Hybrid Human-Machine Information Systems: Challenges and Opportunities

Gianluca Demartini

Information School
University of Sheffield, UK

Abstract

Micro-task Crowdsourcing has been used for different purposes: creating training data for machine learning algorithms, relevance judgements for evaluation of information systems, sentiment analysis, language translation, etc. In this paper we focus on the use of crowdsourcing as core component of data-driven systems. The creation of hybrid human-machine systems is a highly promising direction as it allows to leverage both the scalability of machines over large amounts of data as well as to keep the quality of human intelligence in the loop to finally obtain both efficiency and effectiveness in data processing applications.

Such hybrid approach is a great opportunity to develop systems that are more powerful than purely machine-based ones. For examples, it is possible to build systems that can understand sarcasm in text at scale. However, when designing such systems it is critical to take into account a number of dimensions related to human behavior as humans become a component of the overall process.

In this paper, we overview existing hybrid human-machine systems presenting commonalities in the approaches taken by different research communities. We summarize the key challenges that one has to face in developing such systems as well the opportunities and the open research directions to make such approaches the best way to process data in the future.

Keywords: Human Computation, Crowdsourcing, Database, Semantic Web, Information Retrieval

Email address: `g.demartini@sheffield.ac.uk` (Gianluca Demartini)

1. Introduction

With the rapid growth of data available in enterprises and on the Web, the need for effective and efficient data processing systems gets stronger. Data is a key asset in business and it has become key to support decisions. While machine-based solutions for large-scale data processing exists, they are limited in the type of data processing tasks they can do. Examples of tasks where machine-based systems perform poorly include image understanding, detecting opinions or sarcasm in text, etc.

To alleviate these problems, hybrid human-machine systems leveraging human intelligence at scale in combination with machine-based algorithms have been proposed. These systems make use of crowdsourcing by asking data related questions to a crowd of human individuals available to answer them. Thanks to such human intelligence component, this type of information systems can perform tasks which are otherwise not possible to accomplish. The machine-based pre-processing or post-processing enables scalability over large amounts of data (e.g., thanks to scale-out architecture like Map/Reduce [18]).

Data chunks with related questions which are sent to the crowd by the system are usually called Human Intelligence Tasks (HITs) as they require human intelligence to be completed. A variety of task types is commonly published on these crowdsourcing platforms varying from audio transcription to general population surveys (see [31] for a classification).

Crowdsourcing is a very general term covering topics from innovation [50] to citizen science [35]. Popular crowdsourcing examples include Wikipedia, a free on-line encyclopedia that anyone on the Web can edit; GalaxyZoo, a platform where any user can annotate large amounts of scientific images obtained with telescopes or from experiments [35]; and Recaptcha, used originally to correct OCR errors in a large book digitalization project [64].

In this paper we focus specifically on systems that leverage *paid micro-tasks crowdsourcing.* Commercial platform like Amazon MTurk [36] have been built to support the exchange of HITs between *requesters* who need tasks to be completed and *workers*, that is, members of the crowd, who are willing to complete tasks moved by a financial incentive.

In this paper we describe hybrid human-machine systems that crowdsource many small tasks to a crowd of human workers who complete them in exchange of a small monetary reward. We describe most popular hybrid systems, their characteristics, and the main challenges that need to be faced when building a system with a crowd component inside. Aspects to be dealt with include controlling latency, data quality, and crowd motivation. Finally, we present a set of research directions in the area of hybrid human-machine systems. These include long-term use of crowds, complex hybrid data pipelines, and crowdsourcing efficiency improvements.

The rest of this paper is structured as follows. In Section 2 we overview existing hybrid human-machine systems proposed and evaluated by different research communities including database, information retrieval, social networks, semantic web, and data-driven sciences such as biomedicine and astrophysics. In Section 3 we summarize the main challenges that these type of systems have to face when dealing both with large amounts of data as well as with human individuals performing tasks for the system. In Section 4 we describe different open research questions in the area of human computation and crowdsourcing that need to be addressed to improve efficiency and effectiveness of hybrid human-machine systems. Finally, Section 5 concludes the paper.

2. Existing Human-Machine Systems

Because of the ability to effectively process data at scale, a number of hybrid human-machine systems have been recently proposed within different data-related research fields. In this section we provide an overview of such systems.

2.1. Early Human Computation Systems

Early examples of systems that leverage human intelligence in combination of machine-based data processing mostly leveraged the fun incentive rather than the monetary one. Thus, by means of gamification, systems like the ESP game were designed [62]. In this system two human players have to agree on the words to use to tag a picture without possibility to interact with each other. Tags over which an agreement is reached are collected and used to generate a large collection of tagged images that can be used, for example, to train supervised machine learning algorithms. An extension of the ESP game is Peekaboom: a game that asks players to detect and annotate specific objects within an image [63]. A very popular crowdsourcing application is Recaptcha [64], which generates captcha codes that human users have to type to get access to Web content and which contain scanned words (from books) that would be otherwise complex to identify by means of automated OCR approaches. Thus, by entering valid captcha codes, human intelligence helps to digitize large amounts of textual content otherwise only available on paper. Recaptcha is now being used also for other purposes such as transcribing house numbers within pictures.

