SCALABLE DIGITAL VOLUNTEERING:
A DATA FOR SOCIAL GOOD MARKETPLACE

Diego A. Arenas-Contreras
School of Computer Science, The University of St Andrews, KY16 9SX, Scotland, UK

Abstract – Data for Good has receive increasing interest from academia and civil society. Many initiatives have been created with the goal of use data to solve humanitarian and global issues. A lot has been done but a lot needs to be done. In this paper a design of a marketplace for data for good is presented.

Keywords – Data Architecture, Data for Good, Data Science, Marketplace.

1. INTRODUCTION

There has been an increasing interest in Data for Good in the recent years. The application of Data Science and data modelling techniques to solve real world problems with societal impact present itself as an attractive field and is open for innovation.

Data for Good is an open space for collaboration among different actors such as the social sector, the civil society, government, academia, and private sector. It takes at heart, our societal and global issues, and creates tangible applications of knowledge and science by its practitioners.

We will argued in this paper that the development of Data for Good is been good, it has a steady development, but is still not enough. It requires of a wider community of practitioners to make it sustainable and it requires better integration between technologies and parties.

A digital marketplace is suggested as a solution to these problems and a design is outlined to be used as a rogue guideline to build a needed tool for data for good.

2. HOW DATA SCIENCE IS APPLIED FOR GOOD?

There are different opportunities to work in Data for Good. From approaching to an organisation in the social sector and asking them to work with them in their data issues, to participate in weekend-long or week-long events where a handful of nonprofits are invited to participate and present their data challenges. The latter is the most common instance to practice data for good.

DataKind [1], DrivenData [2], the Data Study Groups [3] at the Alan Turing Institute, the University of Chicago's Data for Social Good programme [4] and a number of other institutions have done and are still doing a great job in the data for good field organising events where people can interact with data for good problems and institutions.

People attending these events decide what problem they would like to work until the end of the event, when findings and results are shared and presented.

This type of work has been proven successful and there is an increasing demand to participate in these type of events. Nonprofits and social organisations receive a number of new insights and information when they participate that makes the event a success. These opportunities exists and are a good opportunity to put data skills at the service of society pro bono.

Although, this model has proven to be effective has some limitations such as the physical presence of the participants, also that the work must be completed during the duration of the event. Sometimes further collaboration is encourage after the event but this is a rare case.

Data for Good projects are no different from standard data science projects. The ”for Good” is given when the outcome of the project is intended to improves people's lives or to improve the state of the world.

In a data for good projects there are usually four components to consider: 1) Data Processing, 2) Data Analysis, 3) Data Storage, and 4) Data Privacy. Each one of these components has its specific challenges and this paper intends to give some lights in the design of a system that is designed for data for good from its conception. In Section 5 this will be covered.

3. THE PROBLEM

The work so far in Data for Good has been done project-wise. Organisations from the social sector connect with data volunteers through events organised in cities.

The data for good events have a duration of one to five days most of them, with the exception of the competitions that last for several weeks with one winner in the end; the data for social good summer programme of the University of Chicago that last for three months.

The limited number of participants, the limited number of organisations participating, and the duration of the events add some constraints to the scope of what can be done in Data for Good events. This calls for a more scalable solution to allow more people to get involved and get in contact with causes and institutions for good. That is why a marketplace platform is suggested as a sound solution.

We are trying to expand the opportunities to collaborate with good causes and explore the opportunities for collaboration with nonprofits in data science projects.

The design of a digital marketplace, or simply marketplace further on, is a sound solution to overcome the limitations of attending events in person and to grow and scale the participation of interested people in collaborating in good causes.

This paper presents the design of a system that intends to scale the good work around Data for Good. The momentum that data for good events have built allow to suggest a new type of platform to scale and allow to start collaboration in projects at anytime and from anywhere.

3.1 Motivation

In the data science community there is an increasing interest in collaborate and participate in data for good projects [5]. Unfortunately there are scarce opportunities until now and the entrance barriers are high for people interested in participate. They need to be available at specific dates to participate in one of the events organised in their local communities, and second there is a misconception in the interested people that it is required to know about the domain of the problem before hand which is desirable but not mandatory. Also the work in the project is most of the time restricted to the duration of the event.

