Charting the Data for Good Landscape: Principles and Methodology

In creating the landscape of Data/AI for Good, we recognized that we were not just classifying initiatives as being “Dataset Providers” or “Data Talent Providers”, but also taking a stance as to what those categories should be. While we could have just used our experience to name a few categories, that would have been perpetuating the same terminology issues with “Data for

Good" that we are trying to combat. Instead, we wanted to create an ontology for organizations in the Data for Good landscape so that they naturally clustered according to similarities in their missions, activities, and operations. This ontology was developed through a combination of interviews with Data/AI for Good practitioners, research, and the author's experience applying data science to social impact problems. We feel the ontology is as useful a tool as the landscape that it generates, and we hope it will help drive discussions about the specifics of what it means to "do data for good".

Note: Like the landscape, this ontology is also a work in progress, so stay tuned for updates as we receive feedback on it.

Creating the Ontology

Before we describe how we arrived at the ontology, let's show you the finished product. The ontology is a simple four dimensional questionnaire that can be used to classify Data/AI for Good initiatives:

Which outcome is your initiative committed to?	Reducing harm			Increasing benefit from		
What system(s) are you influencing?	For-profit	Non-profit	Government	Research	Other	
Which aspects of creating or using data does your initiative seek to influence?	Problem design	Data collection	Data storage and access	Data scientists	Data outputs	End user use
Which activity does your initiative use to influence it?	Generate research about it	Advocate for changes to it	Regulate it directly	Create more of it / Create differently	Apply it in new ways	Fund it

All of these questions can have multiple values selected. For example, the University of Chicago runs a program to train data scientists to work in the social sector. That program's raison d'etre is that there are not enough data scientists trained to work in the social sector, and that by training them accordingly governments and nonprofits would have access to skilled talent that could help them create new data solutions (increase benefit from data) and use data more responsibly (reduce harms from data). For this University of Chicago program, we'd fill in the questions above like so:

Which outcome is your initiative committed to?	Reducing harm			Increasing benefit from		
What system(s) are you influencing?	For-profit	Non-profit	Government	Research	Other	
Which aspects of creating or using data does your initiative seek to influence?	Problem design	Data collection	Data storage and access	Data scientists	Data outputs	End user use
Which activity does your initiative use to influence it?	Generate research about it	Advocate for changes to it	Regulate it directly	Create more of it / Create differently	Apply it in new ways	Fund it

You could argue that other boxes should be checked for this group (perhaps training more social data scientists will increase staffing in Research as well) - go ahead and debate that! The ontology is not absolute truth, but a tool for helping us discuss what makes groups similar or not. We believe this version gets us in the ballpark of alignment, with the understanding that community feedback and iteration will continue to inform this ontology.

Now that you've seen what we want the end result to be, let's talk about how we got here and, more importantly, what it means to "do data for good".

Making a Data *Something*[1]

The first principle we need to acknowledge is that all of our efforts to use Data for Good or AI for Good revolve around operating on data with a computer to create a data *something* that wasn't there before. That *something* could be a report, a visualization, an economic model, an algorithm, or anything else you could create from data. This sentence may sound clunky, or so obvious as to not need mentioning, but there are two important truths in it that are often not acknowledged explicitly:

1. We are talking about operating on digital data with computers, not working with handwritten information. The digitization of information and ubiquity of computing is what kicked off this "Data Age" in the early 2000s, so while these techniques can also apply to handwritten information, we are focused on their application to digital data.
2. The end goal of all "data" and "AI" efforts is to create a new *something* that allows us to do or know something we didn't before. People often talk about data as an end in and of itself, because it is often conflated with knowledge or facts. The implicit belief is that once the data is in your hands, your problems are solved. You can hear this assumption