2.2. Data Processing

The first crowd-powered database was CrowdDB proposed in 2011 by [30]. This system leverages crowdsourcing to process query operators within more powerful SQL queries that can, for example, retrieve images for a motivational slide show. In this case the crowd is used to tag images on their motivational dimension which is a relatively simple task for humans but a very complex one for machine-based algorithms. After this first foundational work, a number of more specific database problems have been addressed by hybrid human-machine approaches.

One of these is *entity resolution*. That is, detecting that two instances in the database refer to the same real-world entity (e.g., 'IBM' and 'International Business Machines'). In this context, proposed hybrid human-machine systems combine automatic approaches that compute similarity between large number of entity label pairs and crowdsource some entity pairs for manual matching thus obtaining both scalable and accurate entity resolution. To obtain this result it is important to minimize the number of HITs to be crowdsourced by leveraging machine-based algorithm confidence scores to selectively crowdsource entity-pairs to be matched [20]. In [66] authors show how an hybrid human-machine approach performs better than both a purely machine based approach as well as reduces the amount of human work to be done as compared to fully manual resolution. They also show how presenting the task in the form of a table containing multiple entities to be resolved instead of single entity pairs reduces the latency of the crowd. Related to this, [68] observed how allowing workers not to answer a specific entity resolution task improves the overall accuracy of the system. Also focusing on entity resolution, [67] studies how to estimate the accuracy gain obtained by each additional crowdsourced task. This is done to select the HITs that maximize the expected accuracy.

Another database related problem is that of *skyline queries*. These are complex-to-process queries that aim at retrieving optimal results over multiple dimensions. For example, hotels that are best in terms of price and distance to the beach. In this example, some results will always be worse than others in terms of both dimensions and can be safely filtered out in the early stages of query processing.

An hybrid human-machine approach has been proposed for this type of queries as well. In [44] authors focus on selecting which data items to crowdsource to obtain maximum result quality for skyline queries while controlling the cost of paid crowdsourcing. In detail, while finding missing values for all the tuples in a database may be not cost-efficient, by computing Pareto optimality, it is possible to select data to crowdsource that have most impact on the query result.

Top-k queries are another special type of database request that aim at selecting a subset of the result ranked over a certain dimension. While classic top-k approaches need to touch a large about of data points, in [49] authors propose methods for crowd-powered top-k query processing that limits the number of requests to be crowdsourced which would be otherwise prohibitive.

Filtering data is a key activity in data processing and is relevant to basically any domain where the volume of data is notable. In [54] authors propose hybrid human-machine approaches to filter data based on human-processable properties. The main aspect of these novel filtering approaches is the need to estimate expected cost and expected error of the filtering operation.

Graph-shaped data can also benefit from hybrid human-machine processing. In [52] authors consider the problem of hybrid human-machine graph search: for example, asking humans whether a node in the graph is reachable from another one can be sometimes more efficient than machine-based algorithms.

2.3. Information Retrieval

Another area where hybrid human-machine approaches have been used is that of information retrieval. Crowdsourcing has originally been proposed as a mean to generate relevance judgements at scale [2]. Relevance judgements are necessary data to evaluate the quality of information retrieval system effectiveness. To create this data, human assessors need to manually judge the relevance of a retrieved document to a user query submitted to an information retrieval system. Such judgements can be either binary or multi-graded. Different works in this domain have studied how crowdsourcing could be used to obtain such relevance judgements [1, 2, 16, 37]. In [37] authors show how different dimensions such as pay, effort, and worker qualifications influence result quality and find that a higher pay yields to a better output in the relevance judgement domain. In [2] authors compare the assessment done by crowd workers against expert assessors and observe good quality answers from the crowd but still worse than those obtained by experts. In [1] authors observe the importance of having a good task design and instructions to obtain quality results for this type of task. In [16] authors compared expert and crowd assessments observing disagreement in specific cases such as for informational queries.

Another area which shows increased interest within the information retrieval and the database research communities is that of *crowd-powered search*. More than just for relevance judgements, crowdsourcing has been considered as a core component of search systems. The goal is to improve classic Web search systems by focusing either on the query interpretation side (see [53, 21]) or on the result retrieval step (see [8]). Examples include CrowdSearcher: A search system that leverage social networks to forward questions and obtain answers on domain-specific topics thus improving automatic search systems by asking questions to personal contacts. [8, 9, 12]. In [53] authors propose a system in which the human component is used to interpret non-textual queries like, for example, images or videos. In [21] the human component is used to interpret long and complex keyword queries and transform them in structured queries which can be answered over the Web of Data.

2.4. Social Networks

Some hybrid human-machine system leverage social networks to improve system effectiveness. An early attempt to crowdsource micro-tasks over a social network has been proposed by [22] where authors present a framework to post questions as tweets that users can solve by tweeting back an answer.

As discussed above, CrowdSearcher [8] also leverages the social network structure by routing HITs to user personal contacts. Further on HIT routing based on social networks, recent work has studied how to model workers based on their social network activity and assign them HITs accordingly [13, 25].

