The case for Data for Good

Diego Arenas Contreras

da60@st-andrews.ac.uk

School of Computer Science, The University of St Andrews

July 2018

Abstract

Data for Good has gained popularity in the last years due to its relevance to society and doing good in the world. Initiatives from well-intended people such as DataKind and DrivenData have shown the relevance of this subject and the global interest it generates. But it will be argued that these efforts are not enough and it still requires an engaged network of collaborators participating in data for good initiatives. This position paper revise this idea and suggests a solution using the strengths from the work done so far. There are isolated initiatives that doesn't work together and doesn't integrate very well to each other except some tools based on standards. This paper revise what is been done so far in Data for Good and estates the we need new actors and scalable and reproducible solutions.

1 Introduction

In recent years there has been an increasing interest in Data for Good. A growing number of institutions organising projects and events around the world with the purpose of how to use Data for the common good.

In this paper when we refer to Data for Good (DfG) we are referring to the concepts of Data Science for Good, Data for Social Good, Artificial Intelligence for Good, Data for Humanity [21], etc. We are meaning the use of data techniques to solve real world issues, humanitarian problems, concerning the social good of our societies.

The DfG initiatives are allowing more people to participate in these instances to do good. Nonprofits are invited to share their data challenges and datasets. People participating on these events includes a variety of backgrounds such as data scientists, designers, project managers, social workers with relevant knowledge about the specific cause and problems of the nonprofit.

In this position paper it is argued that the efforts so far in Data for Good are not enough and it is required a scalable and integrated approach. Two main problems were identified: 1) The lack of integration among the multiple platforms available for Data for Good. And 2) there is still a scalability problem or how to make data volunteering scalable.

We will link the discussion of DfG with topics such as Open Data, Open Source, and the 17 Development Sustainable Goals or the United Nations [25] for 2030.

Section 2 we will talk what has been done so far in DfG. Subsection 2.1 established the ultimate goal to achieve and in subsection 2.2 explains some of the projects, institutions and people working in these matters. Section 4 discusses the needs and improvements. And finally in Section 5 we will present the conclusions.

2 Data for Good so far

To the best of our knowledge there is no wide coverage of Data for Good activities. Most of the online content is about Big Data, Data Science, Machine Learning, etc but with a business or corporate focus. Data for Good content is slowly appearing as tracks at important Data conferences. Most data for good practitioners are still disconnected to each other and starting to meeting at conferences, and data for good related events.

Some intents of covering Data for Good have been made [3] [2] but these efforts are not enough.

Support from big tech companies, philanthropists, and foundations has never been more important than in these days to the development

of the Data for Good. It is necessary to create new business models to support data for good initiatives. Clicks on ads should no longer be the main support for the sector, but by itself.

2.1 Why Data for Good?

Hans Rosling makes it clear in his postmortem recently published book [30]: "The ultimate goal is to have freedom to do what we want." In the sense that everyone spending their time on activities that fulfil them to the most, having dignified living conditions and access to opportunities to exercise follow their passions and curiosity.

To achieve the ultimate goal we need to work on the right subjects and at the same time make good decisions about them. The starting point is to fight the wide ignorance about the current status of our world. Books such as [30], [27], [23] and websites such as [9] are relevant because they help to better understand the status of the world today. The second step is the use of data driven approaches to make decisions improve the status quo.

2.2 Brief history and relevant institutions

2.2.1 Institutions

DataKind [11] is a nonprofit that organises events where data enthusiasts meet up to work on challenges presented by selected enterprises and nonprofits. It is also possible to work in long-term projects. DataKind was funded in 2011 with the mission of using data to help organisations in the social sector. This mission is shared by many institutions and programs since then.

The University of Chicago started the Data Science for Social Good summer program in 2013 [7] and a conference about Data Science for Social Good in 2016. The summer program selects data volunteers to work on data for social good problems for 3 months during the summer.

DrivenData [13] organises data science competitions for social good. Data enthusiasts can compete for a prize and the hosted competitions have an impact in society. It works in a similar way as Kaggle[1].

Due to the importance of data for good in recent years, more companies are joining the movement of using data for good causes. For example, Kaggle introduced their Data for Good events [20] in 2017. Facebook announced the crisis and disaster maps in 2017 [29] in an effort of using Data for Good.

The UN Global Pulse [28] is a leading organisation in innovation in the use of Big Data technologies to harnessing humanitarian issues.

Partnership in AI [1] funded in 2016, was launched as a nonprofit organisation by big tech companies to work around AI developments. They are funded in six pillars and the sixth one is AI and Social Good.

2.2.2 Events

The Bloomberg's Data for Good Exchange [5] launched in 2014 to discuss subjects about the use of data for good. Is one of the first spaces to talk about data for good and exchange opinions.

The AI for Good Global Summit [17] organised in Geneva since 2017 invites people working in AI and Data for Good projects to participate. Is still in early days is looking to establish collaborations among participants. This year they launched and AI for Good repository [18] and announced an ambitious project of an integrated data sharing platform called Data Commons [22] [19] that we should be hearing more about it soon.

