Reports of the Workshops Held at the Sixth AAAI Conference on Human Computation and Crowdsourcing

Lora Aroyo, Anca Dumitrache, Elena Simperl, Matthew Lease, Pietro Michelucci, Jeffrey V. Nickerson

■ *The workshop program of the Association for the Advancement of Artificial Intelligence's Sixth AAAI Conference on Human Computation and Crowdsourcing was held on the campus of the University of Zurich in Zürich, Switzerland, July 5, 2018. This report includes summaries of four of the workshops.*

The workshop program of the Association for the Advancement of Artificial Intelligence's Sixth AAAI Conference on Human Computation and Crowdsourcing was held on the campus of the University of Zurich in Zürich, Switzerland, on July 5, 2018. There were three full-day workshops in the program: CrowdBias: Disentangling the Relation between Crowdsourcing and Bias Management; Subjectivity, Ambiguity, and Disagreement in Crowdsourcing; Work in the Age of Intelligent Machines. There was one three-quarter-day workshop, Advancing Human Computation with Complexity Science, and one quarter-day workshop on Project Networking.

 ISSN 0738-4602

CrowdBias: Disentangling the Relation between Crowdsourcing and Bias Management, organized by Alessandro Checco (University of Sheffield), Gianluca Demartini (University of Queensland), Ujwal Gadiraju (L3S Research Center, Leibniz Universität Hannover), and Cristina Sarasua (University of Zurich), analyzed existing biases in crowdsourcing, discussed measures and methods to track bias, and explored methodologies to prevent and solve bias. Subjectivity, Ambiguity, and Disagreement in Crowdsourcing, organized by Lora Aroyo (Vrije Universiteit Amsterdam), Anca Dumitrache (Vrije Universiteit Amsterdam), Praveen Paritosh (Google), Alex Quinn (Purdue University), and Chris Welty (Google), brought together a latent community of researchers who treat disagreement (and subjectivity and ambiguity) as signal, rather than noise, to discuss theoretical and empirical methodology to characterize, utilize, mitigate, and derive value from uncertainty, ambiguity, and disagreement. Work in the Age of Intelligent Machines, organized by Jeffrey V. Nickerson (Stevens Institute of Technology), Matt Lease (University of Texas, Austin), Kevin Crowston (Syracuse University School of Information Studies), and Ingrid Erickson (Syracuse University), aimed at promoting research convergence among participants on topics related to future forms of work (that is, humans doing their jobs) with intelligent machines, defined as computing technologies characterized by autonomy, the ability to learn, and the ability to interact with other systems and with humans. Advancing Human Computation with Complexity Science, organized by Pietro Michelucci (Human Computation Institute), aimed to jump-start the application of complexity science methods to human computation research to achieve distributed human/machine systems capable of tackling society's most pressing issues, many of which depend on accurate predictive modeling of dynamic interdependent systems.

This report contains summaries of the four events.

Subjectivity, Ambiguity, and Disagreement in Crowdsourcing

The goal of the first Subjectivity, Ambiguity, and Disagreement in Crowdsourcing workshop was to bring together a latent community of researchers who treat disagreement as signal, rather than noise, specifically in the context of their methodologies to characterize, utilize, and derive value from uncertainty and ambiguity in human computation tasks. With this workshop, we aimed to bring ideas from a variety of perspectives on how to improve our understanding of subjectivity, ambiguity, and disagreement in crowdsourcing.

Ambiguity creates uncertainty in practically every facet of human computation, including the information presented to workers as part of a task, the instructions for what to do in the task, and the information they are asked to provide. Besides the typical lexical ambiguities, ambiguity can be experienced in different content modalities (for example, text, images, videos, sounds) and can be caused for a variety of reasons (for example, missing details, visual or linguistic contradictions, subjectivity, or context of interpretation). Subjectivity may stem from differences in cultural context, life experiences, or individual perception of properties that are hard to quantify. All of these can leave workers with conflicting interpretations, leading to results that microtasks requesters (including the end users of crowd-powered systems) would regard as wrong.

