Towards Collaborative Data Analysis with Diverse Crowds – a Design Science Approach

Feldman, Michael ; Anastasiu, Cristian ; Bernstein, Abraham

Abstract: The last years have witnessed an increasing shortage of data experts capable of analyzing the omnipresent data and producing meaningful insights. Furthermore, some data scientists mention data preprocessing to take up to 80% of the whole project time. This paper proposes a method for collaborative data analysis that involves a crowd without data analysis expertise. Orchestrated by an expert, the team of novices conducts data analysis through iterative refinement of results up to its successful completion. To evaluate the proposed method, we implemented a tool that supports collaborative data analysis for teams with mixed level of expertise. Our evaluation demonstrates that with proper guidance data analysis tasks, especially preprocessing, can be distributed and successfully accomplished by non-experts. Using the design science approach, iterative development also revealed some important features for the collaboration tool, such as support for dynamic development, code deliberation, and project journal. As such we pave the way for building tools that can leverage the crowd to address the shortage of data analysts.

DOI: https://doi.org/10.1007/978-3-319-91800-6_15

Originally published at:
Feldman, Michael; Anastasiu, Cristian; Bernstein, Abraham (2018). Towards Collaborative Data Analysis with Diverse Crowds – a Design Science Approach. In: 13th International Conference on Design Science Research in Information Systems and Technology, Madras, 3 June 2018 - 6 June 2018.
DOI: https://doi.org/10.1007/978-3-319-91800-6_15

Towards Collaborative Data Analysis with Diverse Crowds – a Design Science Approach

†Michael Feldman[1], Cristian Anastasiu[1], and Abraham Bernstein[1]

[1]Department of Informatics, University of Zurich, Zurich, Switzerland

feldman@ifi.uzh.ch; cristiananastasiu@googlemail.com; bernstein@ifi.uzh.ch

Abstract
The last years have witnessed an increasing shortage of data experts capable of analyzing the omnipresent data and producing meaningful insights. Furthermore, some data scientists mention data preprocessing to take up to 80% of the whole project time. This paper proposes a method for collaborative data analysis that involves a crowd without data analysis expertise. Orchestrated by an expert, the team of novices conducts data analysis through iterative refinement of results up to its successful completion. To evaluate the proposed method, we implemented a tool that supports collaborative data analysis for teams with mixed level of expertise. Our evaluation demonstrates that with proper guidance data analysis tasks, especially preprocessing, can be distributed and successfully accomplished by non-experts. Using the design science approach, iterative development also revealed some important features for the collaboration tool, such as support for dynamic development, code deliberation, and project journal. As such we pave the way for building tools that can leverage the crowd to address the shortage of data analysts.

Keywords: Collaborative data analysis, Crowdsourcing, Design Science

1 Introduction

Data analysis is a complex task that touches on many skills. Experts conducting data analysis are, therefore, expected to be proficient not only in the domain of their interest, but also in other disciplines such as statistics, computing, software engineering, and algorithms [1]. These high expectations make data scientist scarce, leaving their valuable services out-of-reach for a big share of public. This also means that the way to become data analysis expert is extremely complex and the specialization cannot be easily gained.

In this paper, we introduce an approach for collaborative data analysis *to allow non-experts to cooperate on data analysis projects*. In contrast to the lack of data scientists, there are many freelancers or enthusiasts that have some basic coding skills obtained either in introductory classes during their studies or self-acquired throughout the course of their life. While those non-experts do not have all necessary skills to perform end-to-end data analysis projects, they can be involved in some parts where their skills are sufficient. Specifically, we argue that non-experts with some coding skills can be especially helpful in the *data preprocessing* stage of data analysis. In this step data scientist transforms raw data into a data suitable for statistical modelling, as it is often inconsistent, incomplete and contains many errors. It is, therefore, likely that prior to statistical modelling, which requires significant knowledge in statistics and computer science, there is a need in "data wrangling" – transforming and editing raw data until it is suitable for data analysis [2].

At the same time, data preprocessing and the following statistical analysis cannot be decoupled. Often, in order to apply certain statistical approaches, the data has to be previously transformed and organized accordingly. For instance, to apply a statistical model that assumes linearity the dependent variable often has to be transformed first. Moreover, data analysis is an iterative process where data preprocessing and modelling are intertwined: the results of data analysis lead to new ideas on how better to analyze data, which in turn leads to additional data preprocessing. Therefore, it is important that experts and non-experts cooperate and efficiently coordinate tasks. Following these considerations, we propose a process where data analysis projects are divided into sub-tasks and each is assigned to a freelancer with limited knowledge in data analysis and (some basic) coding skills. While the participants are assigned to different tasks, they interact through various communication channels in order to draw on their collective knowledge [3], and thus, reach the desired results. Dividing the project into several simple tasks allows project manager – a data analysis expert responsible for the whole data analysis project – to distribute and coordinate the tasks. This way the manager can take advantage of various workers' abilities in order to conduct data analysis. In our experiments, we explore whether the results of non-expert teams orchestrated by manager are comparable to the results produced by experts handling the whole project. Therefore, our goal is to propose a practical solution to the problem of shortage of data scientists and allow non-experts to take part in the process of data analysis.

Our contributions are as follows: First, we present a method for collaborative data analysis in online freelance setting. Second, through a set of experiments, we show that the proposed approach is both cost-effective and can produce results with equivalent quality to those produced by data scientists. Finally, following a design-science approach, we develop a platform that supports collaborative data analysis with mixed-level expertise.

2 Literature Review

In the following section, we introduce prior work on which we based our study. Its subsections review the success factors of online collaboration, describe the existing solutions for collaborative data analysis, and discuss the theoretical underpinnings that informed our method.

