Data Science Project Scoping Guide

**Resources in this post:**  [Blank Project Scoping Worksheet,](http://www.datasciencepublicpolicy.org/wp-content/uploads/2021/09/ProjectScopingWorksheetBlank.pdf)[Scoping Presentation](http://www.datasciencepublicpolicy.org/wp-content/uploads/2020/02/scopingslides.pdf), Sample Curriculum for a Scoping Workshop

***How to use this guide:****This Scoping Guide supplements the Project Scoping Worksheet. For a brief walkthrough, please see the*[*Blank Project Scoping Worksheet*](http://www.datasciencepublicpolicy.org/wp-content/uploads/2021/09/ProjectScopingWorksheetBlank.pdf)*. For more details on any section in the worksheet, please use this guide.*

Over the past several years, Data Science for Social Good has worked on over 100 data science projects and one (unsurprising) lesson we’ve learned over those years is that it is critical to scope these projects well at the beginning.

There are a lot of organizations out there – government agencies, nonprofits, social enterprises, corporations – working on important problems that can have a huge impact on society. There are also lots of talented, passionate, and smart people with data science skills who can help them tackle those problems. Yet, when these two sets of people come together, the results are often mixed because of the challenges associated with formulating a well-scoped project. We have found that it is necessary to have people who can mediate between the two groups and formulate a problem that is both solvable and impactful. It’s even better if both groups have these scoping skills themselves so they can work together more effectively.

After scoping hundreds of projects over the past few years, we’ve learned several project scoping lessons that we wanted to share with the larger data science for social good community. We’d love any feedback and comments on our approach and for people to help us improve it.

Although this is written specifically to benefit people scoping data science projects for social good, the lessons here generalize to socially neutral (and unfortunately, to socially evil) projects as well. In fact, a lot of this is based on lessons we’ve learned from working on data science projects in the corporate world in our previous work lives.

Initial Screening Criteria

We’ll assume here that we’ve already gone over the basic criteria for doing data science for social good projects:

1. **Impactful**: The problem we’re solving is real, important, and has social impact.
2. **Solvable**: Data can play a role in solving the problem, and the organization has access to the right data (See our [data maturity framework](http://dssg.uchicago.edu/2016/04/28/introducing-the-data-maturity-framework/) to help you assess whether you have the right data)
3. **Actionable**: The organization has prioritized this problem, is ready to take actions based on the work, and is willing to commit resources to validate and implement it. (see our [organizational maturity framework](http://dssg.uchicago.edu/2016/04/28/introducing-the-data-maturity-framework/) to assess where you are).

Scoping Overview

Once we’ve gone through the initial screening process, we then start the scoping process. There are many approaches to scope a problem. We focus on projects that are actionable and lead to tangible positive societal outcomes. The approach we describe below is one that has worked for us but we certainly don’t think that is the only way to scope a successful project. Let us know if you have alternative approaches you use that we can benefit from. As always, the scoping process is fairly iterative and the scope gets refined both during the scoping process as well as during the project.

**Who needs to be involved?** The ideal participants in the scoping process include various stakeholders in the project including people who understand the business/policy problem, people who understand what data is available, people who will consume the outputs of the system or take action using it, and the people who will be affected by it.

**Step 0: Problem Understanding**– What is the problem? Who does it impact and how much? How is it being solved today and what are some of the gaps? **Step 1: Goals** – What are the goals of the project? How will we know if our project is successful?  
**Step 2: Actions** – What actions or interventions will this work inform?  
**Step 3: Data** – What data do you have access to internally? What data do you need?  
What can you augment from external and/or public sources?  
**Step 4: Analysis** – What analysis needs to be done? Does it involve description, detection, prediction, or behavior change? How will the analysis be validated?  
**Ethical Considerations**:  What are the privacy, transparency, discrimination/equity, and accountability issues around this project and how will you tackle them?  
**Additional Considerations:**How will you deploy your analysis as a new system so that it can be updated and integrated into the organization’s operations? How will you evaluate the new system in the field to make sure it accomplishes your goals? How will you monitor your system to make sure it continues to perform well over time?

The scoping process is iterative and not strictly linear. While understanding the problem and defining the goals of the projects must come before identifying actions, the process of identifying actions may help discern whether a goal is actionable and must be redefined. Similarly, assessing the data available for the project may lead us to rethink which problems, goals, and actions can be informed by a data science project. Our analysis may lead us to rethink our problem and our goals and start the scoping process anew. So, while each step follows the previous, we should take each step as an opportunity to evaluate earlier steps.

All throughout, ethics should be the center of our scoping process. Ethical issues should be considered in every phase of scoping and executing the project, rather than thought of as a discrete “step” in that process. Developing a clear and explicit understanding of your goals in this regard is fundamental to doing it well. Crucially, the integration of ethical considerations into the project should be neither an afterthought nor a burden, but rather a critical and continuous area of focus that involves all stakeholders, especially the people who will be impacted by this system.

