Research Statement: Human-Powered Information Management

Aditya Parameswaran (www.stanford.edu/∼adityagp)

My research broadly revolves around information management, with special emphasis on incorporating "human computation", together with traditional computation, to improve the process of gathering, understanding, and managing data.

With the amount of data recorded or produced every minute — 100,000 tweets transmitted, 200,000 queries issued on Google, and 48 hours of video shared on YouTube — it is critical to quickly process and understand this data in order to enable data-driven applications. For example, during hurricane Sandy, around 400,000 images were uploaded to Instagram, an image sharing site. These images, if properly understood, could help identify areas that are badly affected, how much water has accumulated, or how many people need assistance.

However, this recorded or produced data can be "messy": either not amenable to algorithmic analysis, or algorithms to fully comprehend the data have not been developed. For this reason, using humans to analyze certain aspects of the data can be crucial. Humans have an innate understanding of language, speech, and images; they are able to process, reason about, and provide solutions to the problems faced often in data management. For instance, humans may be able to identify if two records refer to the same person (entity resolution), or look up phone numbers of a given restaurant (missing data), when it is too hard for a computer to do so. Moreover, the abundance of cheap and reliable internet connectivity throughout the world has given rise to *crowdsourcing* marketplaces, such as Mechanical Turk and ODesk, which enable human computation on demand.

Unfortunately, human computation can be subjective or error-prone, time-consuming (humans take longer than computers), and relatively costly (humans need to be paid). Moreover, these three aspects—*accuracy*, *latency*, and *cost*—are correlated in complex ways, making it difficult for crowdsourcing application developers to manually optimize the trade-offs among them. In my work, I have addressed fundamental problems in human-powered information management, focusing on solutions that optimize across accuracy, latency, and cost. These solutions enable developers to get significant improvements over unoptimized approaches.

My research methodology is to first identify real-world problems to be solved by systems. I use an analytical approach to conceptually model these problems, then an algorithmic approach to come up with principled solutions with worst-case guarantees. Then, I use experimentation and real world systems deployment to test the effectiveness of the solutions. As an algorithms person with a bent towards information management problems, my strength is in coming up with principled solutions to real problems.

In addition to human computation, I have applied my research approach to some other problems in information management, including information extraction and recommendations, described briefly below.

My research in human-powered information management focuses on four (not necessarily mutually exclusive) themes, discussed in more detail next:

- Data Processing [4, 8, 9, 10, 11, 18]
- Data Gathering [1, 3, 5, 6, 19]
- Data Extraction [2, 17]
- Data Quality [7, 12]

Data Processing: A basic form of human computation is when humans act as "data processors", i.e., when human involvement can be abstracted as functions or subroutines applied on data elements. In our work, we developed some basic human-powered data processing algorithms; I call them *crowd algorithms*, for example:

(a) *Max*: We considered the problem of finding the best item out of a set of items, e.g., the best profile photo out of a set of Facebook photos, where humans can compare pairs of items. We modeled the problem as a maximum likelihood problem, proved intractability, and provided practical algorithms [10].

(b) *Filtering*: We studied how we can best apply a filter to a large set of items when only humans can evaluate the filter, e.g., identify all images that contain inappropriate content from a set of user-uploaded images. We designed algorithms to minimize cost, while keeping overall error low. Our algorithms are akin to policies for Markov Decision Processes. Our experiments demonstrated a saving of over 20% in costs over other more obvious approaches [11].

(c) *Deduplication*: We evaluated how humans may assist in developing better classifiers for identifying duplicate records. We provided the first solution for solving this problem with guarantees, leveraging prior work in active learning [8].

We are currently extending our filtering work to an emerging application domain: peer evaluation in massive online courses. In initial experiments on the HCI course at Stanford, we find that our algorithms can reduce the number of peer evaluations by a factor of two while minimally impacting the quality of the assigned scores.

Data Gathering: Human workers can also be regarded as a "data source", providing information that is either already known to them (e.g., What is the capital of Egypt?), or that they can easily look up (e.g., What is the phone number of Amazon customer service?). We are developing a database system, called *Deco* (derived from "Declarative crowdsourcing"), that models and offers for querying the collective knowledge of the crowd as a relational-like database [3, 5, 19]. Moreover, Deco seamlessly combines information crowdsourced on-the-fly with traditional stored data. Using Deco, the application developer can implement crowdsourced computation using declarative queries against the database. In turn, Deco is responsible for transparently managing and optimizing the details of the interaction with the human workers. Deco has a principled foundation as well as a sophisticated query processor with cost-based planning and optimization [1, 6].