Online Collaboration: The advances of communication technology as well as a spread of sociotechnical systems made it possible for workers effectively collaborate within a distributed environment. Rather than meeting face-to-face, workers can rely on various communication channels such as emails, teleconference software or chat tools to cooperate in various tasks [4]. Many domains adopted computer mediated collaboration as a useful tool for reaching goals. Scientists use different online tools to engage in research discussions and activities [5]. Educators take advantage of online collaborative learning techniques to support students in achieving competence and foster skills like team working and group decision making [6]. Moreover, online collaborative tools facilitate marketing and decision making activities by, for instance, allowing better understanding of shopping behavior and predicting demand for products [7]. Previous research has identified multiple factors that impact successful online collaboration. First, a team has to be supported by senior member or manager who is facilitating the progress of the task and provides feedback [8]. Second, the members have to make themselves familiar with each other, which in turn should lower the psychological barrier of estrangement and promote cooperation over time [9]. Third, well-established communication is essential to avoid disagreements about the priorities and strategy to achieve pre-set tasks [10]. Fourth, trust along the group members supports the feeling that all members work towards the same goal and make every effort to achieve the best possible outcome in order to earn trust among team members. Finally, the last element is well established organization of the team. A competent leader will support the team in the process of developing manageable and effective workflow to accomplish the task in short time end with reasonable efforts[8, 9]. We considered all these factors during the design of the artifacts that will support collaborative data analysis with non-experts.

In crowdsourcing literature, a few notable methods to support crowd-collaboration have been proposed. For instance, Turkomatic is a tool that utilizes crowdworkers to plan and execute complex tasks. Requesters can watch workers decomposing and solving tasks in real time, either collaboratively or independently. Requesters can intervene to modify tasks or request new solutions to subtasks as needed [11]. Another framework, CrowdForge, introduces a map-reduce paradigm to split complex work into small parts and solve it in crowdsourcing setting. The task is broken into multiple subtasks that are concurrently solved and verified by other workers, and eventually merged into a cumulative output. However, although the framework relies on a powerful paradigm of parallel work execution, it assumes that complex work can be decomposed into lots of merely dependent micro tasks – an assumption that is often violated [12]. Other notable examples of online collaboration in crowdsourcing are CrowdWeaver – supporting with visual interface for real-time managing both human and machine crowdtasks within an integrated workflow [13], and Soylent – a word processing interface, implementing the Find-Fix-Verify crowd programming pattern, which splits tasks into a series of generation and review stages and utilize the collaboration among crowdworkers through independent voting and agreement to produce reliable results [14].

Existing solutions for collaborative data analysis: One of the most well-known examples of collaborative data analysis is Kaggle [15]. Kaggle is a web platform for data analysis that allows organizations to post their data projects and invite enthusiasts all around the world to participate in contests. Participants experiment with different techniques and compete against each other to produce the best models. For most competitions, submissions are scored immediately, based on their predictive accuracy relative to a withhold test-set of data, and summarized on a live leader-board. Once the deadline is over, the competition host pays prize money in exchange for the winning model [16]. Participants are allowed to team up together to collaborate on projects, and thus improve their chances to win the contest. Other solutions, such as Sense.us [17] or Many Eyes [18], have been proposed for collective data analysis by enabling crowds visually inspect data. For example, [19] presented CommentSpace, a collaborative tool for visual analysis that allows to annotate graphic content with tags and links that reflect the relationship between comments and visualizations. Wisteria and Wrangler are example of two human-in-the-loop systems that involve crowds in data cleaning by inferring the operations performed manually by crowds and extrapolating them to the whole dataset [2, 20]. Collaborative data analysis can be seen as an offshoot of distributed software projects. However, despite the evolution of advanced collaboration and software engineering tools (e.g., GitHub, Jira), software development is still mostly a prerogative of experts and does not involve laymen.

All mentioned solutions fall short on supporting collaborative data analysis by relying on crowds with mixed expertise. While platforms such as Many Eyes or Wrangler appeal to crowds without any prior expertise, platforms like GitHub require substantial skills in order to be able efficiently collaborate using their functionalities. Moreover, web-portals for crowdsourced data science such as Kaggle or TopCoder are rather a meeting point for data scientists and customers and, by and large, do not support the teams with any functionalities throughout data analysis.

Theoretical underpinnings: Tasks can be complex and may involve the coordination of a large number of participants with different capabilities. Therefore, different scientific communities have made efforts to associate tasks by decomposing them into the sub-tasks required to complete the full task [21, 22]. For instance, within the AI community, Chandrasekaran et al. [23] proposed a hierarchical task-method decomposition, which recursively links a task to alternative methods and their subtasks. This method emphasizes modeling of domain knowledge by utilizing tasks and methods as mediating concepts and, therefore fits our scope of the data analysis domain. Stefik [24] proposed an approach of constraint hierarchical planning where the constraints are dynamically formulated and propagated as the process proceeds. Subsequently, these constraints are used to coordinate the solutions of defined sub-tasks. The organizational approach, as presented in the Handbook of Organizational Processes of Malone et al., [25], in contrast, introduces methodologies to represent and codify organizational processes and provides different perspectives on how business processes might be decomposed into sub-activities. A difference between these two approaches lies in their different purposes: while AI is focused on building computer systems that automatically execute processes, the organizational approach advocates building systems to support people to plan and execute processes. Howison and Crowston [26] propose a theory of collaboration through open superposition. Developed in the context of open source software development, this theory emphasizes that tasks that appear too large for individual are likely to be postponed until they redefined such that they can be performed by single member, and that most of the tasks are indeed accomplished with only a single programmer.

These theories inform our solution in a few ways. A) decomposition of ill-defined task has to be tied into domain knowledge. B) the envisioned system should enable experts to decompose the task in efficient manner (e.g. through taxonomy or by utilizing expert's knowledge). C) There is a need for efficient coordination and communication in order to enable unimpeded process of data analysis D) data analysts working on a well-defined task will prefer to work on their own rather than collaboratively in an online team. However, they will be interested to coordinate the outputs of their task, to discuss possible solutions, and to receive feedback to their job. E) every task assigned to a worker should be well adapted to the skills and needs of the worker, with a clear specifications and task manager that can supervise and help with advice and guidance.