Historically, the human computation community has largely attributed disagreement to low-quality workers. This perception has led to mathematical approaches intended to minimize the supposed noise through strategies that consider aggregation of crowd contributions (for example, majority, expectation maximization), linguistic approaches (for example, sense disambiguation, data cleanup, transformation, and reconciliation), statistical filtering, incentive design, and many others. All of these are typically executed after the contributions from the crowd are collected.

Recent approaches apply principles from interaction design and computer-supported collaborative work to refine task designs until disagreement is minimized. This strategy adopts an approach similar to that taken by the methodologies of the linguistic annotation and social content analysis communities, where the task guidelines and instructions are refined using interrater reliability. Here, the focus is on minimizing possible ambiguity or subjectivity before the data has been collected. The goal of the Subjectivity, Ambiguity and Disagreement in Crowdsourcing workshop was to outline the current landscape of approaches and problems when dealing with ambiguity, disagreement, and subjectivity both before and after the data is collected, and to further investigate their role in improving the intelligence in AI systems. The core question driving the workshop was, "How can systems gather and utilize (tolerate!) multiple different answers to a question, or labels for an image?" From that starting point, we discussed whether disagreement signals something useful or acceptable. Whether we can distinguish between good and bad disagreement. Whether we can gather and evaluate better corpora with respect to their ambiguity. And how to deal with the problem that there is typically no ground truth.

The workshop brought together researchers from a variety of subfields of human computation as well as related fields such as computer science, information sciences, law, communication science, and political science. It was a full-day workshop split in two parts. In the morning, the workshop opened with a keynote by Drazen Prelec (Massachusetts Institute of

Technology), who spoke on Bayesian truth serum. In his keynote, Drazen Prelec introduced the term as "an information-theoretic scoring algorithm that rewards respondents for honest reporting of private information, using the reports of other people as the only input (individual honesty is nonverifiable)." He noted that the algorithm can also function as an "objective truth-detector, identifying which answer to a multiple choice question is most likely to be true." He presented the theory of the approach and the results from a wine tasting problem experiment focusing on understanding the reasons for disagreements, for example, ambiguous vocabulary and differences in perceptual experience. The keynote was followed by a session of lighting talks introducing four research papers, which were further discussed in detail during a dedicated poster and joint discussion sessions.

The first research session covered four papers. The connecting theme in this session was the exploration of disagreement and ambiguity in a number of textual and linguistic use cases. The session started with a presentation of the preliminary results on data about anaphoric ambiguity collected using the Phrase Detectives game.[1] This collaboration between Queen Mary University of London and the University of Essex showed that in the analysis about half of the markables labeled during the game have at least two interpretations supported by more players than disagree with them.

Typically, crowdsourcing-based approaches to gather annotated data use interannotator agreement as a measure of quality. However, in many domains, there is ambiguity in the data, as well as a multitude of perspectives of the information examples. Anca Dumitrache presented an ongoing work, "Metrics for Capturing Ambiguity in Crowsourcing by Interlinking workers, Annotations, and Input Data," based in CrowdTruth metrics,[2] which capture and interpret interannotator disagreement in crowdsourcing so as to model the degree of ambiguity in each of these three components.

The paper "Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation," presented by Tommaso Caselli and Oana Inel, also promoted a new annotation approach, combining crowd and experts through the application of CrowdTruth metrics in the context of a crowdsourcing experiment for the annotation of plotlike structures in English news articles. The CrowdTruth methodology and metrics were used here to select valid annotations from the crowd. The paper presented the results of the in-depth analysis of the annotated data and showed a valuable use of crowdsourcing annotations for such complex semantic tasks.[3]

The morning session closed with the presentation by Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng (Google), who proposed a class of annotations that exhibit acceptable variation, which can be positioned between the two extremes, that is, (1) items that have truly only one acceptable response and (2) items that have a number of divergent annotations that are truly of unacceptable quality. They illustrated these two extremes within existing annotated datasets and explored the implications of acceptable variation on the task design of annotations and on the evaluation of the quality of annotations.