2.3 The 17SDGs

The 17 Development Sustainable Goals (17SDGs) launched by the United Nations (UN) in 2015 with goals for the year 2030, makes a perfect framework to apply Data for Good. We believe that aligning with the 17SDGs is a good strategy for any endeavour around Data for Good.

Helping the helpers is a sound approach to data for good. Starting projects with nonprofits that are already working towards the 17SDGs would benefit them and have a real impact on society.

3 Are we in the right path?

We must always keep in mind that we are around 7.6Bn people [24] living in the world today. We require good solutions to be *scalable* and reproducible around the globe. We would like to provide the same opportunities that the top first billion people have access to [30] to the rest of the population. Open sourcing data solutions is a first step. We need to work on guidelines and standards [6] to work in Data for Good. And

[1] http://www.kaggle.com

giving the right tools to nonprofit and social enterprises to collaborate and connect with civil society and governments.

We can find many data for good initiatives but often they will not be related to each other. This is the *integration problem* we are talking about.

Good examples are solutions such as DrivenData, open sourcing the winner projects and working on reproducible data science using container technologies; or platforms like Data.World [12] offering integration with multiple data platforms and at the same time working with standards like RDFs and linked data format.

3.1 The Problem

So far, none of the previously mentioned institutions allow for a scalable model. Some of them encourage to keep in contact with the nonprofit but they don't provide formal mechanisms to do so.

There is a limited number of people and institutions that can participate at each event. Events are organised with months of difference and some of them happen once a year.

There is space and need for new actors in Data for Good. Provider and Consumer is an old model. Some disruption is required. The first barrier is ignorance. The second barrier is dedication. The third barrier is evaluation. The fourth barrier is sustainability.

Data Ignorance. We need better ways to communicate what machine learning, artificial intelligence and big data can do for nonprofit organisations. Understanding the needs of organisation in the social sector and being able to translate it to quantitate and qualitative analysis when required.

Effort. It is difficult for nonprofit organisations to see the value in their own data. They are most of the time consumed on their day to day operations. This has an impact on the amount of time people from the institution dedicates to help data scientists and data volunteers on their problems.

Lack of evaluation methods. How big or small is the impact of a data product or data project as the one we are involved? This kind of questions should be answered in the beginning of a project. Presenting potential gains vs status quo. This work could be done by digital volunteers with the guidance of seasoned volunteer data scientists.

Sustainable solutions. An open problem around Data for Good is how to make them sustainable. We need to think of new ways to fund and support Data for Good projects.

How many people have participated in one of the Data for Good events? My educated guess is less than the amount of interested people. This may sound obvious but we need a scalable solution to lowering the barriers of participation in Data for Good. The limitation in available spaces doesn't make it scalable to most organisations and to many skilled data scientists that could contribute to the projects.

3.2 Open Source & Open Data

The openness philosophy is highly aligned with the principles of Data for Good. Open Data embraces *transparency* and *accountability* as key principles, the same principles that Data for Good should work for. The sharing principle of Open Source software enables collaboration, innovation, re-usability, and reproducibility. Openness allows to the external scrutiny of the solutions, third parties interested in the same problem can get involved in the elaboration of the solutions.

Open Data is a good enabler of Data for Good projects. There are many data portals providing rich datasets [15] [14] [8] [26] [10] [4] to develop Data for Good projects.

When open data is not enough, it is good to have alternatives such as [16] where data from private companies can be used for the common good.

Open Source software adds a lot of value to companies and civil society. Having access to what exactly is the code doing allows, most of time, to check fairness, biases, and scrutiny the algorithms in use for certain tasks.

Open sourcing the projects facilitates reproducibility of the results, and to replicate the projects for other organisations. Civil society can learn from the open sourced solutions.

Many of the data science tools, programming languages, and libraries used in data science are open source.

4 Discussion

Most of the data for good events where nonprofits and civil society is invited to participate in

could be summarised as the following: Nonprofits present and invite the attendees to work on their data challenges. Attendees chose a problem to work on for the rest of the event. The work around the data challenges is performed during the duration of the event. A final presentation is prepared where insights and results are presented to the nonprofit and audience.

The results are often impressive to the nonprofits because they haven't explored the datasets or because they don't have skilled people to perform the type of analysis that that are carried out in the events. From empirical observation participating in type of events I can notice that:

1. Most of the implementation could be done by the senior people on each team. Reducing the number of required participants in the implementation.
2. The ideation process of analysis, hypothesis generation, and expected results from the data is very important and is richer when multiple backgrounds are in the team.
3. New participants benefits from interacting with more experienced participants in this kind of collaborative events.

4.1 A Solution

An online data platform to enable collaborations between civil society and nonprofits is required. This platform should have the following requisits: Should allow data sharing and connection to multiple data sources and platforms with security and privacy settings; it should provide a collaborative data science environment and reproducibility of the solutions. A mentoring scheme with seasoned data scientists guiding the work of junior data volunteers should encourage their participation and work with nonprofits in data for good problems.