3 Research Design

The research design presented here follows a design-science research approach in information systems as presented by Hevner et al. [27]. The authors describe design-science process as a sequence of expert activities that produces a set of artefacts with the following evaluation and feedback in order to improve both the quality of the artefacts and the design process. According to the theory taxonomy proposed by Gregor [28], the proposed research resides within the *theory for design and action* by contributing to knowledge via addressing the considerations of a) the utility to a community of users, b) the novelty of the artefact, and c) the persuasiveness of claims that it is effective. As the goal to define and develop artefact that supports a novel approach of *collaborative data analysis with mixed-expertise crowds,* design can be seen as a search process involving an iterative evaluation and refinement of artefacts [27, 29]. The research methodology we adopted follows Peffers et al. [30] and includes six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication (see also Figure 1).

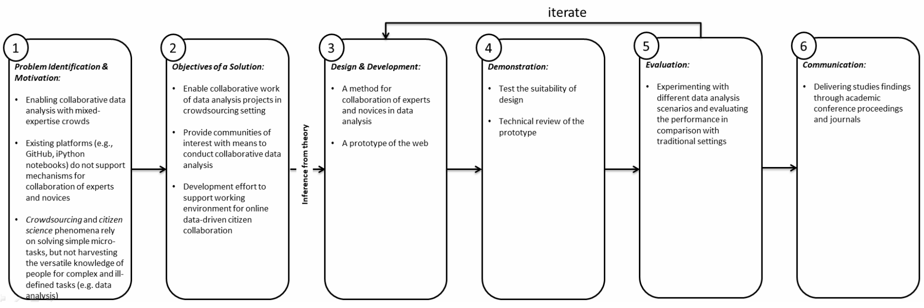
iterate
1
Problem Identification & Motivation:
• Enabling collaborative data analysis with mixed-expertise crowds
• Existing platforms (e.g., GitHub, iPython notebooks) do not support mechanisms for collaboration of experts and novices
• Crowdsourcing and citizen science phenomena rely on solving simple micro-tasks, but not harvesting the versatile knowledge of people for complex and ill-defined tasks (e.g. data analysis)
2
Objectives of a Solution:
• Enable collaborative work of data analysis projects in crowdsourcing setting
• Provide communities of interest with means to conduct collaborative data analysis
• Development effort to support working environment for online data-driven citizen collaboration
Inference from theory
3
Design & Development:
• A method for collaboration of experts and novices in data analysis
• A prototype of the web
4
Demonstration:
• Test the suitability of design
• Technical review of the prototype
5
Evaluation:
• Experimenting with different data analysis scenarios and evaluating the performance in comparison with traditional settings
6
Communication:
• Delivering studies findings through academic conference proceedings and journals

Fig. 1. Research methodology (Following Peffers et al. 2007)

Following the figure, we start by laying out the the research motivation: (a) to enable collaborative data analysis by crowds with different expertise, (b) the lack of platforms that support an efficient environment for data analysis for non-experts in a dynamic manner, and (c) to leverage the crowdsourcing and citizen science phenomena of harvesting knowledge that is hard to reach. We then define objectives of the solution: (a) to enable collaboration on data analysis tasks on web, (b) to provide communities of interest with means to conduct collaborative data exploration, and (c) to propose a web environment for online collaboration. At the design stage, to the best of our knowledge, no dominant method has been identified so far to incorporate people with diverse skills into data analysis. Hence, the major challenge of this paper is defining and evaluating the needs for collaborative data analysis, accounting for the diverse nature of crowd workers. To do so, we start with the top-down approach of expert managing the novices and gradually explore the predominant factors for successful collaboration and tasks' coordination. The results will be demonstrated through the web application prototype built based on the discussed artefacts and set of experiments in which we evaluate the crowd's performance on a series of data analysis projects to check whether the designed prototype satisfies the prerequisites.

4 Conceptualization of the Artefact (Data Analysis Tool)

In this study we present a framework that allows non-experts to work on data analysis projects. Our framework i) supports a project manager in decomposing complex tasks into small and facile sub-tasks, ii) supports coordination and supervision by project manager, and iii) enables an iterative development of the data analysis project. The methodology we propose implies that the project manager defines a project and distributes assignments to workers in a top-down approach. A top-down approach is considered as more appropriate for well-specified, rather than ill-defined problems [31]. However, we decided in favour of this method, as the scenario we envision is of non-experts that are competent to perform preprocessing tasks only with the appropriate supervision. It is, therefore, necessary to impose task decomposition hierarchy to be able to manage the complexity of task on the expense of its flexibility. In addition, as our approach implies iterative exploring of the success factors for the scenario we investigate, the top-down approach is better suited for understanding how strictly hierarchical approach can transition into more collaborative one. For instance, it allows to see throughout the iterative development and evaluation, where the expert oversight can be replaced with peer-review of other novices, how decisions made throughout data analysis can be informed by the broad knowledge of the crowd to enrich expert's decisions, or how to establish effective communication to unleash the untapped knowledge of project members. Following the design science approach, we conducted two iterations of prototype development with consequent evaluations. In the following we first describe the general workflow and then the evolvement of the prototype and of the methodology after each iteration.

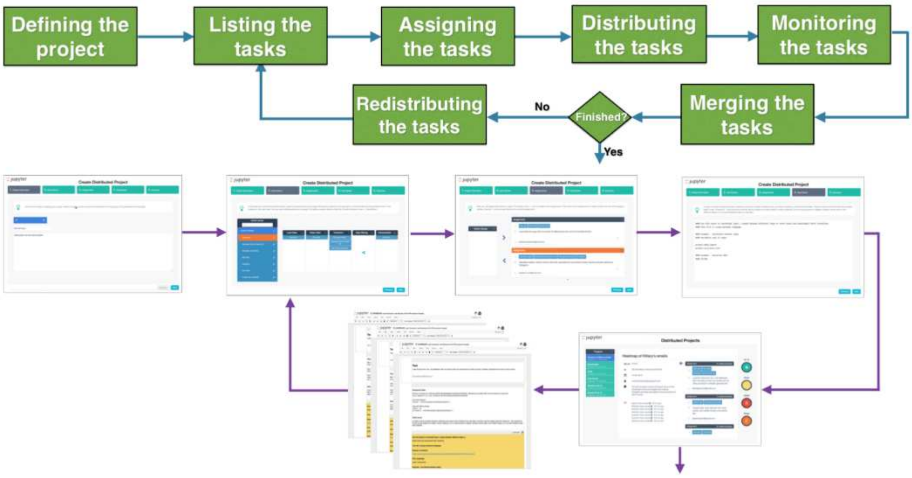

Fig. 2. Process workflow

Figure 2 describes the workflow of the envisioned collaborative data analysis project. The figure presents both schematic workflow of the process on the top and the corresponding print-screens of the prototype on the bottom. The first part of the workflow is focused on the project definition, task decomposition and sub-tasks assignment processes done by the project manager. The second part focuses on the iterative collaboration on the project, enabling the manager and team to refine the implementation and output through multiple iterations.

