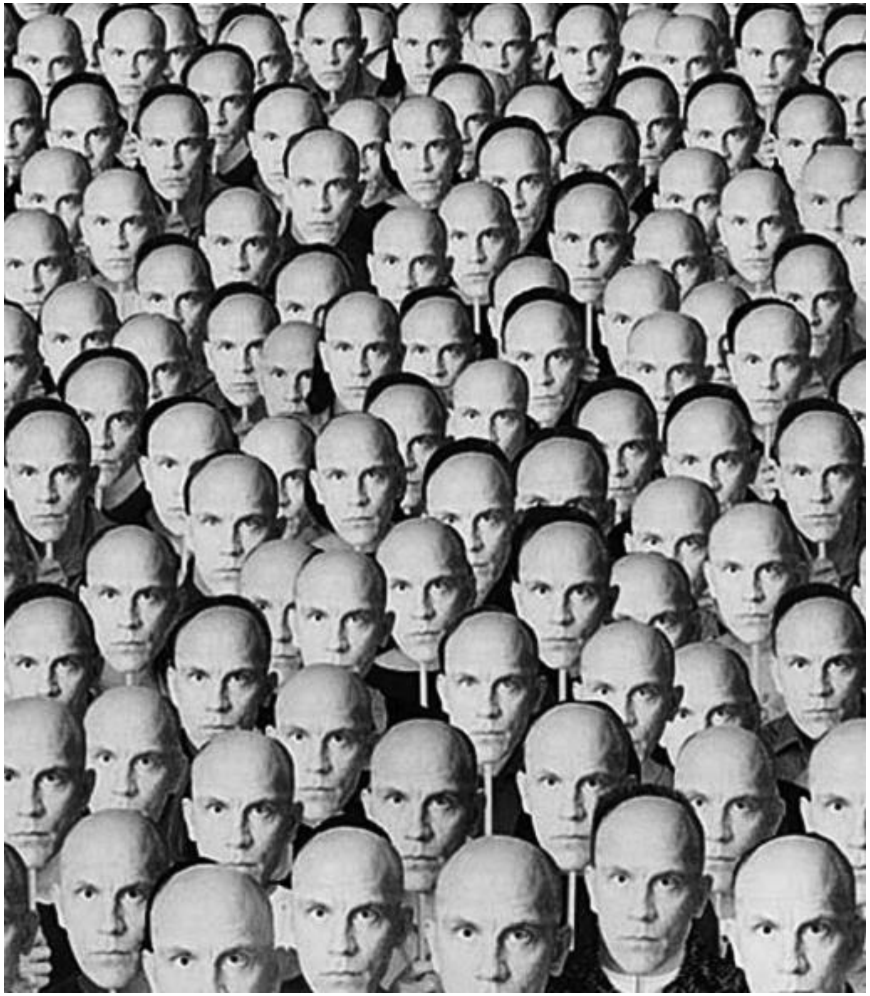
Crowdsourcing



Crowdsourcing



Crowdsourcing Issues



Are all annotators equally reliable?

When people disagree, they don't understand the problem

Crowdsourcing Issues

Domain experts – very few

One expert is enough!
How to gauge expertise?

The Problem

- Labeled datasets are expensive and laborious to produce
- Crowdsourcing => wisdom of crowds
  - Amazon Mechanical Turk
  - CrowdFlower
- Sparsely annotated data
- Long-tail phenomena
- Not all annotators equally reliable