Data Extraction: Extracting information from web pages can be a challenging task, due to their diversity in structure and the lack of accuracy of natural language understanding algorithms. Human input can be very helpful in identifying interesting locations on web pages (e.g., the location of phone numbers of restaurants on Yelp), to enable programmatic extraction from many similar web pages. However, even if humans identify where interesting content is located, the web page structure may change over time, and these pointers may no longer be valid. To deal with web page evolution, we designed extraction schemes that are optimally fault-tolerant: they ask humans to assist extraction only when absolutely necessary. In experiments, our techniques beat the state of the art tenfold [17]. Our extraction schemes were deployed in the internal Yahoo! extraction pipeline.

Data Quality: An important concern in crowdsourcing is the quality of answers provided by humans, determined by the expertise or inherent ability of the humans. Recently, we have started exploring various aspects of human worker quality control, such as determining confidence intervals for individual workers to enable us to get better guarantees on worker and task quality [7], and evaluating workers while they are working on tasks, based on agreement between workers [12].

Selected Projects in Other Areas

Course Recommendations [15, 16, 20, 22, 25]: Early in my PhD I worked on course recommendation systems. Course recommendations pose an especially tricky problem — courses have prerequisites, students need to meet graduation requirements, temporality of courses is important, and students take courses for a variety of reasons: difficulty, interest, lenient grading, or peer pressure. Our course recommendation algorithms were integrated within *CourseRank* [24], a social course recommendation site developed in our lab at Stanford then launched as a commercial product. The CourseRank technology was deployed in over 500 universities in the US.

Concept Extraction and Categorization [18, 21]: Identifying concepts is important for web search because it allows search engines to provide concept-relevant content. For instance, it is important to identify that "Occupy" is a news-worthy concept, so that a web search engine can possibly show snippets from news sources as part of the search results. In collaboration with researchers at Kosmix, I developed a scalable technique to automatically extract new concepts from large data sets — news feeds, tweets, query logs — and then attach them to the taxonomy of concepts. My technique was deployed within Kosmix to find and attach "trending" concepts, keeping the concept taxonomy up-to-date.

Future Research

As in my prior work, my goal is to take a principled, algorithmic approach to systems research, in order to produce systems with formal guarantees. I plan to address some of the remaining unsolved problems in human-powered information management, as well as branch out into other areas of data analytics.

Pushing the Boundaries of Human-Powered Information Management: As more people opt for flexible employment online, crowdsourcing is bound to become increasingly important and increasingly complex in the coming decade. Additionally, a growing number of subfields within Computer Science (e.g., computer vision and bioinformatics) are starting to employ human computation. To further foster and simplify development of applications that harness human computation, here are four important issues I plan to address:

- *How do we reason about interfaces?* The interfaces used for the human worker play a key role in determining not just the quality of the returned results, but also how many questions are required to solve a task. For

instance, to find the best item out of a set of items, we could ask for comparisons among items, or ask for ratings on individual items. The key question I plan to address is: How can we theoretically model interfaces, and how do we choose between them while solving human computation problems?

- *How do we model collaboration?* Some tasks are better done by having humans collaborate than by asking them independently and combining the answers. It is not clear how we should model collaboration, or how we should identify the "optimal" manner of humans working on a given problem, given properties about human workers.
- *How do we move beyond micro-tasks?* Most research in human computation uses the "micro-task" model (pay-per-small-task), which is easy to understand and model. However, a number of crowdsourcing marketplaces use a pay-per-hour model, which is less prone to workers answering questions rapidly (but incorrectly)—instead, workers spend time as they see fit answering questions, while still getting paid. Designing crowd algorithms for pay-per-hour, rather than pay-per-task, necessitates rethinking the basic models of human data processing we have used in our research so far.
- *How do we track expertise and perform efficient matching?* Current crowdsourcing marketplaces use simple methods for matching skilled workers to task requesters. As a result, the quality of worker output is not as good as it could potentially be. I plan to work on the algorithmic problem of identifying and maintaining skill-sets of workers (extracted from past history or manually input), and performing automated recommendations of workers when a new task arrives, while ensuring that there are workers with diverse skills left for subsequent tasks.