Next, we go through each of the workflow steps and explain them. First, the project manager defines the project by entering all relevant details, such as the software language to be used, the project name, and the project description. This step also provides some validations, ensuring that all necessary information is present. Next, the manager lists and defines the actions that need to be done. An action is the smallest unit of sub-task and an assignment is a composition of actions assigned to a worker. An example of an action would be *loadFromCSV*, which receives as input the path of the CSV file and returns a data-frame. Splitting assignments into small actions, especially in the preprocessing part, allows the project manager to distribute them to non-expert workers and supervise their execution throughout the assignment. Further, **tasks are assigned** to suitable workers. The assignment of tasks to workers follows a top-down approach and can be done on the basis of different criteria such as worker or task attributes, or by taking into consideration external factors such project deadlines or budget. The **tasks are then distributed** among the workers by virtue of email invitation to the IPython (or Jupyter) Notebooks that are created and contain all the required information. At this point the workers can work on their personal notebooks stored on their personal cloud storage (Google Drive) and interact with each other through the shared notes-board. They can also review others notebooks and comment on the relevant code using side-comments. All throughout the project, manager can **monitor the progress** of the workers and guide them towards the desired output. Finally, the tasks are merged into one notebook which allows a manager to run the end-to-end implementation. Project managers can then verify that the output meets their expectations and that the interaction between different assignments works properly. Otherwise, if the goal has not been reached, the implementation of the tasks will be changed or new **tasks will be redistributed** and the project will enter a new iteration.

5 Iterative Development-Demonstration-Evaluation

The proposed solution has been developed in two iterations by improving the method and the web-prototype for collaborative data analysis in a consecutive manner. Based on the evaluation of each iteration, we focused on advancing the artifact with respect to the following two criteria:

First, the proposed methodology and web-prototype should enable coordination and successful completion of data analysis projects with diverse crowds. Specifically, typical data analysis projects should be decomposed into subtasks such that they will be simple enough to be performed by non-experts. We evaluate these criteria qualitatively, through a user study by answering the following hypothesis:

> *H1: It is possible to decompose typical data analysis projects into small enough tasks such that the complexity of these tasks is substantially reduced.*

Second, the proposed solution has to be comparable in quality to traditional expert-based data science projects. To answer whether the proposed methodology is feasible and can reach the desired output of collaborative data analysis with mixed-level expertise teams, we propose the following hypothesis:

> *H2: The quality of the results produced by a team of non-non experts is comparable to the one achieved by experts.*

In the following we will present three versions of the prototype and discuss their performance according to these measures. Note that we tested all iterations on real-world examples chosen from Kaggle based on the following criteria: a) the projects should be implemented either in R or Python, as these are the most popular languages in data analysis, b) the projects should contain a relatively large preprocessing part, as that has been found to be a major part of data [32], c) the projects should encompass various types of data analysis such as descriptive statistics, visualization, and prediction, d) the projects should be conducted by individuals that can be considered as experts, either based on their verified biography or because of their high ranking on Kaggle, and e) the projects should not be trivial (i.e., we limited the minimal size of the project to be about 150 lines of code, chose projects with significant number of up-votes, and history of comments such that it can be assumed that the code went through a substantial public review).

5.1 The Pilot Study

Following to literature review we designed the first prototype of our tool. The web-platform is based on the Jupyter Notebook (colloquially known as IPython notebook) and available online. Jupyter is a command shell for interactive computing in multiple programming languages that offers enhanced introspection, media, additional shell syntax, tab completion, and rich history. Using Jupyter, researchers can capture data-driven workflows that combine code, equations, text and visualizations and share them with others. We decided in favor of this platform due to the following reasons. First, it is a browser-based notebook with support for code, text, mathematical expressions, inline plots, and other rich media. These functionalities are essential for collaborative data analysis as they allow participants to exchange results and easily communicate their findings and difficulties. Second, although initially designed for Python, the platform is language agnostic and provides the ability to be extended with additional interpreters such as R and Ruby. Third, this platform supports an interactive data visualization toolkit, often required in data analysis.

To better understand the requirements of the proposed solution, *we conducted a user-study with three graduate students supervised by a PhD student*. As part of their course work, the students conducted data analysis project that involved substantial data preprocessing followed by network analysis. The supervisor was managing the task decomposition and divided the project among the group members with further coordination of the process up to its successful accomplishment (following the process presented in Figure 2).

The goal of pilot study was twofold. First we wanted to reach a proof-of-concept, showing that our approach is feasible and data analysis projects can be successfully accomplished with non-experts. Therefore, we alleviated some constrains such as performing the experiment in real-world setting using freelancers/crowdworkers or assuring that the analysis has been performed exclusively on our platform. Second, we aggregated the feedback to better understand the requirements of the proposed tool and to evaluate the workflow. In addition, the feedback received from this iteration helped us to simplify the coordination process and to resolve some technical issues.

Conclusions/requirements drawn from pilot study and their addressing: First, all participants pointed to the need for collaboration and communication tools. While some can be externally used (e.g., forums, video chats), some tools have to be embedded into the platform to support effective coordination between team members. Especially, since the assignments distributed to workers are often interdependent, it is important to allow team members to comment on the relevant code-blocks of their peers. To address this need, *we developed features that allow workers better to collaborate*. For instance, we presented "sticky notes" – a note that every team member can leave next to the code-box of a Notebook. Second, another point, raised by the manager, is to improve the control over the project by enabling easy access to the notebooks, evaluating the current results, and (re)distributing the tasks. We, therefore, *added a functionality to automatically merge the notebooks into a master notebook that includes all notebooks in predefined order*. This allows to run all distributed assignments at one run and quickly identify bugs and inconsistencies. To redistribute the tasks with new instructions, we implemented a feedback loop (see Figure 2) that allows easily to redistribute the tasks to team members with new instructions and based on the previously submitted code. Third, to improve the collaboration, team members pointed to the need to have access to the instructions every team member received from the manager as well as have the opportunity to intervene in order to clarify what in their opinion has to be done. To address this, *we added a project journal, where all project participants can add their comments*.