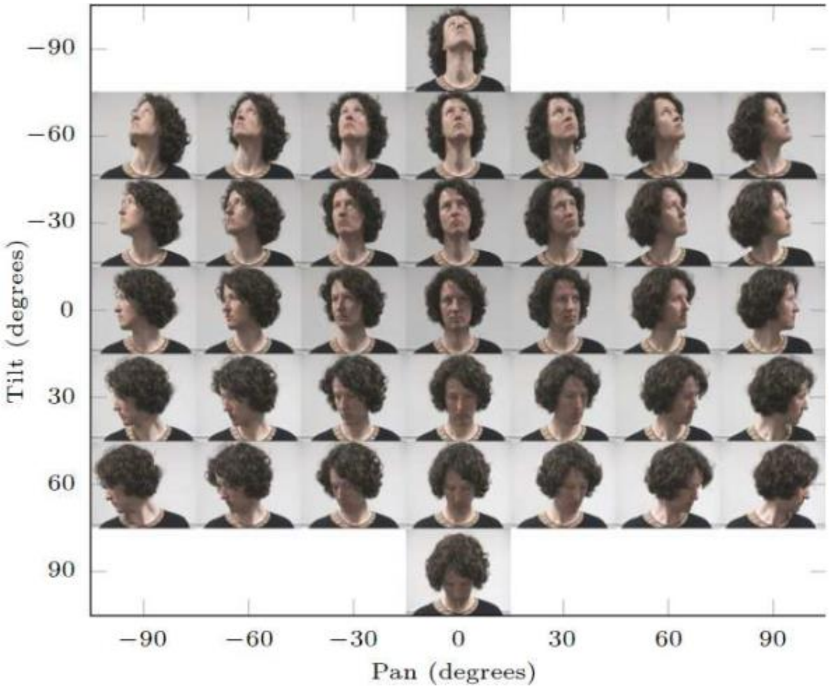
Entity (city)	Value (pop.)	Source
NYC	8,346,794	Freebase
NYC	8,244,910	Wikipedia
NYC	8,175,133	US Census
NYC	7,864,215	BadSource
Urbana	36,395	US Census
Urbana	36,395	Wikipedia
Urbana	34,774	Freebase
Urbana	1,215	BadSource
…	…	…

- **Goals**: Obtain estimated consensus

  Compute annotator’s reliability

(Dis-)Agreement

Inter-annotator agreement: $J = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{\sum_{j \in \mathcal{A}_i} (y_i^j - y_{i_m})^2}{|\mathcal{A}_i|}}$



HeadPose dataset
$J = 0.6807$

Age dataset
$J = 0.4639$

==> Not all annotators agree on the responses of annotated samples

What others have done…

- Inconsistent responses due to various personal and situational aspects such as personality, context, cultural background
- Existing work ignores annotator errors (e.g. low-attention) and outliers (e.g. adversarial behavior)
- Learn annotator behaviours for crowdsourced labeling tasks in [1]
- Review of judgement analysis techniques in [2] => problem difficulty, spammer identification, constrained judgement
- Uncertainty-aware modeling in [3] => estimate kernel density from multiple sources and learn trustworthy opinions
- Non-parametric Gaussian process in [4] => learn behavior of annotators

[1] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, L. Moy, ”Learning from crowds”, *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, Apr 2010.

[2] S. Chatterjee, A. Mukhopadhyay and M. Bhattacharyya, “A review of judgement analysis algorithms for crowdsourced opinions", *IEEE Transactions on Knowledge and Data Engineering*, Mar 2019.

[3] M. Wan, X. Chen, L. Kaplan, J. Han, J, Gao and B. Zhao, “From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach", *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 1885{1894, Aug 2016.

[4] H. Xiao, H. Xiao and C. Eckert, ”Learning from multiple observers with unknown expertise”, *Springer Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining* (PAKDD), vol. 7818, pp. 595–606, Apr 2013.

Problem setup

- Dataset $\mathcal{D} = \{\mathbf{y}_i^1, \mathbf{y}_i^2, \ldots, \mathbf{y}_i^R\}_{i=1}^N$; *N* data samples, *R* annotators
- **Adversariness** a^j :
  - if annotator is adversarial, $a^j = -1$, else $a^j = 1$, thus $a^j \in \{-1, 1\}$
- **Bias** b^j:
  - Normal prior $\mathcal{N}(b^j | \mu_b, s_b)$, $\mu_b = 0$ to favour unbiased annotators and $s_b = 0.05$ to allow for some positive and negative bias
- **Variability** α^j :
  - measures the variance of *j*th annotator, thus lower is better.