Related to this are *social machines* [58] that leverage the interaction between humans and machines on-line. In this case, human interaction with machines is leveraged to produce data or to increase data value. Example social machines where a social network component adds value include Amazon and Facebook.

2.5. Semantic Web

Later in time, crowdsourcing has been leveraged by the semantic web community. Again, this community is highly data-driven and heavily working on structured data applications (e.g., Linked Open Data[1]). In such Web of Data setting, *entities*, such as persons, locations, organizations are considered first citizens of the Web. Accessing information on the Web by means of entities has become very popular [41] and knowledge graphs are used to power semantic search systems [7].

In order to build such structured entity repositories several steps need to be performed. The process of extracting semantic information from unstructured text documents is called *information extraction.* While a variety of purely machine-based approaches exist (e.g., [4, 17, 29]), the quality they can obtain is limited when compared to human extraction. In [34] authors propose an hybrid human-machine method for efficiently and effectively perform extraction of bibliographic citations. After this step, *entity linking* is done to uniquely identify mentioned entities by disambiguating and assigning them a unique identifier taken from a knowledge graph where a structured description of the entity is available. Such descriptions can then be used to support user navigation and sense-making by providing structured entity summaries which have become popular also in search engine results pages. Hybrid human-machine approaches for entity linking have also been proposed [19]. In this case the crowd is used as post-processing step to improve the quality of machine-based entity linking algorithms.

Once such knowledge graphs have been build, another task is to link together separate graphs that describe the same entities. Creating such connections allows for more complex queries that require different data graphs to be joined together (e.g., the query 'British physicists born in Lincolnshire' requires a data graph that knows that Isaac Newton was a physicists born in Woolsthorpe-by-Colsterworth and another dataset knowing that this place is in Lincolnshire). Hybrid approaches for connecting data graphs have also been proposed. In [56] authors propose and evaluate approaches for hybrid human-machine ontology alignment showing how involving humans in the task improves the overall quality of the alignment.

[1] http://linkeddata.org

2.6. Other Data-Driven Domains

More than classic data disciplines, hybrid human-machine systems have been designed for other domains, from the biomedical one to the digital humanities. An example of crowdsourcing applied to cultural heritage is [51] where authors use crowdsourcing approaches to replace professional curators for image annotation of museum content. In the biomedical domain, researchers studied the use of crowdsourcing for the verification of relationships in domain-specific ontologies [48] showing how experts and the crowd could perform at the same level of quality under certain circumstances. Another work related to the biomedical domain is [27] where authors use gamification to extract annotation from medical text aiming at engaging crowds of medical experts. Specifically, this task focuses on the extraction of terms and relations covered by the medical thesaurus UMLS[2] as compared to general text extraction tasks discussed above which do not rely on domain specific resources to support crowd work. They observe how having the possibility of accessing answers from other members of the crowd improves agreement among them.

2.7. Discussion

We have presented an overview of different hybrid human-machine systems applied to different data processing problems. By looking at this overview, we notice that the common aspects of such systems are that they all leverage crowdsourcing of data to improve the quality of machine-based algorithms by either pre-processing data or post-processing algorithmic results. All these systems also leverage machine computation to scale data processing indefinitely. We summarize different properties of the presented hybrid human-machine systems in Table 1.

Looking at Table 1 we can make the following observations. There is a balance across the different data types processed by hybrid human-machine systems with structured data being the most popular one. There is also a balance between systems that use the human component as pre-processing or post-processing of data (11 vs 13). Most of the systems use the monetary

[2] `http://www.nlm.nih.gov/research/umls/`

Table 1: A summary of hybrid human-machine systems over different aspects. Entries are ordered by publication year and incentive type. The columns indicate respectively 1) the year of publication, 2) reference to the work, 3) domain of application of the hybrid human-machine system, 4) type of data processed by the system, 5) the role of the human component in the hybrid human-machine system (i.e, processing data either before or after the machine component), 6) type of incentive used to motivate crowd workers to perform tasks, 7) whether the hybrid human-machine system performs batch or real-time data processing.