Step 0: Understand the problem

Before we start scoping the project, we need to make sure we understand the **problem** and its **impact**. The problem should be a **priority** for the organization that can be addressed using data the organization has or can access. If the problem is not a priority, then even a well-designed model will not help resolve it because the organization will lack the motivation to act on the information resulting from the analysis.

To begin, it is important to understand the scope of the problem. We first ask organizations to describe the problem they are facing, including who or what is affected by the problem, how many are affected, and how much they are affected (i.e., the magnitude of the problem). For example, a school district may be concerned with low high school graduation rates. They should be able to tell us who is affected, perhaps low-income students or students who are otherwise at-risk. They should also be able to tell us the magnitude of the problem: for example, that fewer than 90% of high school students graduate on time and less than half of at-risk students.

We then ask the organization to explain why the problem is a priority now and how they have been tackling the problem. Understanding how the organization has tackled the problem can help us identify ways that data analysis can inform the organization’s actions to achieve its goals. For example, a school district may be concerned about high school graduation rates because they recently found that they were particularly low among at-risk students. The school district may be enrolling at-risk students in after-school tutoring programs that help reinforce in-school learning. If space in after-school programs is limited, then the right analysis could help the school district prioritize students for enrollment who are unlikely to graduate on time.

Finally, it is important to identify the groups or stakeholders inside and outside your organization who need to be involved in scoping and implementing the project. Typically, data science projects need involvement from stakeholders inside your organization, such as policymakers, managers, data owners, IT infrastructure owners, and the people who will intervene, such as health workers. Projects also often require the involvement of people and groups from outside your organization, such as community groups that will be affected by this work. For example, a school district scoping a data science project identifying students who are at-risk of not graduating from high school will want to engage senior policymakers as well as IT and data workers early in their planning. They will also want to engage school administrators, teachers, and tutoring program workers to help understand the problem and scope the project. The district will also want to engage the school board and the parents of students who are likely to be impacted by the project, especially in communities most affected by low graduation rates.

Once the problem is properly understood, we can begin defining the organization’s goals.

Step 1: Define the goal(s)

The objective here is to take the **outcome** we’re trying to achieve and turn it into a goal that is **measurable** and **achievable**. This is the most critical step in the scoping process. In the context of our projects, a goal is a concrete, specific, measurable aim or outcome that the organization will accomplish by addressing the problem. We often come across efforts where an organization will define the goal of their project as building a technical solution, such as a predictive model, dashboard, or map. **We argue that the technical solution (model, dashboard, map) is not itself the goal of a data science project. Rather, the goal of the project should be to solve some policy or an operational problem that impacts the organization’s mission.** Building a data science solution can (and if done correctly, should) help you achieve the goal you’re aiming for.

Most projects start with a very vague and abstract goal (say, improving education), get a little more concrete (increase the percentage of students who will graduate on time), and keep getting refined until the goal is concrete, unambiguous, and achieves the societal aims of the organization. For example,”improving education” is too abstract and vague to be the goal of a specific data science project while “ increasing the number of students who graduate high school on time” is a more appropriate goal. This step is often difficult because a lot of organizations haven’t explicitly defined concrete, analytical goals for many of the problems they’re tackling. Sometimes, these goals exist but are locked implicitly in the minds of people within the organization. Other times, there are several goals that different parts of the organization are trying to achieve.

One approach we find helpful in defining goals is to directly relate it to the problem we’ve identified, and 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) that we want to change.

We will often have (possibly) conflicting goals around efficiency (e.g. help the most number of people in need with limited resources), effectiveness (e.g. maximize the total improvement in outcomes for the people you help), and equity (e.g. allocate resources across groups to achieve equitable outcomes). We should not only define these explicitly during the scoping process but also attempt to prioritize them at this stage.

Usually, goals also include constraints. Constraints are often what make a data science project necessary. For example, if a public health agency could inspect every rental property in the city for housing code violations, they probably would. However, there may be a constraint on the number of properties an agency can inspect within a certain period, so they may want to prioritize the ones most likely to be unsafe to live in. A preventative public health program might want their goal to minimize the number of unplanned Emergency Room (ER) visits. A trivial (and evil) approach to achieving that would be to shut down all ERs, resulting in zero visits. Adding a constraint requiring the solution to *improve*health outcomes, and not just reduce ER visits, will help us identify solutions that have the desired social impact.

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

Let’s take a few social impact problems and see how we can define the goals.

**Example 1 – Lead Poisoning:** In 2014, we worked with the Chicago Department of Public Health on reducing lead poisoning rates among children in Chicago. The initial goal was to reduce lead poisoning by increasing the effectiveness of their limited lead hazard inspections. One way to achieve that goal would be to focus inspections on homes that are likely to have lead hazards. Although helpful, this approach wouldn’t achieve their real goal, which was to prevent children from getting lead poisoning. Finding a home with lead hazards and getting it remediated is only beneficial if there is a high chance that a child is present in the home (currently or in the future) who is likely to get exposed to lead (and develop lead poisoning). The next iteration of the goal was to increase the number of inspections that find lead hazards in homes where there is an at-risk child (before the child gets exposed to lead). Eventually, we got to the final goal: **Reducing the number of children who will get lead poisoning in the future because of lead hazards in their current residence by 1) identifying which children are at high risk of lead poisoning in the future and then 2) targeting interventions at the homes of those children to remove those lead hazards.**

**Example 2 – On-time High School Graduation:** One of the challenges schools are facing today is helping their students graduate on time. They are interested in identifying students who are at risk of not graduating on time and need extra support. When initially talking to most school districts, they start with a very narrow goal of predicting which kids are unlikely to graduate on time. The first step in our scoping process is to go back to the goal of increasing graduation rates and ask if there is a specific subset of at-risk students they want to identify?