The afternoon session opened with the keynote by Brent Hecht (Northwestern University), "Disagreement in Crowdsourcing due to Cultural Context," followed by the second research papers session. In his keynote, Brent Hecht presented a story line around three topics: (1) the influence of cultural context on disagreement, (2) algorithmic bias, and (3) algorithmic diversity. He first discussed the aspects of disagreement caused by the diversity of cultural contexts of the online crowd workers. For this phenomenon, he defined an algorithmic bias and showed that these disagreements can inform a new class of applications powered by algorithmic diversity.

The connecting theme in the second research session (which covered five papers) was disagreement, subjectivity, and ambiguity in different modalities, such as images, videos, and sensor data. "Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis,"[4] presented by Edith Law, reported results from a case study on sleep stage classification, specifically focusing on learning from the expert disagreement in sequential labeling tasks where the interpretation of one case can affect the interpretation of subsequent or previous cases, and then exploring future application scenarios of expert discussions for the training of nonexpert crowdworkers.

Veronika Cheplygina presented "Crowd Disagreement for Medical Images Is Informative," which argued that disagreement between annotators in the process of annotating medical images may be informative in a use case of crowdsourcing the classification of skin lesion as a melanoma or not. A comparison of the mean annotations (illustrating consensus) and the standard deviations (illustrating disagreement) showed that the mean annotations perform best, but that the disagreement measures are still informative.

Lora Aroyo presented "CaptureBias: Using Ambiguity to Support Media Scientists in News Videos Bias Detection,"[5] a human-in-the-loop approach to investigate the role of ambiguity in detection and interpretation of bias. Specifically, this approach explores the presence of ambiguity in textual and visual media and its influence on understanding and capturing possible bias in news (for example, racial and gender bias), as well as framing. The study focuses on supporting media scholars and social scientists in their media analysis.

Finally, Margaret Warren presented "Bounding

Ambiguity: Experiences with an Image Annotation System"[6] in the context of a use case for creating and editing rich metadata descriptions for images. The authors discussed the roles of ambiguity, disagreement, and subjectivity in knowledge formation and their implications for the design of a system for semantic annotation of images.

Lora Aroyo, Anca Dumitrache, Praveen Paritosh, Alex Quinn, and Chris Welty served as cochairs of this workshop. The papers of the symposium were published jointly with the HCOMP2018 CrowdBias workshop as part of the CEUR workshop proceedings series.

Work in the Age of Intelligent Machines

The Work in the Age of Intelligent Machines workshop explored ways in which human work and occupations will be changed as artificial intelligence becomes increasingly prevalent in the workplace. While much media and academic attention has focused on forecasts of the displacement of workers, less attention has focused on ways AI might change the workplace, and in particular, ways AI might generate new jobs or mitigate the displacement of workers. By doing so, AI may help address a large-scale societal problem, a shift in skills needed in the workplace.

This workshop, part of a series of workshops supported by the National Science Foundation, brought together researchers from academia, government and industry.The workshop's goal was the generation of key research questions on the topic that merit further study. Research questions were generated in discussions around two topics: "AI-Human Team Dynamics," facilitated by Kurt Luther (Virginia Tech), and "New Jobs, Education and Training, Unemployment," facilitated by Matthew Lease (University of Texas, Austin).

The AI-Human Team Dynamics group posed the following questions: How can human and AI agents participate in competitive interactions in order to obtain better outcomes? How can AI and human agents effectively interact through negotiation? How can humans represented by AI agents (with potentially different value systems) effectively interact with each other through negotiation? How can we combine them? The first question inverts the commonly held view that humans and machines should complement each other. Instead, organizing friendly competition between humans and machines may help us better understand human and machine capabilities. Such competitions could be the precursor to negotiation, given that it becomes easier to figure out how to trade or align cognitive effort in the service of a shared goal once the comparative advantages between humans and machines are better understood. Answering these questions collectively might help to produce new techniques for product design, which could open up new products and new human jobs around the communication and support of those products. The discussion ranged through the manufacturing, transportation (particularly with respect to autonomous vehicles), and entertainment industries.