Year	Cit.	Domain	Data Type	Human role	Incentive	Time constrains
2006	[62]	Web	Images	Pre-p.	Fun	Batch
2007	[35]	Science	Images	Pre-p.	Community	Batch
2008	[64]	Web	Images	Post-p.	Access	Batch
2011	[52]	Database	Graph	Pre-p.	Monetary	Batch
2011	[30]	Database	Struct. data	Pre-p.	Monetary	Real-time
2011	[5]	Filtering	Video	Pre-p.	Monetary	Real-time
2012	[54]	Database	Struct. data	Post-p.	Monetary	Real-time
2012	[19]	Web	Unstruct. text	Post-p.	Monetary	Batch
2012	[56]	Data Integration	Struct. data	Post-p.	Monetary	Batch
2012	[66]	Entity Resolution	Struct. data	Post-p.	Monetary	Batch
2012	[68]	Entity Resolution	Struct. data	Post-p.	Monetary	Batch
2012	[8]	Search	Unstruct. text	Post-p.	Community	Real-time
2012	[42]	Captioning	Video	Pre-p.	Community	Real-time
2013	[34]	Info Extraction	Unstruct. text	Post-p.	Monetary	Batch
2013	[20]	Entity Resolution	Struct. data	Post-p.	Monetary	Batch
2013	[67]	Entity Resolution	Struct. data	Post-p.	Monetary	Batch
2013	[21]	Database	Struct. data	Pre-p.	Monetary	Batch
2013	[44]	Database	Struct. data	Post-p.	Monetary	Real-time
2013	[48]	Biomedical	Ontology	Pre-p.	Monetary	Batch
2013	[43]	Personal assistance	Unstruct. text	Pre-p.	Monetary	Real-time
2013	[27]	Biomedical	Unstruct. text	Post-p.	Fun	Batch
2014	[53]	Search	Image	Pre-p.	Monetary	Real-time
2014	[49]	Database	Struct. data	Post-p.	Monetary	Real-time
2014	[51]	Cult. Heritage	Image	Pre-p.	Monetary	Batch

incentive. The majority of systems perform batch data processing rather than real-time jobs. This is due to the intrinsic latency of the crowd as discussed in Section 4.1.2. In 2014 we can observe a decreased number of hybrid human-machine systems being proposed. This is explained by the fact that different research communities have started to address specific research challenges related to hybrid human-machine system performance (e.g., focus on improving crowdsourcing efficiency and effectiveness) rather than building new systems. We discuss these open research challenges in Section 4.

Next, we highlight the main challenges that need to be tackled to create such hybrid human-machine systems.

3. Challenges of Hybrid Human-Machine Systems

We now summarize the typical challenges which are common to all the systems which involve a crowdsourced component.

3.1. Quality Assurance

The quality of results obtained from a crowdsourcing platform is affected by multiple aspects. For example, providing detailed instructions on how to complete the task positively affects the quality of crowd work. Moreover, different workers in the crowd perform with different quality levels. Thus, a common way to increase the quality of data received back from a crowdsourcing platform is to assign the same HIT to multiple workers in the crowd. Once this is done, the main challenge is to aggregate the obtained answers in the most effective way. Much work has been already carried on aggregation of crowd answers (e.g., [59, 19, 61, 37]). One of the most recent and advanced approaches is [61] where authors propose an aggregation model where worker trust scores are computed measuring the similarity of their behavior with other workers in the crowd. By identifying communities of workers based on their work patterns, the proposed model weights answers differently and performs well as compared to alternative models also when few answer per worker are available, which is true for the vast majority of workers participating to a crowdsourced task.

Human factors are key in explaining varying quality of data obtained form the crowd. In [38] authors study how HIT properties such as the pay and the effort required to complete the tasks affect the type of workers attracted by the task and thus the quality of the results. Their findings show that the amount of reward has an influence on the quality of the work done. This is different than what earlier work has found [47] where authors show that an higher reward leads to work being completed faster, but not better. The conclusions of [38] are that decisions made by requesters on task design and reward influence the type of workers attracted by the task and, thus, create a biased sample from the crowd which affect the final work quality.

Another way to foster quality is the proactive selection of certain workers in the crowd, also known as, *crowd building*. This is especially important for HITs where certain knowledge or skills are necessary to effectively complete the required task. Examples of work in this direction aim at finding the right workers in the crowd for a certain HIT by modeling workers based on their social network profiles [11]. This requires profiling and harvesting knowledge about worker interests [25].

Another dimension to take into account to improve quality is *trust*. Trust in social networks has been studied (e.g., [33]) and can be leveraged, for instance, to rank social network users based on trust and let only highly trusted workers complete HITs to ensure high-quality results.

3.2. Human Incentives

Along this line, another aspect which is unique of hybrid human-machine systems is the need to design proper incentives to motivate a crowd of human individuals to preform HITs to support the system. While the most common incentive is the financial one, other incentives can be used (e.g., gamification of the task). In [45] authors study the effect of pay over annotation tasks (i.e., planet discovery in telescope images) with varying difficulty. They observe comparable performances between paid and volunteering workers. However, such finding may not generalize to other tasks or settings as in this case volunteers may have an intrinsic motivation in completing the task accurately.

Another relevant work is [55] where authors compare paid workers and volunteers on the same task showing how the quality of the work done is comparable while paying workers can lead to faster results. They also study how different paying schemes can be used as different trade-offs between speed and quality. Another work on payment schemes for crowdsourced tasks is [23] where authors show that appropriate pricing approaches can be used to retain workers longer on the tasks thus improving the overall latency of the human component in hybrid human-machine systems (see Section 4.1.2).

Other ways to improve worker performance on the long-term are described in [39] where authors envision the future of crowd work including long-term career paths for crowd workers with better recognition of experience and expertise.

3.3. Cost/Quality Trade-Off

There is a clear cost/quality tradeoff where better data quality can be obtained by spending more on crowdsourcing resources. Thus, when processing certain data with a hybrid human-machine system based on a paid micro-task crowdsourcing platform like Amazon MTurk, a maximum monetary budget is often defined. The question at this point is how to most effectively allocate the available budget. In [60] authors study algorithms that trade-off cost for accuracy. The proposed approach estimates crowd error rate assuming that assigning the same HIT to different workers improves the quality of the final answer at an higher cost.