Note, while such functionalities exist in professional software development platforms such as GitHub, our goal is to enable *non-experts* to collaborate instantly on data analysis projects in easy and interactive way with no knowledge on the principles of distributed software development. In the following iterations, we qualitatively evaluated the proposed features and extend our platform according to the additional feedback provided by crowdworkers in the real-world setting. Most of the attention in the following two iterations though, is devoted to testing the postulated hypotheses.

5.2 First iteration - Three Data Analysis Projects

For a real-world evaluation, we selected three data analysis projects that represent various types of data analysis. Projects were taken from a large crowdsourcing data science platform, Kaggle. In these experiments, a data analysis expert (also a co-author) assumed the role of the project manager and the workers are recruited through the Upwork[1] platform. As of today, Upwork is the biggest online labor market and contains online freelancers in different domains. Data analysis is one of its most common domains and has a large pool of freelancers with different level of expertise willing to work on data analysis projects [33]. These tasks can be classified as of moderate complexity as they involved mainly data preprocessing and visualization, and did not require any advanced knowledge in data analysis.

[1] www.upwork.com/

Task #1: Earnings Chart by Occupation and Sex[2]*:* The aim of the first project is to create a chart showing the earnings of the population by occupation and gender, using the data of the latest US census from 2014. The original Kaggle project analyzes 24 occupation categories, while in our project we randomly selected 11 categories. The workers had to classify the list of the professions into these 11 job categories (e.g., management, science, military) and plot a chart of the earnings for each occupation with respect to the gender. This project is the easiest and was accomplished in two days.

We split the project into three assignments. The first assignment involved data loading and cleaning with the primary goal of identifying the correct industry code ranges and sub-setting the data. It consisted of five actions. The first was to *Identify Occupation Industry Codes*, and *Subset data* and the output of this task was a file containing the information about the population working in the 11 industries relevant for our chart. The second task focused on the data transformation and had only two actions –*Mean* and *Save results*. The output of this task was an aggregated data set containing the mean earnings of men and women per industry. In the last task, the crowdworker had to plot the data as a bar chart diagram in descending order, showing the distribution of men and women per industry and their average earnings. It consists only of one action – *Bar Chart*, and produced as output a bar-chart similar to the one in the Kaggle project.

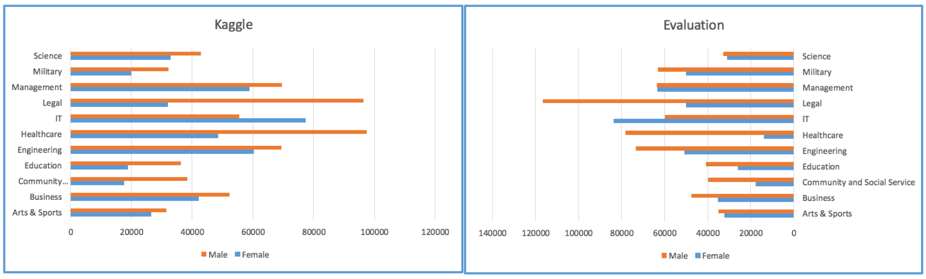

Fig. 3. Pearson correlation coefficient ϱ=0.8

The main focus of this project was to find the right occupation categories and to subset the data accordingly. The project used a random 1% sample of the US census data from 2014. In order to compare the results, we evaluated both implementations (Kaggle's and non-experts' team) on the same data (Figure 3). The team of non-experts managed successfully to finish the project and their results were similar to to those published on Kaggle, resulting in the Pearson correlation coefficient ϱ=0.8.

The differences in the results can be traced back to the nuance that two implementations perform the data subsetting in different way. Each occupation in the data set is identified by a code. The 11 categories used in the project are quite generic, so it is user's responsibility to find the occupations which belong to the respective category. While Kaggle solution identifies only one occupation for each category, the Upwork team's implementation aggregates multiple occupation codes under the same category.

Task #2: Hillary Clinton's Emails[3]*:* This project explores the content of Hillary Clintons emails which were released by her in response to a Freedom of Information Act (FOIA) request, and produces a heat-map of the countries that often appear in the emails. The dataset for this project is available on Kaggle. This project was also split into three assignments. The first assignment focused on data loading and cleaning,

[2] www.kaggle.com/wikunia/d/census/2013-americancommunity-survey/earnings-by-occupation-sex/

[3] https://www.kaggle.com/ampaho/d/kaggle/hillary-clinton-emails/foreign-policy-map-through-hrc-s-emails/code

and consisted of three actions. The output of this task was a cleansed subset containing only the emails sent by Hillary Clinton and a list of all the countries in the world and their alternative spellings and abbreviations. The second task focused on identifying countries in the email data set and contained two actions – *Subset* and *Calculate occurrences*. The output of this task was a country occurrence list, containing the number of times each country is mentioned. The last task focused on the visualization part and consisted of two actions. The output was a sorted histogram and a heat-map in form of a world map, similar to the output of the Kaggle project.

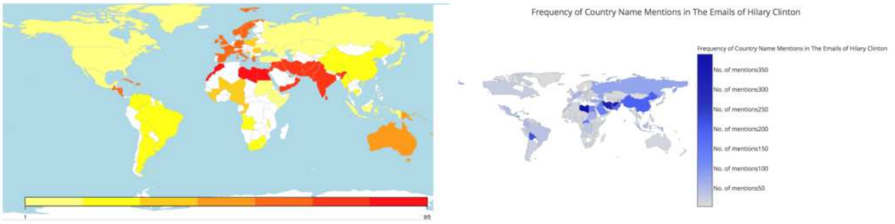
1
95
Frequency of Country Name Mentions in The Emails of Hilary Clinton
Frequency of Country Name Mentions in The Emails of Hilary Clinton
No. of mentions350
No. of mentions300
No. of mentions250
No. of mentions200
No. of mentions150
No. of mentions100
No. of mentions50

Fig. 4. Pearson correlation coefficient ϱ= 0.72

The team of non-experts managed successfully to finish the project and the output of their work was similar to the results published on Kaggle (see Figure 4). In both implementations, the heat-map is based on a country occurrence list. We compared the results by calculating the Pearson correlation coefficient between the two lists with country occurrences which resulted in ϱ=0.72.