Interactive Data Analytics: A McKinsey Global Institute "Big Data" study estimates that to analyze the data being generated today across various sectors, up to a million new analysts will be needed in the US alone in the next two years. However, only a limited set of interactive tools are available for these novice data analysts to process, analyze, and gain insights from data. I plan to work towards understanding the fundamental principles underlying interactive analytics, with the ultimate goal of improving the tools that empower data analysts. Here are four questions I would like to address:

- *How can we help preprocess data?* While there are some tools that help analysts prepare their data for analysis, it is still an unwieldy process, especially when data values are missing, schema information absent, relationships unknown, and when data is scattered across text files. I plan to develop methods that ask an adaptive series of questions to the analyst (or to human workers) to transform the data into a structured form amenable to analysis.
- *How can we help analysts pose queries?* In many domains, e.g., finance or physical sciences, analysts have limited exposure to query languages, preventing them from leveraging standard database technology. I want to explore how we may guide novice data analysts towards queries they have in mind, but are unable to express. Specifically, we could ask a few "targeted" questions to the analyst, then refine the queries and results through further interaction.
- *How can we help analysts visualize query results?* Often, database query results are too unwieldy to peruse, spanning thousands of data elements. Tools like Tableau have demonstrated the utility of visualizations for getting quick summaries of data. However, no current database systems provide automatic recommendations for how to visualize the output of specific queries, highlighting interesting aspects of the query result. To enable visualizations within databases, we will need to leverage work on visualization metrics as well as rethink query optimization, especially when there is a very large number of candidate visualizations.
- *How do we embed machine learning algorithms?* Analysts often use machine learning to look for patterns in data. Currently, most machine learning is performed in an ad-hoc fashion, with imperative procedures being hard-coded for each task. While there have been advances in database systems supporting individual machine learning operations, so far there is no comprehensive system that can support and optimize declarative queries expressing a complex combination of machine learning operations. The system should maintain the learned models (along with the necessary data) and keep them up-to-date as new data arrives.

The research directions I intend to pursue require the expertise of researchers in many fields. I look forward to working closely with researchers who have expertise in systems, HCI, visualization, machine learning, game theory, and algorithms. Additionally, I plan to work with researchers outside computer science to identify typical usage patterns for data analytics in their field, be it history, public policy, or health sciences, and to determine whether human computation can be profitably incorporated. I hope to bring to these collaborations connections to well-established principles underlying information management and human computation, and also knowledge about how large-scale data analysis works in practice.

References

[1] H. Park, R. Pang, **Aditya Parameswaran**, H. Garcia-Molina, N. Polyzotis, and J. Widom. An Overview of the Deco System: Data Model and Query Language; Query Processing and Optimization, *SIGMOD Record*, Volume 41, December 2012.

[2] N. Dalvi, **Aditya Parameswaran**, and V. Rastogi. Minimizing Uncertainty in Pipelines, *NIPS '12: 25th Int'l Conf. on Neural Information Processing Systems*, Tahoe, Nevada, USA, 2012.

[3] **Aditya Parameswaran**, H. Park, H. Garcia-Molina, N. Polyzotis, and J. Widom. Deco: Declarative Crowdsourcing, *CIKM '12: 21st Int'l Conf. on Information and Knowledge Management*, Maui, USA, 2012. Acceptance Rate: 13.4%.

[4] K. Bellare, S. Iyengar, **Aditya Parameswaran**, and V. Rastogi. Active Sampling for Entity Matching with Guarantees, Infolab Technical Report, September 2012.

[5] H. Park, R. Pang, **Aditya Parameswaran**, H. Garcia-Molina, N. Polyzotis, and J. Widom. Deco: A System for Declarative Crowdsourcing (Demo), *VLDB '12: 38th Int'l Conf on Very Large Data Bases*, Istanbul, Turkey, 2012.

[6] H. Park, **Aditya Parameswaran**, and J. Widom. Query Processing over Crowdsourced Data, Infolab Technical Report, August 2012.

[7] M. Joglekar, H. Garcia-Molina, and **Aditya Parameswaran**. Evaluating the Crowd with Confidence, Infolab Technical Report, August 2012.

[8] K. Bellare, S. Iyengar, **Aditya Parameswaran**, and V. Rastogi. Active Sampling for Entity Matching, *KDD '12: 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012. Acceptance Rate: 18%. Invited to **Special Issue of TKDD for KDD 2012 Best Papers**.

[9] A. Das Sarma, **Aditya Parameswaran**, H. Garcia-Molina, and A. Halevy. Finding with the Crowd, Infolab Technical Report, July 2012.

[10] S. Guo, **Aditya Parameswaran**, and H. Garcia-Molina. So Who Won? Dynamic Max Discovery with the Crowd, *SIGMOD '12: ACM SIGMOD Int'l Conf. on the Management of Data*, Scottsdale, USA, 2012. Acceptance Rate: 17%.