3.4. Crowdsourcing Scalability

Another obvious aspect to take into account in hybrid human-machine systems is that not all data can be crowdsourced. When large datasets need to be processed, it is unfeasible to send it entirely to the crowd. Thus, the challenge becomes how to optimally select the data items to be crowdsourced so that both the human intelligence cost and the benefit are optimized.

In other cases data may not be outsourced to the general public due to privacy issues. For example, personal medical data should not be published on micro-task crowdsourcing platforms and sensitive enterprise data has to be kept confidential. For these cases, a dedicated crowd may be used to process data in a hybrid human-machine fashion (see Section 4.3.2).

3.5. Crowdsourcing Efficiency and Termination

It has been shown that large batches of tasks in crowdsourcing platforms attract more workers [28]. This behavior leads to *batch starvation*, that is, batches with few tasks left that attract no worker and thus, remain uncompleted. This is a problem for batch data processing jobs that require all HITs to be completed in order to be solved and thus, this extends their execution time. It then becomes critical to retain workers on a batch of tasks. A possible way to do this is by means of ad-hoc pricing schemes [23].

Real-time crowdsourcing needs fast responses from crowd workers. Research in this direction has shown that collaborative environments help crowd workers obtain results faster as compared to when they work in isolation [43].

4. Research Opportunities

As we have seen so far, Human Computation and Crowdsourcing allow to create hybrid human-machine systems able to do quality data processing at scale. This type of systems open many opportunities for better big data processing pipelines (see Section 2). However, such systems have to be designed carefully taking into account both machine architecture as well as the human aspects at the same time (see Section 3).

In order to make such systems consistently efficient, effective, and scalable in the future, we see a set of open research questions that need to be tackled.

4.1. Improving System Efficiency

In this section we present research opportunities to make hybrid human-machine systems more efficient.

4.1.1. Incentive Design

While the monetary reward is what motivates workers to complete tasks in paid crowdsourcing platforms, this may be not enough as an incentive for good work. Well designed HITs are necessary to retain and motivate workers to work on certain tasks. This is even more important when competing HITs are, in parallel, attracting worker attention on the crowdsourcing platform. A novel set of incentives should be designed as for example, different payment schemes as well as captivating HIT interfaces and gamification techniques embedded into the crowdsourcing platforms.

In the area of gamification for crowdsourcing, Galaxy Zoo is a popular example. This platform is used to manually label space images from a telescope to classify galaxies within a set of predefined categories. Studies have shown that non-expert crowds can perform better then experts and machine-based algorithms at this tack [35]. In [46] authors study the engagement of workers on Galaxy Zoo building prediction models for worker abandonment and design actions to be taken in such cases (e.g., showing interesting tasks to bored workers who are predicted to leave the system soon). This type of research is key to make crowd work better on the long term and has still to be done for paid crowdsourcing as well. This will positively affect both the worker experience as well as the overall hybrid human-machine system quality and latency.

In the area of paid crowdsourcing, pricing schemes have still to be investigated in detail. Early and recent work include [32] where authors design models to set the price of HITs varying it over time to deal with system-side deadlines and budget constrains. Also, pricing schemes seem to have a clear influence on crowd work efficiency [23].

4.1.2. The Latency of the Crowd

Crowdsourced components are obviously the bottleneck in hybrid human-machine systems when we look at data processing speed. While batch data processing done by means of crowdsourcing has no fast execution requirements, real-time crowdsourcing is a necessity for various interactive applications that require human intelligence at scale. Example applications which require fast reaction from the crowd include real-time captioning of speech [42], crowd-powered personal assistants [43], and video filtering [5].

In order to make such systems close to real-time execution we need radically different crowdsourcing platforms which are able to support on-demand work requests using, for example, direct notification to available workers who are redirected to high-priority HITs. Such HITs need to benefit from scheduling algorithms that optimize their execution by assigning them to skilled workers. This will make crowdsourcing platforms much more controlled systems rather than self-organized markets as most of the current crowdsourcing platforms. A first step in this direction is represented by push crowdsourcing platforms [25].

4.2. Improving System Effectiveness

In this section we discuss research opportunities to improve hybrid human-machine system effectiveness.

4.2.1. Worker Career Development and Skill Acquisition

Having a crowd of workers available 24/7 for data processing task is an excellent tool that can be used to build the systems described in this paper. However, on the long term, in order to achieve even more advanced data processing and moving even further from what machine-based algorithms can do, there is the need to involve experts. As involving domain experts can result in costly and slow processes, an alternative option is to develop relevant skills in crowd workers and let them invest on their crowd worker career. Works in this direction have started to look at how to record worker experiences by creating a worker curriculum collecting all completed tasks and qualifications over different platforms [57]. Novel appropriate personal development and training schemes need to be designed for the crowdsourcing setup. This will then lead to research questions related to talent retention as the most skilled workers may become highly requested within a platform and become a limited resource that has an impact on both system efficiency and effectiveness.