Similar to the previous project, the difference in the results is caused by the way two implementations identify the countries mentioned in the emails. The project on Kaggle and the team of non-experts use different approaches to identify countries abbreviations which lead to difference in the results.

Task #3: Reddit Sentiment Analysis[4]*:* The purpose of this project was to create a chart showing which Reddit comments receive the highest scores, based on the sentiment of the comment. Reddit is a large social network where users can submit content. The dynamics of this website is solely dependent on the number of up/down votes that the content receives. The content or comment with the highest number of votes is shown at the top. The categorization into three sentiment categories – objective, negative, and positive – was performed using the designated software package. The initial dataset includes Reddit comments from May 2015 and available on Kaggle.

The goal of *Reddit Sentiment Analysis* is to create a chart showing which Reddit comments receive the highest scores, based on the sentiment of the comment. Three sentiment categories were defined – objective, positive and negative. As in the previous project, we used a random sample of the May 2015 dataset. Both implementations were tested and evaluated using the same dataset. As it can be seen in Figure 5, the results are very similar – the average ranking scores for the positive, negative and objective comment categories are 6.18, 6,78, and 5.96 in the Kaggle project, and 5.75, 6.22, and 6.34 in the Upwork project performed by non-expert team.

[4] https://www.kaggle.com/lplewa/d/reddit/reddit-comments-may-2015/communication-styles-vs-ranks/code

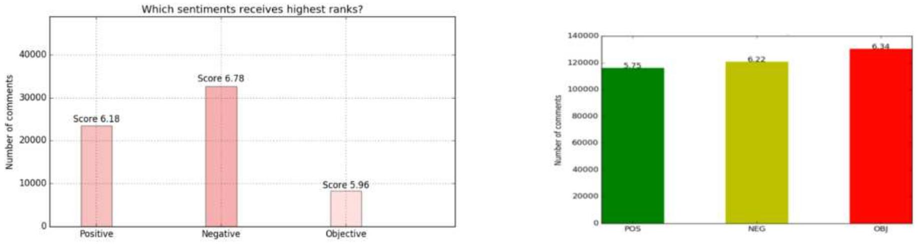

Fig. 5. Equivalence tests: comparison of Kaggle with non-experts results

We also compared the ranking values in each sentiment category by performing equivalence tests on the results of the two projects [34]. The goal of equivalence tests is to statistically test the equivalence of the variables. This was achieved by setting the equivalency region δ and testing whether the calculated confidence intervals for the differences between the two variables are within this region. For each sentiment category, we set the δ to be the average standard deviation of the Kaggle and the team of non-experts results. All the intervals are calculated with 95% confidence:

- Positive: CI_{poss} (-0.21, 1.07) $\subseteq$ (-S.D.$_{poss}$, S.D.$_{poss}$)
- Objective: CI_{obj} (-1.16, 0.4) $\subseteq$ (-S.D.$_{obj}$, S.D.$_{obj}$)
- Negative: CI_{neg} (-0.17, 1.29) $\subseteq$ (-S.D.$_{neg}$, S.D.$_{neg}$)

In all cases the confidence intervals are contained within the equivalency region, meaning that there is no difference between the ranking means in each sentiment category.

Note that the implementation, the classification of the comments into one of the three sentiment categories was done differently. In Kaggle project, the comments are classified by selecting only the comments with values above average (top quartile or top 3/8) for each sentiment, while the in project done by the non-expert team, sentiment scores are first normalized (through division by mean), and only then the comments are classified. Nevertheless, the results are almost identical.

Conclusions: At the end of this iteration, we qualitatively evaluated the features previously developed via a questionnaire, where we asked the participants open-end questions related to the use of the system. Specifically, we asked them to describe the features they found useful, difficulties they experienced in using the platform, and what are the functionalities that are missing or insufficient. We used the feedback received in this iteration to improve our prototype and to add missing functionalities. For example, *we added a notification that the worker has finished his part such that the manager can review the output* and *the worker responsible for the next step can start working with the provisional results*. We also *added a notification to inform the owner of the notebook via email every time a "sticky-note" is attached*.

Regarding H2, all three experiments present substantial similarity between the experts' and non-experts' results. The similarity in the results of task #1 and task #2 is shown through significantly high correlation between the results – 0.8 and 0.72 correspondingly. Similarly, the results of task #3, compared using equivalence tests, indicate equivalence of the results. Altogether, the results of experiments support our hypothesis that crowds with mixed expertise are able to produce outputs comparable with the results produced by experts.

5.3 Third iteration – Fully Autonomous Data Analysis Project

The last experiment we conducted was *Prediction in the Republican Primaries*[5]. The goal of this evaluation was to predict the results of the Republican Primaries 2015 in different counties. This experiment can be seen as full end-to-end data analysis project that includes all elements of data analysis, starting with data preprocessing, visualization, and up to building prediction models. The manager in this project, an expert worker from the crowd, was also responsible for building the prediction model. This setting allows the expert to better define the requirements of the activities, as he will use the processed data to build prediction models. In this project the manager was responsible both for hiring the crowdworkers and defining assignments without intervention. Eventually, the project was split into three assignments performed by manager and two crowdworkers.