- Gaussian model: $\mathcal{N}(\mathbf{y}_i^j | \mathbf{y}_i, \alpha^j, a^j, b^j)$
- Parameters to be estimated: $\boldsymbol{\theta} = \{\mathbf{y}, \boldsymbol{\alpha}, \mathbf{a}, \mathbf{b}\}$

Proposed solution

- Model likelihood:

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{R}\mathcal{N}(\mathbf{y}_i^j|\mathbf{y}_i,\alpha^j,a^j,b^j)\times\prod_{j=1}^{R}\mathcal{N}(b^j|\mu_b,s_b)$$

$$= \prod_{i=1}^{N}\prod_{j=1}^{R}\frac{1}{\sqrt{2\pi\alpha^j}}\exp\left[\frac{-1}{2\alpha^j}\|\mathbf{y}_i^j - a^j(\mathbf{y}_i + b^j\mathbf{1})\|_2^2\right]\times\prod_{j=1}^{R}\frac{1}{\sqrt{2\pi s_b}}\exp\left[\frac{-1}{2s_b}\left(b^j-\mu_b\right)^2\right]$$

- Log-likelihood:

$$\ln P(\mathcal{D}|\boldsymbol{\theta}) = C - \frac{N}{2}\sum_{j=1}^{R}\ln(\alpha^j) - \sum_{i=1}^{N}\sum_{j=1}^{R}\frac{1}{2\alpha^j}\|\mathbf{y}_i^j - a^j(\mathbf{y}_i + b^j\mathbf{1})\|_2^2 - \sum_{j=1}^{R}\frac{1}{2s_b}\left(b^j-\mu_b\right)^2$$

Proposed solution

- Update solution for $\mathbf{y}_i$:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial \mathbf{y}_i} = \mathbf{0} \implies \sum_{j=1}^{R} \frac{1}{\hat{\alpha}^j} \left(\mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1}) \right)^{\mathsf{T}} \hat{a}^j = \mathbf{0}$$

$$\boxed{\hat{\mathbf{y}}_i = \frac{1}{\sum_{j=1}^{R} \frac{1}{\hat{\alpha}^j}} \sum_{j=1}^{R} \frac{\left(\hat{a}^j \mathbf{y}_i^j - \hat{b}^j \mathbf{1} \right)}{\hat{\alpha}^j}}$$

- Update solution for a^j:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial a^j} = 0 \implies -\frac{1}{\hat{\alpha}^j} \sum_{i=1}^{N} \left(\mathbf{y}_i^j - a^j (\mathbf{y}_i + b^j \mathbf{1}) \right)^{\mathsf{T}} \times \left(\mathbf{y}_i + b^j \mathbf{1} \right) \times (-1) = 0$$

$$\boxed{\hat{a}^j = \operatorname{sgn} \left(\sum_{i=1}^{N} \mathbf{y}_i^{j\,\mathsf{T}} (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1}) \right)}$$

Proposed solution

- Update solution for b^j :

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial b^j} = 0 \implies \sum_{i=1}^{N} \frac{1}{\hat{\alpha}^j} \left(\mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1}) \right)^{\mathsf{T}} \times (-\hat{a}^j \mathbf{1}) + \frac{1}{s_b} (\hat{b}^j - \mu_b) = 0$$

$$\boxed{\hat{b}^j = \frac{1}{N + \frac{\hat{\alpha}^j}{s_b}} \left(\sum_{i=1}^{N} \left(\hat{a}^j \mathbf{y}_i^j - \hat{\mathbf{y}}_i \right)^{\mathsf{T}} \mathbf{1} + \frac{\hat{\alpha}^j \mu_b}{s_b} \right)}$$