[11] **Aditya Parameswaran**, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom. CrowdScreen: Algorithms for Filtering Data with Humans *SIGMOD '12: ACM SIGMOD Int'l Conf. on the Management of Data*, Scottsdale, USA, 2012. Acceptance Rate: 17%.

[12] A. Ramesh, **Aditya Parameswaran**, H. Garcia-Molina, and N. Polyzotis. Identifying Reliable Workers Swiftly, Infolab Technical Report, June 2012.

[13] F. Afrati, A. Das Sarma, D. Menestrina, **Aditya Parameswaran**, and J. D. Ullman. Fuzzy joins using MapReduce, *ICDE '12: 28th Int'l Conf. on Data Engineering*, Washington DC, USA, 2012. Acceptance Rate: 24%.

[14] **Aditya Parameswaran**, R. Kaushik, and A. Arasu. Efficient Parsing-based Keyword Search over Databases, Infolab Technical Report, February 2012.

[15] G. Koutrika, H. Garcia-Molina, and **Aditya Parameswaran**. Information Seeking: Convergence of Search, Recommendations, and Advertising, *Communications of the ACM (CACM)*, November 2011.

[16] **Aditya Parameswaran**, P. Venetis, and H. Garcia-Molina. Recommendation Systems with Complex Constraints: A Course Recommendation Perspective, *Transactions on Information Systems (TOIS)*, Volume 29(4), November 2011.

[17] **Aditya Parameswaran**, N. Dalvi, H. Garcia-Molina, and R. Rastogi. Optimal Schemes for Robust Web Extraction, *VLDB '11: 37th Int'l Conf. on Very Large Data Bases*, Seattle, USA, 2011. Acceptance Rate: 18.1%.

[18] **Aditya Parameswaran**, A. Das Sarma, H. Garcia-Molina, N. Polyzotis, and J. Widom. Human-assisted Graph Search: It's Okay to Ask Questions, *VLDB '11: 37th Int'l Conf. on Very Large Data Bases*, Seattle, USA, 2011. Acceptance Rate: 18.1%.

[19] **Aditya Parameswaran** and N. Polyzotis. Answering Queries using Databases, Humans, and Algorithms, *CIDR '11: Conf. on Innovative Data Management (CIDR)*, Asilomar, USA, 2011.

[20] **Aditya Parameswaran**, H. Garcia-Molina, and J. D. Ullman. Evaluating, Combining, and Generalizing Recommendations with Prerequisites, *CIKM '10: 19th Int'l Conf. on Information and Knowledge Management*, Toronto, Canada, 2010. Acceptance Rate: 13%.

[21] **Aditya Parameswaran**, H. Garcia-Molina, and A. Rajaraman. Towards the Web of Concepts: Extracting Concepts from Large Datasets, *VLDB '10: 36th Int'l Conf. on Very Large Data Bases*, Singapore, 2010. Acceptance Rate: 18.4%. Invited to **Special Issue of VLDB Journal for VLDB 2010 Best Papers**.

[22] **Aditya Parameswaran**, G. Koutrika, B. Berkovitz, and H. Garcia-Molina. Recsplorer: Recommendation Algorithms Based on Precedence Mining, *SIGMOD '10: ACM SIGMOD Int'l Conf. on the Management of Data*, Indianapolis, USA, 2010. Acceptance Rate: 21%.

[23] A. Das Sarma, **Aditya Parameswaran**, H. Garcia-Molina, and J. Widom. Synthesizing View Definitions from Data, *ICDT '10: 13th Int'l Conf. on Database Theory*, Lausanne, Switzerland, 2010. Acceptance Rate: 36%.

[24] B. Berkovitz, F. Kaliszan, G. Koutrika, H. Liou, **Aditya Parameswaran**, P. Venetis, Z. Zadeh, and H. Garcia-Molina. Social Sites Research Through CourseRank, *SIGMOD Record*, Volume 38, December 2009.

[25] **Aditya Parameswaran** and H. Garcia-Molina. Recommendations with Prerequisites (Short Paper) *RecSys '09: 3rd ACM Conf. on Recommender Systems*, New York, USA, 2009. Acceptance Rate: 43%.

[26] E. Sadikov, **Aditya Parameswaran**, and P. Venetis. Blogs as Predictors of Movie Success, *ICWSM '09: AAAI Conf. on Weblogs and Social Media*, San Jose, USA, 2009.