What if we could identify students who are only 5% likely not to graduate on time versus those who are 95% unlikely to graduate without extra support? If the goal is just to increase graduation rates, the first group is (probably) easier to intervene with and influence while the second group may be more challenging due to the resources they need. Is the goal to maximize the average probability of graduating, or is the goal to focus on the kids most at risk and increase the probability that the bottom 10% of the students will graduate? Or is the goal to create more equity and reduce the difference in on-time graduation rates between those who are most likely to graduate and those who are least likely to?

All of these are reasonable goals but schools have to understand, evaluate, and decide which goals are most important to them. This conversation often makes them think more deeply about defining what their organizational goals are as well as tradeoffs between them. **A reasonable goal we may end up with after the scoping process is to minimize the disparities in graduation rates across different racial groups while maximizing overall graduation rates.**

**Example 3 – Inspections:** We’ve worked on several projects that involved inspections. Some examples include:

1. the U.S. Environmental Protection Agency and New York State Department of Environmental Conservation, prioritizing which facilities to inspect for waste disposal violations;
2. the City of Cincinnati, identifying properties at risk of code violations; and
3. the World Bank Group, prioritizing fraud and collusion complaints to investigate.

In most inspection problems, there are many more entities (homes, buildings, facilities, businesses, contracts) to inspect than the available resources needed to conduct those inspections.

The goal most organizations start with is focusing their inspections on entities that are likely to be in violation of existing regulations. While this is a good start, most of these organizations can never inspect everything that may be non-compliant. The goal they really want to achieve is **deterrence – reducing the total number of facilities that will be in violation**. An ideal inspection process would then result in reducing the actual number of violations, whether they’re found or not, which may not be the same as an inspection process that is aimed at being efficient and increasing the hit rate (% of inspection resulting in violations).

**Example 4 – Scheduling Waste Pickup**: A few years ago, we worked with Sanergy, a social enterprise based in Kenya. They deploy portable toilets across informal urban settlements and one of their largest costs is hiring people to empty the toilets by collecting waste from each of them. Since toilet usage varies, it is inefficient to empty every toilet every day, as it was done when the project started. In order for them to grow and keep costs down, they needed a more adaptive approach that can “optimize” the schedule for emptying toilets. The goal, in this case, is to make sure that you don’t over-empty the toilet when it’s not full but you don’t let it stay full either because then it’s not usable. This translates to a formulation that **maximizes the number of times toilets get emptied when they’re as close to 100% full as possible without getting to 100%**. A different formulation of the goal could be to **minimize the number of times an individual goes to use the toilet (but cannot) because it was full and not usable with the staff resource constraints they have to empty the toilets**.

**Considering trade-offs while deciding on goals**

As we start defining and prioritizing goals, often around efficiency, effectiveness, and equity, the conversation leads to tradeoffs.

When dealing with students who may need extra support to graduate on time, what do you care about most: finding every single student who may need help (who may not need the support and possibly being inefficient) or prioritizing efficiency and only focusing on students who you’re extremely sure will need the extra support (and thus missing many students) or prioritizing equity to make sure that already-disadvantaged students who may not have other resources available to them do not get left behind?

Would you rather inspect more homes without finding lead hazards in them (which is inefficient), or would you rather miss homes with children who will end up getting lead poisoning?

When dispatching and placing emergency response vehicles, do you want to make sure you can get to every possible emergency within 10 minutes or do you want to make sure that you can get to critical emergencies within 3 minutes and the non-critical within 20 minutes?

**What types of mistakes are you more willing to make?** That is a critical question a good scoping process brings up and attempts to answer based on the priorities of the organization.

In data science terms, would you rather have more false positives or more false negatives? Do you want these false positives or false negatives to be balanced across different racial, gender, age, or socioeconomic groups? Of course, this decision depends on the impact and cost of those errors, which is often hard (and sometimes uncomfortable) to quantify. This is a conversation around social and policy values you want to be embedded in your operations, and policymakers need to decide which policy goals they want to achieve, what resources they have, and which outcomes they want to prioritize. The data science work is then used to support and implement those policy goals. Data science can help explore the impact of those goals and understand the implications better but it’s ultimately a policy decision to decide on what goals to optimize and needs to include all stakeholders affected by this system.

Step 2: What actions/interventions are you informing?