The New Jobs, Education and Training, Unemployment group addressed complementary issues. While AI may displace workers, it may also help retrain people to work in jobs that require skills in short supply. That is, AI may substitute for humans in some tasks, but conversely it can help build human capabilities that work in concert with machine intelligence. The discussion of this group centered around questions such as the following: How can we use AI to reduce skill barriers to jobs, thereby growing job opportunities and the scalability of labor? How can we combat a potential skill-technology gap and so reduce labor market frictions? How can AI be used to simplify highly skilled jobs to make them accessible to a larger part potential workforce?

These questions, in turn, led to a series of questions related to the labor force: What new jobs and/or transformation of existing jobs will come from the advent of intelligent technologies? What job descriptions are emerging on job boards related to AI? Do some industries hire more people as automation increases? In discussing these scenarios, the group considered a potential impact of autonomous vehicles on restaurants: less expensive and more ubiquitous transportation might encourage more nights out.

This discussion branched out into questions about how AI might, in fact, enhance jobs, the leading question being, How can we integrate AI alongside human workers in such a way as to enhance (in some balanced way) productivity, satisfaction, and career growth?

The group addressed what is known about intelligent tutoring, about predicting student failure in advance and providing interventions, and about peer assessment and feedback, especially research informed by MOOCs and crowd-based approaches to skill building and work. A metaquestion was also posed: To work successfully in the age of intelligent machines, what do people typically need to know about AI — and what don't they need to know?

Overall, the workshop raised a variety of important questions that necessitate further research, so that we can be proactive in addressing human work and occupational changes as AI becomes increasingly ubiquitous in the workplace. With the series of NSF workshops on the future of work, and the growing interest in the topic, the community can expect these ideas to be further discussed and explored not only in these workshops but in other community events as well. This workshop was supported by the National Science Foundation under grant IIS-1745463.

Advancing Human Computation with Complexity Science

The goal of the Advancing Human Computation with Complexity Science workshop was to consider the potential role of complexity science in the design of distributed human and machine systems that could address complex societal problems. The workshop served to jump-start the application of complexity science methods to human computation (HCOMP) research to achieve distributed human and machine systems capable of tackling society's most pressing issues, many of which depend on accurate predictive modeling of dynamic interdependent systems.

HCOMP has been applied to the betterment of society through AI methods that leverage the complementary strengths of networked humans and machines in scalable and sustainable participatory systems toward new, high-impact capabilities. This approach has revolutionized the analysis of large, homogeneous datasets, accelerating scientific research by orders of magnitude. However, new methods are needed to tackle today's wicked problems, which involve multisource heterogeneous data and tend to involve interdependent systems, requiring dynamic solutions to address a rapidly evolving problem space.

Complexity science is a rich source for such methods. This transdisciplinary field seeks to make sense out of a wide range of complex adaptive systems through a variety of methods including evolutionary game theory, network theory, nonlinear dynamics, out-of-equilibrium statistical mechanics, information theory, scaling theory of biological and cultural networks, robust design, nontraditional theories of computing, and agent-based modeling.

Applying the methods of complexity science to distributed human/machine systems may be the critical next step in developing human computation systems capable of addressing the existential crises that face humanity in the 21st century (for example, climate change, famine, poverty, disease, war), which are themselves amplified by technology.

We sought to explore relevant opportunities and consider new research directions toward realizing these capabilities through various lines of inquiry: How do we combine citizen science with multiagent systems and mechanism design to generate usable models of complex socioecological systems? How could complexity science help us wrangle the many drivers, consequences, and time scales involved in disaster management and design systems to provide critical feedback loops that improve resilience? How do we design information ecosystems that effectively bootstrap their own evolution in a goal-directed context?