- Update solution for α^j:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial \alpha^j} = 0 \implies \frac{-N}{2\hat{\alpha}^j} + \frac{1}{2(\hat{\alpha}^j)^2} \sum_{i=1}^{N} \| \mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1}) \|_2^2 = 0$$

$$\boxed{\hat{\alpha}^j = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1}) \|_2^2}$$

Proposed solution – EM algorithm

E-step:

$$\hat{\mathbf{y}}_i = \frac{1}{\sum_{j=1}^{R} \frac{1}{\hat{\alpha}^j}} \sum_{j=1}^{R} \frac{\left(\hat{a}^j \mathbf{y}_i^j - \hat{b}^j \mathbf{1}\right)}{\hat{\alpha}^j}$$

M-step:

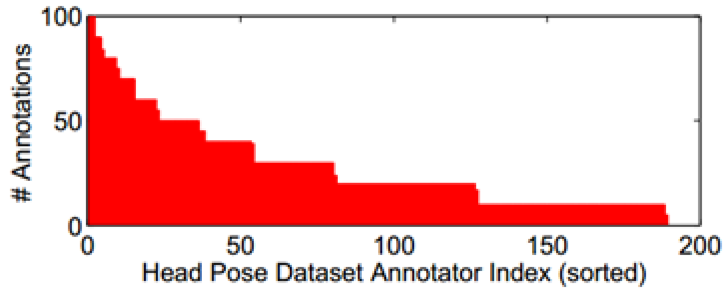
$$\hat{a}^j = \operatorname{sgn}\left(\sum_{i=1}^{N} \mathbf{y}_i^{j\,\intercal} (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1})\right)$$

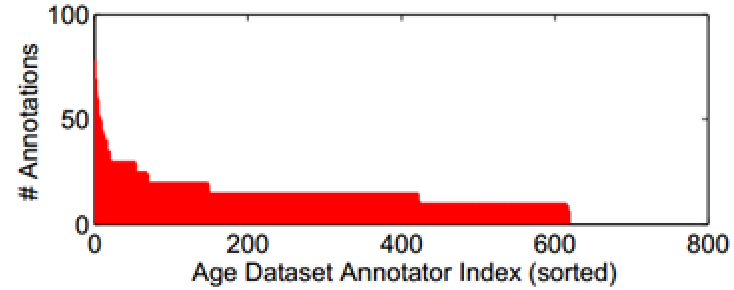
$$\hat{b}^j = \frac{1}{N + \frac{\hat{\alpha}^j}{s_b}} \left(\sum_{i=1}^{N} \left(\hat{a}^j \mathbf{y}_i^j - \hat{\mathbf{y}}_i\right)^{\intercal} \mathbf{1} + \frac{\hat{\alpha}^j \mu_b}{s_b}\right)$$

$$\hat{\alpha}^j = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1})\|_2^2$$

Iterate till $\Delta\|\hat{\mathbf{y}}\| < 10^{-6}$

Long-tail problem





AnnotatorID	#Annotations	Variability ($\hat{\alpha}$)	95% Conf. Int. (CI)
22540655 (HeadPose)	100	0.2491	$(0.2178, 0.3217)$
30507455 (HeadPose)	75	0.3657	$(0.2152, 0.9174)$
30172026 (HeadPose)	**30**	**1.4651**	$(\mathbf{0.7932}, \mathbf{3.2845})$
7837812 (HeadPose)	**5**	**1.6796**	$(\mathbf{0.2583}, \mathbf{6.8521})$
17525614 (Age)	78	0.3615	$(0.2718, 0.3217)$
20730328 (Age)	50	0.3529	$(0.2436, 0.4173)$
4711962 (Age)	**20**	**0.5342**	$(\mathbf{0.2076}, \mathbf{2.6819})$
22201476 (Age)	**6**	**0.5618**	$(\mathbf{0.1738}, \mathbf{4.6931})$

$$CI_{1-\beta} = \left\{ \frac{|\mathcal{R}_j|\hat{\alpha}^j}{\chi^2_{(1-\frac{\beta}{2}, |\mathcal{R}_j|)}}, \frac{|\mathcal{R}_j|\hat{\alpha}^j}{\chi^2_{(\frac{\beta}{2}, |\mathcal{R}_j|)}} \right\}$$