The work we do can only have an impact if it’s actionable. What actions can the organization take to achieve these goals? How can we inform those actions to make them more effective? These actions often need to be fairly concrete: home inspections, enrolling a student in one of three after-school programs, targeted emails for fundraising or advocacy, dispatching an emergency vehicle, or scheduling a waste pickup.

A well-scoped project ideally has a set of actions that the organization is taking that can be better informed using data science. For example, if a public health department is inspecting properties for lead, data science can help inform which homes to inspect.

You don’t have to limit this to making existing actions better. Often, we end up creating a new set of actions, as well. However, it’s generally a good strategy to first focus on informing existing actions instead of starting with completely new actions that the organization isn’t familiar with implementing.

Enumerating the set of actions that exist (or can be easily added) allows the project to be actionable. If the analysis that will be done later does not inform an action, then it usually will not help the organization achieve its goals.

Let’s go back to our example of increasing on-time graduation rates for high school students. Schools have a variety of actions they can take including:

1. Creating new support programs,
2. Enrolling students in existing support programs, and
3. Scheduling a discussion with a guidance counselor or mentor.

All of these actions vary across several dimensions and need several key pieces of information to execute.

1. Creating new support programs: We need to know which programs need to be created, what capacity they should have, and who to enroll in those programs.
2. Enrolling students in existing support programs: Each program will have some capacity constraints and we need to know which students should be prioritized for which program.
3. Scheduling a discussion with a guidance counselor/mentor: This is a resource-intensive intervention, and we will need to know which students should be prioritized for this intervention.

Breaking down actions

When choosing an action to inform, it’s important to keep your goals in mind and think about different aspects of the action. First of all, you should create a list of actions an organization is taking or may take to achieve its goal. Then, you should think about who will take this action and who is impacted by the action. For example, in the case of lead inspections, a lead inspector inspects houses to look for evidence of lead that can be remediated to avoid a child being exposed. You should also consider how the organization currently chooses priorities and how often they do. For example, is the list of properties to inspect selected every day? Every week? How many are selected?

You should also consider what channels an action can be taken through. The channel an action can be taken through can have major implications for capacity constraints. For example, it generally takes fewer resources to send an email, SMS message, or even a letter than it does to make a live phone call or do in-person outreach. Capacity constraints can help you decide how valuable a data science project can be for prioritizing your actions, as well as how many actions the organization can complete at a time.

Finally, you should always consider the ethical implications of the actions you will inform with your data science project and whether prioritizing the action using a data science analysis could exacerbate inequities or other undesirable outcomes.

Step 3: What data do you have and what data do you need?

You’ll notice that so far in the scoping process we haven’t talked about data at all. This is intentional since we want these projects to be problem-centric and not data-centric. Yes, data is important and we all love data but starting with the data often leads to analysis that may not be actionable or relevant to the goals we want to achieve.

Once we understand our problem and have determined our goals and actions, the next step is to figure out what data we have access to inside and outside the organization that will be relevant to solve the problem, as well as what data sources we may need to get access to.

You first want to make a list of data sources that are available inside the organization. This is an iterative process, since many organizations may not have a comprehensive list of their data sources. Sometimes, (if you’re lucky) data may be in a central, integrated data warehouse, but even then you may find individuals or departments who have additional data or different versions of the data warehouse. Check out our [**Data Maturity Framework**](http://www.datasciencepublicpolicy.org/home/resources/datamaturity/) to assess your organization’s readiness for a data science project.

It is good practice to collect detailed information on each data source. For example, you will want to know its:

* **Granularity or detail:**The level at which the data is collected. For example, some data may be collected about individual students, some may be collected about schools, and some may be collected about neighborhoods or school districts.
* **Frequency:**How often the data is updated. For example, data may be updated annually, monthly, weekly, daily, or even immediately (in real-time). Some data may only be updated on an ad hoc basis, for example, whenever the organization receives an updated dataset.
* **History:**How far back the data goes. Having more historical data will improve the analysis.
* **Identifiers:**Usually, you will want data to include reliable and unique identifiers that allow you to link to other data sources, such as Social Security Numbers, insurance numbers, student ID numbers, or addresses.
* **Owner:**The organization, department, group, or persons who control access to the data. For example, school districts may own some student data, but other student data may be owned by individual schools or local education agencies.
* **Storage:** How the data is stored. Some data is stored in databases, while other data may be stored in spreadsheets, pdfs, data stores, or even hardcopy files.
* **Ethics:**Ethical issues that may be associated with using the data sources. For example, using some data sources may require consent from the people whose data are being used. There may also be certain security protocols that are required for accessing and using the data.

Matching the Data to the Actions

Before you start a data science project, it is important to determine whether the data you have access to matches the actions you need to inform. For example, if the actions we want to take to achieve our goals are at the individual level, then you most likely need data at the individual level. If the actions need to be decided on once a day, then you need your data to be updated every day. It’s important to match the granularity, frequency, and time horizon of the actions to the granularity, frequency, and time horizon of the data you have.