This process led to an initial mapping of complexity science concepts to both HCOMP system dynamics and candidate problem domains. Concepts such as swarm theory were related to social network analysis and the design of viral recruitment strategies to fuel sustainable crowd-powered systems. Scaling analysis was suggested for anticipating reciprocal effects between individual and system behaviors. Phase shifts, in dynamic systems theory, were discussed as a method for anticipating sudden changes in collective behavior resulting from "critical mass," as well as threshold-based runaway processes in application domains such as climatology. The more general notion of feedback loops was applied to predictively modeling the real-world impact of candidate solutions generated by HCOMP systems.

The variety of relevant methods and potential applications identified in this initial exploration seem to support the workshop thesis that complexity science has much to offer the scientific and practical advancement of HCOMP. We hope this high-level mapping will inspire follow-on interactions between complexity science researchers and HCOMP scientists that lead to substantive HCOMP advancements that materially improve quality of life and survivability for humans and the earth system.

Pietro Michelucci served as chair of this workshop and wrote this report.

Project Networking

The Project Networking workshop was the first of its kind to be held at HCOMP. The workshop brought together participants from the worldwide HCOMP community with contributions from all types of projects, from academic grants and partnerships between academe and industry, to large funded projects and networks worldwide. The aim of the workshop was to establish connections between people and organizations working in similar or complementary areas both nationally and internationally.

Ten projects participated in the networking workshop. Each presented a short pitch introducing the project and participated in a number of breakout roundtables focusing on discussing open challenges related to sharing datasets, sharing task designs and code, and defining benchmarks for different tasks. Every 30 minutes, projects swapped roundtables to discuss another topic so that everyone had a chance to provide feedback and meet as many projects as possible. Each project also had the opportunity to present a poster and a demo at the official HCOMP 2018 poster and demo session, which allowed them to achieve a greater outreach and receive feedback from the broader HCOMP2 018 audience.

The project networking workshop provided participants with the unique opportunity to get a compact overview of the broad research landscape in human computation through the lenses of academic and industry projects. It provided a suitable occasion for the project teams to meet in an informal and inspiring setting and to learn about their work, network,

and establish links between relevant projects worldwide, to explore synergies, and to identify areas for knowledge sharing, data and technology transfer, and other collaboration opportunities across different countries. The research and funding environment is organized differently across different countries and continents, and this workshop allowed for participants to understand national and international funding practices, exchange experiences, and discuss future joint bids across countries.

The projects represented a number of research areas. For example, mobile crowdsensing and citizen science were represented by the Dusk2Dawn project (www.idiap.ch), which was funded by the Swiss National Science Foundation (SNSF) through the Sinergia interdisciplinary program. The project studies the night behavior of young Swiss people through data gathered by volunteers using smartphones (for example, smartphone sensor data, drink photos, location videos, and interviews). Such data is of practical relevance for city and public health officials in their process of understanding the role of public and private spaces in young people's nightlife, identifying the features that characterize nocturnal habits.

Two examples of projects focusing on smart city solutions were QROWD and SocialGlass. QROWD (qrowd-project.edu) offers local government and transportation businesses innovative solutions to improve mobility, reduce traffic congestion, and make navigation safer and more efficient. To achieve this, QROWD will integrate different sources of data — maximizing the value of big data in planning and managing urban traffic and mobility. SocialGlass (www.socialglass.org) develops methods and tools for human-enhanced social data processing for the analysis of human activity dynamics. It focuses on the understanding of cities through big social data, with applications spanning from crowd management during city-scale events, to the characterization of energy consumption lifestyles. SocialGlass provides a platform to integrate heterogeneous and dynamic geo-social data from geo-enabled social media (such as Twitter, Instagram, and even Sina Weibo) and LBSNs (such as Foursquare), and from publicly available urban data from governmental portals, sensor feeds, and crowdsourcing platforms.