Updated solution – EM algorithm

E-step:

$$\hat{\mathbf{y}}_i = \texttt{wMedian}\left(\hat{a}^j \mathbf{y}_i^j - \hat{b}^j \mathbf{1}, \frac{1}{\hat{\alpha}^j}\right)$$

M-step:

$$\hat{a}^j = \operatorname{sgn}\left(\sum_{i=1}^{N} {\mathbf{y}_i^j}^\mathsf{T} (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1})\right)$$

$$\hat{b}^j = \frac{1}{N + \frac{\hat{\alpha}^j}{s_b}} \left(\sum_{i=1}^{N} \left(\hat{a}^j \mathbf{y}_i^j - \hat{\mathbf{y}}_i\right)^\mathsf{T} \mathbf{1} + \frac{\hat{\alpha}^j \mu_b}{s_b}\right)$$

$$\hat{\alpha}^j = \frac{1}{\chi^2_{(\frac{\beta}{2}, |\mathcal{R}_j|)}} \sum_{i \in \mathcal{R}_j} \|\mathbf{y}_i^j - \hat{a}^j (\hat{\mathbf{y}}_i + \hat{b}^j \mathbf{1})\|_2^2$$

Iterate till

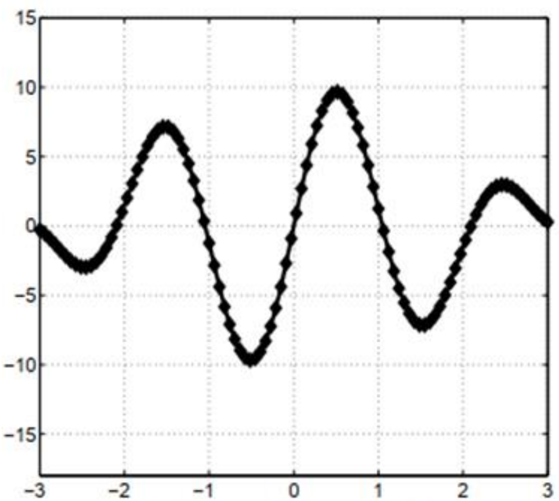
$\Delta \|\hat{\mathbf{y}}\| < 10^{-6}$

Validation

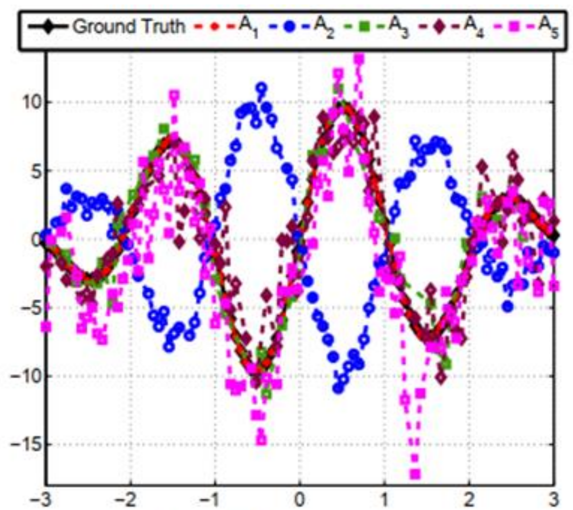
- 5 **simulated annotators**; data of 100 samples from $y_i^j = f(x_i) + \epsilon^j$

  $f(x) = 10\sin(3x)\cos(\frac{1}{2}x),\ \epsilon^j \sim \mathcal{N}(0, \alpha^j),\ \boldsymbol{\alpha} = \{0.1, 0.8, 1.5, 2.2, 3\},$