External and/or Public Data

Once you’ve determined what data you need and what data exists inside the organization, you then want to figure out what external or public data you can get that fills the gaps. Each domain often has commonly used data sources that you can incorporate into your project. The American Community Survey is a good example of a data source you want to use in most projects that involve some geographic component in the United States. Open data portals (at federal, state, and local levels) also have data that can be used to augment your internal data. 311 call data, 911 call data, and fire data are examples of some commonly found local data sources. You also want to take a look at commercial data sources you can buy to augment your internal data. Examples include buying data from organizations such as Acxiom and Experian about purchase behavior or media buying habits from Nielsen.

There may also be other data you would like to include in your analysis that you may not currently have access to or that may be difficult to access. It is important to think beyond available sources of data and consider data that *could*be helpful for your project, even if it’s not currently available. Identifying useful but unavailable data can help you identify potential data sources that may improve your analysis. There may also be proxies for that data that exist. This can include data that is very secure or difficult to access like CCTV videos, phone records, or DNA records. It can also include data that would require additional effort to gather, such as survey results, or data that would require system changes to collect, such as data that is updated more frequently or collected at a different level of granularity.

While it is important to think big when considering what data may improve your analysis, it is also important to keep ethical considerations in mind, especially when accessing sensitive data sources. External data sources that are not public may require additional legal agreements and requirements to access.

Step 4: What analysis needs to be done?

Once we have the goals, actions, and the available data identified, the final step in the scoping process is to determine the analyses we will do to inform the identified actions, using the data we have, to achieve our goals. It is worth reiterating that analysis is not the goal of a data science project, and it’s advisable to leave this piece aside until the goals, actions, and available data are clearly defined.

Analyses can use methods and tools from different areas: computer science, machine learning, data science, statistics, and social sciences. One way to think about the performed analysis is to break it down into five types:

**Description:** Primarily focused on understanding events and behaviors that have happened in the past. Typically all projects contain a descriptive analysis component for gaining an understanding of the data. However, it is rare that a descriptive analysis is the sole analysis component of a project.

E.g.:

* Exploring data around students that haven’t graduated on time in the past, potentially revealing “types” of students that suffer from the adverse outcome when compared to others with respect to highly correlated factors.

**Detection:** Less focused on the past and more focused on ongoing events. Detection tasks often involve detecting events and anomalies that are currently happening.

E.g.

1. Using text data from legislative bills to identify their topic areas
2. Detecting fraudulent credit card transactions

**Prediction:** Focused on the future and predicting future behaviors and events.

E.g.

* Predicting the likelihood of a student graduating high school on time at the time they enter 9th grade,
* Predicting the likelihood of a legislative bill passing into law at the time it’s introduced to the legislature.

**Optimization**: Focused on taking the outputs of other types of analysis and using them to allocate resources or make decisions.

E.g.

* Identifying where to locate ambulances such that their coverage is maximized,
* Given a list of students ranked by their probability to not graduate on time, identify the subset of students for intervention that would have the most impact

**Behavior Change:** Focused on causing a change in behaviors of people, organizations, neighborhoods. Typically uses methods from causal inference and behavioral economics.

E.g.

* Given a student is at risk of not graduating on time, and a set of possible interventions, identifying the intervention that would maximize their likelihood of graduating on time

There are of course many more types of analyses but we’ll keep the focus on these five here.

For each analysis that we do in the project, we should look to answer the following

* *What type of analysis needs to be done and what’s the purpose?*

Is this a descriptive analysis, a predictive model, or a detection or behavior change task? Often, the project involves several of the types of analysis we described above, each designed to inform specific actions and achieve specific goals.

* *How will the analysis inform the actions?*

Each analysis should inform one or more of the identified actions. Some analyses could inform multiple actions and sometimes one action could be informed by multiple analyses.

* *How will the analysis be validated? What validation can be done using existing, historical data?*

It’s important to think about how we can validate each analysis using the available, historical data. We should choose the validation set up to reflect the deployment scenario (how a user would use this model for example) as closely as possible. Typically, for prediction tasks, we would create multiple train validation sets and choose a metric that accurately reflects how we use the model, and observe how our models perform over time. One critical thing we would need to think about is the baseline(s) to which we compare our models. If the organization is performing this task today in some way, a good baseline is the performance of the “existing system”. Additional good baselines include simple heuristics (based on expert knowledge or prior research). We want our new analysis to be “better enough” than existing or simpler approaches to make it worth deploying the more expensive-to-build-and-maintain data science system.

1. *What are the ethical issues associated with the analysis?*

Are there any equity implications that we are concerned about?  How would errors in your analysis impact individuals and society?

We’ll demonstrate this with a few examples.

**Reducing Lead Poisoning Rates**

**Goal:** Reduce the number of children who will get lead poisoning in the future due to lead hazards in their current residence

**Actions:**Proactive home inspections constrained by limited inspection resources.

**Analysis:** We need to answer the key question: “Which homes should we prioritize for proactive inspections?” In other words, we want to identify the homes of kids who are at risk of lead poisoning in the future.