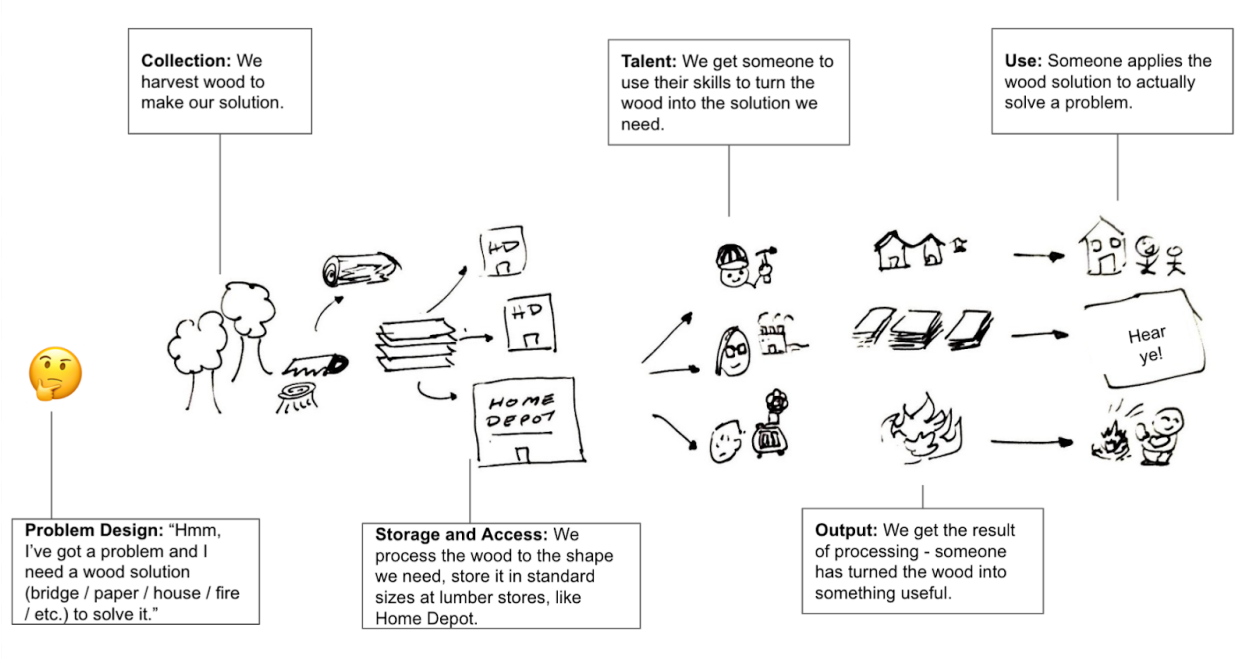
[1] We needed a stand-in here for the types of things you can create with data and, out of a frustration of language, we're writing *something*. We tried "data product", "data solution" and "data output", but those seemed to confuse people further. If you're interested in getting nuanced with what types of outputs one can create with data, check out The Three Uses of Data (yes, there are only three!)

in the ubiquitous phrases “we need data”, “if we can get the data to show X”, “we released the data so therefore...”. However, data is almost always just the material for gaining knowledge, modeling the world, or building an algorithm. There is a process to making it useful and conditions under which it will be successful when it’s applied.

The creation of that new data *something* is the center of all activity. Every Data for Good or AI for Good initiative is focused on changing some manner of how those data *something*s are created or used.

The Data / AI Production Cycle

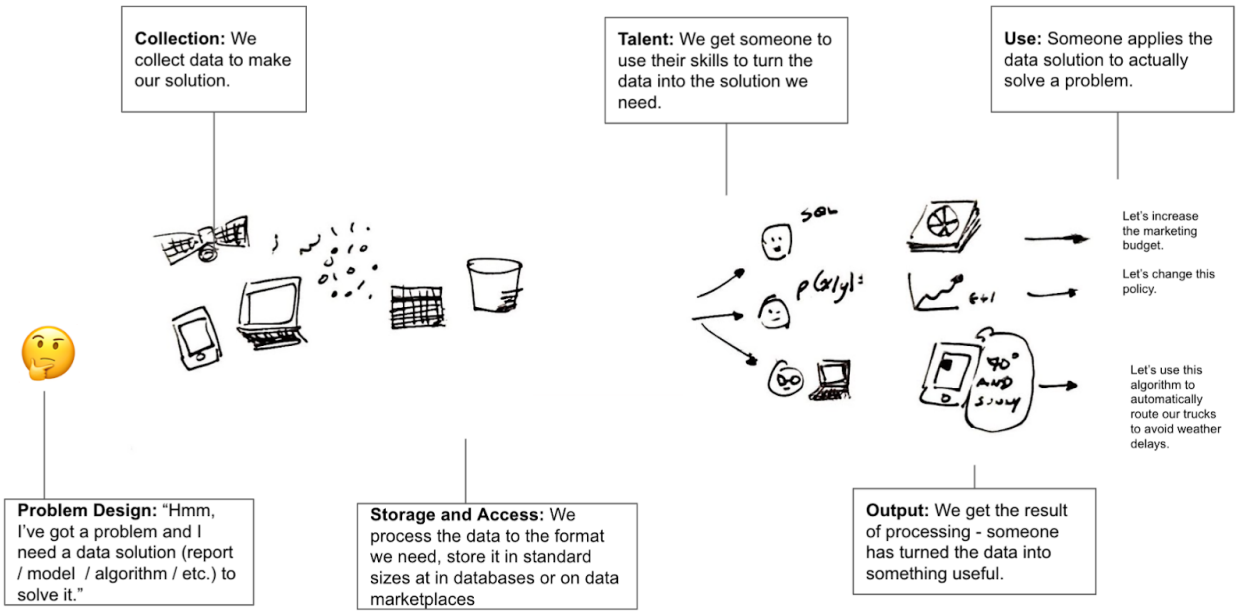
If we’re going to be creating a data *something*, let’s talk about how that data *something* gets made. We often joke that saying “Data for Good” is about as meaningless as saying “Wood for Good”, because it focuses you on the material (wood) instead of what you’re doing with it (e.g., building bridges, making paper, burning fuel, etc.). That analogy is also a useful, albeit imperfect, analogy for thinking about what it means to apply data to problems. Imagine a simplified version of what it takes to build something out of wood to solve a problem:



This graphic describes a simplified creation process for wood solutions. Even if you’re not a carpenter or a paper maker, you can probably relate to these steps for making something out of wood. What is most important to take from this analogy is that, while there are uses for raw wood (a stump can make a great stool), most things you use that are made from wood went through a process of refinement. Moreover, the type of talent you needed to create your wood solution depended on what the solution was (e.g., a carpenter isn’t going to be good at making paper, and someone designing wood-burning stoves isn’t going to know how to build a house). Lastly, how you applied that solution probably depended on what goals you had and what the

context was. A wooden structure can be used for storage, used as a home, a piece of art, or a prison. It can be seen as a boon, an eyesore, or an evil. The context of the final use of the wood product at least in part informs how useful, appropriate, and ethical it is.

Well, the process of creating with data is not all that different, and can be (over)simplified to match those six stages:



Though imperfect, this model can be useful for hitting home our points in the first section, as well as explaining the rough set of steps one needs to use to apply data:

- **Problem Design:** One needs to know that a type of data solution can solve a problem they have.
- **Data Collection:** One needs to be able to collect or obtain data to make their solution.
- **Data Storage & Access:** One needs to either store their data or access existing data to create their solution. They also need to keep it private and secure, as needed. As any data scientist will tell you, this is also where the everpresent step of cleaning and organizing the data takes place.[2]
- **Data Talent:** One needs data talent to transform the data into something useful, and the type of talent depends on what you're building. Like sitting on a stump of wood or burning wood in a fireplace, some uses of data may not need deep expertise. Others, like writing an algorithm or building a predictive model, will.

[2] You could argue that transformation and cleaning of the data actually happens after it's stored and accessed. Indeed the term ETL (extract-transform-load) describes exactly this process of manipulating data from a database to make it ready for analysis or other use. For the sake of this landscape, however, we are lumping it in under "storage & access" because we have yet to find an initiative that intervenes just on that stage. It is still a very personal, internal process that most organizations have to suffer through on their own.

- **Data Output:** By having someone apply their data skills to the data, we've created some new output. Just like wood, these outputs can vary greatly, from written reports full of pie charts to Siri-esque chatbots. They are the key data *something* we seek to create to solve our problem. Everything in the Data for Good and AI for Good revolves around this key moment of creation.[3]
- **Data Use:** We're not done just because we've created something new. How it's used, in what contexts, and under what constraints affects whether the solution is effective and ethical or not.

Affecting the Pipeline - The Actions that Define Data for Good

Every Data for Good or AI for Good initiative seeks, as its mission, to somehow change the way that process above works today so that more "good" is done (more on "good" in the next section). Either that process is happening in a way that is suboptimal or harmful today, or not enough of that process is happening in ways that are helpful. Initiatives tend to focus on one or a few stages of the pipeline to take action on. For example, the Humanitarian Data Exchange is a large data marketplace for open data sets about humanitarian issues. They are thus affecting the Storage & Access stage by providing more data, so that it can be applied to solving pro-social humanitarian problems. AI Now's report on Litigating Algorithms is a piece of research focused on the Use stage of the pipeline, highlighting uses and abuses of algorithmic decisionmaking tools in government. A few initiatives are general enough to talk about the whole process, but most focus on one or two stages in particular.

In addition to identifying the stage(s) that an initiative affects, we can also identify the actions they take to make a change. For example, if we just take the stage of "Data Collection", it's clear that there are many different strategies for affecting data collection. For example, Open Data Kit provides software to increase nonprofits' ability to collect data, thus creating more data availability. The Engine Room's Responsible Data Guide is a toolkit for helping civil society actors thoughtfully and responsibly collect data. In the US, the Federal Trade Commission sets laws and policies for data privacy during collection. There are a standard set of actions that an initiative can take to change a stage of the pipeline, which we've listed as:

- **Research it:** Create knowledge about that stage to educate the public and others.
- **Advocate for/against it:** Advocate for changes to that stage of the pipeline through changes to standards and laws.
- **Regulate it:** Legal bodies can change laws directly for that stage of the pipeline.
- **Create more:** Increase the supply of that stage of the pipeline
- **Apply existing:** Apply that existing stage of the pipeline in a new way
- **Fund it:** Fund activities in that stage of the pipeline

[3] As a reminder, the Three Uses of Data goes into more detail about what these possible outputs can be. [insert link]

Having defined the data/AI production cycle and the actions we can use to intervene on it, we now have two of the dimensions of the ontology: a Data/AI for Good initiative will focus on one or more stages of the pipeline and try to change it/them with an action. Think of the two dimensions as a pair of **action** and **pipeline stage**, for example data science training programs **create more data talent**, while foundation machine learning portfolios **fund data outputs**. Some initiatives may have more than one pair, for example data science consultancies **apply existing data talent** to **create more data outputs**.