The first assignment focused on activities of *data loading, subsetting, and aggregating data* from different sources, such that the resulting data can be used for further analysis. The output of this task was a data-frame that included information about the primaries winner in every county and state as well as the demographic data of regions extracted from different data sources. This task required significant efforts and took about 5-7 hours of work. The second assignment was mainly about visualization of the data and descriptive statistics and resulted in various visualizations describing the relationship between population features of counties (e.g., residents' ethnicity or education, population density) and candidates' voting patterns. The duration of this task was about two hours. The last assignment was to build models predicting vote rates of each candidate. This task included training prediction models and testing them, similarly to Kaggle solution, on the test-set with reporting prediction qualities, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The overall results of the prediction errors of the crowd and the experts are very similar. The mean absolute error of the Kaggle solution is $MAE_{Kaggle} = 6.5\%$ while the solution of the non-experts team yields $MAE_{Upwork}=6\%$. The root mean square errors of both solutions are almost identical with $RMSE_{Kaggle}=8\%$ in the Kaggle solution and $RMSE_{Upwork}=7.7\%$ in the model produced by crowdworkers.

Regarding H2 we can, hence, again conclude that the results produced by non-experts are comparable in their quality to those produced by data scientists.

5.4 Summary and Discussion of Results

Evaluation of H1: We tested the first hypothesis by reviewing the task decomposition output. Specifically, we aimed to ascertain whether it is possible to decompose the selected data analysis projects into sub-tasks such that the complexity of the sub-tasks is reduced compared to the overall complexity of the project. We asked the crowdworkers to report about the perceived complexity of the project and the sub-tasks. Following, we aggregated the results and analyzed them.

It was possible to split all projects into actions. Also, all of the workers were able to successfully complete their assignments. They rated the complexity of their assignment with an average of 2.25 (S.D.=0.96) out of 5. The project, on the other hand, was rated higher than the assignment complexity, with 2.42 out of 5 (S.D.=0.67). Despite the lack of significance (possibly due to the small sample size), we believe the results indicate a trend, that the method might work. Based on our evaluation and echoed by the literature review, we conclude that data analysis can be split into less complicated sub-tasks and accomplished by non-experts.

[5] https://www.kaggle.com/apapiu/d/benhamner/2016-us-election/predictions-in-the-republican-primary

Evaluation of H2: To test the second hypothesis, we statistically compared the results of the projects conducted by experts with the results of non-experts that used our platform. As the data analysis projects we used for evaluation are publicly available on Kaggle, we explicitly asked the participants to not search and browse for the solutions on Kaggle. We also compared the code and the solutions' logic to assure that the code has not been inspired by the original solution. As already described in the iterations above, we attempted to cover a range of typical data analysis projects with complexity that meets real-world scenarios. Moreover, in order to ensure that the similarity is not a result of naturally limited space of solutions (which could lead to highly correlated results), we compared the results of other authors to see whether there is a natural variance in results.

Discussion: Both hypotheses have been empirically supported, meaning that data analysis projects can be effectively decomposed and accomplished with good quality. However, we found that the success of a project also greatly depends on other factors. The decomposed-tasks have to be effectively coordinated and timely adapted for the changing needs of data analysis. This is due to the dynamic/iterative nature of data analysis, where new insights, resulting from intermediate results, inspire new ideas on how to proceed with analysis. This, in turn, often requires additional data wrangling and sparks new iterations of work. While this work is performed in distributed way by non-experts, there is a need to support such process with appropriate coordination tool that will facilitate the process.

Moreover, the total cost of the experiments excluding manger was about 120 USD per project (the projects were split between three crowdworkers), where every worker has been paid 40 USD to accomplish her part, and each project required on average about 12 hours of work. In the project that involved the freelance manager, additional cost of 100 USD was paid for about 8 hours of manager's work. This makes the projects economically competitive, especially in the light of the soaring data scientists' wage.

We also collected information about the background and skills of the crowdworkers that participated in our experiments. Most of them are bachelor or master students in their twenties, studying IT, computer or exact sciences and working part-time as freelancers (13 hours per week on average). The workers perceive themselves mildly proficient in coding (self-rated with 3.2 out 5) and have basic background in data analysis, usually limited to introductory class in statistics or online course. Even though we have not conducted in-depth study on the demographics of online freelancers working in data analysis, our strong impression was that most of them can be characterized as part-time workers with average coding skills and very limited statistical/data analysis education with expected remuneration similar to the one in our experiments. This can be seen as evidence for the existence of sufficient talent to support the scenario we propose.

6 Limitations and Future Work

The proposed methodology has the following limitations. First the proposed top-down approach is not necessarily the optimal structure and other alternatives might be explored. For example, to allow workers to pick a task they want to work on in a self-managed manner and accompany the execution with managerial oversight. Second, we showed that the tasks can be decomposed into multiple simple sub-tasks. However, were not able to confirm this statistically. It is unclear whether this is due to a small sample of respondents (12). Future work might explore this by increasing the sample size and with recording additional data indicating the complexity of tasks. Third, to better evaluate the proposed platform, additional evaluation of the proposed scenario with other systems can be performed. For instance, the experiment where the coordination is done through a version control system that is used for software development such as GitHub[6].

[6] https://github.com/

Lastly, further research is needed to better understand the trade-off between the managerial overhead and saved costs due to outsourcing to non-experts.

7 Conclusion

This paper presents an approach of collaborative data analysis that involves data analysis novices with initial coding skills to participate in the process. We propose and evaluate the scenario where teams of non-experts are guided by expert throughout the process of data exploration and preprocessing. The proposed framework was evaluated with an especially data designed tool and by virtue of multiple experiments, where the constraints are gradually released: first a pilot study where we control for both the workers and the manager, then three experiments, where only the project manager is controlled, and ultimately, a data analysis project, where both the project manager and the workers are hired and perform the task without any external interference. The results demonstrate the feasibility of the proposed approach and support the hypothesis that the output of teams with mixed-level expertise is equivalent to the results achieved by experts. Moreover, through various data analysis projects we show that it is possible to decompose them into simpler sub-tasks that can be then successfully accomplished by non-experts. Additionally, we found that the following features were valuable for collaborative data analysis with crowd workers: support for dynamic development, code deliberation, communication, and a journal with decisions made throughout the project.

In summary, we believe that our study paves the way for including non-expert crowd workers in data analysis tasks. As such, we hope to contribute to the research studying the requirements for building tools that can leverage the crowd to address the shortage of data analysts.

8 Acknowledgments

This work was supported by the Swiss National Science Foundation under contract number 14341.

