Data Science Project Scoping Worksheet

Updated: October 1, 2021

This worksheet is designed for social good organizations (government agencies, nonprofits, social enterprises, and others) to scope actionable data science projects. Additional resources, including the Data Science Project Scoping Guide, are available here.

1. Project Title:

2. Organization Name:

3. Problem Description

*A **problem** is typically an observed, adverse outcome that is real, important, and has social impact. The problem should also be one that is prioritized by the organization and can be addressed using data the organization has access to.*

3.1. What is the business or policy problem you are facing? (e.g. adverse health impacts among at-risk children due to low rates of vaccination, low graduation rates among high school students leading to un- or underemployment, etc.)

3.2. Who or what is affected by this problem? (e.g. people of a certain type, organizations, neighborhoods, the environment, etc.)

3.3. How many of these people/organizations/places/etc. are affected by the problem, and how much are they affected (order of magnitude is fine)? (e.g. only 90% of high school students graduate on time, each organization loses $1M each year to tax fraud, etc.)

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

3.5. How have you tried tackling this problem and what has been the outcome of your efforts?

3.6. What other groups or stakeholders in your organization and outside need to be involved in scoping and implementing this project?

Typically, data science projects need involvement from stakeholders inside your organization (such as policymakers, managers, data owners, IT infrastructure owners, the people who will intervene such as health workers) as well as people and organizations from the outside (such as community groups that will be affected by this work).

4. Goals

*A **goal** is a concrete, specific, measurable aim or outcome that the organization will accomplish by addressing the **problem**. Building a technical solution, such as a predictive model, dashboard, or map, **is not itself the goal of a data science project** even if one of these tools might help you achieve your goals.*

4.1. What are your social, policy, or business goals, and what constraints do you have?

Goals should directly relate to the problem you've identified, and will typically *improve/maximize/increase* or *decrease/mitigate/reduce* a relevant outcome or metric (e.g. increase the percentage of high school students who graduate on time).

Goals often need to balance efficiency (e.g. help the most number of people in need with limited resources), effectiveness (e.g. maximize the total improvement in outcomes from the help you provide to people), and equity (e.g. allocate resources across groups to achieve equity in outcomes).

Common goal-related constraints are limited budget, people and/or time; legal restrictions or lack of political will; or lack of social license.

List goals below in order of priority.

	Goal	Goal Type (Efficiency, Effectiveness, or Equity)	Constraints Around This Goal
1			
2			
3			

4.2 What trade-offs exist across these goals?

Some of the goals above may be conflicting or have trade-offs across other goals. What are these tradeoffs? Which of these goals would you want to place more emphasis on to achieve a competing goal?

5. Actions

*An **action** is an activity, intervention, or program that your organization has, or will perform, to reach the **goal(s)** you've outlined. Actions are generally performed routinely and often involve allocating resources, such as providing preventative services, outreach attempts, or after-school programs to people, or prioritizing inspection of certain homes or facilities.*

The data and the analysis in steps 6 and 7 should inform these actions to help achieve our goals.

5.1. What actions will your organization take to address the problem?

	Action 1	**Action 2**	**Action 3**
What is the action? *e.g. inspect a house for health hazards, enroll a child in an after-school program*			
Which goal does this action help achieve? *e.g. reduce rates of lead poisoning, increase graduation rates*			
Who is executing this action? *e.g. home health Inspector (Department of Inspections), school administrator (school district)*			
Who or what is the action being taken on? *e.g. house, child*			
How often is the decision to take this action made? *e.g. weekly, quarterly*			
What channels are or can be used to take this action? *e.g. in person, digital channels*			

Are there any resource or capacity constraints with this action? *e.g. only 100 inspections can take place every month, or only 50 children can be enrolled in a support program at any time*			
What are the ethical issues associated with this action? *Will acting on someone who does not need this intervention have adverse consequences? Are there ethical issues around excluding someone?*			
Can you provide any other useful information about this action? *Has it been tested to be effective? Are there any approvals necessary before an action can be taken? How long does it take for an action to have an effect?*			

6. Data

***Data**, coupled with **analysis**, should inform the **actions** you will use to achieve your **goals**.*

Many data science projects in governments and non-profits use administrative data as a primary data source, augmented by secondary, publicly available data sources (e.g. the US Census). Partnering with a private sector or nonprofit organization is a way to obtain data you might not have internally.

6.1. What data sources do you have internally?

The data you use to perform your analyses should be updated frequently and granular enough to reliably inform the actions you've identified. For example, if your actions prioritize individuals for help, your data should be at the individual level.