Because we want to intervene before a child is exposed to lead, we would predict the likelihood of every kid under the age of 2 being exposed to lead in their homes. We could use our predictions to rank homes based on the risk that a child will be exposed to lead in them and prioritize them for inspection. If we consider the five types of analyses we mentioned above, this would be a prediction task.

**Validation:**  Let’s say, the number of homes the agency can inspect in a month is 100. Given this, we can optimize the analysis to predict the 100 homes where a child is most likely to be exposed to lead each month, a metric we call Precision at K (or P@K).

In this example, the causal link between the action (lead inspection and removal) and the goal (protecting children from lead exposure) is well established. That might not be the case in every project. Let’s take the example of timely high school graduation:

**Improving graduation rates**

**Goal:** Increase the percentage of high school students who graduate on time, while reducing the disparity in graduation rates across racial groups

**Actions:**Provide additional support to students who are identified as at-risk of not graduating on time

**Analyses:**

1. As with the lead hazards project, we have to perform a predictive analysis; predicting which students are least likely to graduate on time. We can use these predictions to prioritize a subset of students for additional outreach and support.
2. Once the risk is predicted for each student, selecting the subset of students for intervention might not be as straightforward as prioritizing students with the highest predicted risk. Given the nature of the interventions and the metric you are interested in improving, the best students for intervention could change. For instance, if the goal is to improve the average graduation rate with minimum effort, prioritizing students with a more middling risk score could prove to be more effective than prioritizing those who are very likely not to graduate on time. An optimization analysis could help identify the optimal set of students to be prioritized for intervention.
3. Even if we select the “optimal” set of students for intervention, there could be multiple additional support types available (e.g., after-school tutoring, counseling, help with transportation, financial help). Identifying the most effective intervention(s) for individual students could require a causal inference analysis.

Ethical Issues:

As a reminder, consideration of ethical issues should be embedded in every phase of the scoping and execution of a project, rather than thought of as a discrete “step” in that process. Often the initial conversation around ethics in the scoping phase will focus on the ethical and societal values we want to embed in the system. Note that this conversation is not specific to applications of data science or AI, but rather about the values that broadly need to be embedded in the decision-making processes that affect people’s lives. As such, these decisions cannot be made unilaterally by the data scientists or technologists tasked with building the system. While they have an important role in the discussion, it is important for it to be inclusive of all the relevant stakeholders: policymakers, action-takers, system developers, data owners, and the community being affected by the system will all have important perspectives on these issues.

As you work through the scoping process (and the project itself), being explicit about how your work reflects those values can help act as a guiding principle in the course of the many decisions that need to be made to see a project through. To highlight their central nature, we’ve consolidated many of them here, posing some motivating questions to help you explore these issues. This list is far from exhaustive, but rather provides a starting point for that conversation:

**Privacy, Confidentiality, and Security**

Concerns about privacy and data security are common in applications of data science to socially impactful problems, particularly where high-stakes decisions are involved such as healthcare, criminal justice, and education. In some cases, there may be governing legal requirements about how data can be used and steps that must be in place to protect it, but ethical considerations may extend beyond what is legally required. Particularly important here is how people might feel about the data about them that is being used, as well as their expectations about how publicly available that data is. Questions to ask here include:

* What are the privacy considerations (legal as well as ethical)?
* How is the privacy of the individuals in the data being protected?
* What about confidentiality?
* What are the security considerations and protections? Who has access to which parts of the data? For what purposes? What is the security audit process?

**Transparency**

Transparency considerations for the building and deployment of data science systems can involve many different stakeholders: internal actors and decision-makers, individuals whose data is being used, individuals who will be affected by the decisions the system informs, and the public at large. As you consider the following questions, keep the perspectives of each of these stakeholders in mind:

* Which stakeholders should know about the project?
* Do the people whose data you’re using know that you’re using it?
* How will the actions you’re taking based on this analysis affect people? What are the costs or benefits of this action for the people affected?
* Do the people you’re prioritizing know if and why they’re being prioritized?
* What recourse do people have to challenge or change a decision informed by your analysis that has impacted them?

**Discrimination, Equity, and Fairness**

Understanding and improving the fairness of machine learning systems has been the subject of much recent writing and research, both in scholarly works and the popular press. During the scoping process, it is important to understand the kinds of disparities that could result from your project and how they might impact people, accounting for the perspectives of people impacted by your analysis and any historical and ongoing discrimination that might affect them.

For instance, in the context of screening child welfare hotline calls, disparities in false negatives (not following up a report with an investigation when you actually should have) would mean more harm to children in one community relative to another. By contrast, disparities in false positives (making an unnecessary investigation) could result in over-policing some communities resulting in broader societal harms. Two tools from our work that may be helpful here are the Fairness Tree, which we use to help facilitate these conversations, and Aequitas, which helps audit model outputs for disparities – both can be found [here](http://www.datasciencepublicpolicy.org/projects/aequitas/). Some of the key questions to ask with regards to discrimination and fairness are:

* Which specific groups could be impacted unequally by your analysis, and how should you account for this?
* From the perspective of each of these groups, how do they define fairness in this context? How well aligned are these definitions?
* Over what time horizon are you trying to improve equity in this work?
* How will you detect biases or inequities in your system?
* What disparities exist in the current decision-making process?
* If inequities do exist, how will you approach reducing them in your system or mitigating their impact in downstream decisions and interventions?
* Are there broader sources of inequities, either historical or ongoing, that affect the outcomes you are trying to improve? How should your system take these into account?