An online data platform with these characteristics would allow to escalation of data for good work to any location in the world, helping local nonprofits and social enterprises. A network of data volunteers with reproducible work for new nonprofits would advance the field of data science for good in many levels.

4.2 Moral reflection

The fact that a startup does well in venture capital doesn't mean that it's purpose is right. Getting money is validation from the environment, which can be manipulated and highly influenced by the return of investment. A more open question would be: **Are we transforming the world in a positive way?**. The current economic system will put pressure on the economic results of the ideas but additional ethical checks must be performed on new ideas. Is this worth doing, is no longer referring to an economic value perspective but to a societal value one.

Scope and improvement of people's lives should be taken into account. New proposals and solutions should be viewed as collaborators, not as competitors. Knowledge share will be to innovation what open source has been to building high quality and better software.

5 Conclusion

In this position paper we present two problems with the current status of data for good. The lack of integration between most of the plataforms available and the limitation on scalability of the solutions.

We propose a system for data science for good and presented arguments of the foundations related to Open Source and Open Data. We suggest the work on Data for Good should be aligned to the 17SDGs.

To achieve the ultimate goal defined by Hans Rosling we need scalable and reproducible data science solutions. Allowing collaborative work and integration among applications and platforms used for data science.

A network of digital volunteers should be built around this principles that will help to solve our social problems through the use of data science.

Never before support from foundations and governments have been so active. There are interest in social enterprises and social good. Never been a better time to work in Data for Good.

References

[1] Partnership on AI. *Partnership on AI*. 2016. URL: https://www.partnershiponai.org (visited on 07/08/2018).

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

[3] Diego Arenas. *What is Data Science for Good?* 2018. URL: https://opendatascience.com/data-science-for-good-part-1/ (visited on 07/08/2018).

[4] The World Bank. *World Bank Open Data*. URL: https://data.worldbank.org (visited on 07/08/2018).

[5] Bloomberg. *Data For Good Exchange*. 2014. URL: https://www.bloomberg.com/company/d4gx/ (visited on 07/08/2018).

[6] Raja Chatila et al. "The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]". In: *IEEE Robotics & Automation Magazine* 24.1 (2017), pp. 110–110.

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

[8] Open Corporates. *Open Corporates*. URL: https://opencorporates.com (visited on 07/08/2018).

[9] Our World in Data. *Our World in Data*. 2015. URL: https://ourworldindata.org (visited on 07/08/2018).

[10] Data4SDGs. *Global Partnership for Sustainable Development Data*. URL: http://www.data4sdgs.org (visited on 07/08/2018).

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

[12] Data.World. *Data.World*. URL: https://data.world (visited on 07/08/2018).

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

[14] Human Data Exchange. *The Humanitarian Data Exchange*. URL: https://data.humdata.org (visited on 07/08/2018).

[15] Google. *Google Public Data Repository*. URL: https://www.google.com/publicdata/directory (visited on 07/08/2018).

[16] The GovLab. *Data Collaboratives*. URL: http://datacollaboratives.org (visited on 07/08/2018).

[17] ITU. *AI for Good Global Summit*. 2017. URL: https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx (visited on 07/08/2018).

[18] ITU. *ITU AI Repository*. 2017. URL: https://www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx (visited on 07/08/2018).

[19] ITU. *Roadmap Zero to AI and data commons*. 2018. URL: https://news.itu.int/roadmap-zero-to-ai-and-data-commons/ (visited on 07/08/2018).

[20] Kaggle. *Data Science for Good in Kaggle*. 2017. URL: http://blog.kaggle.com/2017/11/16/introducing-data-science-for-good-events-on-kaggle/ (visited on 07/08/2018).

[21] Frankfurt Big Data Lab. *Data for Humanity: An Open Letter*. URL: http://www.bigdata.uni-frankfurt.de/dataforhumanity/ (visited on 07/08/2018).

[22] MultiMedia LLC. *MS Windows NT Kernel Description*. 2018. URL: https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be (visited on 07/08/2018).

[23] Nan Maxwell. "The Great Escape: Health, Wealth, and the Origins of Inequality". In: *American Economist* 59.1 (2014), p. 92.

[24] World Meters. *Current World Population estimation*. URL: http://www.worldometers.info/world-population/ (visited on 07/08/2018).

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

[26] OD4D. *Open Data for Development*. URL: http://od4d.net (visited on 07/08/2018).

[27] Steven Pinker. *Enlightenment now: the case for reason, science, humanism, and progress*. Penguin, 2018.

[28] UN Global Pulse. *UN Global Pulse*. URL: https://www.unglobalpulse.org (visited on 07/08/2018).

[29] Facebook Research. *Facebook Disaster Maps: Methodology*. 2017. URL: https://research.fb.com/facebook-disaster-maps-methodology/ (visited on 07/08/2018).

[30] Hans Rosling, Anna Rosling Rönnlund, and Ola Rosling. *Factfulness: Ten Reasons We're Wrong about the World–and why Things are Better Than You Think*. Flatiron Books, 2018.