9 References

1. Davenport, T.H., Patil, D.J.: Data_Scientist-the_Sexiest_Job_of_the_21St_Century.Pdf, (2012)
2. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive Visual Specification of Data Transformation Scripts. Human factors in computing systems. ACM. 3363–3372 (2011). doi:10.1145/1978942.1979444
3. Bernstein, A., Klein, M., Malone, T.W.: Programming the global brain. Communications of the ACM. 55, 41 (2012). doi:10.1145/2160718.2160731
4. Sere, F.C., Swigger, K., Alpaslan, F.N., Brazile, R., Dafoulas, G., Lopez, V.: Online collaboration: Collaborative behavior patterns and factors affecting globally distributed team performance. Computers in Human Behavior. 27, 490–503 (2011). doi:10.1016/j.chb.2010.09.017
5. Van Noorden, R.: Online collaboration: Scientists and the social network. Nature. 512, 126–129 (2014). doi:10.1038/512126a
6. MacDonald, J.: Assessing online collaborative learning: Process and product. Computers and Education. 40, 377–391 (2003). doi:10.1016/S0360-1315(02)00168-9
7. Yadav, M.S., Pavlou, P.A.: Marketing in Computer-Mediated Environments: Research Synthesis and New Directions. Journal of Marketing. 78, 20–40 (2014). doi:10.1509/jm.12.0020
8. Tseng, H., Wang, C.-H., Ku, H.-Y., Sun, L.: Key Factors in Online Collaboration and Their Relationship to Teamwork Satisfaction. The Quarterly Review of Distance Education. 10, 195–206 (2009)
9. Salehi, N., McCabe, A., Valentine, M., Bernstein, M.S.: Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing. (2016)
10. Yukl, G.: Leadership in organizations. Personnel Psychology. 7th, 542 (2001). doi:10.1016/1048-9843(95)90027-6
11. Kulkarni, A., Can, M., Hartmann, B.: Collaboratively crowdsourcing workflows with turkomatic.

Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12. 1003 (2012). doi:10.1145/2145204.2145354

12. Kittur, A., Smus, B., Kraut, R.: CrowdForge Crowdsourcing Complex Work. Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11. 1801 (2011). doi:10.1145/1979742.1979902
13. Kittur, A., Khamkar, S., André, P., Kraut, R.E.: CrowdWeaver : Visually Managing Complex Crowd Work. Scenario. 1033–1036 (2012). doi:10.1145/2145204.2145357
14. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. Proceedings of the 23nd annual ACM symposium on User interface software and technology. 313–322 (2010). doi:10.1145/1866029.1866078
15. Carpenter, J.: May the best analyst win. Science (New York, N.Y.). 331, 698–699 (2011). doi:10.1126/science.331.6018.698
16. Dissanayake, I., Zhang, J., Gu, B.: Virtual Team Performance in Crowdsourcing Contests : A Social Network Perspective. ICIS 2015 Proceedings. 1–16 (2014)
17. Heer, J., Viégas, F.B., Wattenberg, M.: Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization. Communications of the ACM. 52, 87–97 (2009). doi:10.1145/1240624.1240781
18. Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J., McKeon, M.: Many Eyes: A site for visualization at internet scale. IEEE Transactions on Visualization and Computer Graphics. 13, 1121–1128 (2007). doi:10.1109/TVCG.2007.70577
19. Willett, W., Heer, J., Hellerstein, J.M., Agrawala, M.: CommentSpace: Structured Support for Collaborative Visual Analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 3131–3140 (2011). doi:10.1145/1978942.1979407
20. Haas, D., Krishnan, S., Wang, J., Franklin, M.J., Wu, E.: Wisteria : Nurturing Scalable Data Cleaning Infrastructure. Proceedings of the 41st International Conference on Very Large Data Bases. 8, 2004–2007 (2015). doi:10.14778/2824032.2824122
21. dos Santos, F., Bazzan, A.L.C.: An ant based algorithm for task allocation in large-scale and dynamic multiagent scenarios. Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09. 73 (2009). doi:10.1145/1569901.1569912
22. Campbell, A., Wu, A.S.: Multi-agent role allocation: Issues, approaches, and multiple perspectives. Autonomous Agents and Multi-Agent Systems. 22, 317–355 (2011). doi:10.1007/s10458-010-9127-4
23. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: Ontology of Tasks and Methods. Knowledge Acquisition. 1–25 (1998)
24. Stefik, M.: Planning with constraints (MOLGEN: Part 1). Artificial Intelligence. 16, 111–139 (1981). doi:10.1016/0004-3702(81)90007-2
25. Malone, T.W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., Quimby, J., Osborn, C., Bernstein, A., Herman, G., Klein, M., O'Donnell, E.: Tools for inventing organizations: Toward a handbook of organizational processes. Management Science. 45, 425–443 (1999)
26. Howison, J., Crowston, K.: Olla boration through open superposition. (2013)
27. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS quarterly. 28, 75–105 (2004). doi:10.2307/25148625
28. Gregor, S.: The nature of theory in information systems. MIS Quartely. 30, 611–642 (2006). doi:10.2307/25148742
29. Reinecke, K., Bernstein, A.: Knowing What a User Likes: A Sesign Science Approach to Interfaces that Automatically Adapt to Culture. 37, 427–453 (2013)
30. Peffers, K.E.N., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. Decision Sciences. 24, 45–77 (2008). doi:10.2753/MIS0742-1222240302
31. Redmiles, D.: Software Requirements for Supporting Collaboration through Categories. (2000)
32. Krishnan, S., Wang, J., Franklin, M.J., Goldberg, K., Kraska, T., Milo, T., Wu, E.: SampleClean: Fast and Reliable Analytics on Dirty Data. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 59–75 (2015)
33. Agrawal, A., Horton, J., Lacetera, N., Lyons, E.: Digitization and the Contract Labor Market: A Research Agenda. NBER Working Paper. 37 (2013). doi:10.3386/w19525
34. Mascha, E.J.: Equivalence and noninferiority testing in anesthesiology research. Anesthesiology. 113, 779–781 (2010). doi:10.1097/ALN.0b013e3181ec621