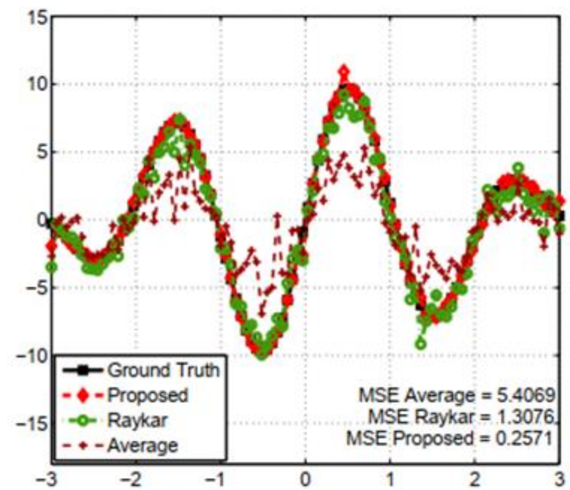
  $|\mathcal{R}| = \{90, 95, 40, 70, 85\}$
- 2nd annotator adversarial and 5th annotator biased



(a) Ground truth data



(b) Data from annotators



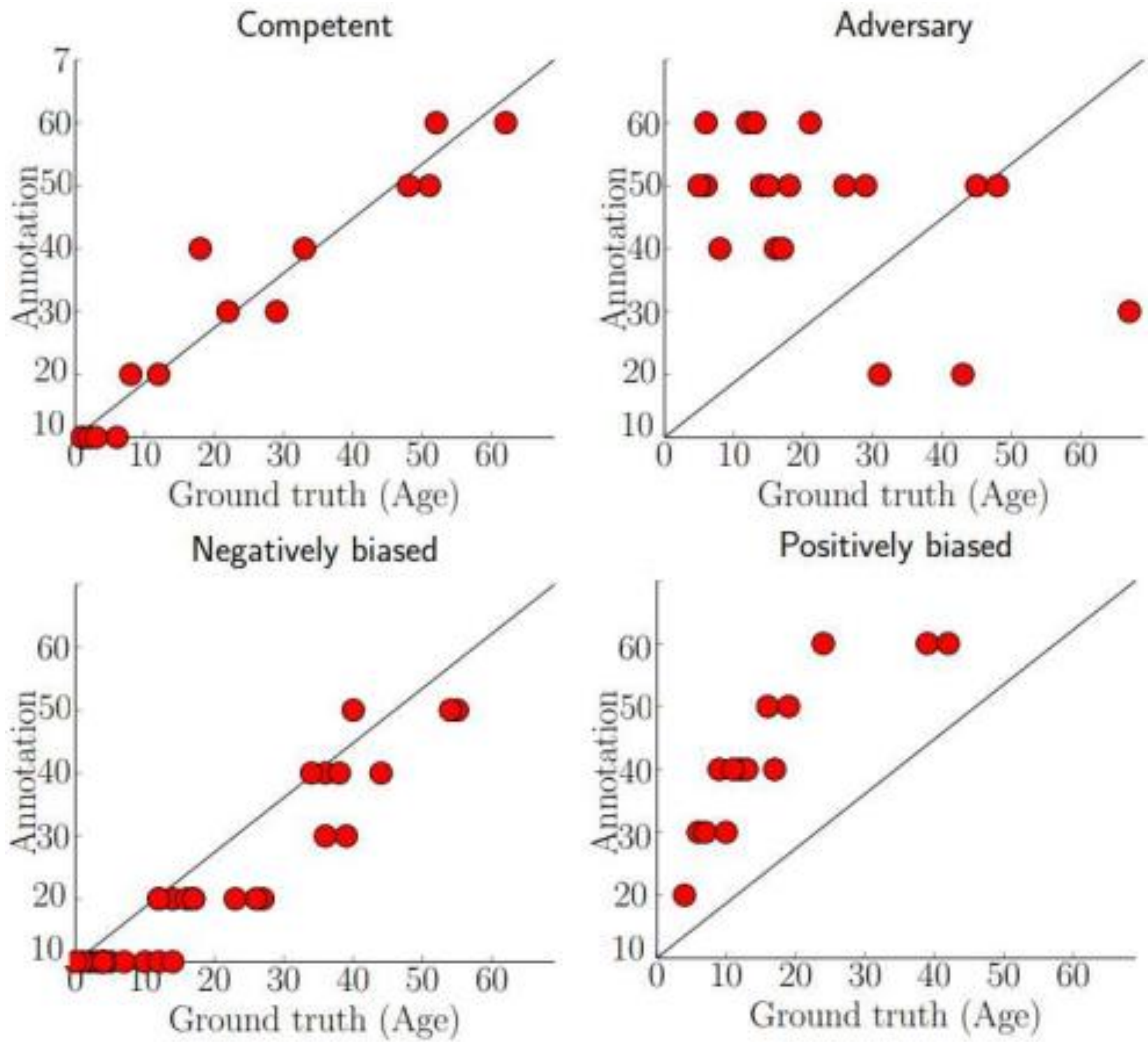
(c) Estimated Consensus

Benchmark datasets

- Housing => 506 data samples, 'MEDV' is ground truth; simulated annotators
- Population [5] => Wikipedia edit history of city population (1,124 samples and 2,344 annotators)
- HeadPose [6] => headpose images of 15 people, different tilt and pan orientations, 555 samples and 189 annotators
- Age [7] => age of different people by looking at images, 619 samples and 1,002 annotators

[5] J. Pasternack and D. Roth, "Knowing what to believe (when you already know something)", *Proc. of the 23rd International Conference on Computational Linguistics*, pp. 877–885, Aug 2010.
[6] N. Gourier, D. Hall and J. Crowley, *Head Pose Image Database*, 2018.
[7] Face and Gesture Recognition Working group, *The FGNet Aging Database*, 2018.

Benchmark datasets



Results

- Proposed algorithm achieves lower MSE due to its ability to identify **adversarial** as well as **biased** annotators.

	Average	Raykar *et al.*	Proposed
Simulated	5.4069	1.3076	**0.2571**
Synthetic	0.5631	0.3872	**0.1436**
Housing	0.6548	0.4317	**0.2391**
Population	126, 198	8, 513	**7, 154**
HeadPose	0.7082	0.4924	**0.2342**
Age	21.6679	15.4278	**13.1816**

Conclusions

- EM algorithm proposed to model the varying behavior of annotators
