Data Science Project Scoping Worksheet

Updated: July 29, 2020

This worksheet is designed for social good organizations (government agencies, non-profits, social enterprises, and others) to scope actionable data science projects.

1. Project Name:

2. Organization Name:

3. Problem Description:

3.1 What is the problem you are facing?

3.2 Who/what is affected by this problem? (people of certain type, organizations, neighborhoods, environment)

3.3 How many people/organizations/places/etc and how much are they affected? (e.g. mean wait time for surgery, number of students dropping out of school, cost due to tax fraud, etc.)

3.4 Why is solving this problem a priority for your organization?

4. Goals: What are the business/policy goals that will be accomplished by solving this problem and what constraints do you have? (in order of priority)

- The technical solution that will be built (e.g. predictive model or dashboard or map) is not the business/policy goal - that is the tool that will achieve your goal
- The goal should be specific and measurable
- Achieving the goal should help solve the problem you're tackling
- Typical goals include improving/maximizing/increasing or decreasing/mitigating/reducing some outcome or metric (such as school graduation rates or unemployment rates) .
- Typical constraints include budget, lack of human capital, legal restrictions, political will and social license.

- Consider tradeoffs between conflicting goals.

	Goal	**Constraints**
1		
2		
3		

5. Actions

- Actions are activities or programs that institutions are doing/will do to address a problem. Actions can involve allocating resources, such as inspecting facilities, providing preventive services, outreach, etc.
- Actions should improve when the institution has the information that is generated in the project.
- Ideal actions should help you achieve the goals defined above.

	Action 1	**Action 2**	**Action 3**
Action: *eg. inspection for compliance with fishing quotas for boats in ports*			
Who is executing the action? *eg. Inspector, Department of Inspections*			
Who/what is the action being taken on? eg. *fishing boats*			
How often is the decision to take this action made? eg. *Daily*			

What channels are/can be used to take this action Eg. *In person*			
Other useful information about the action			

6. Data

- The data has to connect to the actions its informing so the organization can achieve its goal
- Typical data science projects use administrative data as the primary data source, and enhance it with publicly available data sources (Census, other open data). Partnering with the private sector or non-profits could be a way to obtain data you might be missing internally.

A. What data sources do you have internally?

	Data Source 1	**Data Source 2**	**Data Source 3**
Name e.g. Hospital Admissions database			
What does it contain? Describe the attributes included in the data source. *eg. admission and discharge records for hospitals nationwide, including patient sociodemographic data, insurance type, medical doctor information, etc.*			
What level of granularity? *eg. transaction, person, organization, location*			
How frequently is it collected/updated after it's captured? *eg. real time, daily, weekly, monthly, yearly, one off*			

Does it have reliable and unique identifiers that can be linked to other data sources? *eg. SSN, National identifier, patient identifier, insurance number, etc*			
Who's the internal owner of the data? *eg. Hospitals*			
How is it stored? *eg. in a database, pdfs, excel, spss*			
Additional comments			

B. What data can you get from external, private or public sources?

	Data Source 1	**Data Source 2**	**Data Source 3**
Name *eg. Air Quality database*			
What does it contain? Describe the attributes included in the data source. *eg. distinct pollution's particle concentration*			
What level of granularity? *eg. geolocalized hourly sensor data*			
How frequently is it collected/updated after it's captured? *eg. daily*			
Does it have unique identifiers that can be linked to other data sources? *eg. sensor identifier*			

Who's the internal owner of the data? *eg. NOAA*			
How is it stored? *eg. API endpoint from an open data portal*			
Additional comments			

C. In an ideal world, is there additional data you would want to get/gather that would be relevant to his problem? (surveys, CCTV, phone records, DNA, different frequency or granularity for currently available data, etc)

D. **What analysis will the existing data not support?** For example, if we don't have access to outcomes for students in wich case any analysis predicting the outcomes will not be feasible until the outcome data (or a reasonable proxy) is collected

7. Analysis

- Typical data science projects include a combination of analysis, typically including description, detection, prediction, optimization, and/or behavior change.
- Again, the analysis is not the goal of the project - the **analysis** helps you use the **data** you have to inform the **actions** you have access to in order to achieve your **goals.**
- Choose the right set of analysis for each problem
- You must validate the analysis and use a validation process that matches how your analysis will be used in practice

	Analysis 1:	**Analysis 2:**	**Analysis 3:**
Analysis type *e.g. Description, Prediction, Detection, Behavior Change*			
Purpose of the analysis *eg. understand historical behavior of individuals, estimate risk of disease of patient, identify which actions will diminish overfishing in the region*			
Which action will this analysis inform? *eg. inspections of compliance regarding fishing quotas*			
How will you validate this analysis using existing data? What methodology and what metrics will you use? How will you compare against existing baselines? e.g creating multiple train and test sets based on time, using precision at top 10% as a metric, and comparing against a random and an			

"existing system" baseline			
What limitations will this analysis have, either based on available data or choice of methodology?			

8. Ethical considerations

Privacy Are you working with personal and/or sensitive data that is individually identifiable? Mention them.	
Transparency Which stakeholders should know about which parts of the project? *Stakeholders typically include policymakers, frontline workers, people who will be affected by the actions, etc*	
Discrimination/Equity Are there any specific groups for whom you want to ensure equity of outcomes? *eg. groups of interest defined by gender, age, localization, social class, educational level, urban/rural, ethnicity*	
Social Licence If the entire population of the country finds out about your project, will they be ok with it? Why?	
Accountability Who are the people responsible for all the things above?	
Any other considerations such as consent, legal, etc	

9. What field trial or randomized controlled trial can you design to validate the project in the field? The outcomes you will measure should match your goals. Define the population in which the model will be tested. Define the duration of the trial. Specify the baseline. You should measure the impact in different population subgroups (see section 8)

10. Who are the external organizations and internal departments that will need to be involved?
(Typically, data science projects need involvement from data owners, IT infrastructure owners, problem owner, analytics people)

Organization/Department	**Description of desired involvement**	**Name/role of counterpart**
IT department	*Provide Data infrastructure*	*Head of IT department*
Statistics agency	*Provide population data*	*Head of Department of Statistics*

This worksheet is currently being maintained at Carnegie Mellon University. Please email dssg@andrew.cmu.edu for any questions.

This worksheet was originally developed by the Center for Data Science and Public Policy at the University of Chicago. This version has been extended through a collaboration between GobLab UAI and Carnegie Mellon University.