**Accountability**

Accountability should be considered with respect to the entire process, including both the technical decisions made while working with the data and the decisions made about how the system will be used in practice and described to the public. Ensuring there is clarity upfront about who is accountable and responsible for each aspect of the system, particularly with regards to any potential ethical aspects and considerations, can help reduce both the risks of issues arising (by setting boundaries and making oversight somebody’s explicit responsibility) and potential harms if these issues do arise (by having contingency plans in place). For example, our group has worked on several projects seeking to reduce the risk of jail incarceration by providing assistive interventions like mental health supports or social services. However, the same model predicting risk of future arrest could also be put to use in more concerning, punitive ways that our partner agreements need to carefully define and guard against. Some questions to ask around accountability include:

* If sensitive data is used for the project, who is responsible for keeping it safe? What contingency plans are in place in case there is a leak and who will be accountable?
* What limitations are in place on how your analysis will be used? Who bears responsibility if it’s used for harmful or unintended purposes?
* How will you monitor for unintended consequences (for instance, if students predicted at risk for low performance internalize this prediction)? What processes will be in place to reduce any potential harm from them?
* How will you determine if the system is increasing disparities over time? Who is responsible for monitoring this and making decisions about how to respond?
* Who bears responsibility for determining what to disclose about the system to the public and how to describe it? If people object to it, who is accountable and what should be done?
* What recourse do people impacted by the new system have if it makes mistakes about them or is making recommendations based on inaccurate data? Who is responsible for responding to these issues and making any necessary improvements to the underlying model or analysis?

**Social License**

The concept of social license is related to ideas of transparency and accountability, but with a different framing that can provide a helpful thought exercise. Regardless of your actual plans around transparency described above or how many people might learn about the details of the project in practice, here you want to think about how people might respond to your project if it did receive widespread coverage. Would it be uniformly cast in a positive light or a negative one, or to what extent would different people react differently to it? A few questions to help you consider social license for your project include:

* If the entire population of the country finds out about your project, will they be ok with it? Why?
* If it was on the front page of the newspaper, what would the headline be? Would it be positive or negative?
* Even if you believe the population might support your project overall, are there any groups who might object? Who are they and what concerns would they have?

**Other Ethical Considerations**

Finally, it’s important to keep in mind that the list here is only a starting point and there may be other ethical considerations relevant to your specific context. Are there additional legal, policy, or organizational requirements not covered above? Do you need to gather informed consent for individuals potentially affected by this work? Are there specific oversight, reporting, or ethical review procedures? As is the case with the other considerations above, it can be helpful here to think about the project through the perspectives of different stakeholders and explore whether any considerations or concerns might arise that don’t fit well into the other categories here.

Additional Considerations

**Evaluation**

Once your analysis is complete and you have validated it against available data, it is important to evaluate it in the field to make sure that it helps achieve your goals. Field trials are a complicated topic and can be fairly specific to the project and its goals, so we will not go into them in detail here. However, your evaluation should be an application of the analysis to the actions they were proposed to inform and should be measured against your pre-defined goals.

Such an evaluation often requires the buy-in of people inside and outside the organization and a significant commitment of the organization’s time and resources. Even before embarking on a new project, the organization should be certain that they can commit the necessary time and resources and should engage anyone who will be involved in the evaluation and application of the analysis. These conversations and commitments should happen at the beginning of the scoping process to ensure the project will be successful.

The evaluation may show that the analysis is or is not successful at achieving the organization’s goals. The evaluation should not be seen as an opportunity to prove that the analysis is successful, but an opportunity to test its effectiveness and perhaps improve upon it iteratively by re-evaluating other aspects of the project’s scope, such as the analysis itself, the data used for the analysis, or the actions to which the analysis is applied.

**Deployment**

If the analysis proves effective in the evaluation, then you will want to deploy it as a new system on the organization’s infrastructure so that it can update regularly and be incorporated into its operations. Among the most important considerations when deploying a new system is whether or not it was built on the organization’s infrastructure. If the system was built on the organization’s infrastructure, it will often (though not always) be easier to deploy, because it should already be consistent with the organization’s technical infrastructure. Data scientists working within the organization will usually have access to the organization’s internal infrastructure and should work closely with production engineers to make sure their system integrates with it. Volunteers and other external data scientists may also be able to access the organization’s infrastructure through volunteer or business agreements, making for a more seamless deployment process.