4.2.2. Influencing Human Behavior

As humans are integral part of hybrid human-machine systems, it is important to make them perform effectively. Thus, a better understanding on how to positively influence crowd worker behavior is needed to improve the overall quality of hybrid human-machine systems. Example research questions include the understanding of how worker behavior is influenced by external signals like agreement rate, other worker answers. Moreover, understanding how, in such cases, workers would tend to agree with the majority or rather maintain their own point of view about which is the correct answer to the HIT is necessary. As previously shown in [27] human behavior may be influenced by such signals and it is still unclear how this affects the overall quality of hybrid human-machine systems.

Moreover, leveraging existing social networks is a way of influencing working patterns. For example, by showing most popular HITs among peers in the social network, it could be possible to motivate workers better and direct them towards certain high priority tasks.

4.2.3. Malicious Workers

Poor quality in hybrid systems is due either by the inaccuracies of machine-based algorithms or by low quality crowd answers. While the crowd may perform poorly for a variety of reasons (poor task instructions, lack of knowledge) a big concern comes from those workers who maliciously perform paid tasks to obtain the monetary reward attached to them without completing the task with care [24].

Another open question then is how to effectively identify such workers and remove them from the system. Possible options include, worker monitoring, e.g., with mouse tracking, and the use of honeypots (i.e., questions for which a ground truth is available) [6], qualification tests [3], or screening questions [26] to check the trust level of workers. Anyway, such basic approaches seem highly vulnerable to attacks of organized teams of workers.

4.2.4. Task Design

Previous work has shown that task design has an impact on crowd work quality and, thus on hybrid human-machine system effectiveness [15]. An open research question is how to best design HITs in order to optimize for crowd work quality. To this end, user studies and experimental comparisons of different designs for different task types are necessary.

4.3. Improving System Scalability

In this section we discuss research directions that aim at making hybrid human-machine systems applicable at scale over large datasets and diverse sets of problems.

4.3.1. Complex Workflows and Models

When multiple crowdsourced tasks interact with different machine-based components, the orchestration of human and machine processing has to be properly managed to avoid delays. With this goal in mind, human computation workflows have been proposed. For example, in [10] authors propose workflow patterns decomposing complex tasks into simpler ones. Another example is [40] where authors describe a system to design crowdsourcing workflows where complex tasks are decomposed in simpler ones and assigned to workers in the crowd. When the requester participates in the workflow design the quality of the result increases.

These systems however do not yet include the design of hybrid human-machine workflows where several machine-based data processing components are interleaved by crowdsourcing steps. Such complex hybrid workflows open questions of resource optimization, latency, and overall quality at a bigger scale.

Another current research direction is the design of hybrid human-machine systems in a model-driven manner. In [14] authors define a general model for human-machine systems based on user modeling work and social networks. They describe three instantiations of the model in different application domains: 1) multimedia content processing and querying leveraging social networks to retrieve content; 2) general search over social networks [8]; and 3) on-line game event notification.

4.3.2. Enterprise Crowdsourcing

Because of the impossibility of releasing data publicly, large data-driven companies have started to run crowdsourcing tasks internally [65]. They have developed and deployed crowdsourcing platforms and leverage employees as a crowd of people. Such crowd is knowledgeable about the specific business domain and performs data analysis task in combination with data analytics techniques in a hybrid human-machine fashion.

In this context we can find many open research opportunities as the setting is different than the traditional one: on one hand we can assume the absence of malicious workers as company employees will not play an adversarial role; on the other hand there is the need to rethink crowd incentives beyond the financial ones as employees will need to dedicate some time for micro-task completion in addition to their standard job tasks.

4.3.3. Hybrid Human-Machine Systems Applied to Data-Driven Sciences

Data is nowadays a critical asset to do fundamental science work. Examples include physics, biology, health, chemistry. In all these cases, data is the means to scientific discovery and the data volume available to scientists is growing at an extremely high rate. In such context, high data quality is key. Better data understanding can lead to new scientific discoveries. For these reasons, it would very much make sense to design and use hybrid human-machine systems to support research and discovery in fundamental sciences. While expertise may be necessary to deal with scientific data, previous work [48] has shown how, with the appropriate support, anonymous crowd of non-experts can effectively perform micro-tasks related to the health domain. One main reason why this is has not been done yet for other domains is the need for domain knowledge in designing the system. However, the current trend towards interdisciplinary research plays in favor of hybrid human-machine systems being successfully designed also for other sciences.

4.4. Open Research Questions

As discussed in Section 2.7, after developing a variety of hybrid human-machine systems, different research communities have started to focus on improving the weak points identified while developing and evaluating such systems. In the following we summarize the open research questions related to the improvement of hybrid human-machine systems identified in this paper:

- Which pricing schemes are most appropriate to attract and motivate crowd workers in the long term?

- Can task routing and worker notification improve efficiency of real-time hybrid human-machine systems?