- Confidence-interval based estimated consensus is derived for the continuous target task
- Proposed work useful when ground truth is not available and only crowdsourced continuous annotations are available
- Proposed technique can identify adversariness, biasedness, and variability of each annotator through behavior modeling and simultaneously learn the unknown ground truth

References

[1] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni and L. Moy, ”Learning from crowds”, Journal of Machine Learning Research, vol. 11, pp. 1297–1322, Apr 2010.

[2] S. Chatterjee, A. Mukhopadhyay and M. Bhattacharyya, “A review of judgement analysis algorithms for crowdsourced opinions", IEEE Transactions on Knowledge and Data Engineering, Mar 2019.

[3] M. Wan, X. Chen, L. Kaplan, J. Han, J, Gao and B. Zhao, “From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach", Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 1885{1894, Aug 2016.

[4] H. Xiao, H. Xiao and C. Eckert, ”Learning from multiple observers with unknown expertise”, Springer Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), vol. 7818, pp. 595–606, Apr 2013.

[5] J. Pasternack and D. Roth, ”Knowing what to believe (when you already know something)”, Proc. of the 23rd International Conference on Computational Linguistics, pp. 877–885, Aug 2010.

[6] N. Gourier, D. Hall and J. Crowley, Head Pose Image Database, 2018.

[7] Face and Gesture Recognition Working group, The FGNet Aging Database, 2018.