This physical presence requirement can be avoided and projects could start at any moment in time, not necessary just for the events, also the participation of people can be scaled to anyone with internet access.

We believe that a marketplace is a sound solution to these problems and that it will trigger the participation of a broader community in data for good projects.

The identified problems are the lack of communication and instances to connect among data scientists and social sector. Start developing new projects and engage in collaborations and projects. This problem is intended to be solved with the design of a Marketplace to connect data scientists and organisations and institutions working for the common good. The design is presented to anyone who would like to implement it or help in the implementation of it.

A marketplace can be used for collaboration among academia, social sector and civil society is proposed. The system has a highly modular architecture with separated components that are designed to not interfere between them. This allows a greater number of data volunteers to join the Data for Good spirit and collaborate with the institutions that matter the most to them and to apply their knowledge and learn in the process.

There are many data sharing platforms. We can find from Open Data portals to websites where one can buy datasets. There are also many alternatives of data science environments with collaborative features and team management. We can also find data processing platforms in cloud providers easy to use and to deploy solutions and projects. We can find freelancing websites with project and tasks description for data scientists. We can find all these platforms in an indenpent way, what is missing is a system able to integrate them for good, to be used for Data for Good.
We believe that systems like to one it is presented in this paper would help to achieve the UN 17 Sustainable Development Goals [6] defined by the United Nations in 2015 by 2030.

An open space for collaboration rather than competition is required. This paper presents the design of a Marketplace for Data for Good.

4. DIGITAL MARKETPLACES FOR GOOD

Not a single but several marketplaces are required. There are no restriction in their construction but should allow integration and easy ways to migrate projects and profiles between them. This integration requisite is key to ensure a broader participation. Think of this like Amazon and EBay would allow you to publish your product from platform to the other in a single click. That is expected in the Data for Good ecosystem where the ultimate goal is to improves people's lives.

When we talk about a marketplace we are referring to the front-end platform but the system it made by multiple components.

1. A marketplace. Organisations in the social sector can create profiles and people from the civil society and academia can also have their profiles. The products that organisations create in the platform are data challenges, where they specify the knowledge domain they have, the challenges they have around data analysis, and a description of the data they have been collecting. The datasets are not public by default but only when an organisation agrees to work with a digital volunteer.

2. A collaborative data science environment. Providing the right tools to digital volunteers to collaborate and write the code, processes, and pipelines to analyse the datasets provided. It has to be a collaborative environment to allow integration between collaborators, peer reviewed code checking, comments and suggestions on the code, creation of wikis or similar to explain the project is about and communicate results. Several environments can be offered to start a new project and the volunteer can choose from them. The environment should provide means of integration with other platforms. Should allow to export or import code and to connect with other collaborative environments.

3. A data sharing platform. To upload and store the datasets for the projects. Data should be stored and shared in a secure way in the system. The data sharing platform also should provide integration with other data sharing platforms. Open data should be treated as first-class citizen in the platform to access, consume, and publish it.

4. Data storage instractructure. To connect with the data sharing platform and to store the logs of the system. All the collected of traffic and usage of the platform is open data by default. The datasets of the organisations have restricted access with high levels of control to the users. Notifications and report usage should be available for the users.

5. A data processing platform. Infrastructure to process the analysis should be separated and modular. Multiple options should be offered and also the chance to process the project somewhere else and connect with the results once they are completed. This layer is also responsible to implement the monitoring processes of the system.

4.1 The Process

In the marketplace we want to connect data scientists with the third sector. Organisations can submit their challenges and datasets, and data scientists can search and reach those organisations working on the causes that matters the most to the them, and help them to analyse their data.

On of the main targets are organisations without access to data scientists or data analysis projects. But everyone is welcomed and we think this has potential to be used in academia as a coordinating tool in research groups around data for good projects.

The process:

- An NGO registers in the marketplace. Filling their profile will give people looking for projects a clear understanding of the causes and work they are doing. Organisations then can submit a challenge request which is open to anyone who would like to join and work with them. They fill a form trying to collect as many information as possible to match it with data potential data scientists.

- Data volunteers sign up to the website at any time from everywhere. They fill a form indicating the causes they care about, their skills and their aspirations. The platform then offers a matching service.