Six projects focused on bringing the power of machines and people together. Two of them, WDAqua and CyborgCrowd, provide a European and a Japanese pespective on bringing machines and crowds together to effectively solve problems like question answering and natural disasters. The WDAqua (wdaqua.edu) project is a European-funded Marie Sklodowska–Curie Innovative Training Network (ITN) that aims to advance the field of data-driven question answering through a combination of training, research, and innovation. This is a timely example, as question answering is becoming more and more mainstream on various platforms, and as such is quite relevant to a diverse range of end users. The project provides demonstrations in e-commerce, public sector information, and publishing. The JST CREST–funded CyborgCrowd project is a collaboration between three Japanese universities to optimize the integration of crowd and machine processing in a flexible and reusable way. Some of the applications that they use to exemplify the problems concern microvolunteering in natural disasters, library crowdsourcing, and world heritage preservation. They implement their research results on all-academic crowdsourcing platform Crowd4U.

The next two projects in this group, CrowdTruth and AdHum focus on providing human-machine platforms for data quality assessment. The CrowdTruth (crowdtruth.org) framework aims at capturing ambiguity and its role in quality assessment. For this, it encourages disagreement while gathering and analyzing crowdsourced data to provide a more continuous representation of ground truth. The disagreement between annotators is used as a signal for low-quality workers, ambiguous input text, images, or videos, or semantically confusing annotation semantics. The AdHuM project, on the other hand, aims to optimize the combination of machine learning, domain experts, and nonexperts, and to maximize the quality of data and the accuracy of machine-predicted outcomes based on those data, through an Adaptive Human-Machine (AdHuM) analysis platform.

The last two projects in this group are two examples where framework like the CrowdTruth framework has been applied, namely CaptureBias and ReTV. In the CaptureBias project (capturebias.wordpress.com), the focus is to detect and capture ambiguity so as to support media scientists in their bias detection analysis for news videos. In the ReTV project (retv-project.edu), the multitude of perspectives captured through CrowdTruth is used to provide an adequate and suitable adaptation and personalization in video summaries of broadcast material for social media.

Finally, the workshop setting was quite suitable to provide a comparative view of the participating projects with respect to their vision and goals, the resources they produced, (such as datasets, methods, software), and the availability of these resources to the HCOMP community. For this, each project submitted an extended abstract introducing the project by briefly describing its main activities and goals and explaining what the project would show or demonstrate in the networking session. Each project presented also a list of project results that the project is offering to other participants in the session. An important part of the workshop was to outline the expectations for all the projects participating in the session in terms of what each project is looking for (such as partners for testing of results) and what each project is offering (for example, new use cases, mul-

tilingual aspects, disseminating results to other communities). Finally, each project also indicated whether they were national (if so, which country) or international (listing the countries involved) and named the funding agency supporting it.

Lora Aroyo and Elena Simperl served as cochairs for this workshop. All the papers from this workshop were extended abstracts published online on the workshop website. We would like to thank the QROWD and ReTV projects for sponsoring the event.

Notes

1. catalog.ldc.upenn.edu/ldc2017T08.
2. data.crowdtruth.org.
3. aclweb.org/anthology/W18-4306.
4. crowdeeg.ca.
5. capturebias.wordpress.com.
6. www.imagesnippets.com.

Lora Aroyo is a full professor in the Department of Computer Science at the Vrije Universiteit Amsterdam, The Netherlands.

Anca Dumitrache is a PhD student in the Department of Computer Science at the Vrije Universiteit Amsterdam, The Netherlands.

Jeffrey V. Nickerson is a professor in the School of Business at Stevens Institute of Technology, USA.

Matthew Lease is an associate professor in the School of Information at the University of Texas at Austin, USA.

Pietro Michelucci directs the Human Computation Institute and is a visiting professor in the Department of Biomedical Engineering at Cornell University, USA.

Elena Simperl is a full professor in the Department of Computer Science at the University of Southampton, UK.