What's your "Good"?

The first two dimensions give us a little more specificity around the phrase "Data" or "AI" in the terms "Data for Good" and "AI for Good", because we can now talk about the stage of the data/AI process they care about and how they want to change it. However, that framing applies to any data science or AI initiative, good, bad, or otherwise. Malevolent governments want to **create more data** about people they want to control. More benignly tech companies **fund data talent** initiatives that create more workers for the country, as well as for their companies. So how do we distinguish between these efforts and "for good" efforts?

Sidestepping a much deeper philosophical question about what "good" means, we observe that "good" varies based on the ***system*** the data solution is being created in. The system the data solution is created within defines the larger set of rules it plays by. Within each system, "good" initiatives seek to make data solutions coming from that system reduce human suffering and/or increase human wellbeing.

There are likely many systems we can identify, but in our research we found three broad camps of "data for good" and "AI for good" vis-a-vis the systems they affect and what "for good" means in each:

For-profit data products: These are the commercial data products you know like Google Maps, Spotify, Siri, etc. “Data for Good” in this camp focuses on reducing the harms that result from making and misusing these products.

Intelligent data products: This camp cares about data/AI products that are becoming too “smart”. “Data for Good” efforts in this camp focus on regulating what data/AI is allowed to do, in an attempt to reduce harms that a superintelligent AI could create.

Social impact data products: These products are designed for civil society’s goals. “Data for Good” in this context applies to either creating more ways for people to make data products in civil society, or reducing the harms they could cause.

There may be other camps, but these three highlight the last two dimensions of our ontology.

- What system does your initiative seek to affect?
- To what end does it seek to affect it, broadly grouped as “reducing harms” and “increasing benefits”

Every Data for Good initiative is focused on changing or improving one or more systems - for profit, non-profit, government, etc. - and they seek to reduce harms from it, increase the productivity of it, or both. What is important to take away is that “for good” depends heavily on the goals and incentives of the system we’re working within.

Putting It All Together

We can now put all of our ontology dimensions together into one:

Which outcome is your initiative committed to?	Reducing harm			Increasing benefit from		
What system(s) are you influencing?	For-profit	Non-profit	Government	Research	Other	
Which aspects of creating or using data does your initiative seek to influence?	Problem design	Data collection	Data storage and access	Data scientists	Data outputs	End user use
Which activity does your initiative use to influence it?	Generate research about it	Advocate for changes to it	Regulate it directly	Create more of it / Create differently	Apply it in new ways	Fund it

Read from the bottom up, you can see this ontology as filling in the blanks of this sentence:

"Our Data/AI for Good initiative (acts) on (a stage of the data science / AI pipeline) for data solutions built (in a certain system) so that we (outcome)."

In this form, the ontology reads like a small logic model - we take a certain action to drive a near-term impact within a system that leads to a long-term impact. By then classifying initiatives based on how they fill in this ontology, we can cluster them and group them based on the similarity and differences in their answers. Combined with the initiative's size, operating country, organization type (e.g. business, nonprofit), we can create a very rich analysis of the initiatives acting to create "Data/AI for Good".

Building the Groupings: Methodology

To build the initial landscape, we selected 115 initiatives identified as "data for good" or "AI for good" initiatives. They were classified using the ontology described, with input from a group of 30 international advisors. Given that all features are (technically) independent, there are 2^27, or 134,217,728, possible combinations of features an initiative could demonstrate. In our dataset, only 83 unique sets of features were observed in practice. The initiatives were then clustered using hierarchical clustering based on the features they presented. Using a combination of human knowledge and the clustering results, 11 categories of data for good initiatives emerged[4]. You can see these clusters here.

[4] The number and size of these clusters is a function of the dataset. That means that running this exercise on a larger dataset could (and likely will) produce new clusters, either by capturing types of groups we didn't capture this time or by creating a cluster so heterogeneous that a new distinguishing feature(s) is needed to split that cluster into sub-clusters. As with any data project one must ask what data wasn't included and why. This first version of the landscape is heavily influenced by the author's network, though we hope the feedback on this project and ever-expanding advisors list will change that.