- Both parties can request meetings which are performed in person whenever is possible or through a video call. Once both parties agree to connect they start conversations. If both parties agree to start a project, the volunteer will have access to the available data facilitated by the organisation.

- The volunteer starts planning, defining milestones, and a timeline for the the project. After a revision and agreement the project starts with touch-points defined.

Collaboration is key so the data volunteer will have a mentor who will lead and solve doubts about the analysis, the project, tools, etc. There will be up to two peers checking and making comments about the work is being done. The work can happen in collaborative environment with notebooks or in a git repository for the project. A template of guideline analysis can be created from the results of the project and make available to the rest of the users of the platform.

4.2 Why a Marketplace

A Marketplace for Data for Good will allow nonprofit and social sector organisations to publish their data challenges. The system includes a data storage layer allowing the users to upload and import their datasets with security measures and restricted access to non authorised users. They will control the access to their datasets and contact information.

The system includes a collaborative data science environment allowing the data volunteers to start their collaboration from the system or to connect with their favourite data science environment. Integration between systems is key allowing to work in the preferred languages and libraries.

A marketplace will allow to connect data volunteers searching for projects to collaborate with. Organisations in the social sector can upload their work and datasets allowing to digital volunteers to access the summary of their projects. To collaborate an extra step will be necessary and that is an request of collaboration between them, after a meeting or an statement letter. Once both agree upon the work to be done, then the organisation will grant access to the necessary datasets for the problem at hand.

A marketplace is necessary to coordinate efforts among interested parties. It can be used as a mean of information exchange of available resources and challenges in social and data science. A marketplace would allow to group users by communities giving them visibility of the work their peers are doing and following the work other people in their communities are doing. The data collected from the usage of the system can also help to understand how the efforts are deployed and what needs to be done in the future.

A marketplace idea can be used in academia for example. A research group can collect a series of potential data for good projects and put them in the marketplace. Then visibility and coordination can be managed through the marketplace as it is where all can meet discuss and ask about the projects.

All the transactions including web searches, data processing, data accesses, etc. will be logged and stored in the platform. These logs and data will be available to as open data. Accountability and scrutiny should be part of the design of the platform.
The goal of a marketplace is twofold: to make digital volunteering scalable allowing organisations and digital volunteers to join at any time and start collaborations from anywhere, and also to provide the right tools and resources to make sure that the application of data science is been done.

In summary, to facilitate collaboration among practitioners, researchers, social workers, and civil society in general. And allowing institutions to reach out skilled data people to collaborate with them.

4.3 Benefits of a Marketplace

The aim is to have a fully transparent system. Open source software should be used whenever is possible granting accountability, making possible to review the code and understand what is been done.
Giving access to the logs and usage of the platform will help to understand successful projects and replicate or recommend actions to new projects and users.
The data processing and modelling can be published as open source code building up a knowledge repository of past experiments and analysis. The code could be reused with proper rights and recognition.
A marketplace allows collaboration from anywhere in the world and is not limited to specific locations where events are held. Collaboration in data for good can be an ongoing process and not limited to the occurrence of data for good events.
Interaction between people in the data for good world has not reached an optimum. They are still connecting and meeting at data for good events around the world. There is a low level of interaction and integration between the multiple groups working in data for good and the third sector or social sector is still apart from the data experts that want to give some of their times to contribute to the causes of nonprofits and responsible enterprises.

5. ARCHITECTURE & DESIGN

The system is composed by multiple components. It is designed as highly modular and planned to interface and integrate with other platforms that data scientists and data engineers would require to use in the future.

Data privacy should be considered from the beginning [7]. Giving for example options to share or not the collected data about the projects the users are working on. The system should be explicit on what data is collecting and what is its purpose and gives control to the user on the personal data. Personally identifiable information shouldn't be stored in the system but in exceptional cases. All the collected data from the system usage should be open data by default unless there organisations or data volunteers have an issue with it and in those cases the options will be in place to limit the access to the published data.

The system can be thought in terms of independent layers providing specific services. Integration with other systems and layers is at the core of the designing principles which are:

1. Provide a collaborative data science environment.

2. Provide storage capacity or integration to the necessary datasets.

3. Provide a way to connect and visualise the data challenges in the system.

4. Provide a simple way start working in a project after a successful conversation with the organisation that publishes the data challenge.