- What is the best method to track worker achievements, port them across platforms, and to develop worker profiles and skills over time?
- Which external information should be provided to workers to positively influence their work?
- How can we automatically identify malicious workers in crowdsourcing platforms?
- How can we define optimal task design guidelines for different task types?
- Can we automatize the design of hybrid human-machine workflows?
- Which are the most appropriate incentive, task designs, and task routing approaches for enterprise crowdsourcing?
- Which information should we to provide to non-expert workers when crowdsourcing domain-specific tasks?

5. Conclusions

In this paper we have presented an overview of recent works from different Computer Science areas looking at the design of hybrid human-machine systems to process data in both a scalable as well as effective manner. We have summarized recent efforts in the domains of databases, information retrieval, and semantic web also looking at examples from other disciplines. We have highlighted the challenges that hybrid system designers have to face when building such novel systems and outlined open research directions that will make such hybrid human-machine systems improve the quality of available data and foster progress in different data-driven sciences.

6. References

[1] Alonso, O. and Baeza-Yates, R. A. (2011). Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 153–164.

[2] Alonso, O. and Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manage.*, 48(6):1053–1066.

[3] Alonso, O., Rose, D. E., and Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

[4] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*, pages 2670–2676.

[5] Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. (2011). Crowds in two seconds: enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 33–42, New York, NY, USA. ACM.

[6] Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., and Tran, D. T. (2011). Repeatable and reliable search system evaluation using crowdsourcing. In *SIGIR*, pages 923–932.

[7] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

[8] Bozzon, A., Brambilla, M., and Ceri, S. (2012a). Answering search queries with crowdsearcher. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 1009–1018, New York, NY, USA. ACM.

[9] Bozzon, A., Brambilla, M., Ceri, S., and Mauri, A. (2012b). Extending Search to Crowds: A Model-Driven Approach. In *SeCO Book*, pages 207–222.

[10] Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., and Volonterio, R. (2014a). Pattern-based specification of crowdsourcing applications. In *Web Engineering, 14th International Conference, ICWE 2014, Toulouse, France, July 1-4, 2014. Proceedings*, pages 218–235.

[11] Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., and Vesci, G. (2013). Choosing the right crowd: Expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 637–648, New York, NY, USA. ACM.

[12] Bozzon, A., Brambilla, M., and Mauri, A. (2012c). A Model-Driven Approach for Crowdsourcing Search. In *CrowdSearch*, pages 31–35.

[13] Bozzon, A., Catallo, I., Ciceri, E., Fraternali, P., Martinenghi, D., and Tagliasacchi, M. (2012d). A framework for crowdsourced multimedia processing and querying. In *Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, April 17, 2012*, pages 42–47.

[14] Bozzon, A., Fraternali, P., Galli, L., and Karam, R. (2014b). Modeling crowdsourcing scenarios in socially-enabled human computation applications. *J. Data Semantics*, 3(3):169–188.

[15] Carvalho, V. R., Lease, M., and Yilmaz, E. (2011). Crowdsourcing for search evaluation. *SIGIR Forum*, 44(2):17–22.

[16] Clough, P. D., Sanderson, M., Tang, J., Gollins, T., and Warner, A. (2013). Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing*, 17(4):32–38.

[17] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *ACL*, pages 168–175.

[18] Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

[19] Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. (2012). Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 469–478, New York, NY, USA. ACM.

[20] Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. (2013a). Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *VLDB J.*, 22(5):665–687.

[21] Demartini, G., Trushkowsky, B., Kraska, T., and Franklin, M. J. (2013b). Crowdq: Crowdsourced query understanding. In *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings.*

[22] Diaz-Aviles, E. and Kawase, R. (2012). Exploiting Twitter as a Social Channel for Human Computation. In *CrowdSearch*, pages 15–19.

[23] Difallah, D. E., Catasta, M., Demartini, G., and Cudré-Mauroux, P. (2014). Scaling-up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[24] Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2012). Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In *Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, April 17, 2012*, pages 26–30.

[25] Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2013). Pick-a-crowd: Tell me what you like, and i'll tell you what to do. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 367–374, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[26] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system?: Screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2399–2402, New York, NY, USA. ACM.

[27] Dumitrache, A., Aroyo, L., Welty, C., Sips, R.-J., and Levas, A. (2013). "Dr. Detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *CrowdSem*, pages 16–31.

[28] Faradani, S., Hartmann, B., and Ipeirotis, P. G. (2011). What's the right price? pricing tasks for finishing on time. In *Human Computation, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011.*

[29] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

[30] Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). Crowddb: Answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 61–72, New York, NY, USA. ACM.

[31] Gadiraju, U., Kawase, R., and Dietze, S. (2014). A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 218–223, New York, NY, USA. ACM.

[32] Gao, Y. and Parameswaran, A. G. (2014). Finish them!: Pricing algorithms for human computation. *PVLDB*, 7(14):1965–1976.

[33] Golbeck, J. A. (2005). *Computing and applying trust in web-based social networks*. PhD thesis, University of Maryland at College Park, College Park, MD, USA. AAI3178583.

[34] Goldberg, S. L., Wang, D. Z., and Kraska, T. (2013). Castle: Crowd-assisted system for text labeling and extraction. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[35] Hand, E. (2010). Citizen science: People power. *Nature*, 466(7307):685687.

