

# 4th Crowd Science Workshop — CANDLE: Collaboration of Humans and Learning Algorithms for Data Labeling

Dmitry Ustalov\*

Toloka  
Belgrade, Serbia  
dustalov@toloka.ai

Niels van Berkel  
University of Aalborg  
Aalborg, Denmark  
nielsvanberkel@cs.aau.dk

## ABSTRACT

Saiph Savage  
Northeastern University  
Boston, MA, USA  
s.savage@northeastern.edu

Yang Liu  
University of California Santa Cruz  
Santa Cruz, CA, USA  
yangliu@ucsc.edu

Crowdsourcing has been used to produce impactful and large-scale datasets for Machine Learning and Artificial Intelligence (AI), such as ImageNET, SuperGLUE, etc. Since the rise of crowdsourcing in early 2000s, the AI community has been studying its computational, system design, and data-centric aspects at various angles. We welcome the studies on developing and enhancing of crowdworker centric tools, that offer task matching, requester assessment, instruction validation, among other topics. We are also interested in exploring methods that leverage the integration of crowdworkers to improve the recognition and performance of the machine learning models. Thus, we invite studies that focus on shipping active learning techniques, methods for joint learning from noisy data and from crowds, novel approaches for crowd-computer interaction, repetitive task automation, and role separation between humans and machines. Moreover, we invite works on designing and applying such techniques in various domains, including e-commerce and medicine.

## CCS CONCEPTS

•Information systems→Crowdsourcing; • Computing methodologies → *Neural networks*; • General and reference → Evaluation.

## KEYWORDS

crowdsourcing, data labeling, human-in-the-loop

## WORKSHOP TOPICS

We believe that the mainstream studies of crowdsourcing

do not pay enough attention to the four vital currently underrepresented facets presented below (but the list is non-exclusive). Generally, we welcome the studies on developing and enhancing of crowdworker-centric tools, that offer task matching, requester assessment, instruction validation, among other topics. We are also interested in exploring methods that leverage the integration of crowdworkers to improve the recognition and performance of the machine learning models. Thus, we invite studies that focus on shipping active learning techniques, methods for joint learning from noisy data and from crowds, novel approaches for crowd-computer interaction, repetitive task automation, and role separation between humans and machines. Moreover, we invite works on designing and applying such techniques in various domains, including e-commerce and medicine.

*Imperfections of Crowdsourcing.* Improving the annotation schemes and machine learning algorithms to handle unexpected inconsistencies made by the non-expert human annotators because of the topic polarization, gaps in task instructions, and differences in backgrounds of requesters and workers due to the imperfection of the crowdsourcing methodology.

*Human-Machine Collaboration.* Improving the efficiency of the human annotation process by understanding the best ways to introduce the collaborations between crowd annotators and machine learning algorithms. Exemplary topics include building human-in-the-loop learning pipeline.

*Fairness and Inclusion of Crowd Annotators.* Providing organizational and technical means to improve the welfare of crowd workers by using AI for automating the repetitive tasks, process simplification for different stakeholders, and additional worker-centric tooling.

*Reducing Biases in Crowd Annotation.* Reducing biases in crowd annotation, which is suggested by the fact that machine learning models memorizing and reinforcing biases in the training data.

**Acknowledgements.** This work was partially supported by NSF grant FW-HTF-19541.