However, often data science systems are built outside of the organization’s infrastructure, possibly because of security or privacy issues that limit access. When this happens, deployment may be more complicated. The organization’s internal data and engineering professionals who work with and maintain their technical infrastructure should be incorporated into the scoping process early to ensure that the system built by partner data scientists fits the organization’s infrastructure and can be integrated and deployed fairly easily. The project should be a priority for the data and technical professionals in the organization who are necessary for deployment, and there should be time held in their schedules to devote to this project, just like any other. If the organization’s own data and technical professionals are not involved in the project and do not have time budgeted for the project, then it will be doomed to fail. The same data and technical professionals who help deploy the new system may need to maintain it in production, and this should be figured into their schedule and workload.

Data science systems also require computing resources for deployment. A good understanding of the organization’s computing capacity should be achieved before starting the project. In deployment, data science systems will use memory and take time when they are updated. This should be accounted for within the organization’s infrastructure, and internal data and technical professionals should make sure that regular updates of data science systems do not conflict or compete for resources with other important technical processes.

Data science systems also rely on regular and consistent updates to underlying data sources. Data and technical professionals in the organization should make sure that regular updates of the new system are consistent with updates of the underlying data. If the data is old, delayed, or otherwise inconsistent with updates to the new system, then the value of that system will be reduced and may even deteriorate over time.

Finally, it will be important to consider who in the organization will use the output of the system and how they will access it. The key to this is thinking about how the end-user will actually use the information. Will the end-user be looking at individual cases one by one, or will they be creating a list to prioritize? If the end-user is considering cases one-by-one, then a visual user interface that provides them the output alongside other information about the case may be helpful. If the end-user is instead creating a prioritized list (of places to inspect, for example), then it may be better to deliver them a list of the top cases, perhaps in a visual user interface or even as an email or printable document! These decisions should take into account the end-user’s current operational processes, their access to technology in their work, and their literacy (digital and otherwise), among other considerations. Training will also be a key part of the deployment, both for the technical maintenance of the system and the end-users who will use it to make decisions.

**Monitoring**

Once the data science system is deployed and integrated into the organization’s operations, it will need to be continuously monitored to ensure that it continues to perform well and provide information that improves decision-making. The performance of data science systems can deteriorate over time for a variety of reasons, including changes to the underlying data or data quality, changes in the organization’s policies or operations, or even changes in real-world trends. To ensure the information that the system provides remains useful, organizations should have a plan and commit resources to monitor the system’s performance over time as a regular part of their operations.

First, you should create a plan for how the system will be monitored, including what data and metrics should be used to assess its performance over time. This data will likely be future updates to the same data that was used to develop the original analysis and may use some of the same metrics and diagnostics that were used to validate it. When monitoring performance, you should be careful to ensure that data used to assess the system does not leak into the data used to update it. This could lead to overestimating the system’s performance and masking any decay.

The organization will also need to commit technical staff to monitor the system’s performance and managing it if it breaks or if performance deteriorates. This requires understanding not only the system but the data that is used to update and evaluate it. Technical staff may need to be trained to understand, evaluate, and update the system, and some redundancy is recommended so that the organization has the skill depth to maintain the system, even if a staff member leaves.

The organization should also plan to run additional field evaluations (like the one described above) at regular intervals to ensure that the system is still providing information that is actionable and improves operations. This requires the commitment of not only data professionals but also end-users to the evaluation of the system. The frequency of these evaluations may vary with the organization’s capacity, how often the data and the system are updated, and the frequency of changes in the organization’s operations and the overarching policy environment. However, it is important that the organization plans to devote resources to the continued monitoring and evaluation of the system. After all, a system that provides poor information could be worse than no system at all!

Finally, you should plan what you will do if the system does show signs of decay. This plan should consider what the sources of decay may be, including changes to input or evaluation data, changes to the organization’s policies or operations, or changes to real-world trends. Responses will differ depending on the sources of performance decay.

You should also consider what the organization will do with the system if they change their infrastructure, data, policies, or operations. The system should be evaluated under the new changes and it should be determined whether it still provides useful and actionable information, or if it will need to be updated or retired. This is important because changes to the input data or the infrastructure that support the system could impact its performance, while changes to policy or operations may mean the system no longer provides information relevant to those new processes.

Summary

Hopefully, this gives government agencies and nonprofits an overview of how to scope data science/analytics projects and what questions they need to answer before launching the project.

**Step 1: Goals** – Define the goal(s) of the project

**Step 2: Actions** – What actions/interventions do you have that this project will inform?

**Step 3: Data** – What data do you have access to internally? What data do you need? What can you augment from external and/or public sources?

**Step 4: Analysis** – What analysis needs to be done? Does it involve description, detection, prediction, optimization, or behavior change? How will the analysis be validated?

**Ethical Considerations**:  How have you thought through privacy, transparency, discrimination/equity, and accountability issues around this project?

We’ve also created a [Blank Project Scoping Worksheet](http://www.datasciencepublicpolicy.org/wp-content/uploads/2021/09/Project-Scoping-Worksheet-Blank-10-2021.pdf) you can use. We’d love feedback as you use these as we frequently run workshops and training on project scoping for students as well as government agencies and nonprofits. If you’re interested in being part of this training, [let us know](mailto:dssg@cmu.edu).