[36] Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. *ACM Crossroads*, 17(2):16–21.

[37] Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 165–176.

[38] Kazai, G., Kamps, J., and Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.*, 16(2):138–178.

[39] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA. ACM.

[40] Kulkarni, A., Can, M., and Hartmann, B. (2012). Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1003–1012, New York, NY, USA. ACM.

[41] Kumar, R. and Tomkins, A. (2010). A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 561–570, New York, NY, USA. ACM.

[42] Kushalnagar, R. S., Lasecki, W. S., and Bigham, J. P. (2012). A readability evaluation of real-time crowd captions in the classroom. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '12, pages 71–78, New York, NY, USA. ACM.

[43] Lasecki, W. S. (2013). Real-time conversational crowd assistants. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 2725–2730, New York, NY, USA. ACM.

[44] Lofi, C., El Maarry, K., and Balke, W.-T. (2013). Skyline queries in crowd-enabled databases. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 465–476, New York, NY, USA. ACM.

[45] Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M. E., Lintott, C. J., and Smith, A. M. (2013a). Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *HCOMP*.

[46] Mao, A., Kamar, E., and Horvitz, E. (2013b). Why stop now? predicting worker engagement in online crowdsourcing. In *HCOMP*.

[47] Mason, W. and Watts, D. J. (2009). Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85, New York, NY, USA. ACM.

[48] Mortensen, J., Musen, M. A., and Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA 2013 Annual Symposium.*

[49] Nieke, C., Güntzer, U., and Balke, W. (2014). Topcrowd - efficient crowd-enabled top-k retrieval on incomplete data. In *Conceptual Modeling*

- 33rd International Conference, ER 2014, Atlanta, GA, USA, October 27-29, 2014. Proceedings, pages 122–135.

[50] Oliveira, F., Ramos, I., and Santos, L. (2010). *Definition of a crowdsourcing innovation service for the European SMEs*. Springer.

[51] Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G.-J., and Aroyo, L. (2014). Crowdsourcing knowledge-intensive tasks in cultural heritage. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 267–268, New York, NY, USA. ACM.

[52] Parameswaran, A., Sarma, A. D., Garcia-Molina, H., Polyzotis, N., and Widom, J. (2011). Human-assisted graph search: It's okay to ask questions. *Proc. VLDB Endow.*, 4(5):267–278.

[53] Parameswaran, A., Teh, M. H., Garcia-Molina, H., and Widom, J. (2014). Datasift: A crowd-powered search toolkit. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 885–888, New York, NY, USA. ACM.

[54] Parameswaran, A. G., Garcia-Molina, H., Park, H., Polyzotis, N., Ramesh, A., and Widom, J. (2012). Crowdscreen: Algorithms for filtering data with humans. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 361–372, New York, NY, USA. ACM.

[55] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. (2011). An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011.*

[56] Sarasua, C., Simperl, E., and Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 525–541, Berlin, Heidelberg. Springer-Verlag.

[57] Sarasua, C. and Thimm, M. (2014). Crowd Work CV: Recognition for Micro Work. In *Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation (SoHuman'14).*

[58] Shadbolt, N. R., Smith, D. A., Simperl, E., Van Kleek, M., Yang, Y., and Hall, W. (2013). Towards a classification framework for social machines. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 905–912, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[59] Sheshadri, A. and Lease, M. (2013). SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, pages 156–164.

[60] Tran-Thanh, L., Venanzi, M., Rogers, A., and Jennings, N. R. (2013). Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 901–908, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[61] Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 155–164, New York, NY, USA. ACM.

[62] von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM*, 51(8):58–67.

[63] von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 55–64, New York, NY, USA. ACM.

[64] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.

[65] Vukovic, M. and Natarajan, A. (2013). Operational Excellence in IT Services Using Enterprise Crowdsourcing. In *IEEE SCC*, pages 494–501.

[66] Wang, J., Kraska, T., Franklin, M. J., and Feng, J. (2012). Crowder: Crowdsourcing entity resolution. *Proc. VLDB Endow.*, 5(11):1483–1494.

[67] Whang, S. E., Lofgren, P., and Garcia-Molina, H. (2013). Question selection for crowd entity resolution. *PVLDB*, 6(6):349–360.

[68] Whang, S. E., McAuley, J., and Garcia-Molina, H. (2012). Compare Me Maybe: Crowd Entity Resolution Interfaces. In *Technical Report*. Stanford InfoLab.

Gianluca Demartini is a Lecturer in Data Science at the Information School of the University of Sheffield, UK. Previously, he was post-doctoral researcher at the eXascale Infolab at the University of Fribourg, visiting researcher at UC Berkeley, junior researcher at the L3S Research Center, and intern at Yahoo! Research. His research interests include Web Information Retrieval, Semantic Web, and Human Computation. He obtained a Ph.D. in Computer Science at the Leibniz University of Hannover in Germany focusing on Entity Retrieval. He has published more than 50 peer-reviewed scientific publications and given tutorials about Entity Retrieval and Crowdsourcing at research conferences.