Design criteria for the system:

- The system provides a marketplace to connect organisations and digital volunteers. They can start working on projects from inside the platform. Having access to a secure data storage and using a collaborative data science environment. The system should integrate forward and backward with useful technologies and platforms. The system shouldn't cause at any time a lock-in in the platform.

- The components of the systems are. A data storage layer where data can be stored or referenced. A collaborative data science environment, where data volunteers can work, write code, link to their projects, learn and contribute to the system. And finally an e-commerce platform to publish ”data challenges” where in this case the products are ”data challenges” and the vendors are NGOs and third sector institutions that can publish their problems around data. Data volunteers can search offer their help to the institutions that matter to them the most.

- The data storage layer should be distributed and fault-tolerant. The system should be designed for multiple small datasets and with the capability of handling big datasets eventually. A fast format file should be use to store the data.

- The storage layer should provide a good performance for analysis. A column oriented format such as Apache Parquet in Apache Hadoop would make a good choice.

- The data analysis layer should provide a collaborative data science environments or a selection of them. Also, the possibility to connect with a preferred data science tool. Integration with other data science environments is a concern for the success of the system.

- The data processing layer should provide the necessary infrastructure to compute the analysis designed in the previous layer. The system should allow to connect to multiple cloud providers to send the processing there or using a standard internal resources provided for the projects.
- Data Privacy should be incorporate privacy as a non functional requirement [7] of the system. Policies in all layers should be in place to design the system.
- The system should provide resource allocation, security, and tools to publish projects and data challenges, as well as tools to work in data science projects.
- The system is not designed yet to handle highly sensitive data and information. Highly sensitive data projects should be managed in a different environment with higher security policies.
- The system should be designed to store and share common datasets to facilitate data exchange between volunteers and organisations.

5.1 Collaboration & Education

Collaboration is key to enable a good ecosystem of helpers helping the people working in the causes. It is important to connect them and provide the right data science tools to solve the problems that charities and nonprofits present.

Education should be an important part of the system. A mentoring scheme should be put in place to start accepting digital volunteers. So not only experienced people can collaborate but neophytes in data science with the aim of learning and collaborate. The platform will encourage the use of open source software and open sourcing of the analysis. Analysis templates code can be published and reused so new projects doesn't necessary to start from zero, they can be based or reuse code from a similar project or cause and adapt it for their own purposes.

The recommendation in collaborative work is that projects should be handled by one data volunteer having two peers that can review and check the code and analysis that is been produced. A mentor is meant to be assigned to guide and make recommendations from experience in data science.

A mentoring scheme is recommended for the system. Seasoned data scientists could lead from experience to the new data volunteers into the application of data science techniques to specific problems.

6. CONCLUSION

The design of a system to scale digital volunteering is presented. A marketplace to connect data volunteers with nonprofits and organisations working in the social and third sector. This would help to grow the number of data volunteers helping organisations with data science projects.

The idea of building a system with a marketplace to connect data scientists with organisations in the social sector was discussed.

The architecture and design criteria for the system was presented for anyone to build it upon these ideas. The designed system provides the necessary resources to start collaboration projects between data experts and organisations requiring them. The system provides then a data storage for the datasets of the organisations and a collaborative data science environment.

I'm currently working in the implementation of this system. If you want to collaborate please contact me.

REFERENCES

[1] DataKind. DataKind. 2011. URL: http://datakind.org/ (visited on 07/08/2018).

[2] DrivenData. DrivenData. URL: https://www.drivendata.org (visited on 07/08/2018).

[3] The Alan Turing Institute. Data Study Groups. 2016. URL: https://www.turing.ac.uk/collaborate-turing/data-study-groups (visited on 07/08/2018).

[4] The University of Chicago. Data Science for Social Good. 2013. URL: https://dssg.uchicago.edu (visited on 07/08/2018).

[5] Diego Arenas. Data Science for Good Repository. 2018. URL: https://github.com/darenasc/data-science-for-good (visited on 07/08/2018).

[6] United Nations. Sustainable Development Goals. 2015. URL: https://sustainabledevelopment.un.org/?menu=1300 (visited on 07/08/2018).

[7] Jeroen Van den Hoven. "Value sensitive design and responsible innovation". In: Responsible innovation: Managing the responsible emergence of science and innovation in society (2013), pp. 75–83.