	Data Source 1	Data Source 2	Data Source 3
What is the name of the data source? *e.g. hospital admissions database*			
What does it contain? Describe the attributes included in the data source. *e.g. admission and discharge records for hospitals nationwide, including patient sociodemographics, insurance type, and physician information.*			
What level of granularity/detail is the data? *e.g. inspection level, student level, patient visit level*			
How far back does the data in this data source go? *Is it sufficient for the problem being scoped?*			
How frequently is the data collected or updated after it is captured? *e.g. immediately (real-time), daily, weekly, monthly, yearly, ad hoc*			
Does the data have reliable and unique identifiers that can be linked to other data sources? *e.g. SSN, national identifier, patient identifier, insurance number*			
Who is the internal owner of the data? *e.g. Sacred Heart hospital*			
How is the data stored? *e.g. in a database, in pdfs, in excel, in a SAS data store*			

What are the ethical issues associated with using this data source? *e.g. do you need consent from the people in the data to use their data? are there security protocols that need to be in place? does the data collection process systematically result in any type of known collection biases?*			
Can you provide any other useful information about this data source?			

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

	Data Source 1	Data Source 2	Data Source 3
What is the name of the data source? *e.g. air quality database*			
What does it contain? Describe the attributes included in the data source. *e.g. particle concentration of each type of pollution in the air*			
What level of granularity is the data? *e.g. zip code level, daily records.*			
How frequently is the data collected or updated after it is captured? *e.g. daily*			
Does it have unique identifiers that can be linked to other data sources? *e.g. sensor identifier number*			

Who is the internal owner of the data? *e.g. NOAA*			
How is it stored? *e.g. API endpoint from an open data portal*			
What are the ethical issues associated with using this data source?			
Can you provide any other useful information about this data source?			

6.3. In an ideal world, what additional data would you want to have that is relevant to this problem? (e.g. survey results, CCTV videos, phone records, DNA, currently available data more frequently updated or at a different level of granularity, etc.)

7. Analysis

*The objective here is to specify a set of **analysis** the project will do that use the **data** we have to inform the **action(s)** that will achieve our **goals**.*

*The analysis is **not the goal of a data science project**. Data science projects typically include a combination of analysis types, such as description, detection, prediction, optimization, and/or causal inference.*

This section is typically not filled out in the earlier iterations of the scoping process until the problem, goals, actions, and data have been figured out.

7.1. What analyses will you complete to inform your actions?

An analysis can involve 1) better understanding and describing the past, 2) detecting new events as they're happening, 3) predicting future outcomes, 4) selecting among various strategies using optimization techniques, or 5) influencing or changing future behavior.

Each set of analysis will likely need to be validated. Initially, this may be through historical data, and eventually, through some type of a field trial.

	Analysis 1	**Analysis 2**	**Analysis 3**
What is the type of analysis? *e.g. description, prediction, detection, causal inference*			
What is the purpose of this analysis? *e.g. understand historical behavior of individuals, estimate risk of disease, identify which actions will increase graduation rates amongst students*			
Which action will this analysis inform? *eg. inspections of compliance regarding handling of hazardous materials*			
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 or positive predictive value at top 10% as a metric, and comparing against random and "existing system" baselines*			

What are some ethical issues associated with conducting this analysis?			

8. Ethical Considerations

Ethical issues should be considered continuously, in every part of the scoping process as well as during the project. This section provides a set of questions to answer as a starting point for those discussions through the project scoping, design, and execution phases.

8.1. Privacy, Confidentiality, and Security

Are you working with personal and/or sensitive data that is individually identifiable? What are the legal as well as ethical considerations for privacy and confidentiality with the data being used? What type of protections need to be in place? How are these data protections being audited, and how often?

8.2. Transparency

Which aspects of the project do different stakeholders need to be informed about? Stakeholders typically include policymakers, frontline workers, people who will be affected by the actions, the general public, etc. What should each of them know about this project? Do the people who "own" the data know how you're using it? Do the people being prioritized for intervention know why they're being prioritized?

8.3. Discrimination/Equity

For which specific groups do you want to ensure equity of outcomes (e.g. groups of interest defined by gender, age, location, social class, educational level, urban or rural residency, ethnicity, etc.)? How might each of these groups define equity in outcomes in this context? How will you detect biases in your system and reduce them or mitigate their impacts? How should you take into account any broader sources of inequities that affect the outcomes you're seeking to improve?

8.5. Accountability

Who is responsible for ensuring that each of the above ethical considerations are made? What accountability lies with the people building the data science system, the people acting on them, and the policymakers defining the goals and objectives? If there are data leaks, misuses of the system, unintended consequences, or other harms arising from this work, who is accountable?

8.4. Social License

If the entire population of the country finds out about your project, will they be ok with it? Why? Are there any specific groups who might object, and what concerns would they raise? If it was on the front page of the newspaper, would the headline be positive or negative?

8.6. Are there any other ethical considerations that should be made prior to or during the data science project?

e.g. legal issues, informed consent, etc.

This worksheet is currently being maintained at Carnegie Mellon University. Please email **dssg+scoping@cmu.edu** for any questions or suggestions.

This worksheet was originally developed by the Center for Data Science and Public Policy at the University of Chicago and has been extended through a collaboration with GobLab at Adolfo Ibanez University.